University of Southampton

Faculty of Environmental and Life Sciences

Biological Sciences

Bioinformatic Analysis of Human Next Generation Sequencing Data; extracting Additional Information, Optimising Mapping and Variant Calling, and Application in a Rare Disease

Volume 1 of 1

By

Roshan Kumar Sood

A thesis presented for the degree of

Doctor of Philosophy

May 2019

University of Southampton

Abstract

Faculty of Environmental and Life Science

Biological Sciences

A thesis presented for the degree of Doctor of Philosophy

Bioinformatic Analysis of Human Next Generation Sequencing Data; extracting Additional Information, Optimising Mapping and Variant Calling, and Application in a Rare Disease

by Roshan Kumar Sood

With the increased application of Next Generation Sequencing (NGS) to medicine it is important to test and develop approaches to extract the optimum information from datasets. In this thesis five aspects of NGS are investigated ranging from quality control to variant calling. Firstly a method to estimate contamination from a VCF file was developed which would be useful in cases where no BAM file was available to use existing tools. Unmapped reads were investigated to extract additional information from NGS samples and were able to detect the abundance of oral microbes from saliva samples relative to blood collected samples, but failed to identify differences between inflammatory bowel disease patients and controls. For a familial trio with a reported rare case of Sedaghatian-type spondylometaphyseal dysplasia (SSMD) sequenced both by whole exome (WES) and genome (WGS) sequencing it was shown that nearly all coding variants from WES were called in WGS despite differences in mean depth of coverage. This comparison highlighted that as sequencing costs decrease WGS will offer the greatest diagnostic value with potential for future re-analysis of cases currently unable to be resolved. Using the familial trio attempts were made to identify causal variant(s) in the gene currently implicated in causing SSMD – Glutathione peroxidase 4 (*GPX4*). However no variants either small SNPs or large structural were identified over the *GPX4* gene and no plausible candidates were identified from the trio. Finally variant calling of the FCGR low affinity locus was performed using targeted NGS. FCGR genes have been highly duplicated and so by using customised references it was possible to infer the combinations of alleles across homologous sites. Using this approach it was possible to predict SNPs in the FCGR3B gene and predict human neutrophil antigen haplotypes involved in the immune response to treatments such as monoclonal antibodies.

Contents

1 Introduction **25**

1.1 Next generation sequencing 25

1.1.1 First generation 25

1.1.2 Second generation 27

1.1.3 Third generation 33

1.2 Sequencing projects 35

1.2.1 Human Genome Project 35

1.2.2 dbSNP 36

1.2.3 HapMap 36

1.2.4 ENCODE project 36

1.2.5 1,000 Genomes 38

1.2.6 Other sequencing projects 39

1.2.7 100,000 Genomes project 40

1.3 Software pipelines 41

1.3.1 Pre-processing 42

1.3.2 Read mapping or assembly 44

1.3.3 Human reference sequence 45

1.3.4 Post-mapping 46

1.3.5 Variant calling 46

1.3.6 Variant annotation 48

1.3.7 Structural & copy number variants 50

1.4 Aims 53

2 Estimating contamination levels in exome sequencing using alternate allele frequencies and variant zygosity **55**

2.0.1 Contamination estimation tools 57
2.0.2 Alternate allele frequency changes with contamination 59
2.0.3 Applications of machine learning with NGS 63
2.0.4 Regression models 63
2.1 Aims 65
2.2 Materials & methods 66
2.2.1 Contamination simulations 66
2.2.2 Alignment and variant calling pipeline 68
2.2.3 Alternate allele frequency profiles 69
2.2.4 Measurements used for contamination estimation 71
2.2.5 Investigating relationships of measurements 72
2.2.6 Principal component & clustering analysis 72
2.2.7 Regression analysis - model selection and training 72
2.2.8 Application of regression models 73
2.3 Results 74
2.3.1 Alternate allele frequency profiles and measurements used in contamination estimation 74
2.3.2 Principal component and clustering analysis 82
2.3.3 Training regression models 84
2.3.4 Application of regression models 93
2.4 Discussion 105
2.5 Conclusions 111

3 Unmapped reads provide insights to potentially clinically important information and can distinguish collection methods 113
3.1 Introduction 113
3.1.1 Unmapped reads 113
3.1.2 Cross-species contamination 114
3.1.3 Microbiome information from unmapped reads 115
3.1.4 Analysis programs 118
3.1.5 MEGAN6 119
3.1.6 Sequencing methods 119
3.2 Aims 120

3.3 Materials & methods 121
3.3.1 Samples 121
3.3.2 Extraction of unmapped reads 123
3.3.3 FastQC 124
3.3.4 BLAST classification of sequences 124
3.3.5 Creation of taxonomic trees 126
3.3.6 Clustering analysis 127
3.3.7 Tandem repeat finder 127
3.3.8 Depth of coverage across bacterial genomes 127
3.3.9 Calculating sequence similarity between species 127
3.3.10 Plotting of bacterial genomes 128
3.4 Results 129
3.4.1 Extraction & classification of unmapped reads from exome case and control samples 129
3.4.2 Comparing collection methods 140
3.4.3 Unmapped reads of RNA-SEQ and WES from the same individual 145
3.4.4 Comparison of 16S rRNA sequencing with unmapped WES reads mapped to bacteria 154
3.4.5 Investigating *cronobacter sakazakii* read matches 155
3.4.6 Mapping unmapped reads back against the human genome . . . 161
3.4.7 Unclassifiable reads 164
3.5 Discussion 165
3.6 Conclusions 174

4 Comparing variant calling using a whole exome and genome sequencing trio 175
4.1 Introduction 175
4.1.1 Non-coding variant annotation 176
4.1.2 Limitations of WES compared to WGS 179
4.1.3 Variant calling with GATK HaplotypeCaller 182
4.1.4 ACMG genes 183
4.2 Aims 183
4.3 Materials & methods 184

4.3.1 Samples & sequencing 184
4.3.2 Quality control 185
4.3.3 Sample coverage 187
4.3.4 Variant calling 188
4.3.5 Annotation of VCF variants 188
4.3.6 Annotation of non-coding variants 190
4.3.7 GEMINI variant analysis 190
4.3.8 Structural and copy number variant pipeline 191
4.3.9 Comparison of exome and genome variants 194
4.4 Results 195
4.4.1 Quality control 195
4.4.2 Variant calling 202
4.4.3 Non-coding variants 211
4.4.4 GEMINI variant calls 214
4.4.5 Structural & copy number variants 217
4.4.6 Comparing whole exome with whole genome sequencing 226
4.5 Discussion 234
4.6 Conclusions 241

5 Prioritisation of trio called variants 243
5.1 Introduction 243
5.1.1 Using NGS to identify disease causing variants 243
5.1.2 Sedaghatian-type SpondyloMetaphyseal Dysplasia 243
5.1.3 Short-Rib polydactyly syndromes 246
5.2 Aims 251
5.3 Materials & methods 252
5.3.1 Family & sequencing 252
5.3.2 Variant calling 253
5.3.3 *GPX4* variant, transcription factors and binding sites 253
5.3.4 *GPX4* coverage 254
5.3.5 GEMINI variant analysis 254
5.3.6 Non-coding variants 257
5.3.7 Splicing variants 257

5.3.8 Structural variant calling 258
5.3.9 Copy number variation 258
5.3.10 Cross-dataset compound heterozygotes 258
5.4 Results 259
5.4.1 *GPX4* coverage 259
5.4.2 *GPX4* variants 261
5.4.3 *GPX4* transcription factors and binding sites 261
5.4.4 GEMINI variants 263
5.4.5 Splicing variants 272
5.4.6 Non-coding variants 273
5.4.7 Structural & copy number variants 274
5.5 Discussion 284
5.6 Conclusions 291

6 Approaches to call SNVs and CNVs amongst high similarity FC-gamma receptors 293
6.1 Introduction 293
6.1.1 Fcγ receptor functions 293
6.1.2 FCGR Genes - Function & SNVs 295
6.1.3 Copy number variants 301
6.1.4 Mapping and variant calling FCGR genes 302
6.1.5 HaloPlex capture kits 303
6.2 Aims 303
6.3 Materials & methods 304
6.3.1 Samples & sequencing 304
6.3.2 Quality control & mapping pipeline 304
6.3.3 Sample coverage 307
6.3.4 Customised reference genomes for variant calling 307
6.3.5 Variant analysis pipeline 312
6.3.6 Copy number variant calling 313
6.4 Results 315
6.4.1 Quality control 315
6.4.2 Sample depth of coverage 321

6.4.3 Homology between FCGR genes 325
6.4.4 VCF pairwise identity matrices 328
6.4.5 Variant analysis 330
6.4.6 Calling FCGR variants using customised reference genomes . . . 334
6.4.7 Copy Number Variation - CNVkit 339
6.5 Discussion 350
6.6 Conclusions 357

7 Conclusions 359

8 Appendices 391

8.1 Appendix A 392
8.1.1 VerifyBamID analysis of control samples 392
8.1.2 Analysis scripts 393
8.1.3 Contamination estimation program results for 200 contamination simulations 397
8.1.4 Contamination estimation program results for 245 germline samples 403
8.2 Appendix B 410
8.2.1 Unmapped reads per sample 410
8.2.2 Unmapped reads by kingdom or domain 417
8.2.3 *CABIN1* Average coverage per exon 428
8.3 Appendix C 429
8.3.1 Extended 430 candidate gene list for skeletal trio 429
8.3.2 CNVkit WGS filtered variants 436
8.3.3 LUMPY - Large heterozygous variants 437
8.4 Appendix D 438
8.4.1 Genes and promoters targeted by HaloPlex capture kit 438
8.4.2 Initial mapping statistics for HaloPlex samples 440
8.4.3 Estimating contamination of HaloPlex samples 441
8.4.4 Merging of overlapping reads using Pear 443
8.4.5 HaloPlex mean sample coverage 445
8.4.6 Custom calling using per gene references to determine HNA-1 haplotypes 447

8.4.7 Custom calling using per gene references to determine *FCGR2C* X57Q variants 449
8.4.8 Comparing HaloPlex - CNVkit references against MLPA CNV calls 451
8.4.9 Comparing HaloPlex - CNVkit segmentation algorithms against MLPA CNV calls 452
8.4.10 HaloPlex - grouping CNVkit calls with known CNRs in the FCGR locus 453

List of Figures

1.1 Cost per megabase for DNA sequencing since 2001 to 2017 26

1.2 Summary of the Illumina sequencing method 28

1.3 Cost per genome for DNA sequencing for 2001 to 2017 32

1.4 Structural variant types 51

2.1 Cross-sex contamination changes of chromosome X heterozygosity . . . 62

2.2 Pipeline used to generate *in silico* contamination simulations with known percentages of contamination 67

2.3 Novoalign hg19 variant calling pipeline 68

2.4 Contamination estimation program pipeline used to obtain measurements of the alternate allele frequencies of variants. 70

2.5 Alternate allele frequency profile for a representitive contamination simulation set created from VCF files 75

2.6 Scatter matrix of all nine measures and contamination levels 80

2.7 Clustering of contamination simulations 82

2.8 Evaluation of linear regressions using 1-50% training data with all features 86

2.9 Evaluation of non-linear regressions using 1-50% training data with selected features 87

2.10 Regression analyses trained using all features for contamination levels up to 10% 89

2.11 Non-linear regression models using the 1-10% training range with selected features 90

2.12 Evaluation of linear regressions using 1-20% training data with all features 91

2.13 Evaluation of non-linear regressions using 1-20% training data with selected features 92
2.14 Polynomial regressions tested with 245 germline samples with training set 1-20% 94
2.15 OLS regressions with different ranges of contamination used in training 96
2.16 Alternate allele frequency histogram of sample 4 99
2.17 Alternate allele frequency histogram of sample 217 100
2.18 Alternate allele frequency profiles for the highest five predictions from OLS but not VerifyBamID 103

3.1 Unmapped reads BLAST analysis pipeline 125
3.2 Violin plots of unmapped reads by sample group 129
3.3 Percentage matches to databases per exome sample split by controls and cases 130
3.4 Clustering samples using databases matches comparing cases and controls 131
3.5 Comparison of unmapped read totals and percentages by collection method 141
3.6 Clustering of exome samples by collection method 144
3.7 Exome and RNA-seq sample unmapped reads by database 146
3.8 Clustering of exome samples from individuals also with RNA-seq 148
3.9 Clustering of RNA-seq samples by collection site and disease sub-type . 153
3.10 Cronobacter sakazakii read G-C content 156
3.11 Comparison of quality of *Cronobacter* reads with the same reads mapped only human 158
3.12 Visualisation of *cronobacter* species matches 160
3.13 *CABIN1* Re-mapped reads 161
3.14 RNA-seq re-mapped read locations 163
3.15 Sequence duplicaiton levels of all unclassified reads 164

4.1 Flowchart of the BWA-GATK variant calling pipeline used for the trio WES and WGS samples, including GEMINI and LUMPY analysis steps 186
4.2 MultiQC visualisation of per base quality of WES data 195
4.3 MultiQC visualisation of per base quality of WGS data 196
4.4 Exome skeletal dysplasia sample coverage histograms 198
4.5 Plot of decreasing cumulative frequency for WES skeletal dysplasia samples 199

4.6 Histogram of decreasing cumulative frequency for WGS skeletal dysplasia samples . . . 200
4.7 Cumulative coverage histogram for WGS of skeletal dysplasia samples . 201
4.8 Heatmap showing the similarity of Sketeal dysplasia WES samples for the trio. . . . 202
4.9 Heatmap showing the similarity of Sketeal dysplasia WGS samples for the trio. . . . 203
4.10 Visualisation of variant rs2298628 in exome sample SD003 . . . 204
4.11 SNP fingerprint for Skeletal dysplasia hg38 sample . . . 205
4.12 Comparison of variant calls by type for WES and WGS trios . . . 207
4.13 Comparison of coding and splice variant calls by consequence for WES and WGS trios . . . 210
4.14 Non-coding variants called by sample . . . 211
4.15 Distribution of non-coding variants annoated as pathogenic by CADD . 212
4.16 Distribution of non-coding variants annoated as pathogenic by FATHMM-XF . . . 213
4.17 GEMINI variants by tool comparing WES and WGS . . . 215
4.18 LUMPY variants by type per WGS sample . . . 218
4.19 LUMPY variants called from WGS samples split by type and plotted per MB for each chromosome . . . 219
4.20 LUMPY inversions per chromosome . . . 220
4.21 Total segment calls per sample by CNVkit . . . 221
4.22 CNVkit segment calls split by chromosome and type per sample . . . 223
4.23 CNVkit agreeing CNVs between WGS and WES . . . 224
4.24 CNVkit disagreeing CNVs between WGS and WES . . . 225
4.25 Total overlaps of WES and WGS variants . . . 227
4.26 Classification of remiaing high-quality exome only variants . . . 228
4.27 Fraction of exome covered at depths for WES and WGS . . . 230
4.28 Fraction of exome covered at depths for WES and WGS from 1-20x . . 231
4.29 ACMG genes covered at 10x depth of coverage . . . 232
4.30 WGS sample coverage for *BRCA1* . . . 233
4.31 WGS sample coverage for *TP53* . . . 233

5.1 Isoforms of *GPX4* . . . 245

5.2 Chromosome 15q13.3 homology map 249
5.3 Pedigree diagram of skeletal dysplasia case 252
5.4 *GPX4* coverage 260
5.5 Transcription factor binging site in *GPX4* UTR 262
5.6 *WDR60* frameshift deletion called by LUMPY 277
5.7 CNVkit calling of the 15q13.3 micro-deletion 280
5.8 15q13.3 micro-deletion visualisation for SD003 281

6.1 Copy number regions reported in FCGR genes 301
6.2 HaloPlex bioinformatic analysis pipeline flowchart 306
6.3 Examples of allele ratio calculations 309
6.4 *FCGR2B* and *FCGR2C* Q57X allele ratios 311
6.5 Boxplot FCGR inital Fastqc total sequences 316
6.6 Insert sizes of HaloPlex samples 317
6.7 HaloPlex amplicon and inster sizes histogram 318
6.8 Average depth of coverage across all samples 322
6.9 Median depth of coverage across the FCGR Locus 324
6.10 Alignment of FCGR genes 327
6.11 HaloPlex sample matrix of related samples 328
6.12 Plotting total variants against depth and Het:Hom ratio for HaloPlex samples 331
6.13 Examples of changing CNVkit bin size 342
6.14 CNVkit Gains, losses and neutral segments per HaloPlex sample 344
6.15 Grouping of CNVs called over FCGR locus with known copy number regions. 348
6.16 CNR-6 sample CNV visualisations 349

8.1 CNVkit compared with MLPA results 451
8.2 CNVkit segmentation algorithms compared with MLPA results 452

List of Tables

1.1 GRCH genome comparisons 45

2.1 Features measured from the VCF files 71
2.2 Summary of measures used in regressions obtained from 200 simulations dataset 78
2.3 PCA variance explained by components for contamination measures . . 83
2.4 PCA variance explained by each component for contamination measures 83
2.5 Regression R^2 values when changing training sample ranges and features 84
2.6 Regression R^2 values when changing training sample ranges to above 10% 85
2.7 Samples predicted by OLS regressions above thresholds by training range. 97
2.8 VerifyBamID predicted samples with above 1% contamination 98
2.9 Regression coefficents 101
2.10 Highest five predictions from OLS but not VerifyBamID 102
2.11 Summary of measures used in regressions obtained from application . . 104

3.1 RNA-seq samples 122
3.2 16S rRNA samples available to compare with exome unmapped reads . 123
3.3 BLAST databases obtained from RefSeq 126
3.4 Bacterial unmapped reads mapping to the phylum Acintobacter 132
3.5 Bacterial unmapped reads mapping to the phylum Bacteroides 133
3.6 Bacterial unmapped reads mapping to the phylum Firmicutes 134
3.7 Bacterial unmapped reads mapping to the phylum Proteobacteria . . . 135
3.8 Number of samples with ten or more reads for selected bacterial species for the phylum *Proteobacteria* 136
3.9 Top matched viral species from exome samples 137
3.10 Top plant species matches from unmapped exome reads 138

3.11 Top metazoan species matches from unmapped exome reads 139
3.12 Total unmapped reads for seven WES and all 17 RNA-seq samples . . . 145
3.13 Exome species with 10 or more unmapped reads 147
3.14 RNA-seq species with above 500 read matches 150
3.15 RNA-seq bacterial species with above 10 read matches 151
3.16 Unmapped read totals from exome sequencing for samples with 16S sequencing also 154
3.17 Summary results of supplied of 16S rRNA heatmap 154
3.18 *Cronobacter* read match samples and totals 155
3.19 Tandem repeat finder results for *Cronobacter sakazakii* matches 157
3.20 Depth of reads mapped to *Cronobacter sakazakii* 157
3.21 *Cronobacter* sub-species identity 159
3.22 Re-mapping reads to human 162

4.1 Summary of databases for ANNOVAR 189
4.2 CNVkit thresholds 193
4.3 WGS and WES read sequence totals 196
4.4 Adapter sequences detected in skeletal dysplasia WGS samples 197
4.5 Skeletal dysplasia read alignment statistics 197
4.6 Skeletal dysplasia WES and WGS trio results from QC program 203
4.7 LUMPY genotypes of variants per WGS sample 217
4.8 Clustering of genes with coding variants in WES only 229

5.1 SRPS genes listed on OMIM 247
5.2 Chromosome 15q13.3 breakpoint liftover locations from hg17 to hg19 and hg38 250
5.3 Variants in a region covering 1kb either side of *GPX4* from WGS data . 261
5.4 Genome and exome read sequence totals 263
5.5 Autosomal recessive variants - Tier-2 HGMD variants 263
5.6 Autosomal recessive variants - Tier-3 extended candidate gene variants . 264
5.7 Autosomal recessive variants - Tier-5 Filtering variants 265
5.8 Non-Mendelian - Tier-3 extended candidate gene variants 265
5.9 Non-Mendelian - Tier-5 Filtered variants 266
5.10 Non-Mendelian - Tier-5 Filtered variants gene descriptions 267

5.11 Compound heterozygotes Tier-2 variants 268
5.12 Compound heterozygotes Tier-3 variants 269
5.13 Compound heterozygotes Tier-5 variants 270
5.14 Compound heterozygotes Tier-5 variants : Gene and phenotype descriptions 271
5.15 Compound heterozygotes Tier-5 ACMG gene variants 271
5.16 Splicing variants in *GPX* genes 272
5.17 *De novo* splicing variants 273
5.18 FATHMM-XF non-coding variants 273
5.19 LUMPY homozygous alternate variants 274
5.20 LUMPY homozygous alternate varaint gene descriptions 275
5.21 LUMPY *de-novo* variants 276
5.22 CNVkit SD003 exome sequencing non-neutral variants overlapping genes 278
5.23 CNVkit genome variants overlapping the 15q13.3 micro-deletion 279
5.24 LUMPY variants called in the region of the 15Q13.3 microdeletion . . . 282
5.25 *WDR60* Splicing variant 283

6.1 Types of Fcγ receptor produced 294
6.2 Types of FCGR genes 295
6.3 Haplotypes of *FCGR2B* 298
6.4 *FCGR3B* HNA SNPs 300
6.5 HNA-1 Haplotype alleles and amino acids 310
6.6 Low data FCGR samples excluded from further analysis 315
6.7 Trimmed HaloPlex sample with above 20% reads lost 315
6.8 Summary of merging overlapping read pairs with PEAR 319
6.9 Initial sample mapping summary statistics 320
6.10 HaloPlex samples with mean depth of coverage below 20 321
6.11 Global Sequence alignment of FCGR genes 325
6.12 HaloPlex samples from same individuals summary 328
6.13 Excluded HaloPlex samples 329
6.14 HaloPlex FCGR gene variants 333
6.15 Determination of HNA-1 Haplotype 335
6.16 Summary of haplotypes predicted based upon allele ratios between six homologous sites 336

6.17 HaloPlex and MLPA HNA-1 haplotype predictions differing in copy number calls . . . 337
6.18 Disagreeing HaloPlex and MLPA HNA-1 haplotype predictions . . . 338
6.19 Summary of MLPA comparsion for each of the references tested in CNVkit 340
6.20 CNVs falling over the *FCGR3B* gene . . . 345
6.21 CNRs tabulated with copy number . . . 346
6.22 CNR breakpoint locations estimated from CNV calls . . . 347

8.1 Samples used for generating contamination simulations VerifyBamID results . . . 392
8.2 All 200 contamianation simulations used to train regressions . . . 402
8.3 Contamianation predictions from OLS for germline samples . . . 409
8.4 Unmapped reads per sample . . . 416
8.5 Unmapped reads from all exome samples at kingdom levels . . . 427
8.6 *CABIN1* mean exon coverage . . . 428
8.7 430 candidate genes used in prioritising variants for the SSMD trio . . . 436
8.8 CNVkit WGS variant calls for SD003 . . . 437
8.9 LUMPY-SV large heterozygous variants called in sample SD003 WGS . 437
8.10 HaloPlex capture kit targeted genes and promoters . . . 439
8.11 HaloPlex inital mapping statistics . . . 441
8.12 HaloPlex VerifyBamID results . . . 442
8.13 HaloPlex Pear merging reads . . . 444
8.14 HaloPlex mean coverage per sample . . . 446
8.15 HNA-1 haplotypes predicted from using single gene references . . . 448
8.16 *FCGR2C* haplotype predictions compared with MLPA for the X57Q variant . . . 450
8.17 HaloPlex samples matching known CNRs . . . 453

UNIVERSITY OF Southampton

Research Thesis: Declaration of Authorship

Print name:	Roshan Kumar Sood

Title of thesis:	Bioinformatic Analysis of Human Next Generation Sequencing Data;extracting Additional Information, Optimising Mapping and Variant Calling, and Application in a Rare Disease

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;

2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;

3. Where I have consulted the published work of others, this is always clearly attributed;

4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;

5. I have acknowledged all main sources of help;

6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

7. Either none of this work has been published before submission, or parts of this work have been published as: [please list references below]:

Signature:		Date:	26/05/2019

Acknowledgements

I would like to express my deep gratitude to Dr Jane Gibson, Dr Rob Ewing and Dr Mark Coldwell, my research supervisors. I would particularly like to thank Dr Gibson, for her patient guidance, encouragement and useful critiques of this research work.

I would also like to thank Prof. Sarah Ennis, Dr Enrico Mossotto, Dr James Ashton, Dr. Chantal Hargreaves, Prof. Jonathon Strefford, Dr David Hunt for their help and assistance in collaborations throughout my studies. I am also grateful to all the patients and families who provided samples for analysis.

I would like to thank the University of Southampton for funding my studies with the award of a Mayflower studentship and the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

List of abbreviations

Abbreviation	Full
1KG project	Thousand Genomes project
AAF	Alternate Allele Frequency
ACIN1	Apoptotic Chromatin Condensation Inducer In The Nucleus
ACMG	American College of Medical Geneticsand Genomics
ATP	Adenosine TriPhosphate
BAM	Binary Alignment Map
BED	Browser Extensible Data
BGI	Beijing Genomics Institute
BI	Broad Institute
BLAST	Basic Local Alignment Search Tool
bp	Base Pairs
BP(X)	Break Point(X)
BWA	Burrows Wheeler Aligner
BWT	Burrows Wheeler Transform
C. Sakazakii	Cronobacter Sakazakii
CADD	Combined Annotation Dependent Depletion
CCD	Charge-Coupled Device
CD	Cluster of Differentiation
CD	Crohn's Disease
ChIP-seq	Chromatin histone-Immuno-Precipitation sequencing
CNR	Copy Number Region
CNV	Copy Number Variant
CONPAIR	Concordance/Contamination of PAIRred samples
ContEst	Contamination Estimation
Contigs	Coniguous sequences
Dev A	Deviation measure A
Dev B	Deviation measure B
Dev Metric	Deviation Metric
DM	Deviation Metric
DMHR	Deviation Metric X Heterozygote -Homozygote ratio
DMs	Deviation Metrics
DNA	DeoxyriboNucleic Acid
DOC	Depth Of Coverage
ENCODE	ENCyclopedia Of DNA Elements
EXaC	EXome aggregation Consortium
FATHMM	Functional Analysis through Hidden Markov Models
FATHMM-MKL	Functional Analysis through Hidden Markov Models Multiple Kernel Learning
FATHMM-XF	Functional Analysis through Hidden Markov Models eXtended Features
FC	rystalline fragment
FCGR	Fc gamma receptor
FFPE	Formalin Fixed Paraffin Embedded
GATK	Genome Analysis ToolKit
GATK DOC	GATK Depth of Coverage
GATK HC	GATK Haplotype Caller
GEMINI	GEnome MINIng
GPX4	Glutathione Peroxidase 4
GRSF1	G-Rich RNA Sequence Binding Factor 1
GVCF	Genome Variant Call File
GWAS's	Genome Wide Association Studies
HapMap	HaplotypeMap
Het:Hom ratio	Heterozygote to the Homozygote variant ratio
HGMD	Human Gene Muatation Database
HGP	Human Genome Project
HNA	Human Neutrophil Antigen
HR	Homologous Recombination
HRC	Haplotype Reference Consortium
IBD	Inflammatory Bowel Disease
IBDU	Inflammatory Bowel Disease Unclassified
IgG	Immunoglobulin G

Table 1 List of abbreviations continued from previous page

IMG	Integrated Microbial Genomes
INDELS	INsertsions and DELetionS
JGI	Joint Genome Institute
LASSO	Least Absolute Shrinkage and Selection Operator
LCRs	Low Copy Repeats
LD	Linkage Disequilibrium
Mb	Megabase
mRNA	messenger RNA
MSP	Maximal Segment Pair
NAHR	Non-Allelic Homologous Recombination
NCBI	National Center for Biotechnology Information
ncRNA	noncoding RNA
NFE	Non-Finnish European
NGS	Next Generation DNA Sequencing
NHGRI	National Human Genome Research Institute
NHS	National Health Service
NIH	National Institutes of Health
OLS	Ordinary Least Squares
pc	Percentage
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
phiX174	Enterobacteria phage phiX174 sensu lato
pre-mRNA	precursor messenger RNA
RNA	RiboNucleic acid
RBF	Radial Basis Function
rRNA	Ribosomal RNA
SAM	Sequence Alignment Map format
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SOTON	Southampton
SQL	Structured Query Language
SRTPs	Short Rib Polydactyly Syndromes
SSMD	Sedaghatian-type SpondyloMetaphyseal Dysplasia
SV	Structural Variant
SVM	Support Vector Machine
SVR	Support Vector Regression
TIRF	Total Internal Reflection Fluorescence
TLRs	Toll-like receptors
UC	Ulcerative colitis
UCSC	University of California, Santa Cruz
USA	United States of America
UTR	UnTranslated Region
VAR	Variant annotation format
VCF	Variant Call File
VEP	Variant Effect Predictor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WTCHG	Wellcome Trust Centre for Human Genetics
WTSI	Welcome Trust Sanger Institute
WUGSC	Washington University Genome Sequencing Center
ZMWs	Zero Mode Waveguides

Chapter 1

Introduction

1.1 Next generation sequencing

1.1.1 First generation

Over the last decade the introduction of Next Generation DNA Sequencing (NGS) methods have made it possible to quickly and affordably sequence the whole genome (WGS) of an organism or only the coding regions known as whole exome sequencing (WES). WES totals approximately two to three percent of a genome, covering around 51 MB sequences, captured by a hybridisation step using probes designed to capture gene exons[1]. In 1977 Fredrick Sanger devised first generation sequencing, known as Sanger sequencing[2]. Sanger's Method was improved throughout the Human Genome Project (HGP), yet the project took annual funding of $200 million for 15 years to compose a high quality sequence for the Human genome[3]. To improve the time, cost and accessibility of DNA sequencing the NGS methods were devised. NGS methods are currently comprised of second generation methods from 2005 onwards and third generation methods introduced from 2010[4]. The benefits of these technologies have resulted in the fall in the cost of sequencing since mid 2000's, illustrated by the NIH in the USA shown in Figure 1.1.

Sanger sequencing was performed by fragmenting DNA and separating into single strands. To each of the single strands nucleotides are added until a labelled, terminating dideoxyribonucleoside triphosphate is added. Termination is random as the dideoxyribonucleoside triphosphates are of lower concentration than the

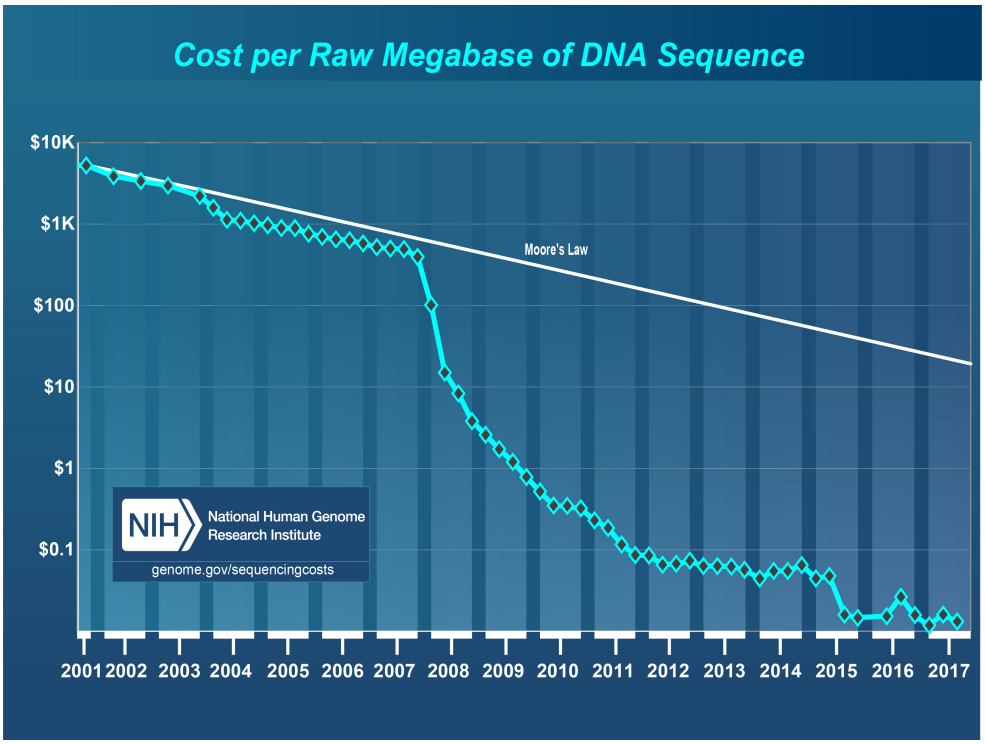

Figure 1.1: Decreases in sequencing costs per megabase from 2001, the graph shows a steady decrease from 2001 to 2007 in accordance with Moore's Law. A large drop in the cost per megabase in 2008 was caused by the commercial introduction of second generation sequencing technology. Estimated costs per megabase were also changed by the NIH from Sanger based capillary electrophoresis methods to second generation technologies. Subsequent improvements with the sequencing technologies further reduced the cost of sequencing until the costs reductions being to plateau from around 2012 with a further reduction in 2015 and 2017. Figure copied from - https://www.genome.gov/sequencingcostsdata/

unlabelled deoxyribonucleoside triphosphates, yielding fragments of varying size. Fragments can then be size separated using electrophoresis, the fluorescence from labels are also recorded to obtain the fragment sequence. This method formed the basis of the HGP which by completion increased the amount of DNA that could be sequenced from one of the project centres from 1000 base pairs per day in the 1980's to 1000 base pairs per second. Part of the improvement due to the use of capillary electrophoresis which added some parallelisation to the method compared to the initial gel based method. As of the end of the HGP the milestone of 1,400 MB per year was achieved[3]. However this technology was far from able to return a sequence for a human genome in single run. To sequence a genome newer methods are now used, though Sanger sequencing has remained useful for the sequencing of smaller sequences, owing to the high accuracy at 99.999%[5].

1.1.2 Second generation

Second generation sequencing methods introduced from 2005 have become the choice methods for WGS or WES due to their massively parallel design. Common amongst second generation systems is the requirement for a DNA sample to be split into small fragments as with Sanger sequencing. To the ends of these fragments universal adapters are attached, though adapters differ in the sequences between the systems. Following adapter attachment the systems also differ in terms of the chemistry used for amplification and sequencing[6,7].

Illumina has become the choice method for NGS due to the cost and time to return results. The original method released under the Solexa name used a similar library preparation as the 454 pyrosequencing. Fragments are then bound to the surface of a flow cell using the adapters at either end. On the surface of flow cells there are complementary sequences, attached via a flexible linker to the flow cell surface as shown in Figure 1.2 by blue and green circles. The surface sequences are complementary to either type of universal adapters attached to the 5' or 3' end of the DNA reads. Therefore both forward and reverse read orientations are allowed to hybridise to the flow cell. Following hybridisation the process of bridge amplification begins where the fragments of DNA are amplified for the sequencing stages. DNA polymerase is used to form a complementary strand using the oligonucleotide sequence on the flow cell surface as the primer for the extension. The original strand is then denatured so only the copy remains attached. The new fragment copy bends to form an arc or bridge shape, this then allows the adapter at the non-attached end of the DNA fragment to hybridise to the complementary sequence on the flow cell. DNA polymerase then acts on the newly hybridised oligonucleotide sequence from the flow cell to form a double stranded bridge. A further denaturation yields two single strands, a forward and reverse of the same strand. The result of this process is around 1,000 copies of a DNA fragment all located spatially close to the original fragment, the copies and original fragment appear as a cluster on the flow cell. Due to the large number of fragments likely generated from the original sample of DNA there will be millions of clusters on the flow cell[6,8,9].

The Illumina method allows for paired end sequencing, when the forward and reverse

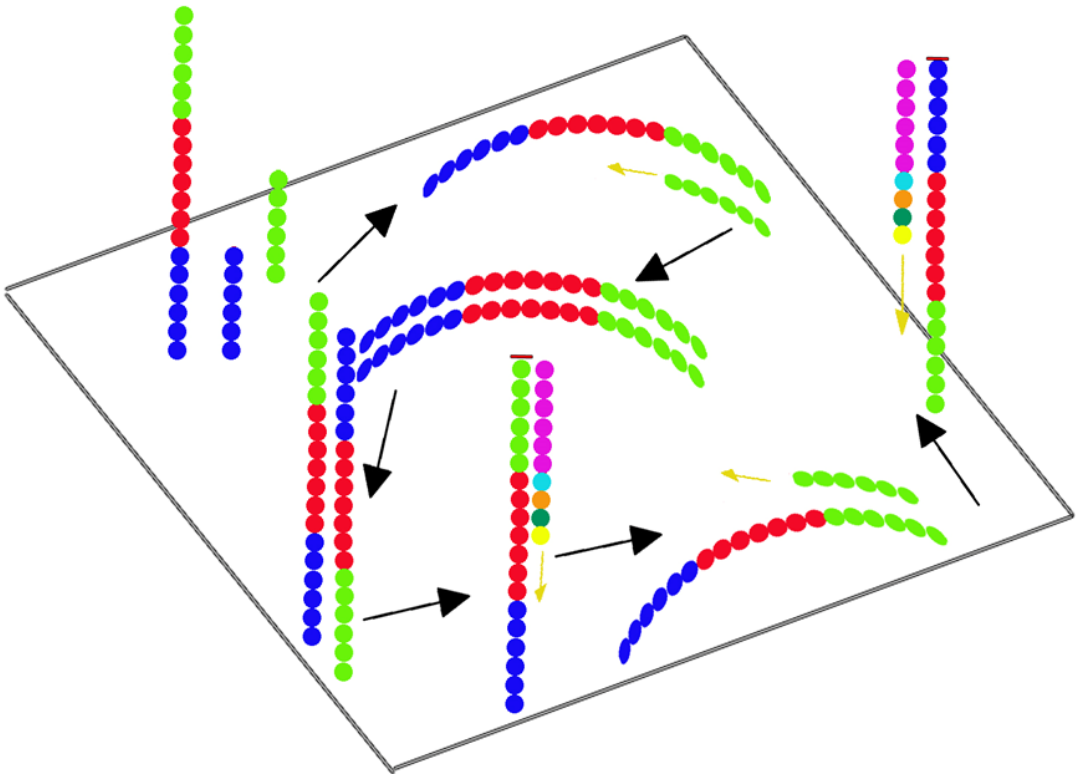

Figure 1.2: Summary of the Illumina sequencing method. The DNA adapter sequence on the fragment therefore hybridises with the complimentary primer (Blue). A DNA polymerase will then extend the complementary strand to form a double stranded molecule of DNA which forms a bridge. One strand is then denatured, the remaining strand is blocked at the 3' end and a sequencing primer added. Fluorescently labelled nucleotides are then added which have the 3' OH group inactivated to allow only one base to be added and the colour detected. The process is then repeated to generate the other strand and then sequence the other strand.

strands orientations of the same original fragment are sequenced. Initially the reverse strands are cleaved from the adapter, attaching them to the flow cell and allowing them to be washed off the flow cell. A chemical block is then placed on the 3' end of the forward strands to prevent any further priming as shown by purple bars in Figure 1.2[9]. A sequencing primer is added which binds to a section of the 3' adapter on the forward strand, shown in pink in Figure 1.2. Addition of nucleotides with chemically inactivated 3'OH groups then occurs. Each nucleotide species has one of four unique colour labels attached which is detected using Total Internal Reflection Fluorescence (TIRF)[6]. TIRF uses two lasers to excite the colour label attached to nucleotide. A charge-coupled device (CCD) then records the coordinates, colour and intensity of the light generated. By using coordinate data a program receiving the CCD results can

then assign the nucleotide type, determined by the light colour, to the correct cluster and hence build up a sequence for the read[6]. To avoid the noise from unbound nucleotides they are washed away prior to imaging. Additionally to avoid the light from previous nucleotide incorporations the fluorophore is removed from the newly incorporated nucleotide in addition to removal of the 3' blocking species to allow additional nucleotide incorporations[6]. The process is then repeated for between 100 and 250 cycles depending upon the length of read specified. Initially Illumina was limited to around 30bp and the method suffered more from the noise in the imaging process causing qualities to drop at the ends of read[6,9]. This method is massively parallel as all of the clusters on a flow cell are visualised per cycle using cyclic reversible termination. Also depending on the instrument used multiple flow cells can be used per run of an instrument. Once the fragment has been sequenced for the forward strand it is denatured and washed away and index primers can be bound to the remaining fragments and new primers generated by polymerase in addition to the removal of the 3' protection allowing the fragment to bend again. With the bridge already formed and the protection removed the reverse strand is now formed, separation follows to generate single strands for the forward and reverse stands. With both forward and reverse strands present on the flow cells the 3' ends are blocked and this time the forward strands removed. Now with the reverse strands the process used for the forward strands can be repeated.

454 pyrosequencing was discontinued as of 2013 due to the cost being uncompetitive with other second generation methods such as Illumina. The method used 28 μm diameter agarose beads to both capture and bind target DNA fragments. On the surface of beads were oligomers which were complementary to the adapters already attached to the target DNA fragments facilitating capture and also acting as primer. To remove all unbound, non-target DNA a denaturing wash is performed. Amplification of bound fragments was performed using an emulsion PCR in which an oil and aqueous mixture were vigorously mixed to isolate each individual agarose bead. Each of the agarose beads were contained in a micelle as a result of the mixing and contained the necessary reagents for PCR to amplify the unique DNA fragment bound to each bead after the transfer to a microtitre plate. Amplification typically proceeds until 1,000,000 copies of the fragment are bound on the bead surface[5,9,10].

Finally beads were transferred to a 454 picotitre plate where sequencing was actually be performed. The actual pyrosequencing process uses two further types beads: a smaller one μm diameter magnetic bead and a latex bead both of which have enzymes Adenosine Triphosphate (ATP) sulfurylase and luciferase attached respectively. The picotitre plate must be used for sequencing as this allowed sequencing reagents to flow through, on one side of the cell nucleotides are supplied whilst above is a CCD. Each of the four nucleotides were sequentially added to the flow cell, when a complementary base was bound and incorporated by a DNA polymerase enzyme a pyrophosphate molecule was released. The sequencing reagent adenosine 5'-phosphosulphate was then able to create an ATP molecule using the ATP sulfurylase enzyme with the pyrophosphate molecule. ATP was then able to react with the a luciferin reagent and produce oxyluciferin and light which was detected by the CCD. The sequencer also used the position and nucleotide cycle to build the sequence for the reads.[5,9,10]. After a nucleotide was imaged the cell was washed using apyrase before repeating the cycle to determine the full fragment sequence. Crucially though this method suffered with homopolymer repeats as multiple of the same nucleotides can be incorporated in a single cycle. Hence the intensity of light is the only measure for the number of nucleotides added which in practice made the method more error-prone[5,9,10].

Applied Biosystems SOLiD (Sequencing by Oligonucleotide Ligation and Detection) method was discontinued in 2016. The method used an emulsion PCR method as with pyrosequencing to amplify DNA fragments bound to the surface of the beads. Beads were $1\mu m$ diameter magnetic beads which following amplification were attached to a glass slide using covalent bonds which enabled sequencing via loading into a fluidics cassette. To perform sequencing primers were first annealed to the adapters and attached to DNA fragments followed by the addition of 8-mer fragments. The correct 8-mer fragments were added by a DNA ligase enzyme. All of the 8-mers have fixed positions at bases five or in later cycles at position two. At each of the fixed positions a fluorescence molecule was attached, the colour of fluorescence being dependent upon bases at positions four and five. After imaging the fluorescence the bonds between base five and six were cleaved, removing bases 6-8 from the 8-mer. The cycle would then be repeated so every five bases are sequenced, once the end of the fragment was reached the process repeated using a new primer with an offset of -1

base so bases 4, 9, 14, 19, 24... until the end of the fragments. The process of changing the offset is performed until the entire fragment would have been covered. As the fluorescence was dependent upon two bases the number of colours required are 4^2 or 16 which when using the primer offset means each base effectively will be sequenced twice and by combining fluorescence colours the base can be decoded. SOLiD therefore can potentially identify miscalls of nucleotides when comparing the colours. Traditionally the limitations of SOLiD were short read lengths, in 2008 a maximum size of 35bp, by 2012 the length had only reached 85bp and the time for sequencing could be upwards of a week. Due to the early limitations of SOLiDs such as short read lengths and time constraints Illumina was used by most labs and have remained using Illumina due to familiarity but also the longer reads, reduced time and costs of sequencing[5,9,10].

Whilst the cost of sequencing rapidly fell following the introduction of second generation methods some systems still cost significantly more than others. As demonstrated by the HGP Sanger sequencing is far too expensive to use with large scale sequencing projects. The costs per genome as estimated by the NIH as shown in Figure 1.3. The estimation as with Figure 1.1 takes into account factors such as: labour, administration, management, utilities, reagents, and consumables. The cost per genome was massive at close to $100 million in 2001 but with the improvements in sequencing by the end of the Human Genome Project costs had fallen to around $60 million per genome. From the commercial introduction of NGS in 2007 sequencing costs fell far faster than Moore's law down to present levels at close to $1000 per genome. Calculations of the theoretical depth returned for a $1000 genome by the NIH as of 2017 suggests that non-Illumina second generation systems would return of an average 10x depth of coverage compared with 30x average depth returned from an equivalent Illumina sequencing.

Second generation systems typically suffer from short read lengths compared to Sanger sequencing. Read length is of great importance when mapping to a reference genome, particularly for challenging regions which contain repeated sequences. Alternatively for species without a reference genome the reads are assembled by sequence overlaps, known as contigs. Contigs can then be used to assemble the *de*

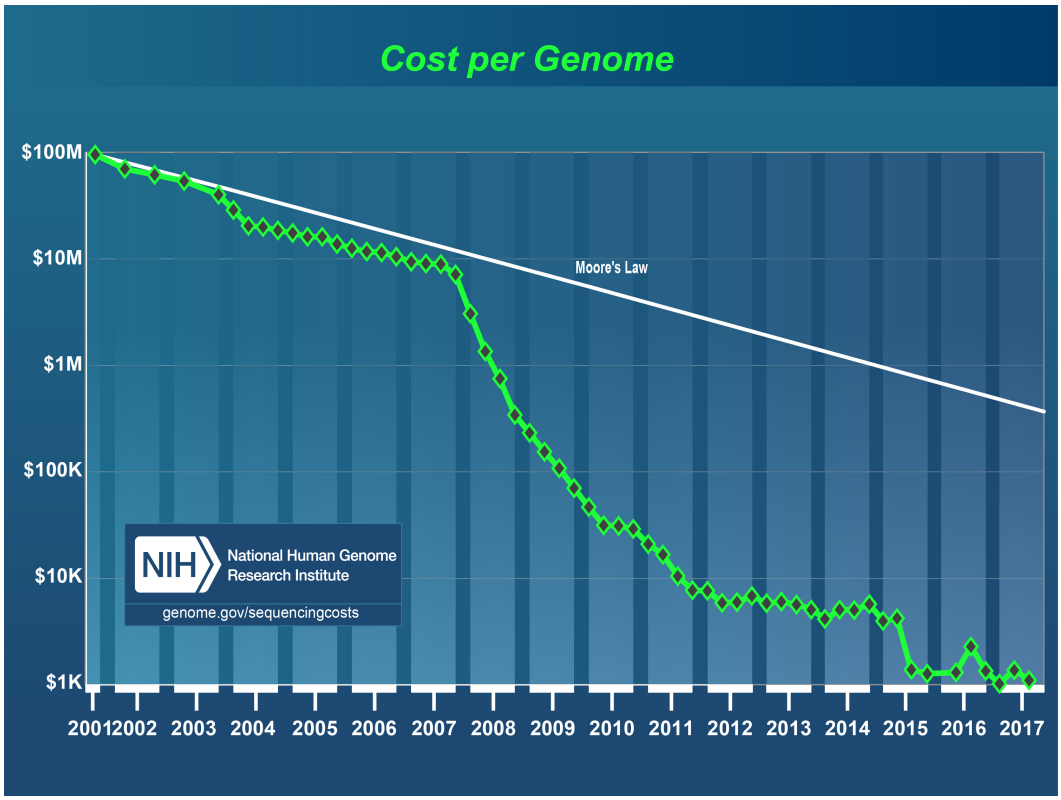

Figure 1.3: Decreases in sequencing costs per genome from 2001, the graph shows a steady decrease from 2001 to 2007 in accordance with Moore's Law. A large drop in the cost per megabase in 2008 was caused by the commercial introduction of second generation sequencing technology and switching of cost measures by the NIH to second generation methods. Subsequent improvements with the sequencing technologies have further reduced the cost of sequencing until cost reductions plateau from 2012. Further reductions in 2015 and 2017 have occurred with the introduction of new Illumina sequencers. Figure copied from - https://www.genome.gov/sequencingcostsdata/

novo sequence. Longer reads reduce the chances of multiple matches when assembling contigs, therefore it should help to reduce mistakes in the assembly process. Illumina sits in the middle of the systems when length is considered at a respectable 300 bp. When the throughput is considered no system can match the data generated from Illumina at maximum 1800 GB per run, the closest being SOLiD at 320 GB followed by 454 at 0.7 GB. To put these figures into some context modern Sanger sequencing can only generate around 1.9-84 kb per run. For large sequencing projects Illumina remains the clear stand-out in terms of the data that can be generated per run[4,5,11,12]. For the Illumina Hi-seq X system a run can be completed within 3 days, over a year of continuous operation generating 1,800 human genomes per machine. Samples can be multiplexed in the Illumina sequencers to make use of the full capacity of the

sequencers allowing multiple smaller samples such as WES to be run[13].

Accuracy is of fundamental importance to a sequencing method. Sanger sequencing has remained a gold standard in terms of accuracy with 99.999%. However a tipping point appears to have been reached with several studies reporting that next generation methods are now more reliably able to detect variants[14]. A comparison of 684 exome (coding exons only) sequenced samples against 2.79 million Sanger reads was performed over 19 genes, some with known pseudogenes, 19 variants were detected by NGS but not all variants were detected in the initial Sanger sequencing and required additional runs to detect all variants[14].

Each of the second generation methods have an accuracy score between 98% and 99.9%, however due to the chemistry of each system they struggle more with specific sequences, so it may not be appropriate to use a general error percentage[4,5]. Particularly, 454 sequencing which has issues with homo-polymer repeats which may mean the accuracy is below the claimed 99.9%. When also considering the relatively high cost and low throughput of 454 perhaps it is little surprise that 454 is being phased out by Roche[4,5].

1.1.3 Third generation

Third generation sequencing aims to solve the challenges and limitations of short read mapping in repetitive regions and accurate structural variant calling by using longer read lengths. An example of a third generation method is Pacific Biosciences method called Single-Molecule Real-Time (SMRT). This system again uses DNA polymerase and fluorescently labelled nucleotides but does not stop the polymerase after nucleotide incorporation. To achieve this the method makes use of a technology called Zero Mode Waveguides (ZMWs). In SMRT sequencing size prevents the light from the excitatory lasers travelling beyond 30 nm from the bottom. Therefore the polymerase molecule is placed at the bottom of the ZMW using a biotin-streptavidin anchor. Once the DNA polymerase begins to synthesise a complementary molecule labelled nucleotides flood the ZMWs at a specific concentration such that the nucleotides diffuse down to the DNA polymerase and then back up[6,7,15]. As the diffusion down and up occurs on a nanosecond time-frame compared to the microseconds for incorporation of a nucleotide it is possible to then differentiate the

fluorescence. Hence the longer intensity of the bound nucleotide fluorescence allows discrimination of the bound nucleotide before removal of the fluorescence group allowing the detector to return back to baseline value for the next nucleotide incorporation. In theory the real time polymerase monitoring should allow reads upwards of 80kb in length, greatly aiding assembly process. However, the cost of SMRT in 2017 remained relatively high compared to Illumina, with the cost per GB of sequencing estimated in one study as $300 compared to $55 for Illumina[16]. The method also remains hampered by low throughput, with a study suggesting that while SMRT produces one GB a comparable Illumina sequencing would have produced 90Gb. A reported high error rate of up to 14% likely to be due to noise issues from the unincorporated nucleotides during the diffusion process also hampers the method. For these reasons Illumina still remains preferable over SMRT sequencing[4,13,16,17].

One of the more promising sequencing methods that has been developed is nanopore sequencing. The fundamental principle of nanopore sequencing being that a whole single stranded DNA molecule is fed through the nanopore causing optical or electrical current changes depending in the nucleotides passing through[4,12,15]. Currently there are four nanopore sequencers available for consumer testing the: SmidgION, MinION, GridION and PromethION from Oxford Nanopore (https://nanoporetech.com/). The SmidgION is an add-on for smartphones which is smaller than the USB-sized MinION but uses the same chemistry as described previously. PromethION is workbench sequencer with many nanopores which operate in parallel. GridION like the PromethION is a workbench system but is scalable so that multiple GridION sequencers can be used together. Currently the cheapest cost of nanopore sequencing in 2018 is $5 per GB when using a PromethION, which is cheaper than Illumina at $12-27 per GB[13]. Throughput is lower for nanopores with up to 20GB per MinION though a PromethION can be scaled with 48 flow cells each creating 125GB to produce 6TB of data in up to 64 hours. Accuracy is low though compared to Illumina with up to 15% error rate, similar to that of SMRT sequencing[13,16]. Read lengths of up to 1Mb have been reportedly produced in some experimental runs which are far higher than any other method have been able to produce. Recently an update to the nanopore system now uses both strands of DNA to double the depth of coverage called 1D[2]. In the system an adapter is used which

inhibits the entry of the second strand until the first strand has passed, reportedly lowering the error rate to 3% though doubling sequencing time and lowering throughput[4,13]. In April 2018 using a MinION for the sample NA12878 generating 91.2 GB of sequence data with a reported accuracy of 95% but after multiple polishing steps upwards of 99% was reported, highlighting the potential improvements in the technology in recent years[13]. However Illumina remains the system of choice for WGS or WES as exemplified by the selection of Illumina as the preferred partner for the NHS project to sequence 100,000 genomes[18]. It is likely in the future that third generation methods will displace Illumina.

1.2 Sequencing projects

Due to the initial throughput limitations of Sanger sequencing the first large sequencing project was the HGP, which aimed simply to provide a high quality reference for the human genome. As sequencing technologies evolved Genome Wide Association Studies (GWAS's) became a favoured approach due to the ability to capture thousands of SNPs per sample on a single array at a much reduced cost. Most crucially though the commercial availability of second generation sequencing technologies has enabled studies to use 1000's of WES or WGS samples. This empowered the ability to detect differences between populations and also to identify rare and disease causing variants not just in coding but non-coding regions.

1.2.1 Human Genome Project

The HGP was the first large scale genomics project. The project started in 1988 with the aim of compiling a complete sequence for the human genome to improve our understanding of the genome. In 2003 the project was completed having obtained a majority sequence for the human genome[3]. This effort took 20 centres over 15 years with an estimated cost of $1.3 billion. As a result of the project there was clearly the need to improve the time, cost and resources needed to conduct genomic sequencing. By the completion of the project some of these issues had been addressed or partially solved but not to a point where genomic sequencing was feasible to the average lab[3].

1.2.2 dbSNP

The commercialisation of first and later second generation sequencing methods meant the amount of DNA that could be sequenced rapidly rose whilst the cost and time rapidly fell. As a result of these improvements it has become practical to implement large scale sequencing projects to try and elucidate some of the unknowns about the genome and in the process develop a better understand the genetic contributions to a number of diseases. There was a concerted effort to centralise variant data now generated into databases such as dbSNP[19]. This database is a store for detected variants recording information such as the alleles at the locus, flanking sequence of a variant and some experimental details regarding the methods used to generate the data. Each entry recorded into the database is then accessible to the general research community.

1.2.3 HapMap

One of the first large sequencing projects was HapMap which began towards the end of the Human genome project in 2003[20]. The intention of the project was to use genomic data to create a database of common variants seen in the human genome. Using the patterns of SNPs haplotypes were created. A haplotype is the combination of the alleles at a particular region on a chromosome, alleles that are located closer together are more likely to be inherited together. The measure Linkage Disequilibrium (LD) is used to describe the correlation between alleles that are inherited, high LD indicates the alleles are likely to be inherited together. LD is generally high between neighbouring SNPs as recombination tends to occur at specific hotspots. Using the information gained on patterns of alleles being inherited together only a few SNPs are required to test a whole region, these SNPs are termed 'tag' SNPs.[21].

1.2.4 ENCODE project

In 2003 the Encyclopaedia of DNA Elements (ENCODE) project was initiated by the National Human Genome Research Institute (NHGRI) with the aim of uncovering all functional elements encoded in the human genome including: exons, sites of RNA processing and transcriptional regulatory elements such as promoters, enhancers, silencers and insulators[22]. By 2007 the pilot project was completed for a pre-defined 1% of the human genome allowing the project to expand and study the entire genome using the methods developed for the 1%. To annotate functional elements the project

focused on five areas: gene annotation, transcription analysis, chromatin analysis, transcription factor binding and other analyses including methylation.

In 2011 the ENCODE consortium released data describing regions of the genome which are more functionally active using a variety of experimental techniques[22]. Techniques included RNA-seq to identify new transcripts and their relative expressions. Transcript data reveals the possible novel transcripts for each of the genes including novel start and end sites, exons and splice sites aiding annotation of variants. DNA is stored in a compact form by wrapping around proteins called histones, known as chromatin. In particular eight histone proteins form a complex known as the histone octamer consisting of two copies of histone proteins: H2A, H2B, H3 and H4. Chromatin structure can be either closed or in open confirmation caused by modifications to histone proteins. Acetylation of histones is associated with an open conformation allowing the binding of transcription factors, conversely deacetylation leads to a closed conformation. When chromatin is in the open conformation the DNAse-I enzyme is able to bind DNA and cleave specific sequences on a single strand. To measure how accessible DNA is for transcription factor binding techniques such as DNAse-I hypersensitivity assays were used measuring the cleaved fragments using biotinylated adapters or direct sequencing of cleavage sites at the ends of DNAase fragments. FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) separates open and closed DNA using cross-linking, with the open conformation DNA unbound and able to be extracted using a phenol-chloroform based method for NGS sequencing and identification of open regions. CHIP-SEQ (Chromatin-histone immunoprecipitation) works in a similar method to FAIRE but using cross-linked chromatin which is precipitated and the DNA recovered for sequencing and identification of transcription factor binding sites. Chromatin-chromatin interactions were also investigated to capture long range interactions using 5C, a modified version of chromatin conformation capture (3C) where chromatin crosslinks are formaldehyde fixed before digestion by restriction endonucleases. Ligation is then allowed to occur between fragments to identify interacting fragments. Finally methylation in the genome was analysed using bisulfite sequencing to detect methylation at position 5 of the cytosine in CpG dinucleotides which have associations with transcriptional silencing and also imprinting[22,23].

Using experimental data from the ENCODE project the GENCODE project developed methods to identify all evidence-based gene features including genes, coding sequencing and transcripts. Annotations will include the alternate transcripts for genes[24]. GENCODE uses a combination of the ENCODE experimental datasets with *in silico* algorithms, additional experiments to confirm transcripts with manual curation. Experimental approaches are rapid amplification of cDNA ends, real-time PCR or using antibodies to target specific peptides[24,25].

ENCODE remains an ongoing project and is currently in phase four (https://www.genome.gov/26525220/) with areas still under active research. Consequently as new experimental data or refined models and algorithms are produced the number of transcripts, exons and regulatory features contained in the GENCODE database has increased. In 2011 when initial ENCODE results were released the GENCODE database was version seven containing 51,082 genes and 161,375 compared to 58,381 genes and 203,835 transcripts as of GENCODE v28.

1.2.5 1,000 Genomes

Towards the completion of the HapMap project the introduction of second generation sequencing revolutionised the study of genomics. It became feasible in terms of cost and time perform WES or WGS for a sample. Using second generation sequencing the inherent biases of genotyping towards more common variants are avoided. The first large study to utilise NGS was the 1,000 genomes project (1KG)[26]. WGS of 1,092 individuals was performed to try and better understand the natural variation between individuals. At completion the project generated and validated haplotype maps of 38 million SNPs, 1.4 million INDELs and over 14 thousand larger deletions[27].

Analysis of the data has shown that there are differences between the populations analysed in the project. A key difference identified were the frequencies for rare and common alleles. In particular low frequency variants differ considerably by geographic location. With the amounts of data gathered increasing per study there needed to be improvements with the methods of analysing data. To address the data analysis issues part of the brief for the first project phase was the development of new

analytical methods to improve the detection and genotyping of variants. One final aspect also developed was the modelling of genotype likelihoods and performing statistical haplotype integration[27,28]. Subsequent phases of data have been released from the project, the first phase data was low coverage data of the 1,092 samples. For the second phase the sample size was increased to around 1,700 samples with high coverage. For the final phase 2,500 samples were analysed including samples from South Asia and new African samples. The data gathered in phase two allowed building on the phase one data to improve existing methods while developing new analytical methods to process multi-allelic variant sites, structural variants and more complicated variation scenarios. The data from the project is open to access for the wider scientific community for further studies and analysis, phase three data has been made available from mid-2014 using the projects website (http://www.1000genomes.org). Many studies which used imputation based approaches also made use of improved reference panels created from 1,000 genomes project data allowing more accurate genotype imputation and for more SNPs.

1.2.6 Other sequencing projects

Following the 1000 genomes project a larger study called the UK10K project was performed encompassing 10,000 samples[29]. The UK10K project ran from 2010 to 2013 and focused on numerous disorders, of the 10,000 total patients there were 5,500 patients described as having: obesity, autism, schizophrenia, familial hypercholesterolemia, thyroid disorders, learning disabilities, ciliopathies, congenital heart disease, neuromuscular disorders or rare disorders including severe insulin resistance. These 5,500 samples all underwent WES at relatively high depths (average depth of coverage x72). A second cohort of 4,000 phenotyped samples were also added which were whole genome sequenced at low depth (average depth of coverage x6) to allow imputation. From the data which is accessible to approved groups there have been a number of papers published which have identified variants implicated in the phenotypes associated with the data as highlighted on the publications web page of the UK10K (http://www.uk10k.org/publications.html).

From both the 1000 genomes project and the UK10K there have been many improvements in methods used to identify variants and the analysis of these variants.

The application of the new technologies to the study of disease shows much promise as shown by the publications from the UK10K. The Haplotype Reference Consortium (HRC) sought to integrate the sequencing data from the multiple projects to create an improved reference panel of human haplotypes. The improved haplotype map then allowed for increased accuracy of genotype imputation, particularly for low-frequency variants, but also the total number of imputable variants. The initial release from the project contained 64,976 haplotypes with 39,235,157 SNPs[30].

Subsequently a project in Iceland sequenced the genomes of 2,636 individuals in combination with genotyping of a further 104,220 individuals[31]. By integrating information from WGS with genotyped individuals with genetic records the study was able to impute information on nearly 90% of the Icelandic population. Iceland is a founder population and has a statistically low migration rate making the original genome of the founders has been largely preserved[31]. The Icelandic genome was then able to be compared against other regions by comparing the frequencies of both common and rare variants. By comparing the genomes each individuals genetic predispositions for certain diseases can be identified. Most controversially the study claimed that from the data collected they could identify 2000 individuals with a mutation in the gene *BRCA2*, predisposing these people to a 4.6-fold increase in the lifetime risk of developing any cancer. From the 2000 individuals statistically it is likely that least 724 women will develop ovarian or breast cancer. 360 men are also predicted to develop prostate cancer[31].

1.2.7 100,000 Genomes project

As of December 2018 one of the largest genome sequencing projects was completed in England. 100,000 WGS samples were collected from 70,000 patients. Each of the selected patients were diagnosed with rare diseases or forms of cancer for which little is known regarding the genetic contribution or were relatives to provide controls for an affected patient[18]. Rare disease categories investigated included: musculoskeletal, immune, respiratory, cardiovascular disease, endocrine and metabolism, hearing and sight to name a few categories. By targeting rarer diseases which are unlikely to have been included in previous large scale sequencing projects there is hope that in the coming years new targets for treatments can be identified to improve patient

outcomes. Studying the mechanisms of disease development and progression for rarer diseases may produce findings which are also applicable to more common diseases, improving our understanding and treatment of more common diseases.

It is hoped that the genomic results from the project will lead to more common use of WGS within the NHS and integration with patient records potentially leading into the era of genomic and personalised medicine. To this end it was announced in October 2018 that there would be seven regional genomic laboratory hubs which would organise resources to provide a national genetic testing service for the NHS. Genomic sequencing is becoming more mainstream within the NHS and also from companies offering testing publicly. Based on results diseases can diagnosed and implicated genes identified, from which treatments may be designed. Alternatively results could be used to inform potential parents as to their status as carriers for diseases such as cystic fibrosis. In these and similar cases using genome sequencing accuracy of results is of paramount importance. To ensure the accuracy of diagnoses effective quality control of samples must be performed. Quality control steps will need to assess important aspects such as: sequencing quality scores, depth of coverage for variants or genes of interest and test for potential contamination of samples. In the study of rare diseases and cancers WGS will call large numbers of variants and so require effective methods to prioritise and filter variants both coding and non-coding to identify the most important variants for causing the disease. From filtered variants interpretation of the variants are also necessary to understand the effects of the variants on the gene and or protein function which can then inform treatments or advice to patients.

1.3 Software pipelines

Second and third generation sequencing provides a greater bioinformatic challenge to process than the first generation due to the high volume of data. Files generated from second generation sequencing are gigabytes in size, several orders of size larger than first generation sequencing. To process the larger files it is important to develop software pipelines which can generate accurate, reliable and repeatable results. NGS has several stages which need to be addressed in any pipeline, these stages are: pre-processing, read mapping or assembly, post-alignment processing, variant calling

and annotation of variants. Many software packages have been developed to perform each of these stages, below is the discussion of some of the more commonly used programs.

In Illumina second generation sequencing samples are often sequenced using paired end sequencing, the information gained from sequencing the reverse orientation greatly aids in the mapping of repetitive regions. Paired end sequencing uses the information from two files in the generation of a single Sequence Alignment Map (SAM) file. Both of these raw sequence fastq files need pre-processing in order to identify any issues in the data. Common examples of such problems are when adapter sequences used in the sequencing process themselves become incorporated into the read or when a significant proportion of whole or positions in reads have low quality scores. Whilst most issues with samples being sequenced are likely to be identified from a report generated by the sequencer it is good practice to check the data before mapping.

1.3.1 Pre-processing

Illumina NGS data in the fastq format is comprised of a repeating four line format: sequence identifier line beginning with @, nucleotide sequence, quality score identifier line consisting of a "+" character and on the final line are ASCII character for each of the nucleotides which describes the PHRED quality of the base. Historically quality encoding has varied between +33 and +64 which will affect the range of PHRED values possible, currently +33 is used from Illumina v1.8. The PHRED quality scores are used to denote the probability that a base has been correctly called. The scores are logarithmic, for example a score of 10 would indicate the base call was 90% probability the base was called correctly, a score of 20 would indicate 99% probability and a score 30 would be 99.9% certain. Therefore a cut-off used in most NGS studies is a minimum PHRED of 20 which for a 100 base read would statistically only contain 1 error. With Illumina sequencing the quality or composition of the first 6 bases can be more variable due to biases caused by the use of the random hexamer primers. Illumina sequencing chemistry also degrades towards the end of reads and consequently PHRED base qualities can be observed to tail off and require trimming to avoid using the poor quality bases in alignments and variant calling.

FastQC is a fastq analysis program which returns a report detailing basic statistics such as: total sequences, file encoding and sequence lengths in addition to more complicated measures of normality[32]. Results from FastQC report on 12 different measures of data quality. Often the most important modules are the: basic statistics, per base sequence quality, per base N content and overrepresented sequences. One of the most important measures is the per base PHRED sequence quality allowing users to visualise the range of quality scores for each position as a boxplot. If the PHRED sequence scores are low (typically below 20) for a significant proportion of the data it may be necessary to filter out the low quality sequences. Also data may require filtering if the program identifies there are number of 'N' calls, where a sequencer was unable to call a base at a position. FastQC uses sampling of the first 100,000 different sequences in the file. Sequences are then matched using a minimal length of 20 bp, if the percentage of matches is greater than 0.1% of the total reads the sequence is described as over-represented. Using the sampled data the program also identifies overrepresented Kmers (5-mers). By counting the enrichment of each k-mer FastQC compares the count against theoretically expected values, this helps to identify overrepresented sequences which would otherwise be missed in long sequences with low quality and partial sequences. The per base N content module reports if there are read positions where an abnormally high number of sequence calls are missing and 'N' calls are used instead.

PRINSEQ is a commonly used quality control program which like FastQC performs checks for artefacts such as adapter sequences, sequence duplication and using a di-nucleotide odds ratio checks the relative abundance of sequences. PRINSEQ has the option to trim sequences to remove 5' and 3' ends and also then filter the sequences on parameters including: length, quality scores, G-C content, ambiguous 'N' calls, sequence duplicates[33]. There are many alternative programs available such as the adapter removal program Cutadapt[34]. The fastx toolkit is a set of tools which are focused at the pre-processing of both fastq and fasta data[35]. The web-based Galaxy suite uses fastq groomer to verify then format fastq data with a read trimmer which applies a quality filter[36]. Currently with the improvements of tools it has been suggested pre-processing steps such as trimming and removal of adapters may not be necessary as tools such as Genome Analysis ToolKit (GATK) can account for regions of poor

quality or adapters[37,38].

1.3.2 Read mapping or assembly

Fastq data needs to be aligned to a reference sequence where available or assembled *de novo* for organisms without a reference sequence. There are a variety of aligners which can perform the mapping including: Novoalign[39], bowtie2[40], velvet[41] and BWA[42,43]. Each of these programs use different algorithms to map or assemble the sequences and will likely give slightly differing results, however all will result in the output of a SAM file.

Novoalign (http://www.novocraft.com/products/novoalign/) has been one of the most popular aligners used with NGS data due in part to the reported high accuracies and total of reads mapped[43]. Novoalign operates by creating hash tables of both the reads and the reference genome where the keys for entries are a subsequence while values are a lists of positions where the subsequence can be found. Novoalign then sub-divides the hash table for reads into overlapping oligomers and tries mapping them to the reference genome using the Needleman-Wunsch algorithm with defined gap penalties to provide the optimum alignment of reads[44].

Bowtie2 uses a text index created from the reference genome, to which it tries to match a read to by mutating the read string within allowed parameters and identify the best match. Bowtie2 allows for gapped alignment, unlike the original Bowtie, by performing an un-gapped alignment followed by a gapped extension step. Both BWA and also bowtie2 use a Burrows-Wheeler Transform (BWT), a method of data compression where strings are rotated by changing the end of line and the string rearranged into alphabetical order and the last letter stored for each rotation, from the compressed BWT file a backward search allows exact and inexact read matching to the reference[42].

BWA is the recommended alignment program for the GATK and consequently has become one of the most popular alignment programs. BWA can be run using the Smith-Waterman mapping algorithm for reads of sizes below 70bp or for larger reads up to 1MB in length BWA-mem is recommended. BWA-mem also uses the BWT

transform and a combination of local and end-to-end alignments to map longer reads. The aligner also allows for more gapped alignments and better capture larger events such as structural variants. Comparisons of BWA-mem against other aligners yielded the second highest percentage of mapped reads and accuracy, beaten only by novoalign[43]. Whilst BWA is an open source program and free to use novoalign requires a licence to be purchased and given the similar performance of BWA many groups choose to use it over novoalign.

1.3.3 Human reference sequence

In 2009 the genome reference consortium released version 37 of the human genome which resolved a reported 255 mapping issues, this release was updated and maintained until June 2013. The current version of the human genome, GRCh38 (hg38), released in December 2013 and added support for: alternate sequences in 261 regions of high variability, updated mitochondrial sequences and reduction in sequence gaps particularly around centromeres[45]. The increase in the covered regions and better capture of the variation between populations via alternative contigs and sequences improved hg38. Annotation and other data sources are increasingly being made available for hg38. A brief comparison between the final version of hg19 (patch 13) and the latest version of hg38 (patch 12) is shown in Table 1.1.

Assembly	Released	Units	Total(bp)	Non N (bp)	Primary N50 (bp)	Regions	Alternate loci
GRCh37.p13	28/06/2013	12	3,234,834,689	2,991,688,216	46,395,641	182	9
GRCh38.p12	21/12/2017	38	3,257,347,282	3,095,978,931	67,794,873	317	261

Table 1.1: GRCH genome comparisons for GRCh37.p13 and the current hg38 release as of March 2018 GRCh38.p12 showing the improvements in capturing alternate loci by using hg38. Genome assembly statistics are available from the genome reference consortium website: https://www.ncbi.nlm.nih.gov/grc/human/data.

The hg38 human reference genome covers 22.5 MB more than hg19 and reduces the percentage of N-bases (No call positions) from 7.5% down to 4.95%. Primary assembly N50 is approximately 46% increased making the process of assembling scaffolds into the final genome more likely to be more accurate.

1.3.4 Post-mapping

Post-alignment processing begins with SAM files which are large in size, typically around 10's of gigabytes for WES, but several hundred for WGS. To make these files smaller and more manageable for other analyses the SAM file is converted to a Binary Alignment Map (BAM) file with an accompanying index file. In the generation of BAM files they are usually also sorted based on the genomic co-ordinates. One of the most commonly used the program is SAMtools[46]. SAMtools also gives options to remove what it detects as PCR duplicates which have arisen from the sequencing stages and can later lead to erroneous variant calls due to the same read fragment being duplicated and therefore disproportionately represented.

GATK is a suite of tools for DNA and RNA-seq analysis which has gained in popularity due to the reported performance improvements relative to SAMtools for variant calling and also due to its scalability[47,48]. GATK also recommends using Picard-tools to convert the SAM file into a BAM file which can then be indexed, sorted and PCR duplicates marked before variant calling[49]. Having created a BAM file with duplicated reads marked GATK first recommends to perform base quality score recalibration on the BAM file. It has been shown that base qualities as generated from Illumina sequencing machines are subject to systematic technical errors leading to over and under estimation of scores[50]. To correct for the errors GATK uses a base recalibration tool which applies a machine learning model to determine the bias in scores. Using a known set of dbSNP variants a model of co-variation is built and compared against the site qualities from the BAM files. It is assumed that all reference mismatches are therefore errors and indicative of poor base quality, using the comparisons the program generates tables based on the covariates (e.g. read group, reported quality score, cycle). From the co-variates table the model calculates the probability of probability of errors in the qualities at the sites and returns a recalibration table from which quality scores can be corrected.

1.3.5 Variant calling

BAM files can be used to generate variant calls, the most commonly utilised programs are SAMtools[46], VarScan2[51] and GATK HaplotypeCaller (replacement for GATK unified genotyper)[47] though many others are available. SAMtools performs variant

calling by creating a pileup of reads at each base, this information is then be compared against a specified reference genome. From the pileup file variants can be called using bcftools[46]. VarScan2 uses the pileup data generated from SAMtools to also call variants but instead uses a heuristic algorithm to determine a genotype for germline and a matched tumour sample.

Tumour samples are heterogenous resulting in many sub-clonal populations[52] and or contain copy number abnormalities. Hence when calling these variants, which can be present with low variant allele fractions, they would be called as homozygous reference by a germline variant caller[51]. Somatic variant calling can also be complicated by the issue of tumour purity when germline DNA contaminates a somatic DNA sample also lowering the variant allele fraction. By using matched germline and somatic samples from the same organism, with the germline as a reference the somatic variants can be detected down to a lower variant allele fraction.

Both SAMtools and VarScan2 use different algorithms in the variants calling process but both store results as a Variant Call Format (VCF) File[53]. VCF files are standardised so mandatory columns are always the same and in a pre-defined order to allow. However in recent years due to the increasing number of samples and larger sequencing methods being used such as whole-genome sequencing a new format used by GATK HaplotypeCaller of a gVCF. This file retains the column structure of a VCF file but contains an entry for all bases covered in reference genome with genomic regions of only reference calls collapsed into single entry blocks. gVCFs scale well with multiple samples when used as part of larger studies. Initial variant calls with GATK are created from the score recalibrated BAM file per sample using the program HaplotypeCaller, which performs calling of SNPs and INDELs reliably up to around 50 bp in size. When performing calling the program performs local de-novo assembly of haplotypes, so in regions of variation the program will re-map reads in order to provide more accurate variant calls particularly with INDELs.

Often in studies it is desirable to have all individuals represented in a single multi-sample file which enables comparisons between individuals but also to perform more advanced analysis such as trio analyses. With SAMtools to generate a

multi-sample VCF all of the sample BAM files are read at the same mpileup step which creates a bottleneck and is computationally demanding in analysis pipelines when scaling with large numbers of samples. To avoid these problems GATK advices users to generate a gVCF per sample which has a description for every base in the reference. An additional program called '*combineGVCFs*' is then able to take advantage of there being information from each file for every base and as gVCFs are only text files in position order they can be read from many files in chunks and avoid the bottleneck of reading raw read data from multiple BAMs simultaneously[48]. The combined gVCF file can then be genotyped using the '*genotypeGVCFs*' tool which calculates the genotypes for variants utilising information from all of the samples to calculate the most likely genotype for each sample and resolve potentially ambiguous sites with low evidence or depth in problematic samples.

Comparisons of variant callers have consistently reported that the most sensitive SNP variant caller is SAMtools[48,54]. INDEL calling performance though has been reported best as from GATK HaplotypeCaller[48,54]. Another study has also suggested that variant calls from GATK HaplotypeCaller are more likely to be reproducible than other variant callers such as SAMtools and the scalability of GATK has meant it is now one of the most commonly used variant callers[55].

1.3.6 Variant annotation

Variants called from NGS data in VCF format alone are not able to be prioritised. Additional annotations are required from programs such as ANNOVAR[56], VEP[57,58] or snpEFF[59]. Each of these programs match variants using the location and alleles from the VCF against multiple annotations databases to add them into the VCF, either in the INFO column (VEP, snpEFF) or as new columns (ANNOVAR).

Over time annotation databases available have improved and grown in complexity such that it is almost now possible to have a pathogenicity prediction for all variants both coding and non-coding. In addition multi-population allele frequency measures also help identify disease causing variants from the tolerated or benign mutations. In earlier days of NGS limited datasets and methodologies meant predictive tools such as SIFT[60] and PolyPhen2[61] were the chosen scores. Both of these tools were limited to

predicting coding variants and as subsequent tools such as FATHMM have shown the accuracy was variable depending on the testing set from 0.65 to 0.74[62]. Therefore it is advisable to use the predictive pathogenicity tools as a piece of supporting evidence rather than a key piece of evidence in making a case for a variant.

In October 2011 the 1000 genomes (1KG) project released allele frequencies for variants seen in the 1,092 WGS samples. This provided a powerful resource to filter and review variants by how common they were as the participants in the 1KG project were all healthy individuals. Variants that were common or seen amongst the 1,092 samples would not likely be pathogenic. Therefore a large number of variants could be eliminated from analyses by applying a frequency filter such as 5%, if investigating rare diseases using the allele frequency annotation greatly aids variant prioritisation. In recent years allele frequency data has become increasingly available from projects such as ExAC which aggregated 60,706 WES samples of healthy individuals and gnomAD containing 123,136 WES and 15,496 WGS samples[63]. These larger datasets provide a framework against which variants can be filtered to identify only variants rare enough to be in keeping with a disease phenotype. However it has also been identified that some loss of function variants actually appear tolerated and some variants previously described as causal were incorrect[63,197]. Annotations for variants which have previously been seen can also be recorded in databases such as dbSNP[19], COSMIC[64] or CLINVAR[65] can also be added which can also help to identify variants previously seen in other disease cases.

Combined with the increased power to filter variants using population allele frequencies there were improvements in the pathogenicity algorithms which made use of data from the ENCODE project to not only provide predictions for coding variants but also non-coding. By using ENCODE data algorithms could be weighted away from just measures such as conservation and amino acids changes to enable the non-coding predictions but also use machine learning with multiple evidence sources to improve predictions. Many of these tools have recently been produced such as: CADD[66], DANN[67], GAVIN[68], FATHMM-MKL[69] and FATHMM-XF[70]. Therefore as more rare diseases and more complex diseases are now being sequenced we are now better able to identify common causal variants but also identify less common variants

with high predicted pathogenicity from multiple tools for both coding and non-coding variants, potentially elucidating the regulatory mechanisms which can cause disease in non-coding regions.

Using ENCODE data non-coding variants were able to be identified in cis-regulatory elements including promoters, enhancers, insulators or silencers. Cis-regulatory elements can be located proximal to genes or as distal elements far from genes[71]. Each of these elements act following binding of transcription factors. Loss, gain or modification of a transcription factor binding site therefore can reduce or increase cis-regulatory functions leading to altered gene activity.

Enhancers act to promote transcription after binding of a transcription factor to a motif by assisting in the recruitment of RNA-polymerase to gene promoters. Currently enhancer variants are the most well described in terms of disease associations. An example of an enhancer alteration is in the gene telomerase reverse transcriptase (*TERT*), which encodes the catalytic subunit of the telomerase enzyme and is often over-expressed in aggressive forms of cancer. Mutations identified in the *TERT* promoter either site G228A and G250A are found mutated in up to: 21% of medullo-blastomas, 47% of hepatocellular carcinomas, 66% of urothelial carcinomas of the bladder, 71% of melanomas and 83% of primary glioblastomas[72]. Both of these mutations are associated with increased *TERT* expression and telomerase activity. Both mutations also alter the binding sequence, allowing the transcription factor GA-binding protein to bind, leading to increased *TERT* expression. Conversely the loss of enhancer sites reduces promoter expression such for *PAX5*, when insulators block interactions occurring from enhancer to a promoter and silencers repress RNA polymerase recruitment at promoters[71,73,74].

1.3.7 Structural & copy number variants

Variant callers such as SAMtools and HaplotypeCaller are limited in their ability to reliably call variants above 50bp, consequently larger structural re-arrangements within genomes were likely missed without further analysis. Because of the limitations of the traditional variant callers specialist programs have been created to identify the larger variants from NGS data. However programs differ in the types of

variants they detect and also the approaches used to identify them. The types of structural variants commonly attempted to identify are shown below in Figure 1.4.

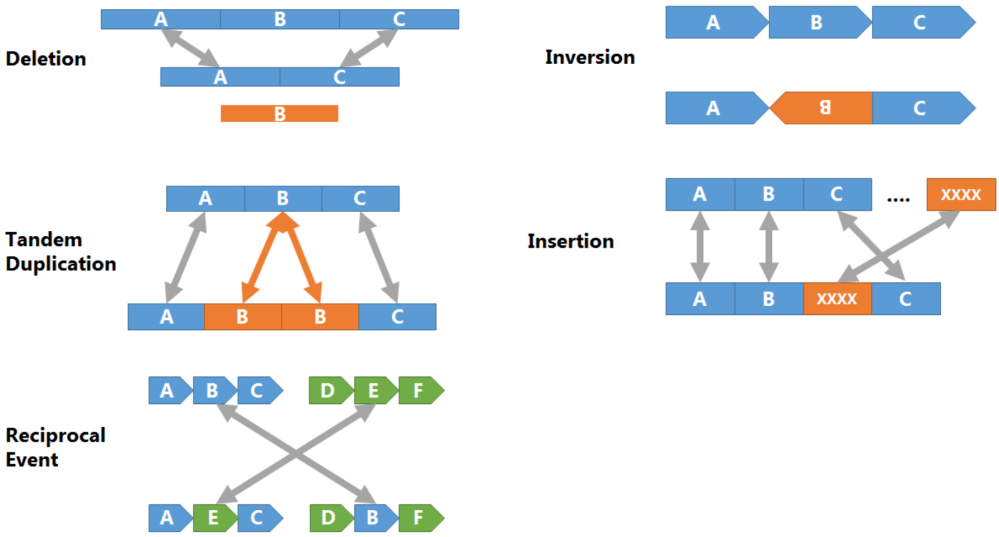

Figure 1.4: Five most common structural variant types illustrated to highlight the structural changes. In deletions a segment of the genome is lost compared to the reference. For inversions a segment of the genome is in a reversed orientation compared to the reference genome. Tandem duplication events occur when a section of DNA is duplicated adjacent to the original segment. Insertions are when a sequence is inserted relative to the reference genome. Reciprocal events are complicated rearrangements which typically occur between distal locations or cross-chromosome where a segment of DNA is swapped between sites.

A large scale study by Sudmant *et al.*[75] analysed 2,504 human genomes for patterns of structural variation. Due to the various algorithms and approaches taken by different programs results were often highly variable for example the program PINDEL[76] identified 9,580 deletions of which only 150 were also called by the program BreakDancer[77] or 28 when compared to Delly[78]. Because of the large variations in the number of called between tools manual curation was attempted. Many of the calls often proved difficult to observe from visualisations.

More recently newer programs have been suggested to more accurately and reliably detect structural or copy number variations. Two of the programs utilised in this thesis are LUMPY[79] and CNVkit[80]. LUMPY differs from previous generations of tools as it integrates multiple sources of evidence including: split-reads, paired-end

read distances and read depths to call variants such as deletions, duplications or inversions. When compared to Delly and PINDEL using data from the sample NA12878 LUMPY was both more sensitive and with a lower false-discovery rate[79].

CNVkit uses the ratio of off and on-target reads to infer copy number changes by comparing against a reference, usually created by pooling samples. This method was initially designed to work with targeted data but has the ability to work on whole genome data by defining the entire genome as target and then calculating ratios for these regions. Initial sample depths are calculated and then binned. Each of the binned files are then corrected against the reference generating copy number ratios normalised for GC content of the reference genome, exon distribution and repeat sequences using a reference fasta sequence file. Using thresholds for the ratio and a circular binary segmentation algorithm it is then possible to call segmental gains or losses of copy[80].

1.4 Aims

This thesis is divided into five projects each focusing on different stages of NGS pipelines.

In the first project a method will be devised to estimate sample contamination of NGS samples from a VCF file by using measures of the variant alternate allele frequencies. The second project will examine the use of unmapped reads from NGS samples to determine if biologically useful information can be gained from their analysis or if they indicate cross-species contamination.

The third project aims to compare WES and WGS sequencing both performed on a familial trio with a case of a rare disease. By comparing the variant calls of the WES and WGS the project will assess additional benefits and potential limitations of each method for identifying causal variants both now and by future re-analysis. The fourth project will apply the latest variant calling and annotation techniques to attempt and identify causal variant(s) in the affected child within the familial trio with a reported case of the rare disease Sedaghatian-type spondylometaphyseal dysplasia by combining both WES and WGS data.

The final project will focus on the FCGR genes involved in modulation of immune responses. Sequencing of the receptors is challenging due to high homology between genes, by using targeted NGS variants and copy number changes over the genes which could be modulating the immune response will be identified. The project will assess the ability of short read sequencing to call variants over the FCGR genes.

Chapter 2

Estimating contamination levels in exome sequencing using alternate allele frequencies and variant zygosity

Due to the high purchase and operating costs of NGS instruments it is common for the outsourcing of sample sequencing. Samples are prepared and sequenced in batches to make efficient usage of sequencing machines, making cross-sample contamination a potential problem. Three types of contamination can potentially occur: cross-species, within-individual and cross-individual. Each type of contamination poses unique challenges to identify. Examples of contamination involving cross-species contamination have been most commonly reported in literature[81–87]. Cross-species contamination tends to be detected in the sequencing procedure due to sequence differences between species being detected[88]. Alternatively post-alignment to a reference genome cross-species reads will be flagged as unmapped[89,90], this will be discussed in more detail in Chapter 3.

Intra-species contamination is less commonly reported in literature as reads will still map to a reference genome and are harder to identify without further analysis. Intra-species contamination poses a serious problem due to the loss of accuracy which could lead to miscalled variants dependant upon the level of contamination, variant

calling procedure and sequencing depth[88,91–93]. Intra-species contamination should be identifiable as there are likely to be significant differences between the crossed individuals assuming they are unrelated. Contamination can occur at any stage from the collection of a sample up to the sequencing stages, in particular during the extraction and purification of the genomic DNA when samples are prepared in batches.

Within-individual contamination poses the greatest challenge to identify computationally as this predominantly occurs with tumour DNA samples[94]. Within-individual contamination can also occur in RNA-seq, where tissue specific patterns exist[95]. Tumour samples may easily be contaminated with the regular germline DNA from the organism, which only has a few somatic variant differences. Contamination potentially leads to inaccurate estimations of tumour sample purity. Tumour purity is the proportion of cancer cells in the admixture, with increased germline DNA contamination the tumour purity decreases.

Decreased purity can lead to the failure to detect all somatic variants and the loss of the ability to delineate the mutations that have occurred in the evolution of the tumour and potentially the sub-clonal architecture of the tumour[92]. In a traditional model of cancer most somatic mutations in cells will passively accumulate throughout the lifetime of an organism. Should somatic mutations provide a competitive advantage over other cells this leads to the development of persistent clones and the origins of cancer[52]. As somatic mutations accumulate passively in cells with age it has previously been difficult to study the mutations present in small numbers of cells. One recent study by Martincorena *et al.*[52] performed sequencing across upper- and mid-esophageal epithelium for nine patients, including four smokers. Using 2mm^2 grids across the tissues a total of 844 samples were obtained each capturing 74 cancer associated genes with median depth of coverage 870x. 6,935 somatic mutations were found over the 844 samples supporting the idea of passive accumulation. 21 of the samples were dominated by large clones, which were then also found by WGS of the same samples. Somatic mutations were able to be detected with a median allele frequency of was 1.6% and down to below 1% due to the high sequencing depths[52,96]. Given the heterogeneity between cells, in terms of the number of somatic mutations, cell size and competitive advantages obtained

from mutations it explains why most mutations may be present with such low allele frequencies. Therefore identifying somatic mutations accurately requires high depth of coverage to distinguish them from sequencing errors. Contamination estimation for tumour-normal pairs is therefore complicated, current tools compare a germline sample with the tumour. From the study by Martincorena *et al.* there is evidence that benign somatic mutations can be present mosaically in germline samples, should depth be high enough these somatic variants will also be called as germline. The mosaicism of the benign somatic variants will also likely differ significantly depending upon the tissue type and factors such as mutagen exposure, for example sun exposed skin or oesophageal tissue from a smoker. The mosaic presence of these variants will reduce the accuracy of somatic variant calling and further complicates the process of contamination estimation.

2.0.1 Contamination estimation tools

Current tools for the estimation of intra-species contamination have been available since 2011. ContEst was released in 2011[91] and VerifyBamID in 2012[88]. VerifyBamID was reported to give similar results to ContEst at lower contamination levels but was suggested to outperform ContEst at higher contamination levels[88]. VerifyBamID has been adopted by more studies than ContEst as highlighted by current literature[97–101]. For intra-individual contamination estimation CONPAIR was released in 2016 which estimates for tumour-normal pairs[94].

VerifyBamID

VerifyBamID is a program that uses a BAM file and or array based genotype data as inputs[88]. When using a combination of both sequence and genotype array data for a number of samples the program assumes genotype data is correct. Should there be an error in any of the sequenced reads it is assumed that the all other nucleotides are equally probable to be erroneously reported because of the sequencing error. Input BAM files are used to calculate the number of matches and mismatches to the known genotypes for all sites in the array and the genotype concordance calculated. As each individual should have a unique combination of genotypes the program then identifies the potentially contaminating sample using the genotype data and estimates the percentage contamination termed 'chipmix'. In this model the program assumes that

the array data is entirely correct and genotype probabilities are calculated using the population allele frequency data, assuming Hardy-Weinberg equilibrium.

VerifyBamID also works without sample specific external genotype data, only requiring the BAM file and a VCF of population allele frequencies. As there are no known genotypes the program does not calculate the swapped percentage between samples. Instead the program models the sequence reads as mixture of two unknown samples based on the allele frequency information in the VCF file and then estimates the probability of each site being correct to estimate overall contamination up to 50% termed freemix.

Validation of VerifyBamID was performed by crossing reads from 1000 genomes samples *in silico*. The levels of contamination were therefore known as a specific percentage of the reads from each file would be crossed. Results indicated that when using sequence data with genotype a Pearson correlation coefficient (r) of 0.9996 was gained and an r value of 0.9840 for sequence data alone. Both of these methods therefore indicated good correlation to the actual levels of contamination[88]. However some degree of caution should be used with these results as the paper is not specific in describing the methods used in generating the contamination simulation files as the read totals in each file should have been maintained.

One method also of interest is the regression based method for array data which analyses minor/alternate allele frequency change for homozygous alternate variants. This identified that minor-allele frequency and contamination do not share a linear relationship but yielded similar results with a linear regression to models which required population allele data. However this method was only tested with 10% contamination and was complicated by noise from heterozygous variants which was not further resolved[88].

ContEst

ContEst identifies sample contamination also using a VCF file of population allele frequency data from the 1000 genomes project and the sample BAM file. The program identifies homozygous alternate SNPs, noting for each site the probability of

seeing a contaminating allele. For each site the program uses a Bayesian approach to calculate the posterior probability of contamination levels to determine the most likely level, assuming other nucleotides are all equally likely to be contaminating. The program then estimates the overall contamination using the summation of the individual sites[91].

There have not been a significant number of papers reporting the results and accuracy of ContEst, in the original publication testing was performed on a public ovarian cancer dataset which identified 12 samples which showed low contamination levels. Testing was then extended by *in silico* simulations based on a single sample termed the primary sample being contaminated at a range of levels up to 25% from other samples which reported good agreement[91]. As all of the samples used for testing were from a single cohort of a public dataset the program may be over-fitted as other validations and studies have not confirmed the reported accuracy of the program.

Conpair - Tumour-Normal contamination estimation

CONPAIR (Concordance/Contamination of PAIRred samples) focuses on detecting contamination of tumour and normal samples[94]. Tumour sample contamination estimation can be challenging due to alterations in copy number which are suggested to cause copy number-driven allelic imbalance frequently seen in cancers and leading to shifting of the expected 50% allelic fraction for heterozygous markers. CONPAIR only analyses the homozygous alternate variants. It implements the same models as VerifyBamID but takes in tumour and normal BAM files for an individual. CONPAIR measures contamination first in the normal and then in the tumour sample, using the genotype information from the normal as the truth dataset but only uses a reduced marker set as opposed to the entire file as with VerifyBamID. Results reported good accuracy with simulations for crossed glioblastoma tumour exomes with known copy abnormalities with a root mean standard deviation of 0.0064 compared with 0.0075 for ContEst and 0.062 or VerifyBamID[94].

2.0.2 Alternate allele frequency changes with contamination

For uncontaminated samples heterozygous variants should have an expected Alternate Allele Frequency (AAF) of 50% as there are one copy each of reference and alternate

alleles. In this chapter the AAF is defined as the non-reference allele as called when using the human reference genome. However in most samples there will be an approximately normal distribution around a peak at 50% due to sequencing depths not being high enough to have equilibrated out at 50%. Variants which are homozygous for the alternate allele should only contain the alternate allele, hence in AAF profiles there will be a peak at 100%. Depending on the dataset and the pipeline used there may be a small tail below the 100% AAF peak caused by a combination of sequencing errors and or sample contamination from another individual causing the deviation from the expected AAF. Alternatively in tumour samples a tail may be caused by sub-clonality[52].

As the contamination level increases there will likely be characteristic changes in the AAF profiles which can be observed and measured. At low levels of contamination below 10% most of the changes should be observed around the homozygous variant AAF peak as the tail below 100% will increase due to lowered variant AAFs. At higher levels of contamination above 10% the tail will cross the calling threshold for a homozygous variant causing the tail to become part of the right side of the heterozygote variant AAF range. The point at which the tail variants become called as heterozygous will be dependent upon where the variant caller used defines the homozygote variant threshold for a dataset. As the contamination level increases above 10% the homozygote tail will reduce and gradually return back to a normal level with a peak only at 100% but with peak height reduced and a growing number of variants on the far right of the heterozygote variant AAF distribution.

Beyond a 10% level of contamination most changes are likely to be detectable in the heterozygote variant AAF distribution. The general effect of contamination is to lower AAFs, therefore some homozygote variants will become called as heterozygote variants. At the same time existing heterozygote variant AAFs should lower with contamination. The heterozygote peak at 50% will likely shift in favour of lower AAFs and the peak at 50% to reduce in height and broaden. As contamination continues to higher levels of 40 - 50% the heterozygote distribution should become trimodal with peaks in the centre, far-left and far-right. Hence at higher contamination levels heterozygote AAFs should become less normally distributed due to the increased

peaks at the extremities of the heterozygote AAFs distribution. All of the information needed to the devise AAF profiles and measure changes can be obtained from a VCF file using the INFO and GT (genotype) fields. From this data measurements of the homozygote and heterozygote peaks can be performed including the: ratio of heterozygous to homozygous variants, skew and kurtosis of the distributions.

Gender of samples can also be used as a supporting piece of evidence to suggest a cross-individual or cross-gender contamination event. The average exome of a female typically reports about 50 - 65% chromosome X heterozygote variants. In the genomic sample of a male there should only be the one X chromosome, theoretically all X chromosome variants should be homozygous, though in rare cases such as Klinefelter syndrome, patients have chromosomes XX and Y or Turner syndrome where females only possess one X chromosome[102]. Additionally at the tips of X and Y chromosomes there have been two short homologous regions identified known as pseudoautosomal regions. Pseudoautosomal regions pair and recombine during meiosis allowing in males the presence of heterozygous variants[102].

Crossing of sequence data will yield noticeable changes in chromosome X heterozygotes. As male contamination is added to female samples the heterozygosity increases as the samples will be made of one male X and two female X chromosomes, heterozygosity increases from the normal female range of 50 - 65% by approximately 10% after the addition of 50% male sample contamination, shown in Figure 2.1. Addition of female samples to male samples causes the heterozygosity to rise rapidly, by 20% contamination male samples would appear in the normal range for females. Female to female contamination will also increase heterozygosity to an abnormally high level close to 75%. Male to male additions make samples appear as two X chromosomes, so also appear as female at contamination levels above 50%. All situations therefore result in the increase of chromosome X heterozygosity level.

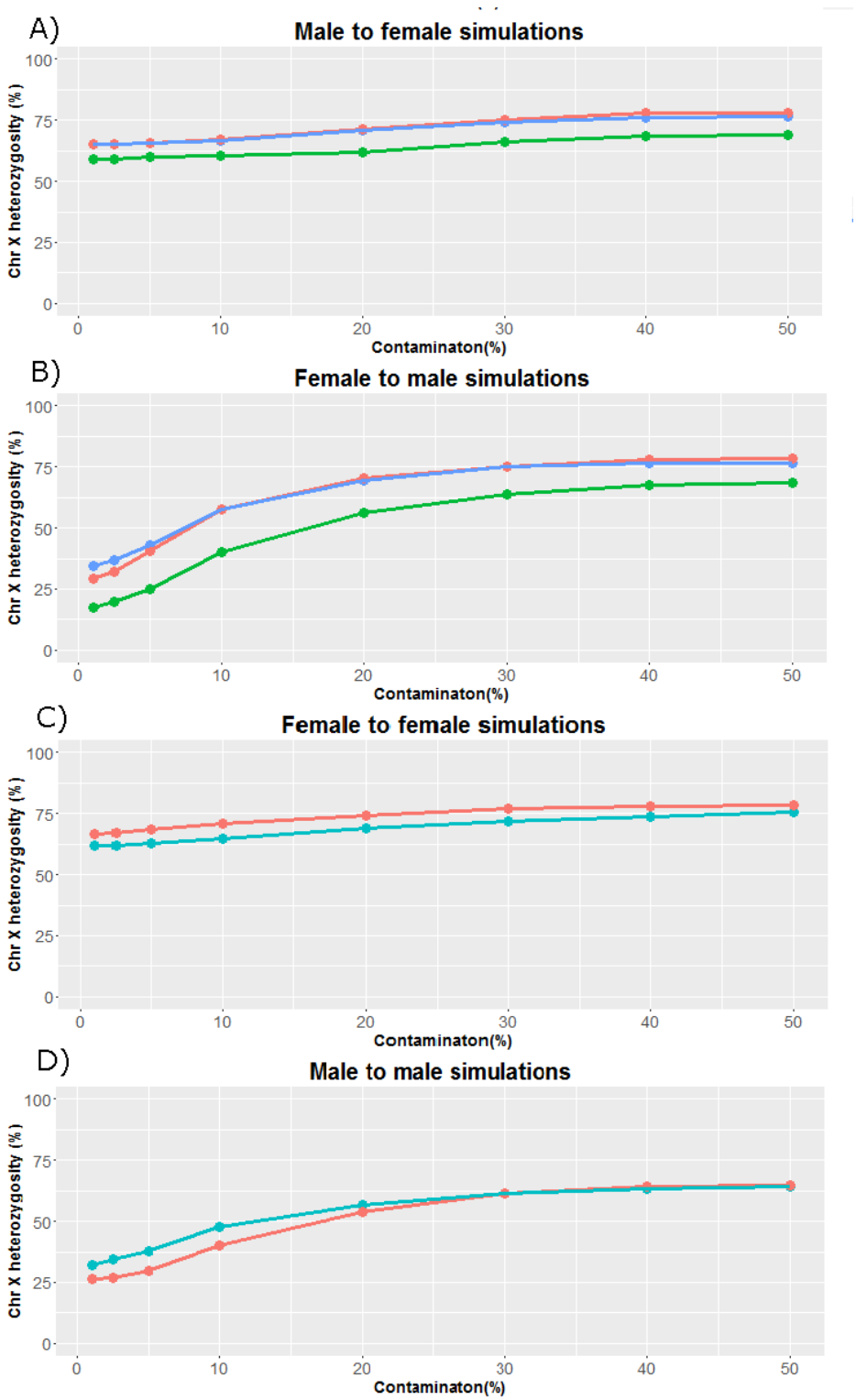

Figure 2.1: Contamination simulation illustrations highlighting the alterations in X chromosome heterozygosity when crosses are simulated. In order the figure shows: **A)** Male contaminating sample added to female **B)** Female to Male **C)** Female to female **D)** Male to male. All contamination events lead to the increase of the percentage of heterozygous of chromosome X variants.

2.0.3 Applications of machine learning with NGS

Machine learning has driven many of the recent improvements in analysis tools and techniques seen in NGS such as the base quality recalibration performed in GATK to correct sequencing quality errors[47] or in the pathogenicity prediction tool FATHMM-XF[70]. Machine learning can be broadly classified as a set of algorithms performing: pattern recognition, classification and prediction[103]. Machine learning algorithms can generally be separated into two groups of supervised and unsupervised learning. In supervised learning algorithms are trained using an example dataset which has labelled features and target values. From the labelled dataset the algorithm learns about the features and their relationships to the targets to then be able to perform classifications or regressions (when using continuous variables). The example dataset can also be sub-divided into a training dataset and a testing dataset to assess the accuracy of a clustering or regression before the application to a non-training/testing dataset. In unsupervised learning algorithms are supplied with a dataset without labels from which the algorithm must interrogate the data to identify groups to perform tasks such as clustering[104].

2.0.4 Regression models

Several regression models are able to be implemented when using supervised learning. One of the most common models is the Ordinary Least Squares (OLS) regression. This regression fits a linear model with the coefficients set to minimise the residual sum of squares between the actual points and the predicted value by the regression[105].

RIDGE regression operates using the same approach as OLS in applying a linear model to minimise the residual sum of squares but with the addition of penalties on the size of coefficients. OLS algorithms can give inaccurate results when several of the independent variables are very highly variable leading to under or over-fitting and consequently poor performance when using other samples. Hence to better control variance penalties are imposed on the coefficients to restrict the variance, effectively regularizing the coefficients with penalties to avoid under or over-fitting[105].

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a linear model that estimates sparse coefficients. LASSO is similar to ridge regression but

regularisation/penalties are performed using an absolute value which does not penalise high value coefficients, but actually setting them to zero if they are not relevant, potentially leading to a regression using few or sparse features[105].

A support vector regression (SVR) regression can also be implemented with a linear kernel. This algorithm looks at data to identify decision boundaries between labelled features or groups. The margins between the groups within the dataset are known as support vectors. Using the support vectors the algorithm is able to convert multi-dimensional data into a 2-D plane using what is known as a kernel trick[105]. Multiple kernels can be implemented with an SVR algorithm including both linear and polynomial kernels[105].

Polynomial regressions are applicable to models where the relationship between the dependent - X and independent - Y variables are non-linear. The degree of the polynomial regression can also be changed to better describe the relationships between the variables.

A Radial Basis Function (RBF) can also be used in a polynomial regression model. RBFs calculate the distance from the origin or a defined point. The RBF kernel can also be used in a manner similar to support vector regression.

2.1 Aims

This chapter describes the development of a method to detect contamination using only data obtained from VCF files. Often VCF files are obtained from public datasets or shared in collaborations due to the smaller file size. Therefore the lack of BAM files prevents estimations of contamination using programs such as ContEst or VerifyBamID. To create a method which estimates contamination from VCF files the patterns of contamination will be in investigated from *in silico* contamination simulations. From the simulations it will then be possible to assess the metrics and measures obtained from VCFs to determine those useful in predicting contamination. Having identified measures for describing contamination the data will be used with machine learning based regression models to select the optimal algorithm to estimate contamination levels. Finally the predicted contamination values from the model will be applied to a larger dataset and compared against the commonly used VerifyBamID tool to compare the similarity of contamination estimates.

2.2 Materials & methods

2.2.1 Contamination simulations

Using WES fastq files contamination was simulated *in silico*. For a desired percentage of contamination, the number of reads equal to the percentage in the the fastq files were calculated. For example a sample of one million reads total, 1% would be 10,000. To simulate 1% contamination 10,000 reads would be removed from both of the paired fastq files and 10,000 reads added from another pair of fastqs from a second sample as shown in Figure 2.2. After crossing of files there will be a known level of contamination in fastq files which are then run through the alignment and variant calling pipeline as shown in Figure 2.3 to create a VCF file.

45 samples were used to generate 25 contamination simulations. Each of the 25 simulations were performed with eight contamination levels: 1, 2.5, 5, 10, 20, 30, 40 and 50%, giving a total of 200 contaminated samples. The samples used in simulations to train regressions are described in Table 8.1. 25 male samples were used and 20 females. BAM files for each of the samples without crossing of reads were also checked by VerifyBamID which indicated contamination less than one percent for any sample as shown in Appendix A, Table 8.1.

Samples chosen to model contamination were from a variety of human exome capture kits and studies in order to simulate a range of possible contamination scenarios in which contamination could occur. By generating crosses between a range of samples both the trends of contamination could be investigated and the model trained on a wide variety of samples to avoid over-fitting to a specific group of samples.

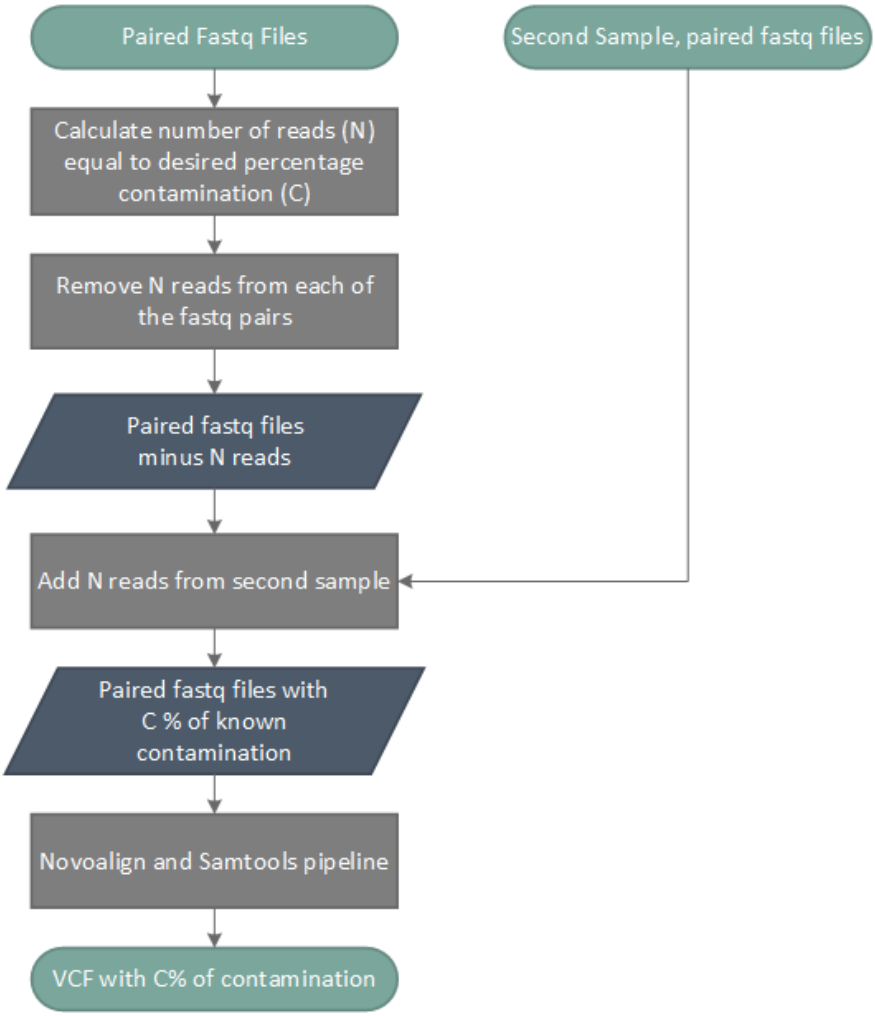

Figure 2.2: Pipeline used to generate *in silico* contamination simulations with known percentages of contamination. Reads were removed from the fastq files corresponding to the desired percentage of contamination and replaced with the same number of reads from another sample. Fastq files crossed between the two samples are then run through the alignment and variant calling pipeline as shown in Figure 2.3 to produce VCF files with known levels of contamination.

2.2.2 Alignment and variant calling pipeline

Contamination simulations were performed *in silico* using WES samples. All samples were processed using the standardised pipeline as shown in Figure 2.3. Alignment was performed to the hg19 human reference genome using novoalign (v2.08.02) for paired-end fastq files. BAM files were processed using picard tools (v1.97). Variant calling was performed using SAMtools (v0.1.19).

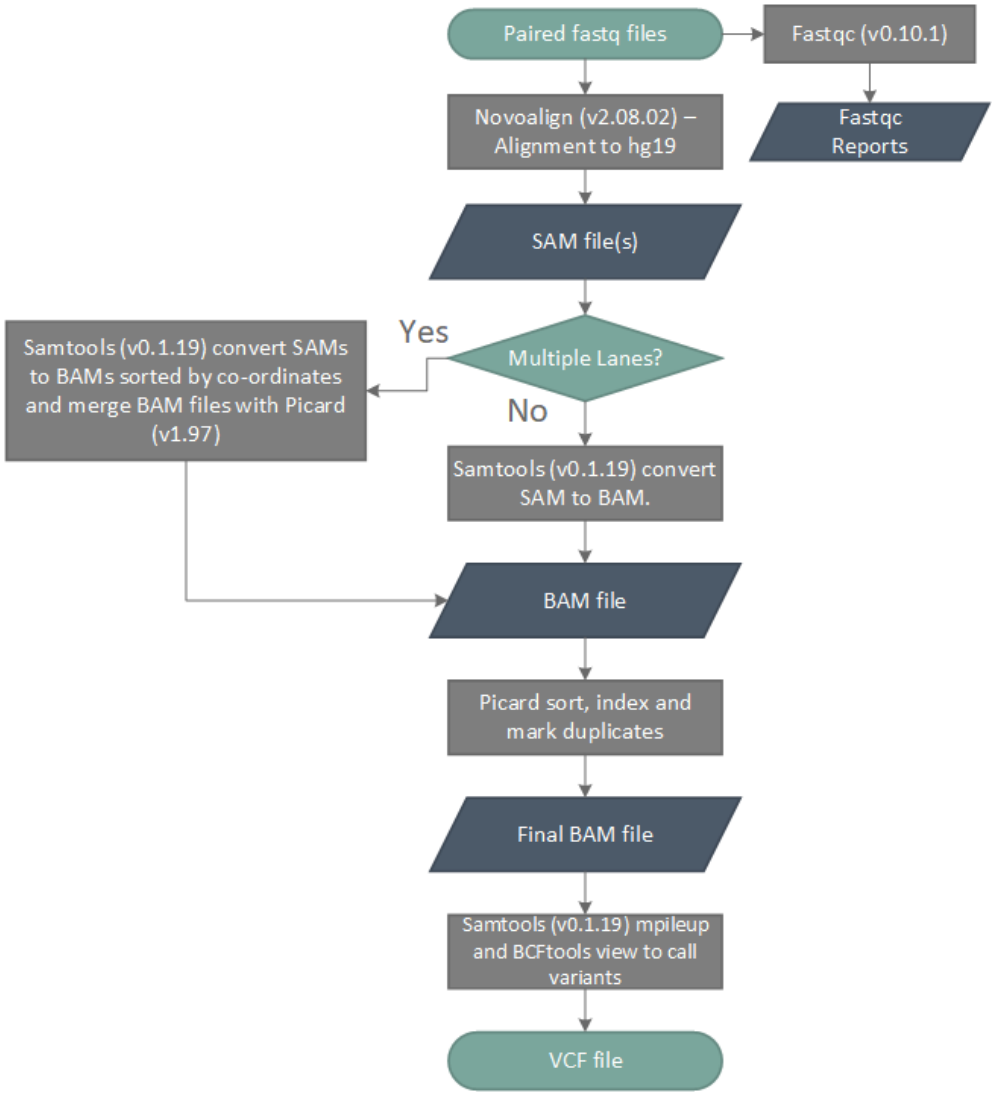

Figure 2.3: Novoalign, SAMtools hg19 variant calling pipeline processing paired fastq files. Alignment was performed using novoalign (v2.08.02) before processing with picard tools (v1.97) and variant calling using SAMtools (v0.1.19) to generate VCF files.

2.2.3 Alternate allele frequency profiles

AAF profiles were created to investigate how variant AAFs are affected by contamination at different levels. Data was extracted from the VCF files using a BCFtools query command extracting the variant location descriptors: chromosome, position, reference and alternate allele along with the variant quality. From the INFO and GT fields of VCF files DP4 values and variant genotypes were also obtained. For each VCF variants were grouped by genotypes and the AAF was calculated using DP4 values. Variants were also filtered to remove those with depths below 10 or a PHRED quality below 20 as described in Figure 2.4. WES samples with below 5,000 variants passing filtering were also excluded. Filtering was applied to reduce the noise from poor quality variants and sequencing errors.

AAFs for variants were split by genotype into homozygotes and heterozygotes before plotting, enabling changes in each of the extremities of profiles to be more easily identified. The boundary for calling of homozygote variants was identified from the AAF histograms as 88% for SAMtools variant calls.

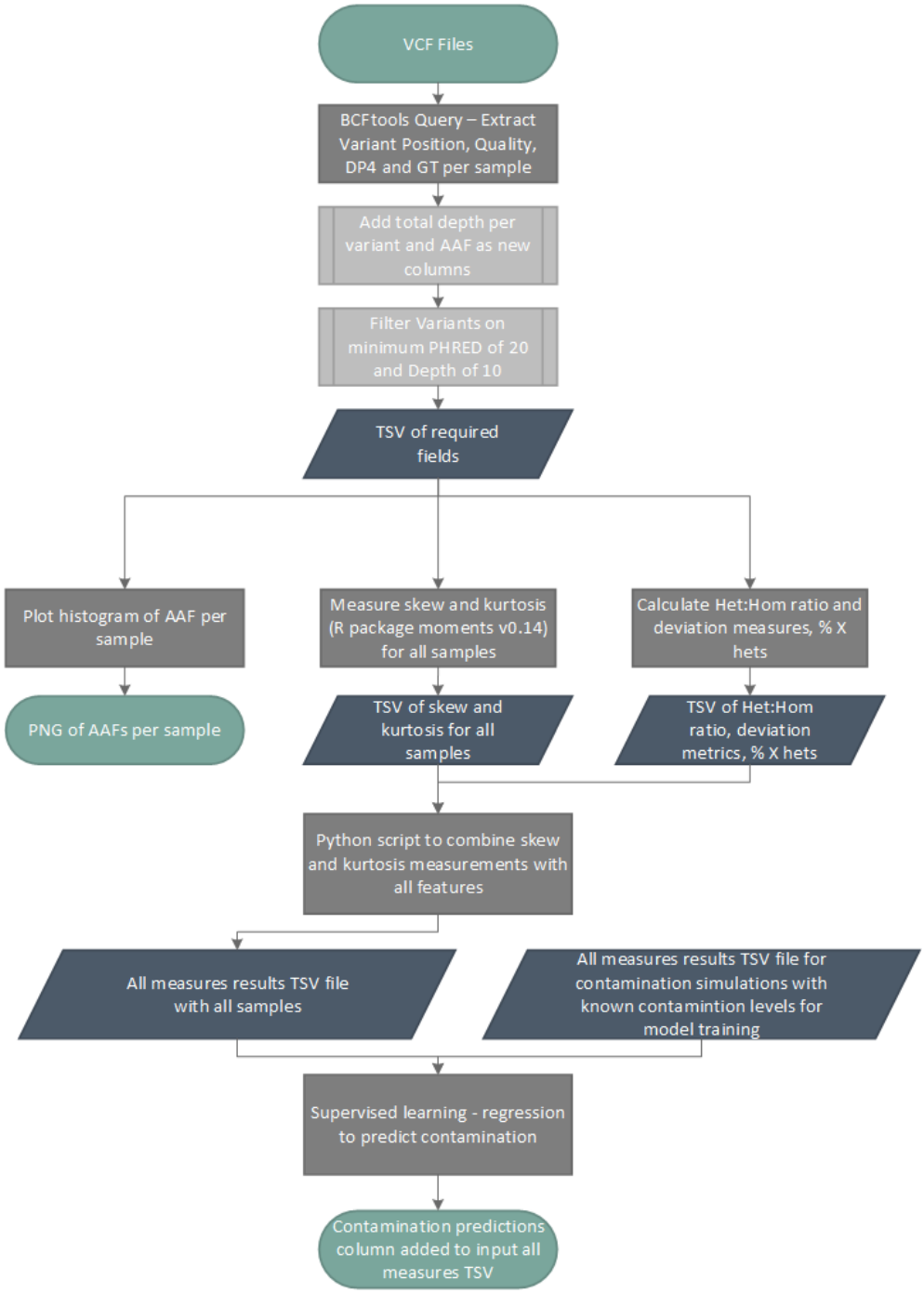

Figure 2.4: Contamination estimation pipeline used to obtain measurements of the AAFs for variants. Relevant fields are extracted from VCF files to create a tab separated file with variant position descriptions along with PHRED quality, genotypes and DP4 values of variants. Using DP4 values the AAFs were calculated and plotted as histograms to visualise profiles and assess suspected samples for contamination. Using the AAFs measurements of skew and kurtosis are also performed. Other measures such as the heterozygote to homozygote variant ratio and deviation measures were also obtained. All measures were finally combined and fed into a supervised machine learning regression algorithm which uses the training set of contamination simulations to generate an estimation of contamination in the samples.

2.2.4 Measurements used for contamination estimation

Nine measurements are taken or calculated from each of the VCF files for the contamination estimation as summarised in Table 2.1. Skewness and kurtosis of AAFs were calculated using the Moments R package (v0.14). Pearson's measure of kurtosis was performed for only heterozygous variants and then for all variants.

Measure	Data Range	Calculated from
Skewness	Homozygous variants	Alternate allele frequencies
Skewness	All variants	Alternate allele frequencies
Pearson's measure of kurtosis	Heterozygous Variants	Alternate allele frequencies
Pearson's measure of kurtosis	All variants	Alternate allele frequencies
Het:Hom ratio	All variants	Ratio of heterozygous autosomal total variants to homozygous
Deviation A	Homozygous variants	Difference from 100% for average AAF
Deviation B	Heterozygous Variants	Difference from 50% for average AAF
Deviation Metric	All variants	Sum of Deviations A and B
Percentage X heterozygotes	Chromosome X variants	Percentage of Chr X variants which are heterozygous

Table 2.1: Nine features were measured from the VCF files to supply into the regression model for prediction of contamination. Measures examined the distribution of AAFs for heterozygous and homozygous variants and deviations from expected AAFs in addition to comparing the number of heterozygous and homozygous variants.

The ratio of heterozygous to homozygous variants were calculated using all autosomal variants with genotypes of heterozygous (0/1) or homozygous (1/1). Deviation measure A (Dev A) measured the difference between the observed average AAF of the homozygous variants from the expected AAF of 100%. Deviation measure B (Dev B) measured the difference between the observed average AAF of heterozygous variants from the expected AAF of 50%. A summation of the two deviation measures A and B was then termed the deviation metric (Dev Metric). Finally the percentage of heterozygote variants on chromosome X were calculated. All five of these measures were calculated using a custom script using bash and AWK to process the tabular results file previously created by BCFtools, scripts used can be found in Appendix A - Section 8.1.2 and a on GitLab (https://gitlab.com/rs91/exome-contamination).

2.2.5 Investigating relationships of measurements

Investigations of all measured features and relationships to the levels of contamination were performed for the simulations dataset. A ten by ten scatter matrix was created using python with a histogram of the data distributions plotted on the diagonal. The scatter matrix allowed exploration of each of the measurements as contamination increased and between each of the measures to see if relationships were linear and correlated with contamination levels used to colour each of the sub-plots.

2.2.6 Principal component & clustering analysis

Principal component analysis (PCA) was performed for the 200 contamination simulations using ClustVis (http://biit.cs.ut.ee/clustvis/)[106]. ClustVis was run using unit variance scaling to standardise features and the PCA method singular value decomposition with imputation. All nine measured features were supplied and the principal components calculated. A clustering visualisation was also generated to see if the contamination levels were able to be differentiated. Component loadings and the variance explained for each of the principal components were also obtained which provided an indication as to which of the measurements were the most informative for potentially differentiating the contamination levels.

2.2.7 Regression analysis - model selection and training

To predict contamination all of the measures were supplied into regression models from the scikit-learn machine learning tool kit. Initial training of the models was performed using the 200 sample simulation dataset used to assess four linear models: OLS, Ridge, LASSO and SVR. Polynomial regressions were also performed using degrees from two to five along with the implementation of a RBF regression.

Simulation data was used to train models as they had known levels of contamination. The dataset was first split into a training set and a testing set to evaluate the performance of the model. Data from simulations was split 70:30 using the package *train_test_split* from the sci-kit learn module *model_selection*. To quantify the performance of each model the R^2 value denoting the accuracy of predictions was then recorded. To investigate if the initial split of the data for testing and training was reliable the models were also cross-validated by shuffling the initial data splits

and re-calculating the R^2 value. To perform cross-validations the scikit learn package *cross_val_score* also from the module *model_selection* was used. For each of the regressions 10 cross-validations were performed.

Each of the regressions were tested using the data split and cross-validations described previously but with either all features or only those describing the homozygote variant distributions with the Het:Hom ratio included. Testing of the model was also performed using specific contamination ranges from 1-50% down to 1-10% along with intermediate ranges to investigate the changes in model coefficients and predictions.

2.2.8 Application of regression models

The final regression model selected was applied to 245 germline exome samples using the pipeline showed in Figure 2.4. Samples were also called using the same novoalign and SAMtools pipeline with the hg19 reference genome as shown in Figure 2.3. All 245 samples with unknown levels of contamination were also analysed using VerifyBamID to compare predictions between the programs.

2.3 Results

2.3.1 Alternate allele frequency profiles and measurements used in contamination estimation

The effects of contamination were observed using AAF profiles. They show that the changes predicted in the introduction of this chapter are visible. A representative simulation is shown below in Figure 2.5 for all eight simulated levels of contamination along with the original sample for comparison. The changes in AAF profiles were repeatable across the simulations but showed variation at each contamination levels between each of the 25 crosses. Measures supplied into the regression analysis were designed around capturing the identified changes.

From Figure 2.5 as shown in Panels B-D, when the level of contamination is below 10%, most changes were observed in homozygote variant AAFs. A tail can be seen in the homozygote variant AAF profile which increases as the level of contamination increases. The measure Dev A calculated the difference between the observed average AAF of the homozygote variants and the expected value of 100%. A second metric was implemented to measure the skewness of the homozygote AAFs to reflect the relative reduction in variants from the 100% peak, as can be seen in Figure 2.5, as the contamination level increases. The value for skewness of homozygote AAFs was initially negative but as the level of contamination was increased homozygote variant AAFs decrease, so reduce the height of the 100% peak. Due to the reduced 100% peak the profile becomes less left-skewed and closer to a value for a normal distribution of zero as shown by Table 2.2 for contamination levels up to 10%.

Measurements of skewness were also extended to the all variant AAF distribution. Contamination has the effect of generally shifting variants to the left with lower AAFs as shown by Figure 2.5. The skew all measure can be seen to increase in Table 2.2 as the contamination level increases indicating a stronger right skew, fitting with the shift of variants to lower AAFs.

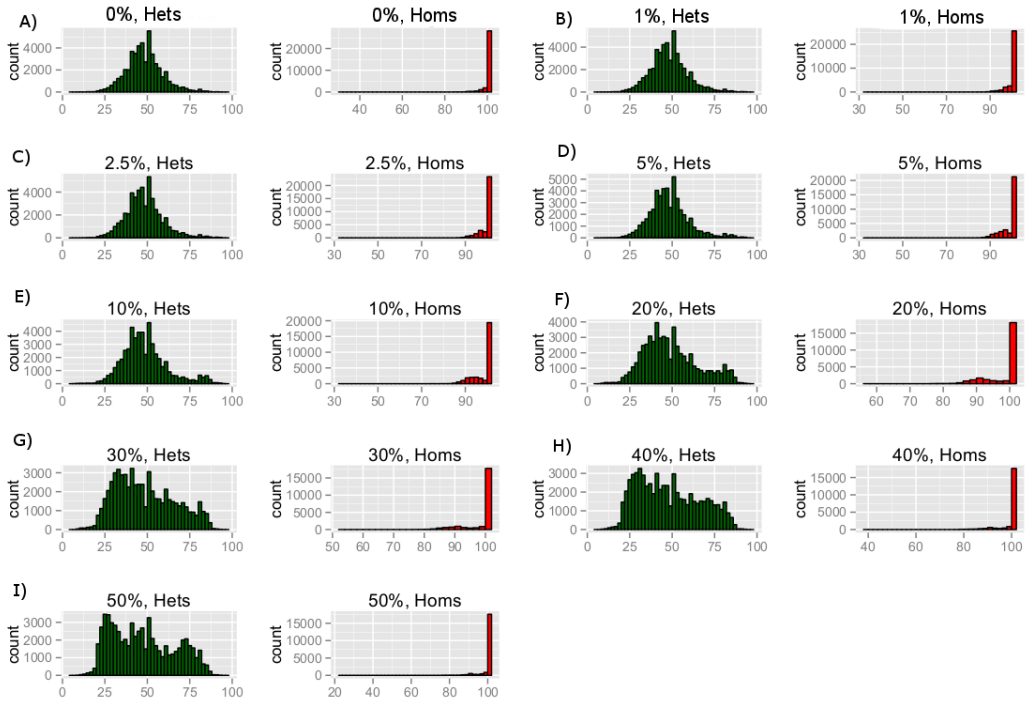

Figure 2.5: Alternate allele frequency profile for a representative contamination simulation set. Each of the panels show homozygote and heterozygotes at contamination levels: **A)** 0% **B)** 1% **C)** 2.5% **D)** 5% **E)** 10% **F)** 20% **G)** 30% **H)** 40% **I)** 50%. With increasing contamination the peak at 100% for homozygotes decreases in both count and AAF such that they become called as heterozygotes. At contamination levels below 10%, as shown in Panels A-D most change can be seen in the homozygote variant distribution as variants reduce in AAF towards the heterozygote range. Between 10 and 20% contamination, shown in Panels E and F, most change in observed on the far right of the heterozygote distribution as the homozygous variants are now called heterozygous. Above 30% contamination, shown in Panels G-I, heterozygous variant AAFs can be seen to reduce leading to a shift of the peak at 50% to the left and lower AAFs at the same time as homozygous variants being called as heterozygous on the right side. By 50% contamination the heterozygote distribution appears three-peaked compared to the normal distribution seen in no or low contaminated samples.

At contamination levels of 10% and above a peak can be seen to develop on the right of the heterozygote variant AAF profile in the range upwards of 75% AAF from Panels E-I of Figure 2.5. Variant calling for these samples with SAMtools used a threshold AAF of around 88%, below which variants are called as heterozygotes and above as homozygotes. Increasing the level of contamination generally caused lowering of variant AAFs as seen by the tail increase and reduction in height of the 100% homozygote AAF peak. As homozygote variants become lowered in AAF by contamination they are no-longer above the approximate 88% threshold and become called as heterozygotes. This highlights that changes at the higher levels of contamination occur increasingly for heterozygote variants. One of the most powerful measures for predicting contamination is the ratio of heterozygous to homozygous variants (Het:Hom ratio). Due to the shift of homozygous variants to be called as heterozygous variants the ratio becomes larger as the contamination levels increase.

Dev B was also used to measure differences between the observed average heterozygote variant AAF from the expected AAF of 50%. The relationship in Dev B can be seen to be non-linear due to the introduction of variants from the homozygote distribution on the far right of the heterozygous distribution at higher AAFs upwards of 10%. At contamination levels below 10% Dev B decreases but rapidly increases from a contamination level of 20% and upwards as shown in Table 2.2.

To better capture the changes that occur in the heterozygous distribution Pearson's measure of kurtosis was used. Kurtosis measurements are effectively a measurement of the extremities or tails for a distribution. A kurtosis value of three is indicative of a normal distribution, which is what would theoretically be expected for an uncontaminated sample. Higher values above three indicate a more asymptotic distribution at the extremities. Therefore as contamination increases the tails grow in size and the distribution effectively becomes flatter as the heterozygote peak at 50% decreases, giving a smaller value for kurtosis. Kurtosis measurements were also performed for all variants. Due to the extremity of the homozygote peak being large and the lack of low AAF heterozygote variants in comparison the measurements displayed the opposite trend. Low scores close to zero for uncontaminated samples were obtained. As variants are called with lower frequencies due to increasing

contamination the distributions become relatively more normal.

The final measure included was the percentage of chromosome X heterozygote variants. Males should have a low level of heterozygotes below 20% whilst female samples will be in the range of 50-65%. As contamination levels increase the percentage chromosome X heterozygosity can be seen in Table 2.2 to increase on average. Samples which appear excessively heterozygous with above 60% chromosome X heterozygote variants or with intermediate levels greater than 20% but less than 40% fall outside the expected ranges for either an average male or female.

		Contamination(%)							
Feature	**Measures**	**1.00**	**2.50**	**5.00**	**10.00**	**20.00**	**30.00**	**40.00**	**50.00**
DEV A	mean	0.60	0.92	1.30	1.71	1.77	1.50	1.23	1.08
DEV A	std	0.13	0.12	0.18	0.33	0.37	0.36	0.38	0.40
DEV A	min	0.44	0.70	0.93	1.11	1.24	0.92	0.53	0.44
DEV A	max	1.03	1.22	1.66	2.34	2.43	2.14	1.86	1.70
DEV B	mean	2.45	2.41	2.31	2.07	2.13	2.80	3.46	3.77
DEV B	std	0.41	0.39	0.38	0.46	0.70	0.81	1.14	1.34
DEV B	min	1.42	1.34	1.23	1.06	0.91	1.04	1.22	1.28
DEV B	max	2.89	2.84	2.86	3.07	3.42	4.12	5.58	6.04
DEV METRIC	mean	3.05	3.33	3.61	3.79	3.90	4.30	4.69	4.85
DEV METRIC	std	0.32	0.31	0.31	0.34	0.50	0.57	0.84	1.02
DEV METRIC	min	2.45	2.56	2.68	2.77	2.63	2.66	2.75	2.73
DEV METRIC	max	3.47	3.75	4.05	4.40	4.86	5.13	6.18	6.59
HET:HOM RATIO	mean	1.64	1.67	1.74	1.92	2.34	2.73	3.01	3.12
HET:HOM RATIO	std	0.08	0.09	0.11	0.13	0.23	0.39	0.52	0.58
HET:HOM RATIO	min	1.47	1.53	1.59	1.73	2.00	2.16	2.25	2.28
HET:HOM RATIO	max	1.81	1.89	2.00	2.16	2.89	3.61	4.05	4.18
KURTOSIS ALL	mean	0.19	0.19	0.19	0.20	0.22	0.25	0.29	0.30
KURTOSIS ALL	std	0.04	0.04	0.03	0.03	0.03	0.04	0.06	0.07
KURTOSIS ALL	min	0.06	0.07	0.09	0.13	0.15	0.16	0.17	0.17
KURTOSIS ALL	max	0.25	0.25	0.26	0.25	0.26	0.35	0.42	0.44
KURTOSIS HETS	mean	3.75	3.83	3.89	3.57	2.76	2.42	2.29	2.25
KURTOSIS HETS	std	0.24	0.26	0.34	0.33	0.14	0.18	0.20	0.22
KURTOSIS HETS	min	3.31	3.40	3.41	3.12	2.51	2.19	2.01	1.96
KURTOSIS HETS	max	4.27	4.39	4.43	4.17	3.07	2.73	2.65	2.63
PC X HETS	mean	34.72	36.09	39.39	47.19	57.42	62.52	64.78	65.33
PC X HETS	std	23.70	22.73	20.55	15.43	9.54	7.57	7.12	7.04
PC X HETS	min	7.67	8.53	11.24	20.72	37.43	45.70	48.34	49.74
PC X HETS	max	64.78	65.03	64.96	66.06	70.08	73.36	74.62	75.30
SKEW ALL	mean	1.38	1.39	1.42	1.47	1.57	1.63	1.67	1.68
SKEW ALL	std	0.03	0.04	0.04	0.04	0.04	0.04	0.06	0.07
SKEW ALL	min	1.33	1.34	1.35	1.41	1.51	1.54	1.56	1.56
SKEW ALL	max	1.46	1.48	1.51	1.56	1.64	1.73	1.80	1.82
SKEW HOMS	mean	-3.67	-2.71	-2.10	-1.75	-1.85	-2.27	-2.73	-2.95
SKEW HOMS	std	0.36	0.21	0.25	0.29	0.19	0.21	0.52	0.66
SKEW HOMS	min	-4.15	-3.11	-2.66	-2.36	-2.24	-2.73	-3.66	-4.03
SKEW HOMS	max	-2.58	-2.29	-1.74	-1.40	-1.56	-1.94	-2.22	-2.24

Table 2.2: Summary of measures used in regressions obtained from 200 simulations dataset highlighting the mean, standard deviation and range for each measure at each of the eight tested contamination ranges.

To ascertain if each of the measures were correlated with each other or the known levels of contamination and if any were potentially more useful in predicting contamination, a scatter matrix was plotted of all measures against each other using the 200 simulations as shown in Figure 2.6. The matrix contains 100 sub-plots with histograms plotted along the diagonals displaying the distributions of the measurements. Analysis of each measure individually against contamination, as shown in column ten and row ten, shows that for each of the measures there are repeatable patterns amongst the simulations. However the range of values in each of the simulations often significantly overlapped between the contamination levels meaning a combination of measures will be required to predict contamination.

From Figure 2.6 the relationships of all measures can be assessed, which also confirm the described patterns in sample contamination. In the first row the percentage of chromosome X heterozygotes are generally split into 2 groups with lower percentages indicating male samples. From the percentage of chromosome X heterozygotes to identify contamination groups would need to be formed outside of the expected ranges for males and females. As contamination is added the higher percentage female samples and lower percentage males intersect with samples all returning in excess of 60% chromosome X heterozygotes from upwards of 20% contamination.

The Het:Hom ratio appears the most powerful measure as a near linear increase in the ratio can be observed when plotted against the actual contamination, though the overlapping range of values at each level of contamination inhibits this from being able to predict contamination alone. Examination of the relationship of the Het:Hom ratio to other measures highlights the low contamination levels below 10% and above 20% can often be differentiated and the relationship with the measure of skewness for all variants is close to linear.

The deviation metric (Dev metric) is a sum of deviations A and B, due to the scale of Dev B being larger than Dev A the Dev metric more closely resembles the patterns seen in Dev B. Some separation is visible when plotting Dev metric against the Het:Hom ratio, skew all and kurtosis all but often contamination ranges blend continuously, making the measure likely to add some information but be insufficient

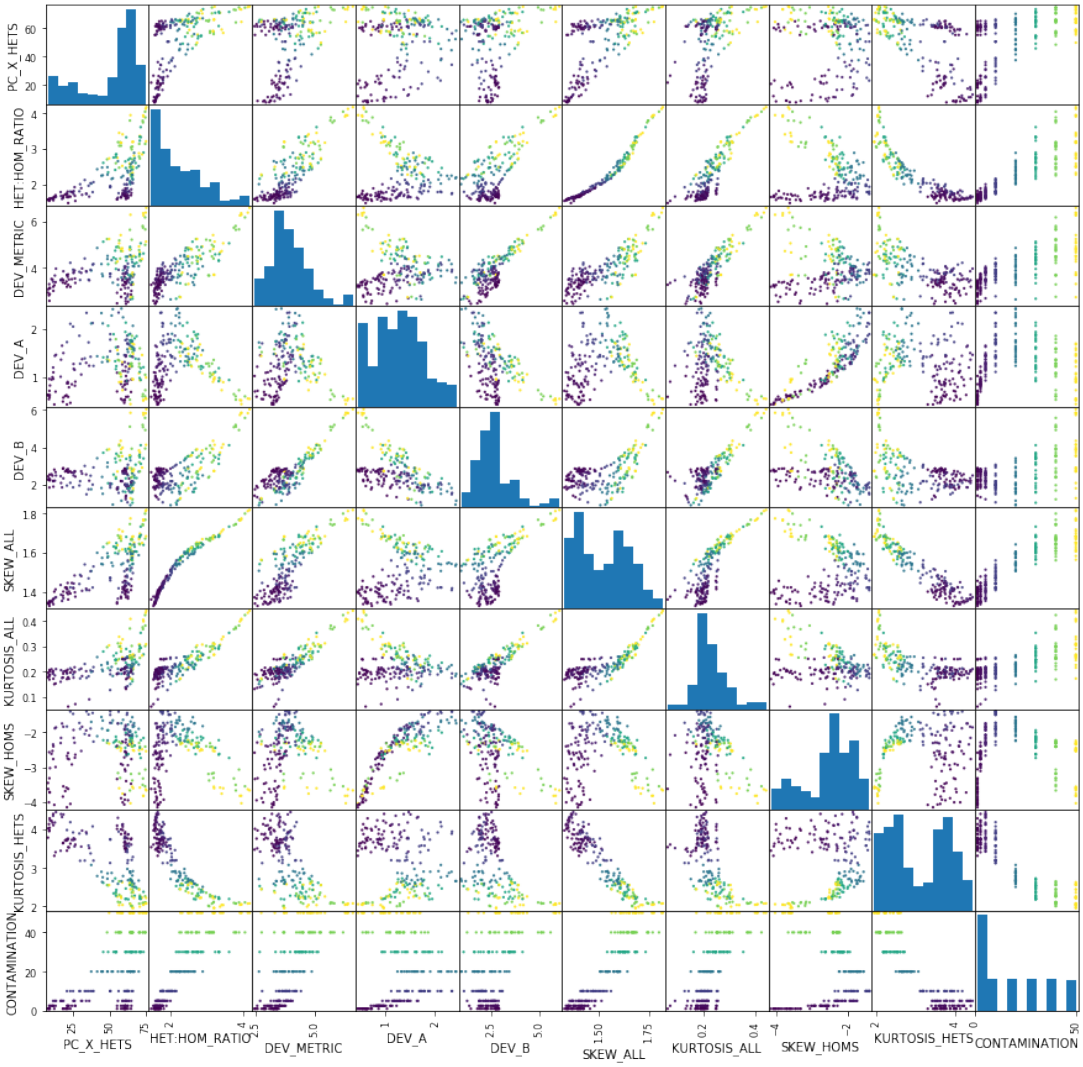

Figure 2.6: Scatter matrix of all nine measures and contamination levels. On the diagonal is a histogram for the measure showing the data distribution. In purple are the lowest contamination levels through to yellow with the highest levels of 50%. From the measures the Het:Hom ratio, skew all and kurtosis heterozygotes appear the most useful measures.

to predict contamination alone. Dev A does not display a clear linear relationship with any other measure, most interesting is when plotted against contamination levels the increase in deviation appears somewhat linear up to 20% before decreasing with further contamination. Dev B appears to be one of the less useful measures with the contamination levels often overlapping as shown particularly when plotting against known contamination. Whilst Dev B appears less correlated with other variables it is often possible to identify clusters, such as when plotted against kurtosis of heterozygotes where levels below 10% contamination separate from higher levels.

Skew over all variants appears to correlate well with the Het:Hom ratio and kurtosis for heterozygotes with a close to linear relationship. All of these relationships suggest that the skew measurement of the AAFs could be an important measurement for contamination prediction. As with the Het:Hom ratio when plotted against the known level of contamination the relationship appears almost linear but for the increased spread observed in 40 and 50 % contamination levels. Conversely the skew of homozygous variant AAFs appears to be able to separate the low levels of contamination below 10% but struggles with levels above 10% as would be expected from the trends in the AAF profiles. This is particularly highlighted when skew of homozygotes is plotted against actual contamination where the measures rise up to 10% before falling back to values seen at low contamination similar to Dev A.

Measurements of kurtosis using all variants appears less informative than just using the heterozygous variants. The all variant measurement displays near total overlap when plotted against the known contamination, particularly with low levels of contamination below 10%. The heterozygote only measure correlates well with most other measures but particularly well with the Het:Hom ratio. The heterozygote measure is also more effective at larger contamination levels above 10%, likely due to the changes in the higher AAF heterozygote variants which join from the homozygote distribution.

Along the diagonal of Figure 2.6 a histogram of each measure is shown. The Het:Hom ratio, Dev metric and Dev B show a right-skewed distribution. Due to Dev B being the largest value in the Dev metric this explains the similar distributions between the two while Dev A is less left-skewed and closer to a normal distribution. Conversely the percentage of X heterozygotes and skew homozygotes display left-skew. Both kurtosis heterozygotes and skew all variants measures are found with a bi-modal distribution.

2.3.2 Principal component and clustering analysis

Having established measures individually could not be reliably used to predict contamination principal component and clustering analysis was performed as a proof of concept to see if the contamination levels could be separated. As shown in Figure 2.7 there is some separation of contamination up to 10%. Most interesting was that clusters are ordered by ascending contamination levels vertically by principal component 2. From 20% contamination and upwards clusters overlap substantially making differentiation problematic.

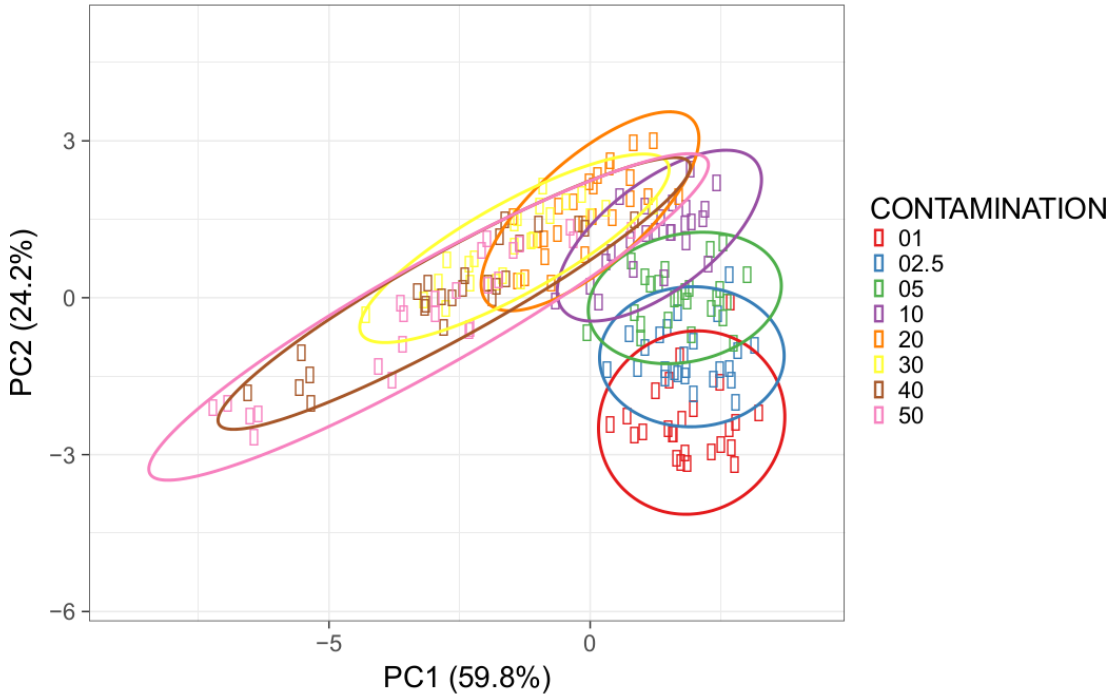

Figure 2.7: Clustering analysis of contamination simulations coloured by contamination level. Separation of contamination up to 10% can be partly observed. 20% contamination and upwards display majority overlap making differentiation problematic. Principal component 1 and principal component 2 explain 59.8% and 24.2% of the total variance, respectively. Prediction ellipses are such that with probability 0.95, a new observation from the same group will fall inside the ellipse.

Nine principal components were generated with the variance explained shown in Table 2.3. 59.8 and 24.2% of variance is explained by principal components 1 and 2. Therefore principal components 3 to 9 only explain 16% of the variance, 8.6% of which comes from principal component 3.

0	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Individual	0.598	0.242	0.086	0.042	0.018	0.010	0.002	0.001	0.000
Cumulative	0.598	0.840	0.927	0.969	0.987	0.997	0.999	1.000	1.000

Table 2.3: PCA variance explained by components for contamination measures. Principal component 1 and principal component 2 explain 59.8% and 24.2% of the total variance, respectively. Therefore principal components 3-9 only explain 16% of variance.

Analysis of the loadings of principal components 1 and 2 was performed as shown in Table 2.4. Principal component 1 has a largest value on the positive axis of 0.36 for kurtosis of heterozygotes and a largest negative axis value of -0.419 for the Het:Hom ratio, suggesting these two features explain the extremities of the dataset. Principal component 2 is loaded such that Dev A scored 0.64, followed by skew homozygotes with 0.59. Whilst the lowest loading was for Dev B with -0.305. Neither of the principal components 1 or 2 are therefore using measurements solely looking at either heterozygote or homozygote distributions. Therefore regression analyses may need to use all of the features to obtain better predictions despite the lack of correlation and separation observed previously for individual measures.

Measure	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
PC X HETS	-0.271	0.263	-0.597	-0.673	-0.108	0.170	-0.082	-0.005	0.000
HET:HOM RATIO	-0.419	0.055	-0.040	0.208	0.090	-0.394	-0.782	-0.044	0.000
DEV METRIC	-0.383	0.017	0.484	-0.122	-0.108	0.463	-0.095	-0.023	0.607
DEV A	0.059	0.643	0.163	0.135	0.455	0.435	-0.126	0.015	-0.354
DEV B	-0.356	-0.305	0.331	-0.171	-0.319	0.178	-0.018	-0.027	-0.712
SKEW ALL	-0.411	0.163	-0.072	0.227	-0.047	-0.120	0.306	0.798	0.000
KURTOSIS ALL	-0.400	-0.087	0.156	-0.245	0.635	-0.340	0.395	-0.270	0.000
SKEW HOMS	0.097	0.591	0.389	-0.221	-0.421	-0.487	0.115	-0.115	0.000
KURTOSIS HETS	0.367	-0.191	0.302	-0.529	0.278	-0.115	-0.306	0.523	0.000

Table 2.4: PCA variance explained by each component for contamination measures. Principal component 1 has a largest value of 0.36 for kurtosis of heterozygotes and a smallest value of -0.419 for the Het:Hom ratio. Principal component 2 is loaded such that deviation A scored highest with 0.64 whilst the lowest loading was for Dev B with -0.305

2.3.3 Training regression models

Training of regression models was performed using the 25 contamination simulation sets each at the 8 levels of contamination with the 70:30 training and testing split. For the regressions the R^2 values were calculated for testing with all features and a subset of features looking only at the homozygote distribution (Skew homozygotes, Dev A and Het:Hom ratio). Also the range of the contamination levels used from the simulations to train models were tested from 10% contamination and upwards to investigate how the spread seen in measures at higher contamination levels affects predictions. R^2 values for each of the combinations are shown below in Table 2.5. Training values used are summarised in Appendix A, Table 8.2 and on a GitLab repository (https://gitlab.com/rs91/exome-contamination).

Features	Contamination	OLS	Ridge	Lasso	SVR	RBF	Polynomial(2)	Polynomial(3)	Polynomial(4)	Polynomial(5)
All	1-10	0.91	0.89	0.8	0.87	0.44	0.28	-1.63	-0.64	-4.8
Selected	1-10	0.9	0.89	0.8	0.84	0.85	0.9	0.94	0.52	-39.26
All	1-20	0.91	0.91	0.89	0.9	0.37	0.91	0.05	0.04	-0.71
Selected	1-20	0.91	0.89	0.88	0.86	0.75	0.91	0.96	0.1	-0.78
All	1-30	0.89	0.89	0.87	0.88	0.36	0.82	-1.02	-1.05	-1.17
Selected	1-30	0.88	0.85	0.85	0.85	0.76	0.9	0.93	0.9	0.42
All	1-40	0.86	0.86	0.84	0.84	0.35	0.79	-120.08	-6	-4.18
Selected	1-40	0.84	0.82	0.82	0.81	0.77	0.89	0.92	0.87	0.61
All	1-50	0.82	0.82	0.81	0.79	0.3	0.72	-18.84	-29.93	-22.47
Selected	1-50	0.78	0.76	0.76	0.75	0.72	0.86	0.88	0.89	0.82
Average	NA	0.87	0.86	0.84	0.84	0.57	0.8	-13.69	-3.43	-7.15
Average - All	NA	0.88	0.87	0.84	0.86	0.36	0.7	-28.3	-7.52	-6.67
Average - Selected	NA	0.86	0.84	0.82	0.82	0.77	0.89	0.93	0.66	-7.64
Maximum	NA	0.91	0.91	0.89	0.9	0.85	0.91	0.96	0.9	0.82

Table 2.5: Regression R^2 values obtained for each algorithm with contamination ranges tested up to 50% with both the selected homozygote features and all features. The highest scoring linear regression was OLS with an average R^2 score of 0.88 and a maximum of 0.91 using the 1-20% training range with all features. The all features regressions were also on average better performing than the selected features for linear models. Polynomial models produced more variable results and performed better when using only the selected features. The top scoring regression was a polynomial of degree three, which was consistently the highest performing regression peaking also in the 1-20% range at 0.96.

Comparing each of the linear regression models shows that the OLS regression was often the highest scoring linear regression as indicated by three averages shown in Table 2.5. On average for the linear regressions when using all features R^2 values were on average 0.01 to 0.03 higher than only the selected features. OLS regressions scored an average R^2 score of 0.88 and a maximum of 0.91 using the 1-20% training range with all features. Performance of non-linear regressions display much greater

variability with regressions performing poorly down to a score of -120, up to the best performing of all regressions at 0.96 for a polynomial of degree three with a training range also of 1-20% using selected features. All of the polynomial regressions performed better with the selected features rather than all features.

Further investigations were also performed to determine if using alternative contamination ranges which do not start with 1% could improve the R^2 scores. As shown below in Table 2.6 by the maximum R^2 scores and averages using these alternative ranges decreases the performance of regressions.

Features	Contamination	OLS	Ridge	Lasso	SVR	RBF	Polynomial(2)	Polynomial(3)	Polynomial(4)	Polynomial(5)
All	10-20	0.76	0.78	0.73	0.77	-0.15	-817.75	0.35	-0.19	-1.21
Selected	10-20	0.76	0.71	0.60	0.43	-0.06	0.85	0.51	-0.52	-15.40
All	10-30	0.77	0.79	0.767	0.79	0.17	0.54	-3.53	-0.58	-0.71
Selected	10-30	0.73	0.71	0.67	0.65	0.50	0.87	0.71	0.63	-53.47
All	10-40	0.74	0.74	0.74	0.71	0.17	0.56	-11.92	-10.41	-13.25
Selected	10-40	0.63	0.62	0.58	0.60	0.59	0.83	0.79	0.32	-2.50
All	10-50	0.69	0.68	0.69	0.65	0.31	0.20	-96.40	-38.87	-31.81
Selected	10-50	0.57	0.57	0.57	0.55	0.56	0.76	0.76	0.64	0.18
All	20-30	0.67	0.61	0.57	0.45	-0.46	-1110.70	-1.42	-1.66	-2.49
Selected	20-30	0.52	0.51	0.48	0.37	0.36	0.59	-0.89	-53.64	-2732.86
All	20-40	0.70	0.59	0.55	0.51	0.10	-0.62	-28.14	-26.13	-26.67
Selected	20-40	0.50	0.50	0.52	0.49	0.44	0.61	0.53	-2.71	-120.00
All	20-50	0.66	0.51	0.46	0.43	0.09	0.07	-181.03	-101.50	-106.19
Selected	20-50	0.46	0.44	0.46	0.41	0.35	0.52	0.45	-0.05	-9.52
All	30-40	0.20	0.18	0.26	-0.05	-0.59	-849.47	-8.73	-10.46	-13.14
Selected	30-40	0.23	0.21	0.19	0.05	-0.02	0.11	-1.03	-54.60	-4436.51
All	30-50	0.30	0.21	0.17	0.13	0.00	-6.20	-117.27	-104.31	-114.30
Selected	30-50	0.26	0.22	0.17	0.15	0.14	0.05	-0.11	-4.30	-358.47
All	40-50	-0.45	-0.05	-0.05	0.30	-0.58	-4359.71	-133.07	-115.17	-118.79
Selected	40-50	-0.05	-0.03	-0.05	-0.31	-0.50	-0.27	-0.32	-238.13	-11859.20
Average	NA	0.48	0.47	0.45	0.40	0.07	-356.91	-28.99	-38.08	-1000.82
Average - All	NA	0.61	0.57	0.55	0.49	-0.04	-309.26	-49.79	-32.74	-34.42
Average - Selected	NA	0.52	0.50	0.47	0.41	0.32	0.58	0.19	-12.69	-858.73
Maximum	NA	0.77	0.79	0.77	0.79	0.59	0.87	0.79	0.64	0.18

Table 2.6: Evaluation of alternate contamination levels which do not use data below 10% to train linear and non-linear regression models. No regressions obtained higher R^2 scores compared to using ranges which start from 1% contamination as shown in Table 2.5.

Training using 1-50% contamination simulations

To investigate the patterns of R^2 values the full range training set of 1-50% contamination was visualised with the predicted contamination levels plotted against the actual contamination. For the linear regression models as shown in Figure 2.8 all linear regressions perform similarly with scores ranging from 0.75 to 0.78. All of the regressions perform poorly at contamination levels above 20% as the predicted contamination values overlap near completely. At low levels of contamination regressions also struggle to differentiate the lower levels of contamination between one and five percent.

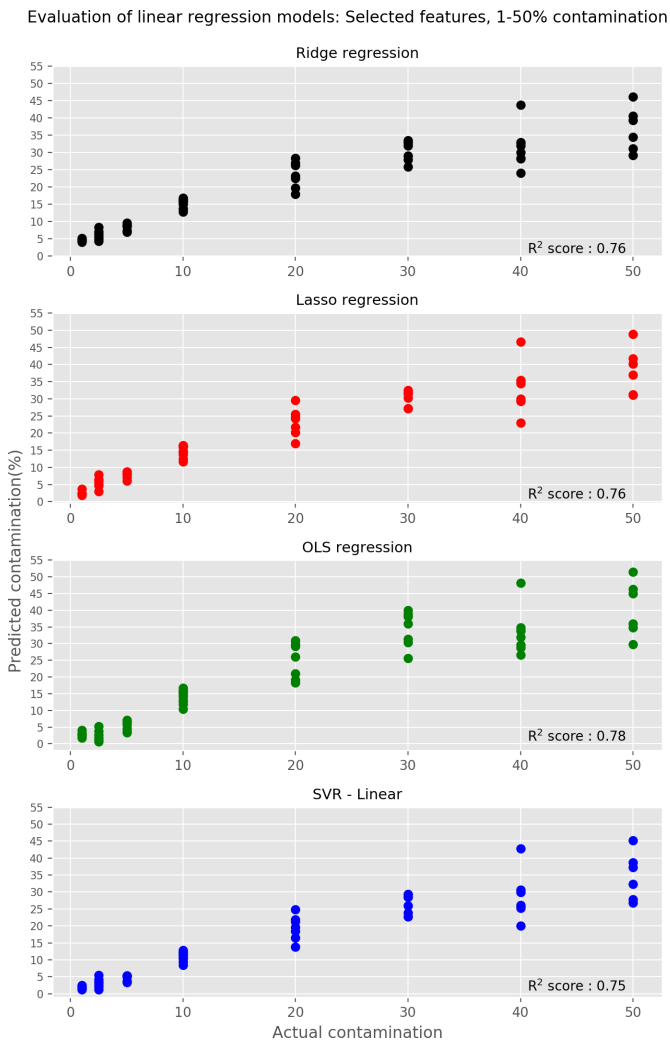

Figure 2.8: Evaluation of linear regressions using 1-50% training data with all features. All linear regressions perform similarly with scores ranging from 0.75 to 0.78 for the OLS regression. Models perform particularly poorly in predicting the contamination from upwards of 20%

Evaluation of non-linear regressions using selected features over the 1-50% training range yields a comparative improvement in scores compared to linear models with all polynomial regressions outperforming the highest linear model. As shown in Figure 2.9 the performance of polynomial regressions are similar at the lower levels of contamination but become better at predicting the higher levels and overlaps between contamination levels appear marginally reduced.

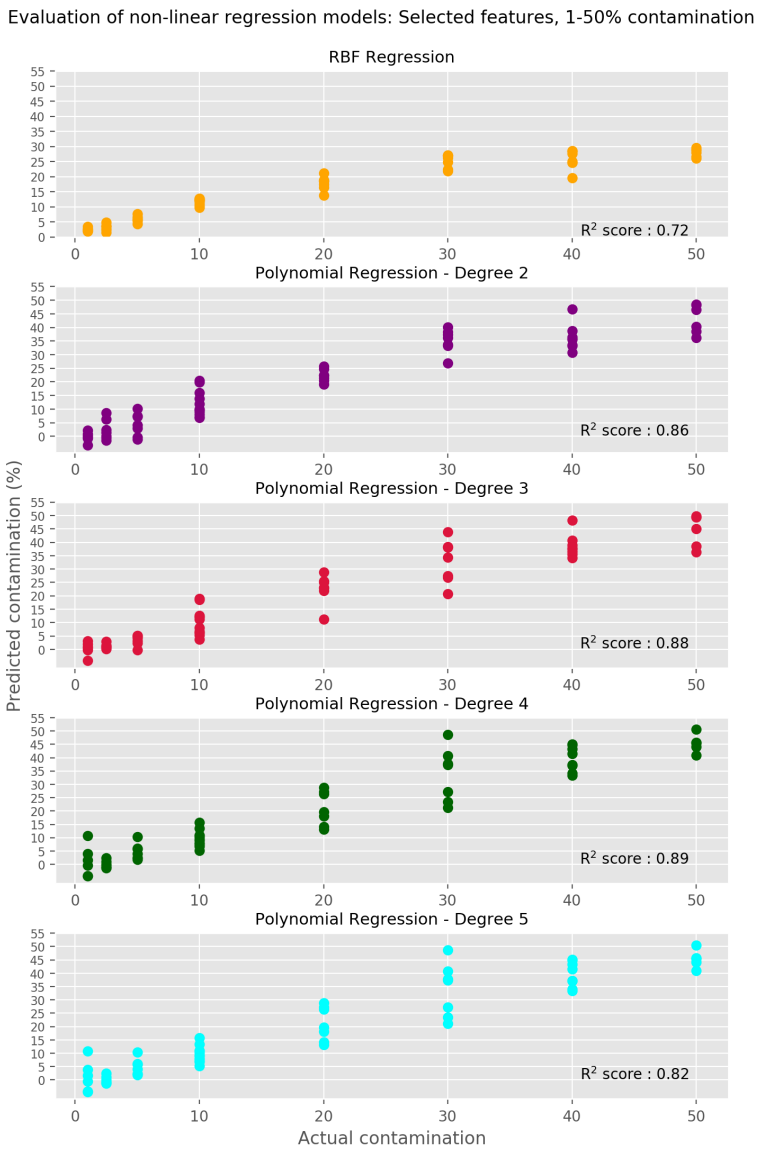

Figure 2.9: Evaluation of non-linear regressions using 1-50% training data with selected features. All polynomial regressions show improvements over the linear regressions based on R^2 values but still show some overlaps between contamination levels.

Training using 1-10% contamination simulations

Regressions were restricted to training with lower levels of contamination from 1-10% for the all and selected features as shown in Figure 2.10. It was expected that restriction of the training data would give increased the R^2 values for the regressions due to contamination predictions from regressions becoming more variable at higher contamination levels. Use of all features retained an R^2 score advantage over the selected homozygote features as shown in Table 2.5 for linear regressions. The OLS regression when using all features appears closest to the actual contamination when comparing the predicted values and shown to be the highest R^2 for the 1-10% trained regression.

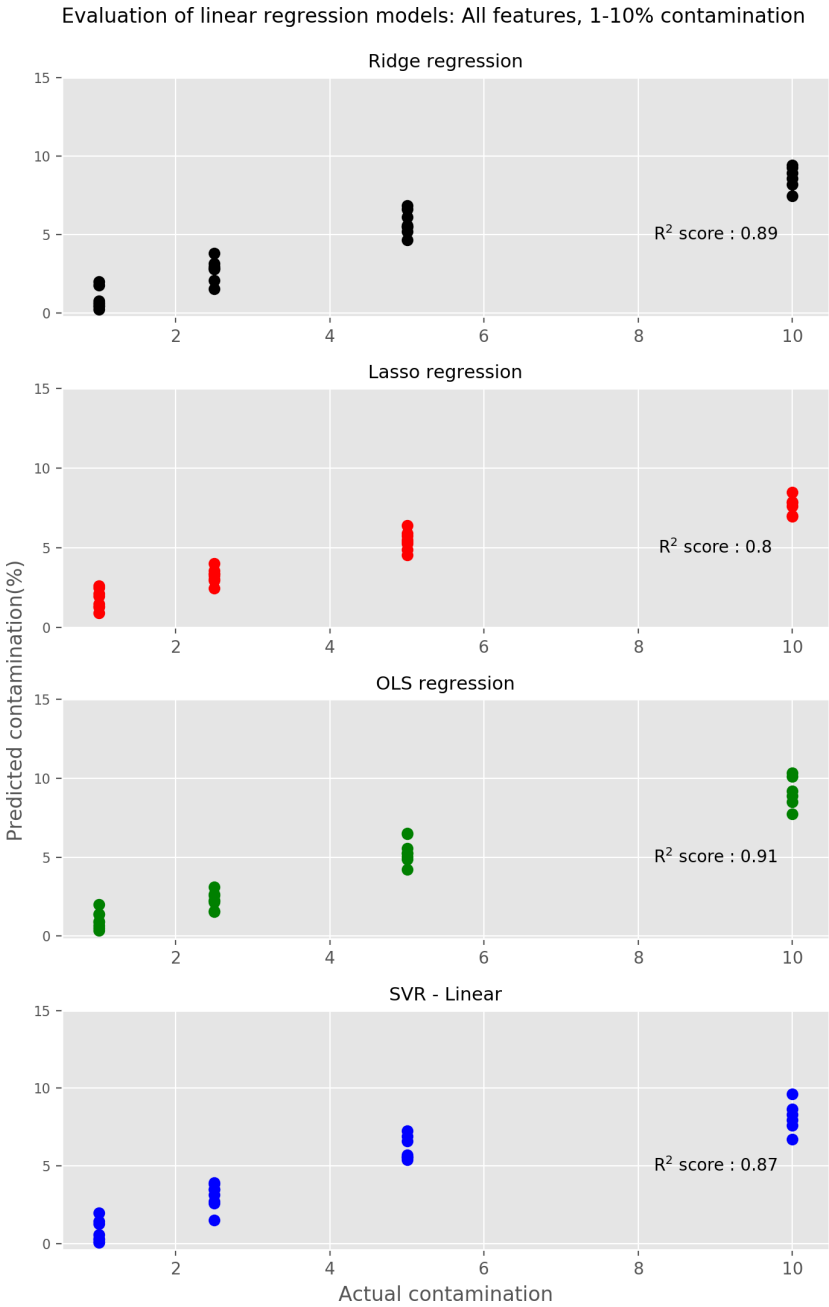

Figure 2.10: [Regression analyses trained using all features for contamination levels up to 10%. Training samples only up to 10% shows the best agreement is obtained using the OLS regression as the range of predictions tend to be centred closest around the actual contamination levels up to 5%. All of the regressions consistently underestimate the predicted contamination for 10% actual contamination. Due to most of the changes in AAF profiles happening in the homozygote distribution below 10% performance between all features and the selected measures is similar.

When non-linear regressions, shown in Figure 2.11, are evaluated using selected features high R^2 scores are obtained of 0.9 and 0.94 using degrees of two and three. Though the 1% contamination for the degree of two does deviate to below 0% contamination into negative values for one sample. Degrees of four and five perform poorly and are worse than the linear regressions. RBF performs similar to the linear regressions but by 10% contamination the regression begins to lose accuracy and shows a wide range of predictions.

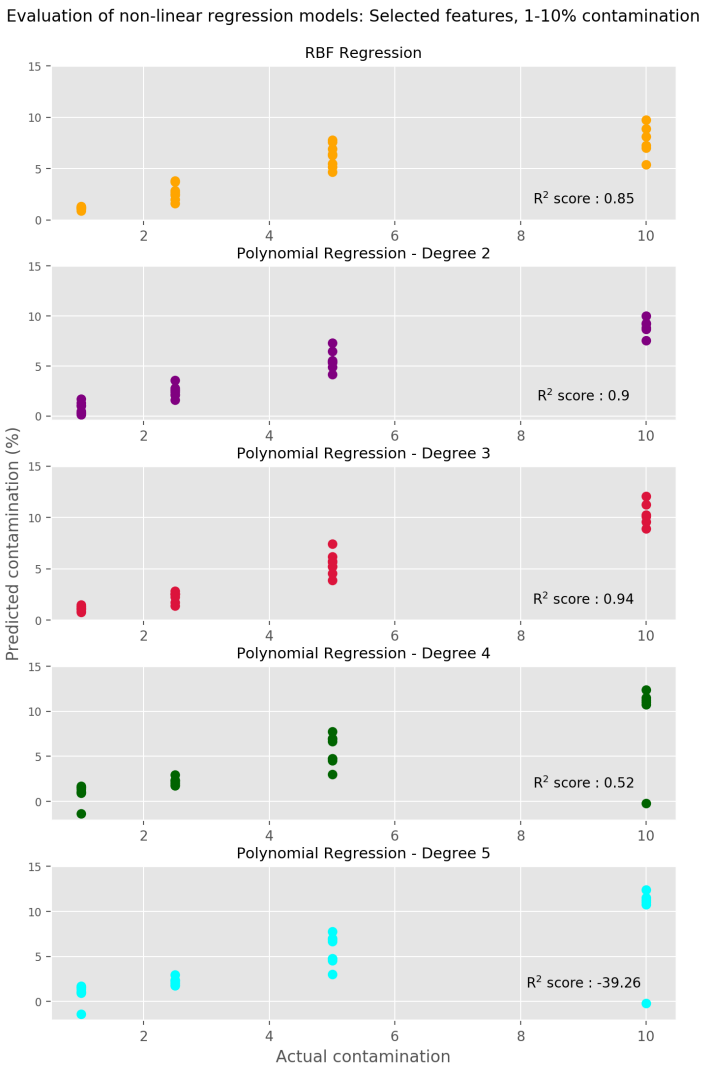

Figure 2.11: Non-linear regression models using the 1-10% training range. R^2 Scores are obtained of 0.9 and 0.94 using degrees of two and three which are the best performing regressions. Some of the predictions for the degrees of two and three. RBF performs similar to linear regressions but poorly at 10%. Higher degrees of four and five perform poorly at predicting contamination.

Training using 1-20% contamination simulations

When evaluating R^2 scores as shown in Table 2.5 the maximum values were consistently obtained when using a range of 1-20% for both linear and non-linear regressions. The highest linear value was 0.91 using OLS, as shown below in Figure 2.12. The highest non-linear was 0.96 when using a polynomial regression with a degree of three, as shown below in Figure 2.13.

For linear regressions all display a variety of predictions at the 20% actual contamination range which can range from down to 15 and up to 26%. The OLS regression appears to have the tightest distribution of predictions around the actual contamination value and explains the marginally highest R^2 score.

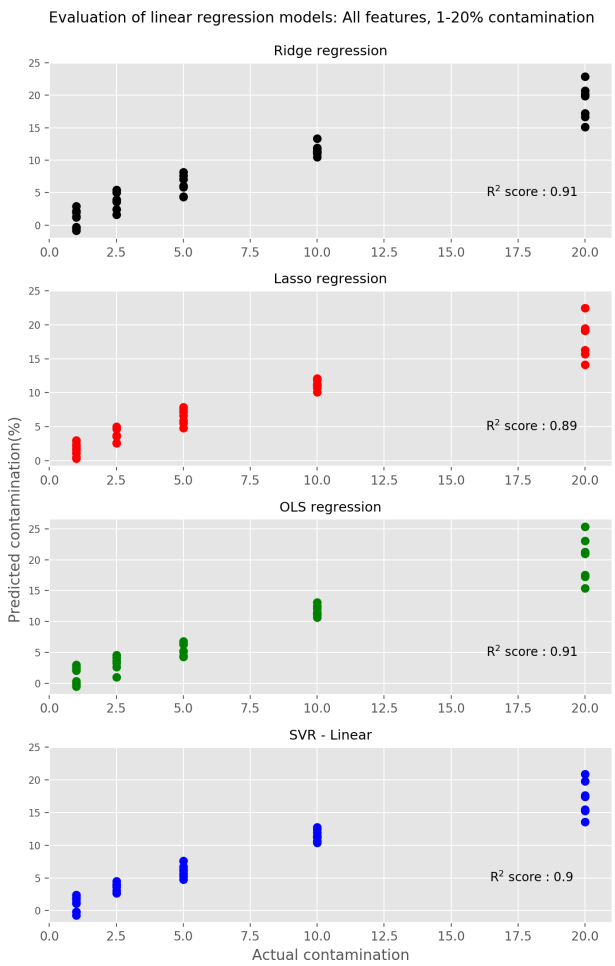

Figure 2.12: Evaluation of non-linear regressions using 1-20% training data with all features. All regressions perform similarly with the highest score of 0.91 being obtained by an OLS regression. Predictions display a large spread at 20% actual contamination for all regressions.

Evaluation of non-linear regressions shows that higher R^2 scores are obtained with polynomial regressions of degree two or three at 0.91 and 0.96 respectively as shown in Figure 2.13. All other polynomial regressions perform worse than the linear regressions. Comparing the best two non-linear regressions the degree of three polynomial displays a tighter prediction range closer to the actual contamination level than the degree of 2. The most visible improvement for the degree of three is at the 20% actual contamination where predictions are closest of any regression.

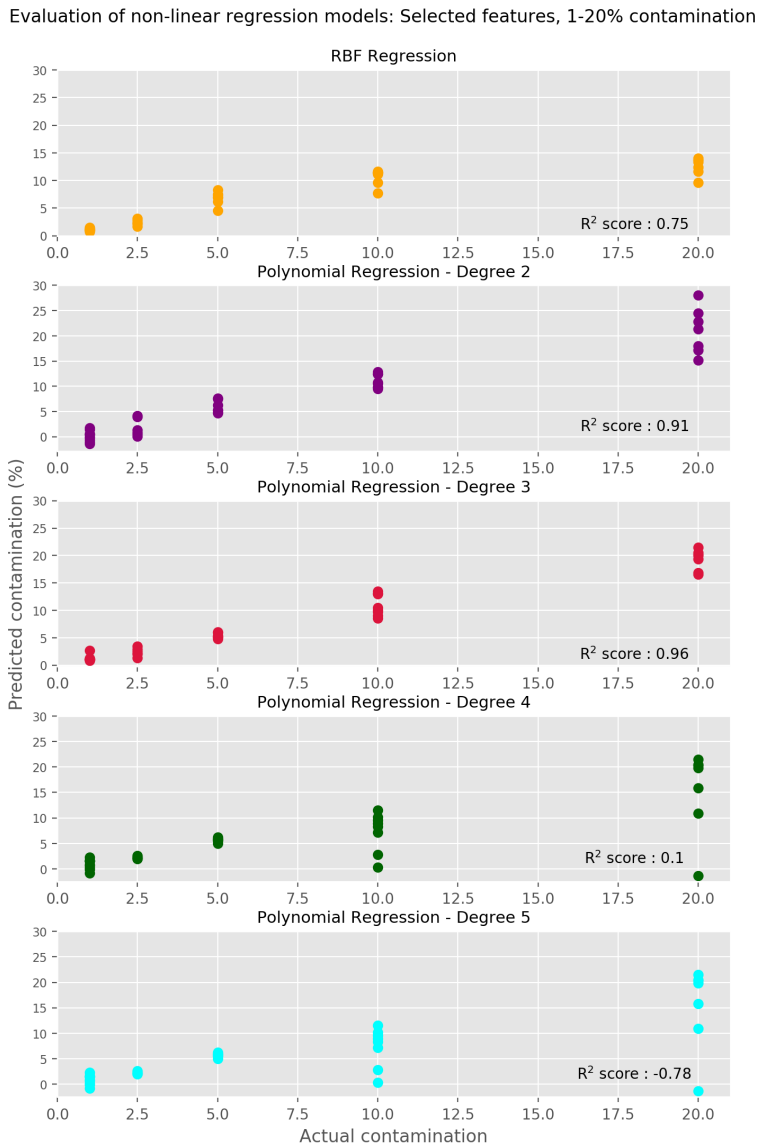

Figure 2.13: Evaluation of non-linear regressions using 1-20% training data with selected features. Polynomial regressions with degrees two and three gave the highest scores of any regressions run. The degree of three regression appears to give the most accurate prediction, which is supported by highest R^2 score of 0.96.

2.3.4 Application of regression models

The best performing regressions from training were applied to the 245 germline samples. All samples were assumed to be be either un-contaminated or with unknown levels of contamination. Estimates of contamination were also calculated using VerifyBamID.

The best performing regressions overall in training were the polynomials of degree two and three with the 1-20% range. Testing with all samples though failed to yield the same results as in training as shown by Figure 2.14. Panel A of the figure shows predictions when using selected features and Panel B using all features. Both of the non-linear regression contamination predictions show the majority of samples are predicted to have implausible negative contamination with all features. Testing was also performed to extend the training range to 1-50% with the polynomial regressions which further increased the range of contamination predictions to higher positive and negative values.

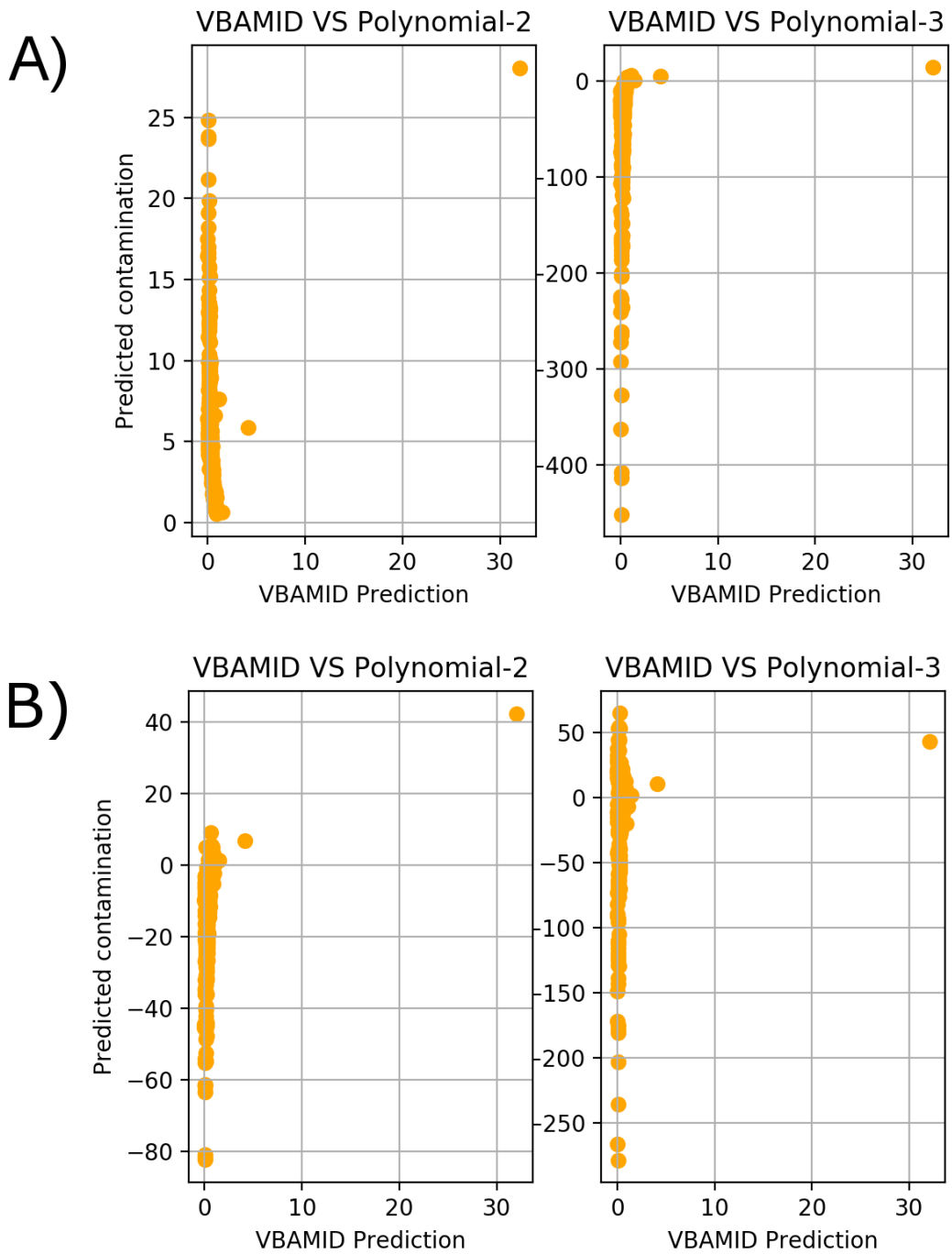

Figure 2.14: Polynomial regressions tested with 245 germline samples with training set 1-20%. **A)** Selected features with a degree of two shows a large disagreement with VerifyBamID predictions and a range of predictions up to 25%. For the degree of three nearly all samples are predicted with negative contamination percentages. **B)** All features regressions show for the degree of two most samples are predicted to have negative contamination. With the degree of three samples are classified over a wide range from 60 to -275 % predicted contamination.

The best performing linear regression model was OLS, from testing of models it was suspected that the regression analyses became compromised when extending the levels of contamination used beyond 10%. To confirm this the predicted contamination levels from OLS regressions when using the different contamination ranges from simulations in regression training were plotted as shown in Figure 2.15, full results are shown in Appendix A, Table 8.3. Initially with 1-10% trained results samples cluster well when plotted against the VerifyBamID predictions in the lower left corner, with good agreement for a sample predicted as 4% contamination by VerifyBamID and just over 5% by the OLS regression. A single sample was predicted as having 32.0% contamination by VerifyBamID which was higher than the predicted regression value at 25.2%. The underestimation of higher contamination levels is in keeping with the trend observed from testing with the 1-10% simulation set.

As the training sets were widened to include higher contamination levels it was clear that the regression becomes more balanced in favour of calling higher contamination levels. The shape of the distribution remains largely identical but appears to stretched vertically upwards on the y-axis. The closest agreement for the sample with 32% contamination with VerifyBamID was obtained when using the 1-40% contamination training set with 33.0% however by this point most of the predictive power at low contamination levels has been diluted and showing large overestimation of the contamination levels 20%.

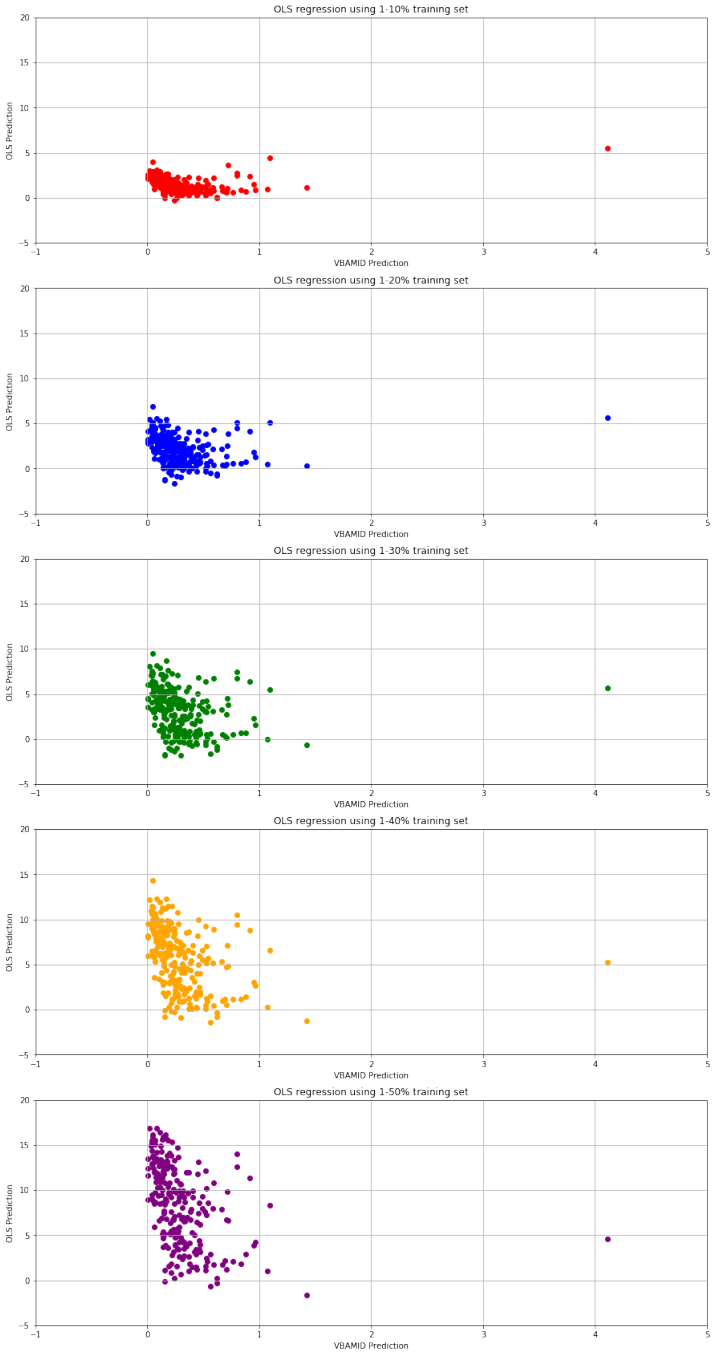

Figure 2.15: OLS regressions using all features with different ranges of contamination used in training restricted to show samples below 5% by VerifyBamID predictions. A single sample of high contamination predicted as 32% VerifyBamID was not shown, each regression A-E predicted values of: 25.2, 27.7, 29.6, 33.0 and 37.22% percent contamination respectively. **A)** 1-10% - Samples tightly cluster at lower levels and show general agreement with VerifyBamID. At higher contamination levels the regression underestimates contamination. **B)** 1-20% - As the training data ranges are increased contamination predictions increase. **C)** 1-30% - Contamination predictions continue to increase and are visibly is stretched upward along the Y-axis. **D)** 1-40% - Contamination predictions further increase **E)** 1-50% displays similar results to 1-40% with contamination now overestimated for suspected low contamination samples.

The number of samples predicted above 1, 2.5 and 5% contamination by the OLS regressions using each of the different training ranges are shown in Table 2.7. As the ranges of contamination used to train regressions were increased the number of samples estimated above 1, 2.5 or 5% increases. While 165 samples are estimated as above 1% contamination and 23 above 2.5% contamination when using 1-10% training dataset. In comparison 172 samples are identified above 1% contamination and 104 above 2.5% contamination when training with a 1-20% dataset.

Training Range (Pc)	Samples above 1pc contamintaion	Samples above 2.5pc contamintaion	Samples above 5pc contamintaion
1-10	165	23	2
1-20	172	104	11
1-30	182	149	67
1-40	219	190	145
1-50	237	216	180

Table 2.7: Samples predicted by OLS regressions above thresholds for contamination by training range. As the contamination range used to train the regression is increased the number of samples predicted at higher contamination ranges were observed to increase.

A comparison was performed between the five highest predictions of contamination from VerifyBamID and each of the OLS regressions are also shown in Table 2.8 along with several features. As the OLS regressions are trained with a wider range of contamination simulations the predictions become more variable with some samples actually closer to the VerifyBamID predictions but more often get further away from the VerifyBamID predictions. The closest values to the VerifyBamID predictions most often come from the 1-10% training set when comparing lower levels of contamination predictions. When training with simulation sets including above 30% contamination the sample 86 is predicted with negative contamination and so would not be in the highest five contamination predictions for these ranges.

SAMPLE	Depth	HET:HOM	DEV A	DEV B	Skew-All	Kurtosis-Hets	VerifybamID	OLS-10	OLS-20	OLS-30	OLS-40	OLS-50
4	28.45	2.95	1.87	0.96	1.67	2.05	32.04	25.27	27.69	29.59	33	37.42
217	115.89	1.72	1.38	2.69	1.42	3.91	4.11	5.53	5.65	5.67	5.27	4.59
86	89.39	1.57	0.7	2.16	1.35	4.03	1.43	1.12	0.32	-0.6	-1.25	-1.63
69	232.77	1.91	0.86	3.04	1.47	4.02	1.1	4.42	5.1	5.53	6.63	8.34
124	94.97	1.61	0.57	2.32	1.37	3.87	1.07	1.01	0.5	0	0.28	1.06

Table 2.8: VerifyBamID predicted samples with above 1% contamination. Five samples were identified with above 1% predicted contamination. Predictions were often best using the OLS regression with the 1-10% contamination training set. Most samples at low contamination levels as predicted by VerifyBamID agreed with the OLS except for sample 69 which had double the mean depth of coverage of the next highest sample and had the largest Het:Hom ratio. Sample 4 with 32% contamination by VerifyBamID was underestimated by the regression analysis when using 1-10% and predicted closest once a set of 1-40% was used which worsened other sample predictions.

Sample 4 was predicted by VerifyBamID as 32% contamination but was under-predicted by OLS regressions unless a training set included levels upwards of 40% contamination. The AAF histogram of sample 4 is shown in Figure 2.16. The most obvious characteristic in this profile is the lack of peaks in the heterozygote range from 1 to around 88%, the distribution also seems shifted towards lower AAFs. Finally there is no peak for the homozygotes below the 100% AAF which would be seen at lower levels of contamination supporting the high contamination call upwards of 20%. The sample also is of low depth of coverage at 28.45x on average.

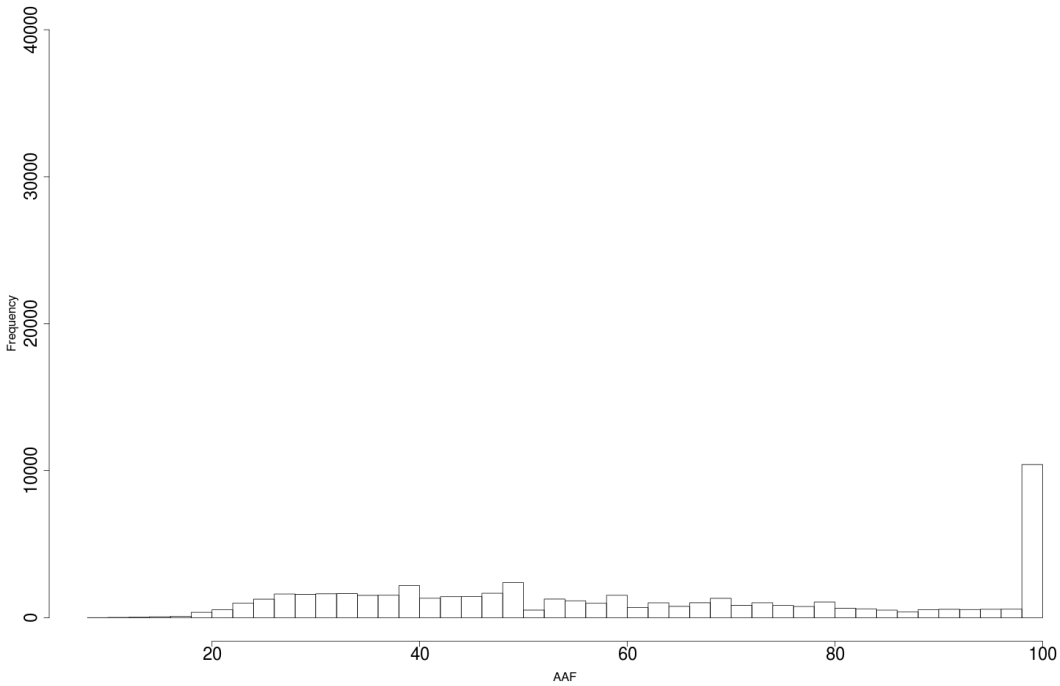

Figure 2.16: Alternate allele frequency histogram of sample 4 predicted to have 32 % contamination by VerifyBamID but only 25.27% by regression. The AAF histogram shows the lack of a normal distribution in the range of heterozygotes and shift towards lower allele frequencies suggesting contamination. The relative normality of the region below the 100% peak also suggests contamination levels upwards of 20%.

The second highest contaminated sample as predicted by VerifyBamID was sample 217. The OLS 1-10% regression predicted 5.53% compared to 4.11% for VerifyBamID. The AAF profile is shown in Figure 2.17 which shows a peak below the 100% for homozygotes which does not return down to the background level for approximately 2 bins, roughly equal to 4% supporting the prediction from the regression. Mean depth of coverage for the sample was 115.89x.

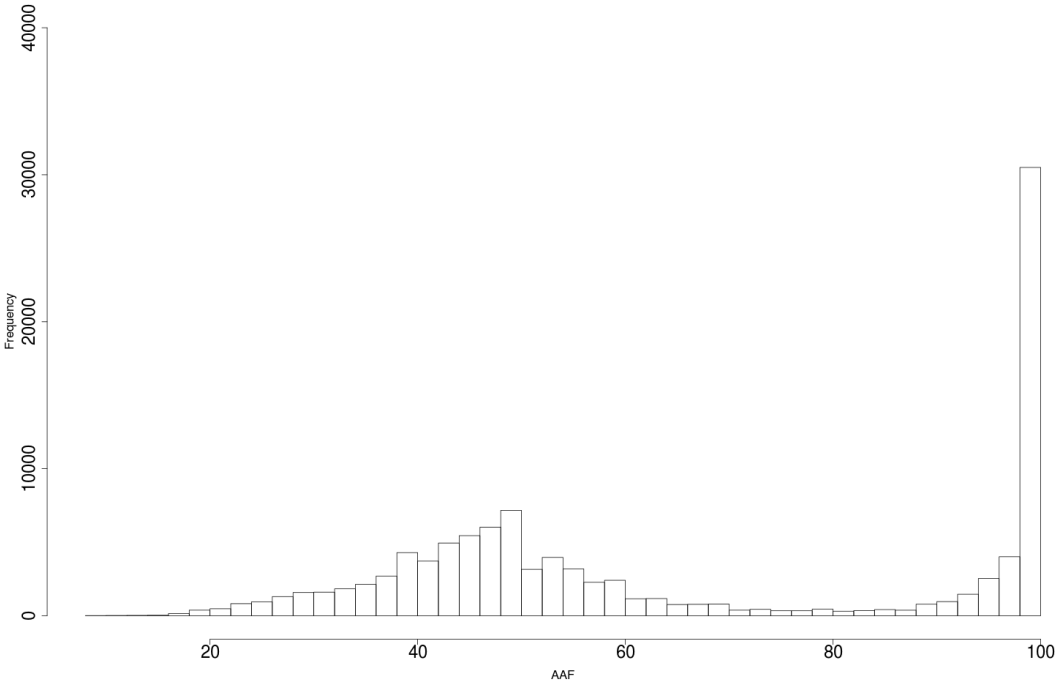

Figure 2.17: AAF histogram of sample 172 with 4.11% predicted contamination by VerifyBamID and 5.53 by OLS regression. From the AAF profile a clear peak below the 100% peak can be seen indicating the contamination. Approximately 2 histogram bins can be seen as clearly elevated indicating contamination in the range of 3-4%.

High prediction values can be explained by looking at the coefficients for the features used in the regressions as shown in Table 2.9. The highest coefficient was the Het:Hom ratio at 17.13 for all features at 1-10% training range. As shown in Table 2.8, sample 86 had the second highest Het:Hom ratio amongst the selected five samples. The mean depth of coverage at 232.77x for sample 69 is double the next highest depth. The prediction is likely not called as high as sample 69 due to the other features such as Dev A and the skew all off-setting the Het:Hom ratio.

Features	Range	SKEW ALL	KURTOSIS ALL	SKEW HOMS	HET:HOM	DEV A	DEV METRIC	DEV B	KURTOSIS HETS	PC X HETS
Homozygotes	1-10	N/A	N/A	-0.21	9.88	5.43	N/A	N/A	N/A	N/A
All	1-10	-19.78	-10.42	-0.75	17.13	4.11	2.23	-1.88	-0.93	-0.01
All	1-20	-21.20	12.60	-1.27	13.88	5.42	2.92	-2.50	-5.10	-0.03
All	1-30	42.04	20.69	-1.65	12.46	6.23	4.91	-1.32	-8.87	-0.01
All	1-40	-19.85	41.18	-2.40	5.46	6.87	4.56	-2.31	-13.14	-0.02
All	1-50	8.15	47.04	-2.52	-0.53	5.45	3.41	-2.04	-17.35	-0.04

Table 2.9: Regression coefficients for OLS regressions when using heterozygote only features at the 1-10% training range and all features from 1-10% to 1-50%. The Het:Hom ratio was the highest coefficient when using the 1-10% or 1-20% training sets. As the ranges in the training sets is increased to above 30% the Het:Hom ratio begins to decrease and other measures such as kurtosis all increase in score as measures focusing of the heterozygote distribution changes increase in score.

The five samples listed in Tables 2.8 were the only five predicted with above 1% contamination by VerifyBamID. 80 of the 245 samples were predicted with below 1% contamination, 103 samples were predicted between 1 and 2% contamination as shown by the full results in Appendix A, Section 8.1.4.

Review of the top five highest predicted contaminated samples by OLS 1-10% trained regression and below 1% predicted by VerifyBamID was performed as shown in Table 2.10. All samples are male with depths above 30x with some of the highest Het:Hom ratios seen across the dataset at between 1.79 to 2.00 .

SAMPLE	Depth	PC X HETS	HET:HOM	DEV A	DEV B	SKEW ALL	KURTOSIS ALL	SKEW HOMS	KURTOSIS HETS	VBAMID	OLS-10
230	43.48	25.30	2.00	0.37	2.80	1.53	0.29	-5.22	3.20	0.05	3.96
68	223.82	21.11	1.87	0.75	2.92	1.46	0.29	-3.69	4.09	0.72	3.62
90	39.11	22.66	1.79	0.33	2.84	1.45	0.19	-5.43	3.06	0.08	3.10
231	45.44	22.61	1.80	0.34	2.84	1.46	0.20	-5.52	3.13	0.02	3.03
102	33.74	25.07	1.85	0.33	2.87	1.47	0.25	-5.60	3.36	0.05	2.95

Table 2.10: Highest five predictions from OLS regression but not VerifyBamID. All samples are male with depths above 30x and some of the highest Het:Hom ratios seen in the dataset at between 1.79 to 2.00. Due to the high coefficient for the Het:Hom ratios in models trained with restricted contamination ranges the predictions appear to be exaggerated by OLS.

The AAFs for samples were visualised to investigate the predictions as shown in Figure 2.18. There was little or no evidence supporting the higher contamination predictions. Only sample 68 has evidence for 1-2% contamination due to the small peak at 99-98% AAF and has high depth compared to other samples.

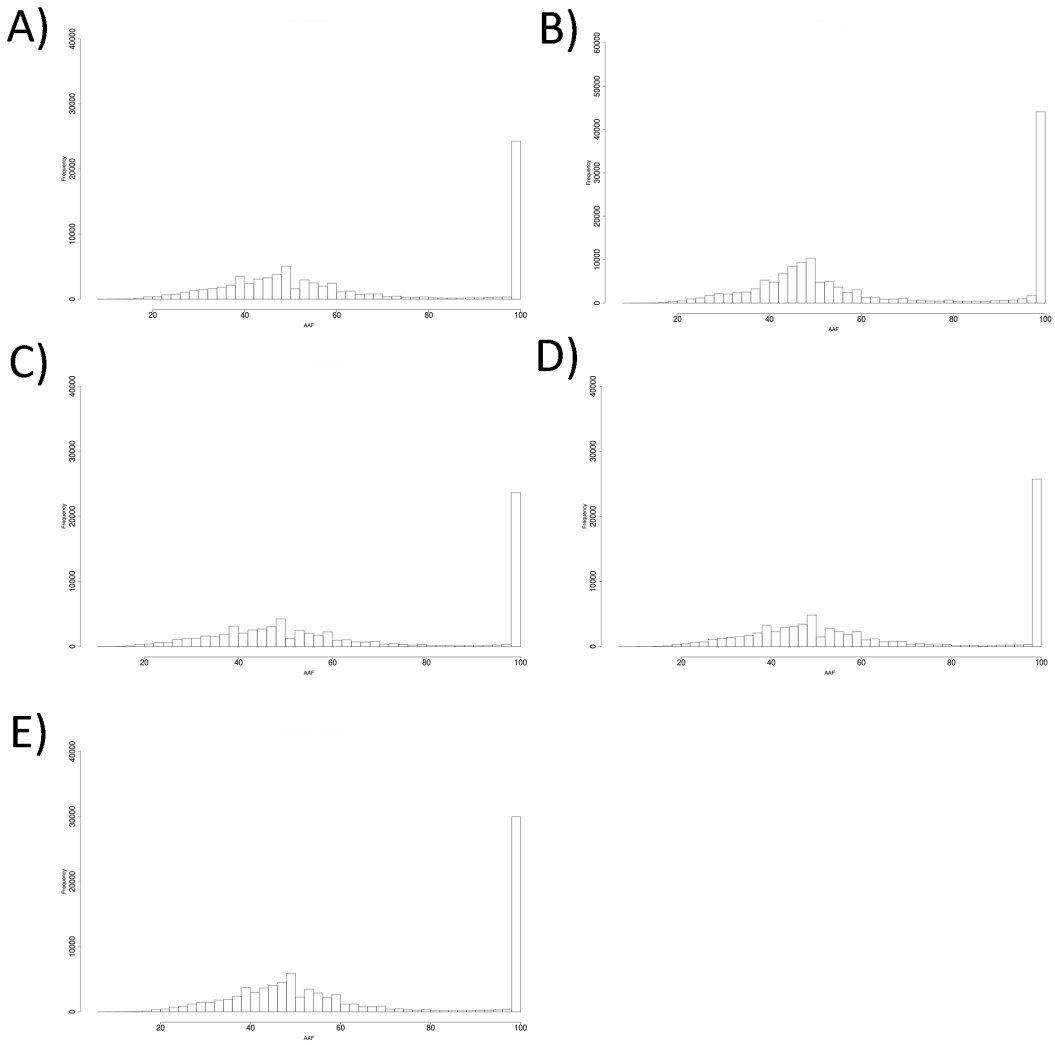

Figure 2.18: Alternate allele frequency profiles for the highest five predictions from OLS but not VerifyBamID. All 5 samples display profiles more consistent with low or no contamination. **A)** Sample 230 has no obvious deviations. **B)** Sample 68 has evidence for 1-2% contamination due to the small peak at 99-98% AAF. **C)** Sample 90 has no obvious deviations **D)** Sample 231 has no obvious deviations. **E)** Sample 102 has no obvious deviations

All of the samples have some of the highest observed Het:Hom ratios at between 1.79 and 2.00 which are likely skewing the predictions due to the high coefficient for the measure. A summary of the typical values for each of the features are shown in Table 2.11 which shows the average Het:Hom ratio was 1.67.

Measure	SKEW ALL	KURTOSIS ALL	SKEW HOMS	KURTOSIS HETS	HET:HOM RATIO	DEV A	DEV B	PC X HETS
Mean	1.4	0.19	-5.79	3.58	1.67	0.31	2.62	33.06
Std. Dev	0.04	0.03	1.01	0.26	0.12	0.16	0.33	22.67
min	1.29	0.04	-8.26	2.05	1.34	0.14	0.96	5.59
0.25	1.37	0.17	-6.37	3.41	1.61	0.23	2.48	10.73
0.5	1.39	0.19	-5.78	3.58	1.66	0.29	2.73	22.47
0.75	1.41	0.21	-5.24	3.78	1.71	0.35	2.81	58.93
max	1.67	0.3	-1.93	4.15	2.95	1.87	3.58	70.63

Table 2.11: Summary of measures used in regressions obtained from application to 245 samples with unknown levels of contamination. For each measure the mean, standard deviation and range for each measure at each of the eight tested contamination ranges.

2.4 Discussion

Contamination of NGS samples presents a serious challenge to accurate variant calling. Whilst cross-species contamination is described more in literature[81–87] than intra-species contamination between individuals[88,91–93] it is likely due to the easier identification of cross-species reads as they do not map to a target reference and therefore not reflective of the relative incidence of intra-species contamination. To identify intra-species contamination specialised programs such as ContEst[91] or VerifyBamID[88] have previously been required to analyse BAM files and predict the levels of contamination when comparing to an external truth dataset. Therefore to estimate contamination the BAM file must be available, which is often not possible when working on shared data when only a small file such as a VCF is available. Also the requirement of an external truth or population dataset is possible for *homo sapiens* but would not be available for less well studied species. To provide a method of estimating contamination this project developed a method using only changes measurable from VCF files without BAM files or external truth sets capable of estimating contamination.

Reductions in variant AAFs were postulated to occur as a result of contamination. Contamination simulations were performed at eight levels: 1, 2.5, 5, 10, 20, 30, 40 and 50% were used to confirm these AAF changes. At 10% contamination or below a tail to homozygote peak at 100% was visible and caused by the lowering of homozygote variant AAFs. These changes were captured by Dev A and the skew for the homozygote variants. The homozygote distribution was left-skewed yielding an average value of -3.67 at 1% contamination which increased to -1.75 by 10% contamination as shown in Table 2.2. Both of these measures become less informative at contamination levels above 10% because lowering of homozygotes variant AAFs leads to them now being called as heterozygotes. The tail of the homozygous peak subsequently merges into the background with increasing contamination level. Therefore both Dev A and skew homozygotes are more useful at predicting low levels of contamination.

Due to the increased proportion of heterozygote to homozygote variants as

contamination levels are increased the Het:Hom ratio can be observed to increase with contamination. However the ratio is affected by natural admixture of samples and by sequencing depths, with lower depths favouring the calling of homozygotes[107–109]. VerifyBamID also attempted to use the Het:Hom ratio to estimate contamination but due to the variability of the Het:Hom ratio per sample models were not as universally applicable as the final methods implemented[88]. Measures including Dev B, Dev metric, skew of all variants, kurtosis of heterozygotes or all variants and the percentage of chromosome X heterozygotes were added to models to identify more repeatable trends between samples in the overall AAF profiles or for heterozygote variants which a machine learning approach would identify.

PCA and clustering analysis were performed using all nine of the features as shown in Figure 2.7. Whilst principal component 1 explained for 59.8% and principal component 2 explained 24.2% of variance. Both principal components had high component loadings from measures for both homozygotes and heterozygotes vindicating the selection of using all features further. From the generated clustering there is at least partial separation of the levels of contamination and a pattern of separation upwards on the y-axis of principal component 2 up to 10% contamination, after which the heterozygote distribution changes begin and leads to large overlaps of contamination levels from 20% upwards. This pattern was indicative of being able to differentiate low levels of contamination from higher levels upwards of 20% when the effects of homozygote measures become more informative.

Using the 200 contamination simulations with known levels of contamination all four linear and five polynomial regression models were trained with all nine features or only selected features of Dev B, skew of homozygotes and the Het:Hom ratio. The highest R^2 values in training were obtained with polynomial regressions of degree two and three when using selected features and training ranges of 1-10 or 1-20%. However as was seen when testing the polynomial regression models with the 245 samples of unknown contamination the models are seen to perform poorly with a degree of three polynomial predicting most samples as negative contamination and the degree of two varying from 0 to 25% for samples VerifyBamID predicts as close to 0 or 1%. To improve this model therefore wider range of samples in the training data set may be

required to help better capture the natural variation in measures which may be distorting the predictions for polynomial regressions.

For linear regressions the OLS model was the best performing when using all features in regression models, over the four linear models all features averaged 0.809 compared to 0.763 for selected features. Performance of linear regressions using all features was less accurate at lower contamination levels but better at higher levels. Opposing trends were seen for selected features with more accurate results at lower contamination levels but less at higher levels which supports the theory that Dev A, Skew of homozygotes and Het:Hom ratio are good predictors of low levels of contamination. Conversely the heterozygote and all profile measures also improve the accuracy of higher contamination level predictions.

Training was also performed using all and selected features but with restricted ranges of contamination from the simulations data. By restriction of the contamination levels used in training the coefficients for each of the features used in regressions were observed to change as shown in Table 2.9 for the OLS regressions. As contamination levels increased the Het:Hom ratio co-efficient decreased while the kurtosis measure looking at all variants increased rapidly. Restricting to 1-10% contamination gave improved R^2 values for all linear regressions, with all features an average of 0.869 and selected of 0.858 were calculated. Increases in R^2 values were expected due to reducing the range of simulations to that which is largely predictable from measures in the homozygote range. It was also expected that using selected features would give more accurate predictions at the lower ranges, but this was not proven meaning some predictive power was obtained by the other features which are also predictors of contamination at low levels, validating the selection of using all features over the selected features. Training using 1-10% was found to be the optimal range, despite the increased R^2 values (Average: All features - 0.902, Selected - 0.883) being obtained when training with 1-20% contamination, a wider spread of predictions obtained at 20% actual contamination is evidence the linear regressions are losing accuracy. From 20% contamination and upwards the changes in AAF are mostly in the heterozygote variant AAF distribution, when trying to find the optimal linear regression, closest to the actual contamination value, regressions becomes a

compromise. The compromise leads to the higher contamination estimations for the samples actually with low levels of contamination below 10% and the wide spread of predictions obtained at the levels of actual contamination of 20%. As seen in Figure 2.8 where linear regressions are trained with contamination simulations of 1-50% the predictions for 20% and 30% actual contamination often overlap. Predictions for 30-50% contamination also display substantial overlap highlighting that using larger ranges does yield accurate prediction of higher actual contamination levels.

Increasing the levels of contamination used in training as shown in Table 2.9 leads to the reduction of the co-efficient for the Het:Hom ratio from 17.13 for 1-10% training down to -0.53 by training with 1-50%. The Het:Hom ratio as discussed by VerifyBamID[88] was one of the most powerful predictors of contamination but also was variable between samples. As was shown in this study and by increasing the training range its effect becomes diluted in predicting contamination. Kurtosis of all variants has one of the lowest coefficients at the 1-10% contamination training range of -10.42 but by 1-20% has risen to 12.60, an increase of 23.02, just 1.28 below the coefficient for the Het:Hom ratio at same contamination range. By the 1-50% contamination training set the kurtosis all variants value had risen to 47.04, the highest of any coefficient with the nearest being skew all at 8.15. As the contamination ranges used to train the model are increased the model becomes too reliant on the kurtosis all variants measurement. The scatter matrix in Figure 2.6 shows that both kurtosis all and skew all variants against the actual contamination level do not separate the higher actual contamination levels upwards of 20%. However both measures do not overlap with the lower contamination levels below 20% as with measures such as the Het:Hom ratio or Dev A, Dev B explaining why when training using the higher contamination levels a model would assign these measures higher coefficients.

All of the results indicated that the linear regression models performed better when using all features than just using the selected homozygote features. Also, higher R^2 values were obtained when using simulation data restricted contamination ranges with the 1-10% range selected. Finally of the four linear regression models implemented OLS was the highest scored regression when trained using all features with a 1-10% contamination range and as shown in Table 2.5 OLS has the highest

average R^2 value for all features over all of the tested ranges.

When applying the OLS regression model to 245 germline exome samples the effect of the training set range is most apparent. When OLS regression predictions are plotted against VerifyBamID predictions, as shown in Figure 2.15, the changes in coefficients for each of the contamination ranges used to train samples are most evident. Using the training set of 1-10% contamination samples which clustered tightest in the lower left of the graph agreeing with the VerifyBamID predictions. Each subsequent increase of the contamination training range causes contaminations predictions from regressions become higher widening disagreements with the VerifyBamID predictions. Inspections of AAF profiles did not support the predictions of high contamination which were made by the models which were trained by the higher contamination ranges.

Given the highest agreement between the regressions and VerifyBamID were obtained with the 1-10% trained regression the top five VerifyBamID samples were compared against the OLS predictions which achieved good correlation with two exceptions. Sample 4 was predicted by VerifyBamID as 32% contamination but by regression was 25.27%, whilst this is lower than the VerifyBamID value it is still sufficiently differentiated from the low contamination cluster below 5% to flag the sample as of concern. The lower prediction is caused due to the training range used as if the training range of 1-40% was used then the prediction rises to 33% hence the coefficients in the 1-10% are more accurate at the lower contamination range at the cost of under-estimating higher contamination levels. Sample 217 was predicted as 4.1% contamination by VerifyBamID and regression predicted this slightly higher at 5.5% but again is higher than any of the other samples in the lower left cluster again distinguishing the sample as of potential interest.

Sample 69 did not agree well with the VerifyBamID predicted level of contamination, the likely cause being the high Het:Hom ratio of 1.91, the third highest observed and well above the average of 1.67 as shown in Table 2.11. Due to the high Het:Hom ratio the sample appears to be called at a higher level than the actual contamination from the AAF profile of 1% suggested by a peak just below the 100% AAF. A similar

pattern of high Het:Hom ratios were also seen in the top five contamination predictions by OLS that were not predicted to be above 1% by VerifyBamID as shown in Table 2.10. The Het:Hom ratio of samples can be affected by a number of factors in addition to contamination such as the degree of admixture which could increase the ratio[88]. Sample depth of coverage would also affect the ratio as low depths below 30x coverage are suggested to favour the calling of homozygous variants in genome sequenced samples[107–109]. To account for the wide range of Het:Hom ratios many samples would be needed from a wide range of origins and additional simulations needed to better train the regressions to account for the variability. Currently the lack of sample diversity and the somewhat limited number of samples are likely the principal factor impairing the accuracy of regression predictions. More sample crosses are need to encompass a wider range of ethnicities and ideally at high depths of coverage above 30x to better capture the range of Het:Hom ratios that naturally occur. Therefore many more samples from other populations would likely need to be simulated using the *in silico* method described in this chapter to improve the training of regressions. Also 1-10% contamination as a training range probably is not the most optimal range, which probably lies between 10 and 20%. To better train the models simulations would probably be needed at smaller intervals than 10%, ideally at 1% intervals. An improvement to the model could also be to not just focus on one of the homozygote peaks, e.g. not only the 100% peak but also look at the peak which forms around the low AAF peak nearer to the 0% as heterozygotes become reference homozygotes. The effect of this peak of reference homozygotes is likely to be small but could give a small improvement in performance.

Using a linear regression model to predict percentage outcomes does however have several limitations. Firstly this assumes that variables are linearly correlated, in this dataset as shown not all of the variables are linear. Secondly linear regressions look at the average of dependent variables and so are more sensitive to outliers and too many variables can reduce the accuracy of predictions which all will significantly affect percentage outcome predictions.

2.5 Conclusions

From investigations of contamination it was possible to identify the characteristic changes that occur in the alternate allele frequency of heterozygote and homozygote variants. The changes in alternate allele frequency allowed multiple measures to be used to try and use these changes to predict contamination. Using simulations of contamination performed *in silico* it was possible to train a regression model to predict the levels of contamination. The prediction model could still be improved further as it is likely limited by the variety and number of simulations used to train the regression model and ultimately a more extensive training set may lead to better performance from polynomial regressions.

Chapter 3

Unmapped reads provide insights to potentially clinically important information and can distinguish collection methods

3.1 Introduction

3.1.1 Unmapped reads

Unmapped reads are reads which cannot be mapped to the reference sequence of the organism being investigated. Two previous studies looking at saliva captured exome samples[89] and a variety of exomes and genome sequenced samples[85] found unmapped reads ranging from one million to ten million. Unmapped reads comprised between 1.9 - 13.2% of the saliva captured exomes and between 1 - 10% of the assorted exomes and genomes[85]. The majority of unmapped reads arise due to sequencing errors or several base pair variations causing the reads to be un-mappable to a reference[85,87]. Once technical artefacts have been excluded unmapped reads have been shown to contain data which can be informative of either sample contamination from another species[85] or a representative snapshot of a patients microbiome[89].

In order to identify cross-species contamination of samples it is necessary to remove both the sequencing errors and low quality data but also technical artefacts. Illumina

sequencing uses the *phiX174* genome as a quality control[110], which is spiked into samples or a parallel lane. During sequencing by synthesis strands may fall out of sync, by using the well known *phiX174* sequence the sequencer can monitor the sync in real time[110]. Software integrated into Illumina sequencers requires lanes with the spiked in *phiX174* to be specified, *phiX174* reads should then not be indexed though some still may be indexed with other libraries, the non-indexed reads will be contained in separate files[110]. *PhiX174* reads which are indexed can subsequently be removed from samples by aligning against the *phiX174* reference sequence. Alternatively *phiX174* reads will be flagged as unmapped reads when aligning the sequencing data. PhiX sequences have been identified in the genomes of more than 1,000 bacteria recently from sequences submitted to the Integrated Microbial Genomes (IMG) database from a variety of sequencing centres across the world[87].

3.1.2 Cross-species contamination

Increasingly studies are now identifying potential contamination within published datasets by unmapped read analyses. Unmapped reads were obtained by running samples through pipelines as normal with unmapped reads then extracted and used to search for matches using BLAST (Basic Local Alignment Search Tool)[111]. Data from the 1000 genomes project was analysed in the paper by Langdon et al. in which 604 billion reads identified 4.8 million matches to *Mycoplasma*[83]. It is estimated that up to 7% of the data in the 1000 genomes project may contain contamination from *mycoplasma*[83].

DNA sequences of human origin have also been shown to contaminate a variety of non-primate species. Screening 2,749 non-primate species from databases: NCBI, Ensembl, JGI, and UCSC found a total 492 to have be contaminated with human sequence. The human sequences used were primate specific SINE, AluY and were detected in species ranging from bacteria to plants to fish[84].

Merchant *et al.*[82] re-analysed the *Bos taurus* reference genome and using metagenomic analysis with Kraken[112] and identified 138 unmapped contigs which best mapped to *Neisseria gonorrhoeae* (TCDC-NG08107), a bacteria which infects humans. A deeper analysis of the matches after assembling to contigs revealed they

only mapped to four locations in the genome, contig sizes ranged from 200bp to 634bp. This prompted the study to review the genome status in GenBank which found discrepancies between the original genome publication and the on-line genome. Gaps in the original publication were reported as a complete genome reference in GenBank but neither source stated how genome gaps were filled. Gaps were then shown to be originating from *Bos taurus* and *Ovis aries*. This lead to the study concluding that the bacterial genome was erroneously uploaded as a finished genome, with contigs concatenated together from *Bos taurus* and *Ovis aries* possibly as cross-species contaminants. GenBank subsequently suppressed the entry for the bacteria TCDC-NG08107 genome[82].

Unmapped reads from samples sequenced at five large sequencing centres were investigated by Tae et al.[85] using BLAST[111] which found the largest difference was caused by the sequencing centre used. Using samples from the 1000 genomes project including both exomes and genomes. The number of unmapped reads showed little variance when dividing samples by ethnicity, most variance was found when comparing samples grouped by the sequencing centre. The study shows that the patterns of contamination occurring in the sequencing and sample preparation stages will likely depend upon the other samples being processed at the centre at the same time[85].

The origins of cross species contamination have also been suggested to be more complicated than handling errors in sample preparation or sequencing. Genes from food have been claimed to be transferred into human blood making them a potential contaminant of blood samples, though a method of transmission which avoids degradation has yet to be identified[113].

3.1.3 Microbiome information from unmapped reads

Cross-species contamination of NGS samples can also be present at the time of collection and so actually represent the microbiome around the site of collection. Bacterial microbiomes were investigated for exomes samples obtained by buccal swabs in a recent publication by Kidd et al.[89]. Unmapped reads were mapped against a database of bacterial sequences using BLAST. The bacterial content of 15 South

African Khoesan samples were normalised and compared against a cohort of buccal samples from North America. The KhoeSan are a group of indigenous, hunter-gatherer and pastoralist people of South Africa[89]. Comparisons revealed that five oral species in the Khoesans were not found in the American cohort. From the unmapped reads the study showed that between populations it was possible to compare bacterial abundances, however within a population at both the genus and species level there was too much variation to identify any patterns[89].

Unmapped reads from RNA sequencing (RNA-Seq) samples were also able to be extracted and mapped to a range of bacterial and archaeal species. In the study 192 samples including: controls, schizophrenia, amyotrophic lateral sclerosis and bipolar disorders all of which were captured by poly(A) enrichment. Increased bacterial levels were found from the unmapped reads for schizophrenia samples[114].

Viral integration can potentially be identified from unmapped reads. Retro-viral integration is problematic to identify with short read sequencing technology due to increased mutation rates and instability either side of the integration site. This prevents reads mapping to the human sequence either side of the virus, inhibiting bridging across of the viral insertion. By mapping unmapped reads to viral species and calling the variants to the reference virus sequence a customised viral sequence can be built. Customisation of the host genome to include the new viral sequence which can then be used to identify common mutations either side of the insertion and viral insertions[90]. This method was therefore able to increase the sensitivity of detecting viral integration sites and is able to more accurately identify variants in rapidly evolving viruses[90].

The human microbiome can be altered in diseases such as Inflammatory Bowel Disease (IBD). IBD is believed to be caused by the combination of: immune system dysregulation, genetics, environmental factors and microbial flora[115]. Intestinal commensal microbes normally aid with nutrient absorption and normal immune system function but in IBD cases the immune system acts against the commensal microbes leading to an inflammatory response. The Human Microbiome Project 2 is one large scale project looking at the microbiome of IBD patients using sequencing

techniques such as 16S from a variety of patient tissues[116]. 16S data is commonly collected from faecal samples to survey the gut microbiome, whilst exome and genome sequencing samples are typically blood or saliva based capture. Some evidence suggests that changes in the gut microbiome are mirrored by changes in the oral cavity using 16S data but results in sites that can vary by orders of magnitude making direct comparisons difficult[117–120].

Previous studies have identified some of the bacterial phyla and genera most reduced in the gut in cases of IBD compared to healthy unaffected individuals. *Faecalibacterium* a genus of intestinal commensal bacteria under the phylum Firmicutes, typically making up around 5% of the bacterial population of a healthy individual[121]. In cases of IBD such as Crohn's disease the population of *Faecalibacterium* were found to be reduced and at low levels of the bacteria correlated with recurrence of ileal Crohn's disease[122].

Prevotella is a genus of Gram negative bacteria with species found in the gut, oral cavity and mucosal membranes which was originally classified under the *Bacteroides* genus before being re-classified under a new genus[123]. In the gut the levels of *Prevotella* have been shown to be influenced by diet. Children in Burkina Faso consuming a high fibre showing increased levels of *Bacteroidetes* and *Prevotella* compared to European children. In particular species with the genes for cellulose and xylan hydrolysis[124]. In IBD cases the levels of both gut *Prevotella* and *Bacteroides* have been shown to present at increased levels accompanied with reduced overall diversity[125]. *Bacteroides* are similar in function to *Prevotella* performing breakdown of a range of plant polysaccharides such as amylose, amylopectin and pullulan[126].

Both Ruminococcaceae and Lachnospiraceae are part of the order of Clostridiales under the phylum Firmicutes which has been shown to be reduced in cases of IBD[127]. Species in both families are generally involved with the butyrate production using anaerobic fermentation[128,129]. *Blautia* is a genus containing a number of species formerly classified under the *Ruminococcus* genera. Species under the *Blautia* genus are Gram positive bacteria also involved with sugar fermentation[130].

Sutterella is a genus containing anaerobic, Gram-negative bacteria. Faecal samples of IBD patients have been shown to contain bacteria coated with the IgA antibody, transferring the bacteria to germ-free mice induced increased intestinal damage upon stimulation. Mice with higher levels of *Sutterella* were able to more effectively to degrade via proteolysis the bound parts of IgA[131].

3.1.4 Analysis programs

BLAST

BLAST as a sequence alignment tool has been extensively used globally for many years for both amino and nucleic acids[132]. BLAST is not used to assemble NGS data as the algorithms used are not optimised for accuracy, especially in poorly aligning, gapped regions[42]. Seed matching is employed by BLAST, where a sequence is taken from the input sequence read and matched to a reference, if no exact seed match is found then no match will be reported. When comparing two sequences it is possible a match could arise by chance. Assuming each of the four nucleic acids are equally likely at each position, for a given score (S) for a MSP of two sequences m and n then it is possible to detect and filter some false positives using a calculated value (E)[132–135].

In 1997 revisions to BLAST improved the speed of alignments and allowed gapped alignments[134]. Gapped alignments can lead to an increased number of matches of varying quality, making score filtering and E-value filtering more important. Improvements were made to the core of BLAST for over a decade then which culminated in the release of the new BLAST+ program[111]. The new iteration was split into three modular sections: query set-up, scanning and trace-back. The set-up module establishes the search parameters and applies filtering before building a lookup table of words from the input query. Low complexity and repeat regions are challenging to assign, typically there are many false positives, to compensate either hard or soft masking can be applied. From the lookup table the scanning phase begins by reading sequence databases and performing word extension with gaps to assemble a list of preliminary matches. Finally trace-back begins with re-analysis of matches to identify insertions and deletions and to resolve ambiguous positions. BLAST has been used by many of the studies looking at unmapped reads[83–85,89]. The

E-value used in this study was chosen based on similar studies[89,112,136].

3.1.5 MEGAN6

The microbiome analysis tool MEGAN6[137] is able to integrate a variety of data types including from: metagenomic, metatranscriptomic, metaproteomic and rRNA data. From the input data a taxonomic tree of results and their abundance can be displayed in a phylogenetic tree using taxonomy mapping files[138].

3.1.6 Sequencing methods

RNA-seq

RNA-Seq applies NGS to RNA, as with DNA the most commonly used NGS method is Illumina. From sequenced RNA the transcripts being expressed in a cell can be quantified and from the number of each transcripts the relative expression can also be estimated. A typical RNA-seq experiment will involve first isolating the RNA for a sample along with species selection to enrich or deplete specific RNA types. Extracted RNA will be comprised of multiple types such as: ribosomal RNA (rRNA), precursor messenger RNA (pre-mRNA), mRNA and various classes of noncoding RNA (ncRNA). The majority component of most cell types has been shown to be rRNA, on average comprising 95% of the total cellular RNA[139]. Hence unless the rRNAs are depleted they will constitute the majority of the sequencing results, reducing the target RNA depths and coverage. Two methods of selection are commonly used, the first of which is polyadenylation in which the mRNAs are targeted by the 3' poly-A tail using poly-T sequences which are attached substrate allowing enrichment of the target mRNAs. A minority of bacterial RNAs have poly-A tails[140]. Alternatively rRNA can be depleted using methods termed ribo-depletion using commercially available kits. After selection of RNA species the target RNAs are then reverse transcribed into complementary DNA[139]. With complementary DNA the sample can be prepared with adapter attachment prior to sequencing using the method for Illumina or other NGS method as outlined in section 1.1.2.

16S rRNA-Sequencing

16S rRNA is used to study bacterial species by using a similar protocol to that of RNA-seq but with the prokaryotic 16S rRNA sequences are targeted to provide a count of the bacterial species present in a sample[141].

3.2 Aims

The aims of this study are to investigate the content and taxonomic classifications of reads which are unmapped after the initial mapping to a reference genome and identify if there are any patterns. Firstly from the unmapped reads of Inflammatory Bowel Disease patients investigations will be performed to see if differences can be detected between the disease sub-types: Crohn's Disease (CD), Ulcerative Colitis (UC) and Unclassified (IBDU).

Secondly the unmapped reads of exome samples which were either blood or saliva collected will be compared. Using taxonomic classifications the numbers and genera detected will be compared to try and identify patterns between the collection methods. Using a large cohort of 358 WES samples consisting of both known collection methods and unknown, it will be investigated if the collection method can be predicted based on the patterns of unmapped reads. Furthermore unmapped reads will be mapped back against the human reference genome to identify the closest matches and ascertain if regions are consistently identified which may be highly variable or may be more challenging to map to due to repeat sequences.

Finally investigations of unmapped reads from IBD samples with both WES and RNA sequencing will be performed to investigate the differences between sites for the same individuals.

3.3 Materials & methods

3.3.1 Samples

In this study we analysed unmapped paired-end reads from human WES of 245 paediatric Inflammatory Bowel Disease (IBD) patients and 113 non-IBD 'control' samples. Access to sample BAM files was provided by Professor Sarah Ennis and her Human Genomic Informatics Group, in the Faculty of Medicine at the University of Southampton. All samples were sent for sequencing over several years and prepared for sequencing by human genomic informatics group members. Collection methods were known only for the IBD samples with 233 blood samples and 12 saliva, no collection information was available for the non-IBD samples. The 113 non-IBD control samples are patients with other diseases including cancers and imprinting. Exome samples were all captured using Agilent Sureselect 51/60 Mb V4/5/6 and were mapped using BWA (v0.7.12) to the hg38 human reference.

For 10 individuals RNA-seq data was collected and prepared by Dr James Ashton and made available pre-aligned using Tophat (v2.1.1) using only the hg19 reference genome. RNA-Seq samples underwent Poly-A targeting of RNA. Seven out of ten individuals were both RNA-seq and whole-exome sequenced as shown in Table 3.1. Seven out of ten individuals were also sequenced at two sites with RNA-seq, two were ileal sequenced only, while nine of the ten were rectal sequenced. RNA-seq samples were sequenced in the Wessex Investigational Sciences Hub laboratory at the University of Southampton on an Illumina Miseq.

Patient	**Diagnosis**	**Ileal RNA**	**Rectal RNA**	**Exome**
311	Ulcerative Colitis	Y	Y	Y
314	Ulcerative Colitis	Y	N	Y
318	Ulcerative Colitis	Y	Y	Y
317	IBDU	Y	Y	Y
310	Crohn's disease	Y	Y	Y
313	Crohn's disease	N	Y	Y
320	Crohn's disease	N	Y	Y
315	Control	Y-Caecum	Y	N
321	Control	Y-Caecum	Y	N
322	Control	Y	Y	N

Table 3.1: RNA-seq samples taken from ten individuals, in total 17 RNA-seq samples were available. RNA-seq samples were collected from either or both ileal or rectal sites, eight ileal and nine rectal samples samples were collected. Seven whole exome samples were only available for Inflammatory Bowel Disease cases and not the three control samples.

Six of the IBD samples were also 16S rRNA sequenced from faecal samples as summarised in Table 3.2, data from the 16S samples was available to compare against the unmapped reads from the blood collected exomes. All six samples were sequenced in the same batch at the Wellcome Centre for Human Genetics (WTCHG), Oxford.

Sample ID	Disease	Gender	Capture Kit	Batch	Ethnicity
SOPR0240	Crohn's	F	Aglient51 V5	WTCHG Sept 2014	African
SOPR0241	Crohn's	M	Aglient51 V5	WTCHG Sept 2014	Caucasian
SOPR0243	Ulcerative colitis	F	Aglient51 V5	WTCHG Sept 2014	Caucasian
SOPR0244	IBD- Unclassified	F	Aglient51 V5	WTCHG Sept 2014	Caucasian
SOPR0245	Crohn's	M	Aglient51 V5	WTCHG Sept 2014	Caucasian
SOPR0246	Crohn's	F	Aglient51 V5	WTCHG Sept 2014	Caucasian

Table 3.2: Samples with 16S rRNA seqencing data information available to be compared against unmapped reads from exomes for the same individual. All 6 samples were sequenced in the same batch at the Wellcome Centre for Human Genetics (WTCHG), Oxford.

3.3.2 Extraction of unmapped reads

A pipeline was developed to quantify and classify unmapped reads starting with BAM files as shown in Figure 3.1. All 358 exomes were provided aligned to the human genome hg38, while RNA-seq BAM files for 17 samples were provided aligned to hg19 using TopHat (v2.1.1). All BAM exome files were aligned using BWA. For a paired read there are three possible combinations for the pair: both reads are unmapped, one read is mapped but the other is not and vice versa. If one of the reads has successfully been mapped to the reference sequence then it is likely that the other read contains a sequencing error or an insertion preventing mapping. To avoid potential sequencing errors being included in taxonomic mapping and causing false positive matches only reads which were both unmapped were extracted using SAMtools (v0.1.19): samtools view -u -f 12 -F 256 input_bam > name_both_unmapped.bam. This command selects only reads which are both unmapped (-f 12) and not in the primary alignment (-F 256). Output from this command is a BAM file containing only both unmapped read pairs. BAM files were then split into fastq files using bedtools (v2.21). Each of the fastq files were then filtered using the fastx-toolkit (v0.0.14) to remove reads with positions of phred quality lower than 20 for 90% of the read sequence before converting the read sequences into fasta format to be used in BLAST (v2.2.28).

3.3.3 FastQC

FastQC (v0.11.3)[32] analysis was performed on extracted fastq reads from unmapped BAM files. FastQC was used to investigate if any particular biases in the per base sequence and GC content could be identified.

3.3.4 BLAST classification of sequences

BLAST matches input sequences against a pre-compiled NCBI or custom databases to return species matched[111,132–135]. A total of six databases were used: Bacteria, Virus, Eukaryota (Fungi), Plants, Mammalian (excluding human) and Human obtained from packaged versions of fasta sequences obtained from the NCBI ftp website (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/).

The human database was separated from the larger mammalian database to a obtain the best match per read for humans which could then be compared to the best match for other mammalian species. As mammalian and human genomes are highly conserved, particularly over genes, keeping them in the same database in testing resulted in reads mapping better to other mammalian species than limiting the analysis of potentially poorly mapped human regions.

Each of the databases were run using BLAST with an E-value of 1e-10, used to score matches on the basis of a sequence match being found by chance and so reduces the number of false positives. BLAST options 'max target seqs 1' and 'max hsps per subject 1' were also both used to only return a single result per read and report the highest scoring in the subject-query pair to remove sub-read matches and ensure the longest sequence matches were returned. BLAST output format six was used to generate tabular results files which maximised compatibility for downstream analyses and also made use of threading to run with 16 processors to reduce processing time.

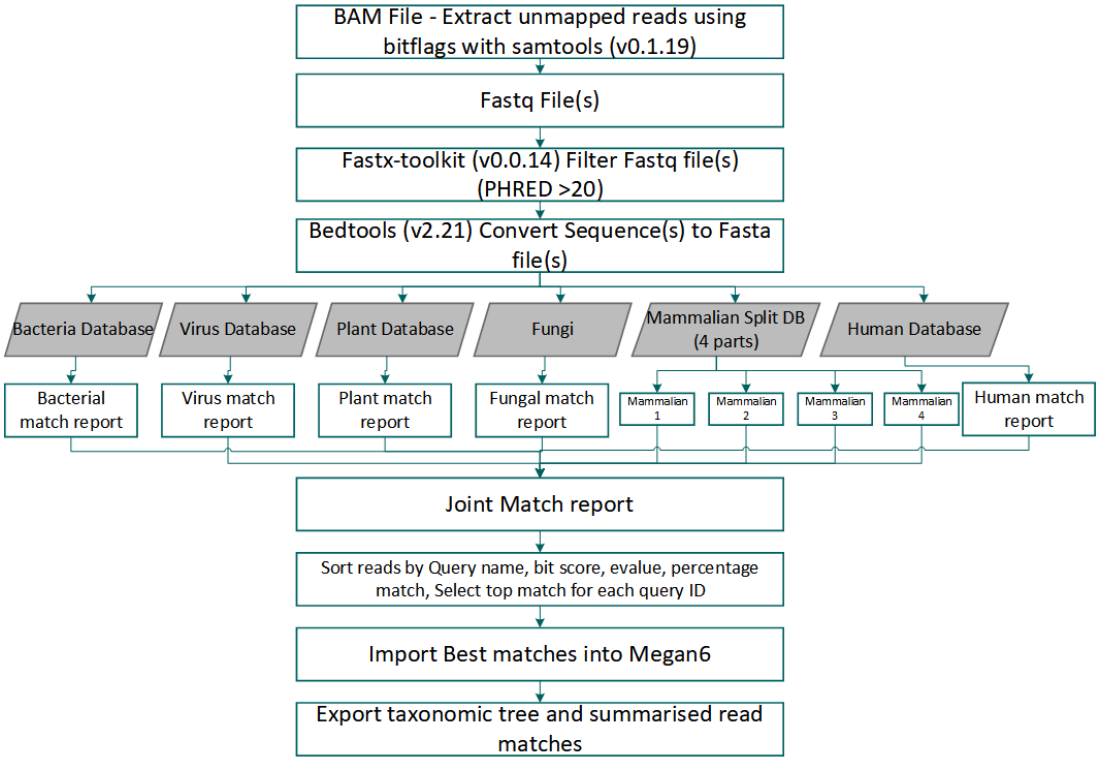

Figure 3.1: BLAST pipeline for the analysis of unmapped reads extracted from BAM files. Fastq reads were filtered to remove low quality reads which are likely to contain sequencing errors. Fasta sequence files created from filtered fastq files are then run through each of the BLAST databases before results are concatenated and sorted to only report the best match per sequence. For each sample results were analysed using the taxonomic classification program Megan6[137].

BLAST was used with six databases as previously described, however as shown in Figure 3.1 the mammalian databases were split into four smaller sub-databases due to processing constraints. Whilst BLAST was run using 16 processors each read sequence still had to be compared against all of the database sequences. As shown in Table 3.3 bacteria, virus and fungi databases are all small at between 0.16 and 9.1 GB. The plant database was much larger at 49GB. The mammalian database obtained from RefSeq totalled 261GB, for which every read was scanned against. This caused mammalian analyses to exceed the maximum processing time allowed of 60 hrs per sample, depending also upon the number of reads to query. Therefore the mammalian database was divided into four to speed up and allow BLAST to complete matching.

Database	Release	Source	Number of Taxa	Size
Bacteria	Jun-15	NCBI Refseq	2785	9.1GB
Virus	Jun-15	NCBI Refseq	4400	167 MB
Fungi	Jul-15	NCBI Refseq	33	584 MB
Plants	Jan-17	NCBI Refseq	87	49 GB
Mammalian	Jan-17	NCBI Refseq	108	261 GB

Table 3.3: BLAST databases obtained from RefSeq, initially bacteria, virus and fungi databases were obtained in 2015. Additional mammalian and plant species were obtained in 2017 to expand classifications. Plant and particularly mammalian databases were much larger than previous databases. The mammalian database required splitting into smaller sub-databases which could then be processed using BLAST.

3.3.5 Creation of taxonomic trees

Each database, including the sub-databases for mammalian species, were run independently with BLAST each generating the top match for all input reads against all of the species in that specific database. Hence each read would be mapped multiple times between the databases. To only analyse the most likely matches results files were concatenated then sorted by the Sequence ID, E-value and percentage match over the entire read query to only keep the best read match. With a single results file for each sample containing only the best match per read the files were loaded into the taxonomic classification program MEGAN6. Each of the results files were loaded using the additional script provided with the MEGAN6 installation '*blast2rma*' script and the mapping file provided on the MEGAN6 website, converting nucleic acid GI codes to the NCBI-taxonomy classifications. From the mapped results imported the creation of taxonomic trees were possible. Trees were then joined and exported into a single CSV file containing per sample information for all samples classified at each level of the taxonomic tree allowing further investigations and analyses.

Human only matches were analysed differently, files were imported to MEGAN using the previously described method but were then extracted using the 'rma2info' command. The command enabled the extraction of locations and contigs/scaffold matched for each of the reads, used later in assessing the locations of matches back to the human genome.

3.3.6 Clustering analysis

Clustering was performed using the program ClustVis[106] which performs principal component and clustering analysis (PCA) to uploaded data via the website https://biit.cs.ut.ee/clustvis/. ClustVis was run using unit variance scaling to standardise features and the PCA method singular value decomposition with imputation. Clustering was performed on the CSV files generated from MEGAN6 containing the total read matches or percentage read matches per sample to: bacteria, virus, fungi, archaea, plants, metazoa, human and unclassified. Also clustering analysis was used to examine the unmapped reads from RNA samples by collection site and disease subtypes.

3.3.7 Tandem repeat finder

Fasta read sequences for identified species of interest were extracted and then checked for tandem repeat sequences using the web program Tandem Repeat Finder (TRF) (v4.09)[142]. It was possible that read sequences were repetitive and so were unable to be initially mapped against the human genome by BWA and so would be detected by TRF.

3.3.8 Depth of coverage across bacterial genomes

For the bacterial genomes of interest read locations were extracted from rma6 files using the '*rma2info*' MEGAN6 command and converted into BED files using AWK. Coverage was then calculated using a single line BED file describing the length of the bacterial genome. Bedtools coverage (v2.21) was then used to calculate coverage over the length of the bacterial genome.

3.3.9 Calculating sequence similarity between species

Mummer (v3.23) was used to compare the similarity between species. The nucleic acid option was used called nucmer to compare the fasta sequences[143].

3.3.10 Plotting of bacterial genomes

Using a genbank (*.gbk) file describing the reference sequence and known or predicted genes. A linear representation of a genome was created using the program Artemis (v16.0)[144]. To the linear genome the read match locations in BED format and were added as an additional track describing the positions overlapped by a read. A combined file of genbank gene information and read matches was then exported prior to import into DNAplotter(v1.11)[145]. DNA plotter generated a circularised visualisation of the genomes with reads from each of the loaded samples added as additional tracks around the genome.

3.4 Results

3.4.1 Extraction & classification of unmapped reads from exome case and control samples

Unmapped reads were extracted for all 358 exome samples after filtering and excluding *phiX174* reads as shown in the box plot in Figure 3.2. Per sample totals are shown in Appendix B - Table 8.4. As indicated by the use of logarithmic Y-axes the number of unmapped reads per sample was variable. Soton control samples were the most consistent sample group unmapped read totals ranging from 300 to 200,000 with an average total of 5,967 reads per sample. IBD samples displayed a range from 1 to 500,000 approximately with the average of 4,151 reads per sample. 50% of samples in the IBD cohort also fall in a smaller range from 700 to 1,000 unmapped reads.

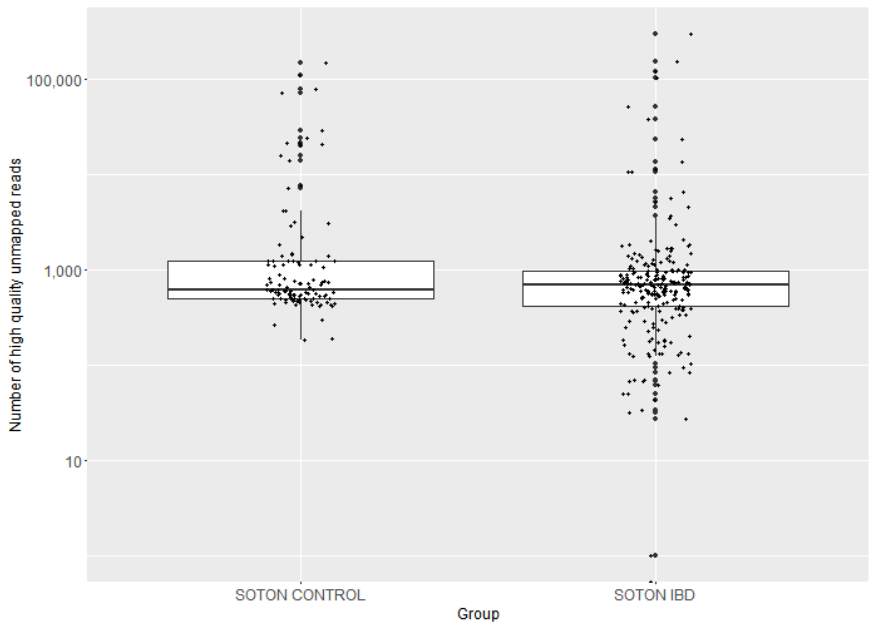

Figure 3.2: Box plots of unmapped reads by sample group. Soton control samples ranged between 300 and 200,000 unmapped reads per sample with the average of 4,151 reads per sample. Soton IBD samples a thicker segment can be seen between 500 to 1,000 with an average of 5,967 reads per sample..

By far the largest matched species from samples was *phiX174* which totalled 2,555,986 million across samples. Detection of *phiX174* is a technical artefact hence it was excluded from sample results shown in this chapter, all values have been reduced to reflect this.

For each sample unmapped reads were matched against databases using BLAST with only the best match per read retained as shown in Figure 3.1. Full results for each sample are summarised in Appendix B, Table 8.5. Results of the best matches were then imported into MEGAN6 to allow calculation of phylogenetic trees and the percentage of reads best mapped to each database as shown in Figure 3.3 which shows the percentage matches per sample to each of the databases for control samples and cases.

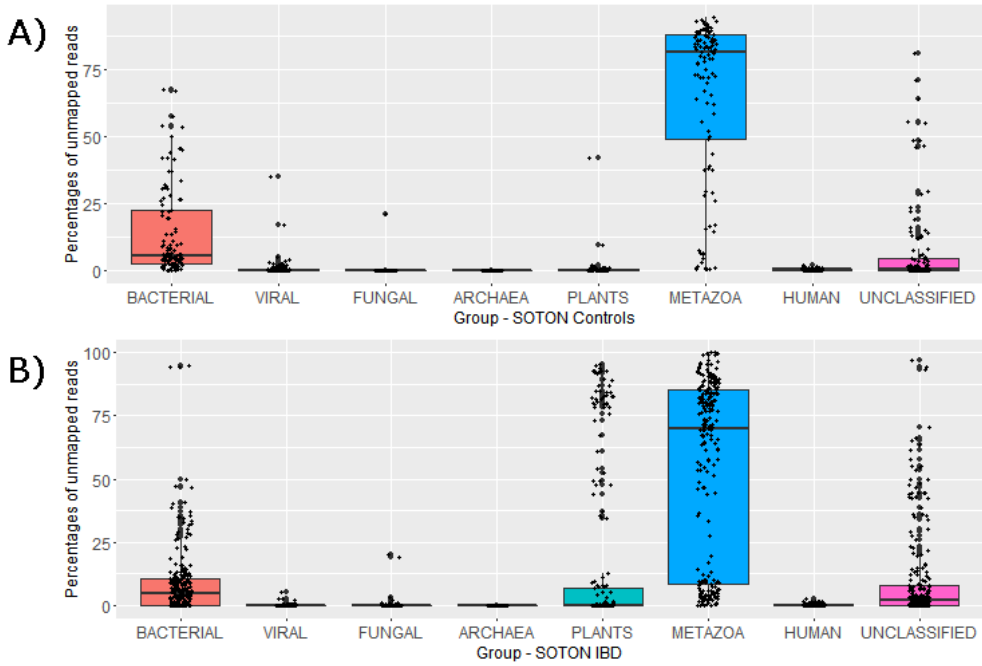

Figure 3.3: Percentage matches to databases per exome sample split by **A)** Soton controls and **B)** Soton IBD samples. The four databases reporting the highest percentage match were: bacteria, plant, metazoa and unclassified. Few human matches were detected when considering the best match per read likely due to the high number of primate matches in the metazoan database. A lower percentage of plant matches were found in the contol samples compared to cases.

The four databases reporting the highest percentage match were: bacteria, plant, metazoa and unclassified. A lower percentage of plant matches were found in the control samples compared to cases whereas the percentage of metazoan read matches are higher in control samples than in cases.

Clustering and PCA analysis was undertaken to investigate if differences could be found between the IBD samples and controls using the unmapped reads as shown by Figure 3.4. Plot A shows PCA analysis using the total read matches to each database, for which only 30.3% of variance is explained by PC1 and 19.7% for PC2

totalling 50%. Clusters shown display majority overlap between case samples and controls. Plot B performs PCA analysis using the percentages of reads mapped to each of the databases for which PC1 explained 26.7% and PC2 explained 16.6% totalling 43.3%. While some of the IBD samples lie outside of the control cluster most samples still overlap.

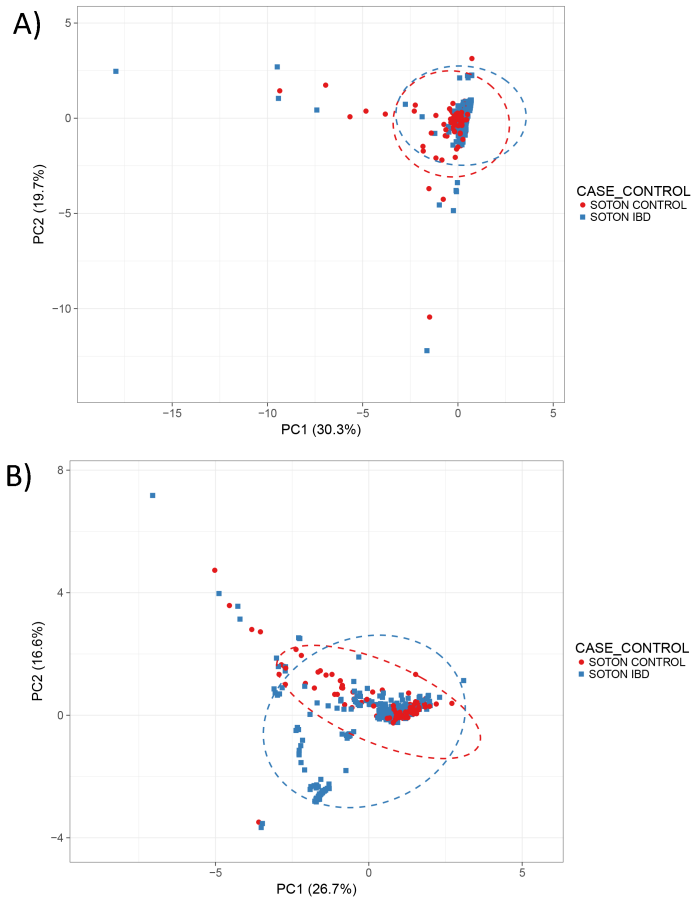

Figure 3.4: Clustering samples using databases matches comparing cases and controls. Clustering by **A)** Total matches to databases, 30.3% of variance is explained by PC1 and 19.7% for PC2 totalling 50%. **B**) Percentage database matches as shown in Figure 3.6 but here comparing the differences between cases and controls. PC1 explained 26.7% and PC2 explained 16.6% totalling 43.3% There was no substantial differences that could reliably differentiate between cases and controls.

Classifying reads down to a genus level identifies 980 entries across the 358 samples. 173 of the 980 species classifications had more than 100 reads across the samples. 100 total reads across all samples for a species was used as a minimum cut-off to reduce spurious matches in the data with low read totals. The 173 species matches were divided as 132 bacteria, 2 viral, 18 metazoa and 21 plants.

Top matched bacterial species - WES samples

Due to the high number of bacterial species returned the bacteria were sub-divided by phylum to analyse which gave six groups: Actinobacteria, Bacteroides, Firmicutes, Fusobacteria, Proteobacteria and Spirochate. Within each of these groups the due to the high number of species the threshold for prioritising taxa was raised to 500 unmapped reads. This reduced the number of species of interest down from 132 to 56.

The Acintobacteria phylum contains Gram-negative, strictly aerobic, non-fermenting bacteria. Five species were identified from three distinct genera: *Rothia*, *Atopobium* and *Propinibacterium* with the totals for each species shown below in Table 3.4. *Rothia* contained two species matches, *Rothia mucilaginosa* & *Rothia dentocariosa*, which totalled the most and 13th most read matches of any of the species at 103,492 and 15,730 respectively across the 358 samples.

Phylum	Species	Total
Actinobacteria	*Rothia mucilaginosa*	103,492
Actinobacteria	*Rothia dentocariosa*	15,730
Actinobacteria	*Atopobium parvulum*	3,660
Actinobacteria	*Propionibacterium propionicum*	1,160
Actinobacteria	*Propionibacterium acnes*	558

Table 3.4: Bacterial unmapped reads mapping to the phylum Acintobacter. Five species were detected with above 500 reads over the 358 exome samples. The two highest matched species were both from the genus *Rothia* and associated with the oral cavity and dental disease as in *Atopobium parvulum*. Both *Propionibacterium* genus species are bacteria usually found on the skin.

The Bacteroides phylum contains Gram-negative, obligate anaerobic bacteria usually found as the a major part of the normal gastrointestinal flora. Nine species were identified as shown in Table 3.5. Five species are from the genus *Prevotella* most associated with the oral cavity, the species *Prevotella melaninogenica* was the third highest matched out of all species.

Phylum	**Species**	**Total**
Bacteroidetes	*Prevotella melaninogenica*	71,118
Bacteroidetes	*Prevotella sp. oral taxon 299*	9,962
Bacteroidetes	*Prevotella denticola*	4,377
Bacteroidetes	*Prevotella intermedia*	3,910
Bacteroidetes	*Porphyromonas gingivalis*	3,427
Bacteroidetes	*Capnocytophaga ochracea*	1,721
Bacteroidetes	*Tannerella forsythia*	766
Bacteroidetes	*Prevotella dentalis*	589
Bacteroidetes	*Bacteroides vulgatus*	553

Table 3.5: Bacterial unmapped reads mapping to the phylum Bacteroides phylum contains Gram-negative, obligate anaerobic bacteria usually found as the a major part of the normal gastrointestinal flora. Nine species were detected, five to the genus *prevotella* which are found at various sites in the human body. *Porphyromonas gingivalis* is also associated with the oral cavity and most commonly with oral diseases such as gingivitis. Both species *Tannerella forsythia* and *Capnocytophaga ochracea* are also associated with oral diseases while *Bacteroides vulgatus* is usually associated with the gut.

The Firmicutes phylum contains Gram-positive bacteria but contains both anaerobic and aerobic species. Whilst 18 species were reported 15 of which fall under the genus *Streptococcus* as shown in Table 3.6. *Streptococcus* strains are usually found in the upper respiratory tract though are capable of colonising mucosal surfaces and found as part of the normal microbiome though some species are opportunistic pathogens. *Veillonella parvula*, *Selenomonas sputigena* and *Bacillus subtilis*, a commensal bacteria of the gut were the other three species identified. Two species of Fusobacteria were identified: *Fusobacterium nucleatum* and *Leptotrichia buccalis* with 3,572 and 1,061 reads respectively.

Phylum	**Species**	**Total**
Firmicutes	*Streptococcus parasanguinis*	26,962
Firmicutes	*Veillonella parvula*	20,656
Firmicutes	*Streptococcus pneumoniae*	16,742
Firmicutes	*Streptococcus mitis*	13,475
Firmicutes	*Streptococcus oralis*	12,218
Firmicutes	*Streptococcus pseudopneumoniae*	6,557
Firmicutes	*Streptococcus sp. I-P16*	5,501
Firmicutes	*Streptococcus sanguinis*	5,280
Firmicutes	*Streptococcus salivarius*	5,204
Firmicutes	*Selenomonas sputigena*	3,170
Firmicutes	*Streptococcus gordonii*	3,032
Firmicutes	*Streptococcus sp. I-G2*	2,224
Firmicutes	*Bacillus subtilis*	2,125
Firmicutes	*Streptococcus oligofermentans*	1,973
Firmicutes	*Streptococcus intermedius*	974
Firmicutes	*Streptococcus dysgalactiae*	973
Firmicutes	*Streptococcus thermophilus*	932
Firmicutes	*Streptococcus anginosus*	517

Table 3.6: Bacterial unmapped reads mapping to the phylum Firmicutes. 15 of the 18 species were from the genus *streptococcus* which predominantly colonise the upper respiratory tract. Both *Veillonella parvula*, *Selenomonas sputigena* are found in the upper respiratory tract as part of the normal flora. *Bacillus subtilis* is found as part of the normal commensal species in the gut.

The remaining phylum was Proteobacteria which includes many Gram-negative pathogens. Due to the opportunistic, pathogenic nature of many of the species in this phylum they are found across a wide range of mucosal surfaces. 22 species were detected with above 500 reads as shown in Table 3.7 sorted by species name.

Phylum	**Species**	**Total**
Proteobacteria	*Aggregatibacter actinomycetemcomitans*	587
Proteobacteria	*Aggregatibacter aphrophilus*	2,672
Proteobacteria	*Burkholderia ambifaria*	1,884
Proteobacteria	*Burkholderia cenocepacia*	4,517
Proteobacteria	*Burkholderia cepacia*	1,060
Proteobacteria	*Burkholderia lata*	2,858
Proteobacteria	*Burkholderia vietnamiensis*	1,023
Proteobacteria	*Campylobacter concisus*	9,410
Proteobacteria	*Cronobacter sakazakii*	5,634
Proteobacteria	*Escherichia coli*	8,094
Proteobacteria	*Haemophilus influenzae*	3,589
Proteobacteria	*Haemophilus parainfluenzae*	35,495
Proteobacteria	*Klebsiella pneumoniae*	1,137
Proteobacteria	*Neisseria gonorrhoeae*	8,644
Proteobacteria	*Neisseria lactamica*	5,986
Proteobacteria	*Neisseria meningitidis*	24,822
Proteobacteria	*Pseudomonas mendocina*	3,631
Proteobacteria	*Ralstonia pickettii*	1,202
Proteobacteria	*Salmonella enterica*	7,224
Proteobacteria	*Shigella boydii*	539
Proteobacteria	*Stenotrophomonas maltophilia*	10,689
Proteobacteria	*Variovorax paradoxus*	8,888

Table 3.7: Bacterial unmapped reads mapping to the phylum Proteobacteria. 22 species were detected with above 500 reads including some highly pathogenic species such as *Cronobacter sakazakii*, *Salmonella enterica* and *Shigella boydii*.

Two *Aggregatibacter* species were identified: *Actinomycetemcomitans* and *Aphrophilus*. Five species of *Burkholderia* were identified, this genus has three documented species most of which are soil saprophytes and phytopathogens. Two *Haemophilus* species were detected: *Influenzae* and *Parainfluenzae* both of which are found as part of the normal commensal species in the oral cavity. The species *Klebsiella pneumoniae* was the highest detected *Klebsiella* species and only with above 500 reads. Three *Neisseria* genus species, *N. gonorrhoeae*, *N. lactamica* and *N. meningitidis*, were identified all of which have are present in a humans as part of the natural flora.

Three further species identified were: *Pseudomonas mendocina*, *Variovorax paradoxus* and *Stenotrophomonas maltophilia*. Two species were identified as being related to IBD or significantly pathogenic. *Ralstonia pickettii* is part of the *Ralstonia* genus, split from the *Burkholderia* genus and *Campylobacter concisus*.

The remaining three species were all associated with diseases spread by contaminated food or water. *Salmonella enterica* has approximately 2500 serovars and typically are orally acquired from contaminated food. *Shigella boydii* is associated with acute diarrhea. *Cronobacter sakazakii* is identified which among infants in associated with necrotising enterocolitis.

Five of the species identified in the Proteobacteria phylum were identified as being particularly pathogenic or associated with IBD. The number of samples and the average number of reads mapping to these five samples are shown in Table 3.8, only samples with above ten reads were counted. *Cronobacter sakazakii* had the lowest number of samples above ten reads but the highest maximum and average per sample.

Species	***Salmonella enterica***	***Cronobacter sakazakii***	***Campylobacter concisus***	***Ralstonia pickettii***	***Shigella boydii***
No. of Samples	60	3	18	15	16
Max	2128	2701	2572	311	49
Average	117	1872	520	72	23

Table 3.8: Number of samples with ten or more reads for selected bacterial species for the phylum Proteobacteria. Also shown is the maximum and average number of reads for each of the samples. *C. sakazakii* had the highest average and maximum of the selected most pathogenic species.

Top matched viral species

Only two viral species were identified with over 100 reads total across samples: *Enterobacteria phage HK629* - 340 reads & *Enterobacteria phage lambda* - 195 reads. Eight samples had more than 10 reads to the *E. phage HK629* species with a maximum of 143 in a single sample. Only three samples had above 10 reads matching to the *E. phage lambda*, all three of which were those with above 10 reads from the *E. phage HK629* only one of the samples was an IBD sample.

Sample	Group	Type	*Enterobacteria phage HK629*	*Enterobacteria phage lambda*
CL018-3868-Cous2	Control	NA	50	29
NG151-2	Control	NA	17	7
PR0098	IBD	IBDU	13	5
PT006	Control	NA	143	93
W1200013 MK	Control	NA	11	3
W1215091 MH	Control	NA	15	10

Table 3.9: Top matched viral species from exome samples. Only two viral species were detected *Enterobacteria phage HK629* - 340 reads & *Enterobacteria phage lambda* -195 reads, with relatively low read totals and per sample matches.

Top matched plant species

21 plant species were matched over the exomes with more than 100 total read matches. The identified species covered a wide variety of genera ranging from commonly consumed foods such as tomatoes or cocoa or the model plant organisms used by many research groups *Arabidopsis thaliana* as shown in Table 3.10. The top three matches were to *Gossypium hirsutum* (upland cotton), *Arabidopsis thaliana* (model plant organism) and *Citrus sinensis* (sour orange) as shown in Table 3.10. When considering any samples with above 10 reads between the top three species 57 samples were identified. 55 of the 57 samples were all from three batches.

Species	Description	Total
Gossypium hirsutum	Upland cotton (South,Central America)	28,951
Arabidopsis thaliana	Model plant organism	23,058
Citrus sinensis	Sour orange	14,291
Citrus clementina	Clementine	6,450
Solanum pennellii	Wild tomato	1,535
Gossypium arboreum	Tree cotton (Asian)	1,197
Solanum lycopersicum	Tomato	1,197
Malus domestica	Apple	1,078
Nicotiana tabacum	Cultivated tobacco	954
Nicotiana sylvestris	Woodland tobacco	715
Gossypium raimondii	Cotton plant (Peru)	648
Nicotiana tomentosiformis	Wild tobacco	598
Brassica rapa	Root Veg subspecies including : Turnip, bok choi	469
Theobroma cacao	Cocoa tree	465
Medicago truncatula	Barrelclover - Model organism	432
Populus euphratica	Euphrates poplar tree	322
Capsicum annuum	Peppers	175
Brassica napus	Rapeseed	150
Fragaria vesca	Wild strawberry	148
Arachis duranensis	Peanut model species	108
Pyrus x bretschneideri	Chinese white pear	101

Table 3.10: Top plant species matches from unmapped exome reads. Three samples were matched with particularly high read match totals: *Gossypium hirsutum* (upland cotton), *Arabidopsis thaliana* (model plant organism) and *Citrus sinensis* (sour orange). All of which have been genome sequenced in recent years and would have been present in some sequencing centres.

Top matched metazoan species

18 metazoan species met the criteria of containing 100 or more reads as shown below in Table 3.11. Whilst a wide variety of species were detected amongst the 18 species the top matched species were all mammalian.

Species	Description	Total
Pan troglodytes	Chimpanzee	80,958
Gorilla gorilla	Gorilla	21,574
Mus musculus	Mouse	20,218
Pan paniscus	Bonobo	10,128
Octodon degus	Common degu	7,368
Pongo abelii	Sumatran orangutan	2,470
Nomascus leucogenys	Northern white-cheeked gibbon	1,581
Rhinopithecus bieti	Black snub-nosed monkey	1,207
Homo sapiens	Human	946
Pantholops hodgsonii	Tibetan antelope	661
Macaca mulatta	Rhesus macaque	318
Bos taurus	Cow	299
Colobus angolensis	Angola colobus (Angolan Monkey)	271
Acinonyx jubatus	Cheetah	251
Nannospalax galili	Mole rats	151
Chlorocebus sabaeus	Green Monkey	126
Bison bison	Bison	109

Table 3.11: Top metazoan species matches from unmapped exome reads. The top matched species were primates or mammals with high similarity and homology to *homo sapiens* and are likely human reads in origin which failed inital mapping and now map better to other metazoan species.

The top four matched species all had above 10,000 read matches were: Chimpanzee (80,958), Gorilla (21,574), Mouse (20,218) and Bonobo (10,128). All of these organisms are amongst the closest related to the human genome and may be due to errors in or variable regions which failed to map against the human genome initially.

3.4.2 Comparing collection methods

In exome samples the top matched bacterial species were predominantly associated with the oral cavity and the upper respiratory tract. To investigate if differences were therefore due to the collection methods the 12 known saliva captured exome samples were compared against the 233 blood and 113 unknown collection method exome samples. Figure 3.5 compares the number and percentage of matches to each of the databases when grouped by the collection method (blood, saliva or unknown).

Comparisons of the total unmapped reads which passed filtering, as shown in Plot A of Figure 3.5, reveal that the blood samples ranged from 0 - 14,000 reads per sample compared with 3,500 - 300,000 for saliva. Only five of the 233 blood samples had above 3,500 reads per sample (1.5% of samples). The median average for blood was calculated as 651 reads per sample compared with 30,509 for saliva.

Viral unmapped reads are shown in Panel B of Figure 3.5 to investigate if more viral reads could be detected in either method. However consistently low totals for virus reads were returned and therefore not plotted at a percentage level.

Plots C & D show the percentage and totals of bacterial reads matched respectively. Comparing the percentages of bacterial reads shows that the median for blood is 4.34% compared with 32.97% for saliva. 1 of the 12 saliva samples had 0.1% bacterial reads which was identified as the sample shown in plot E for saliva which had excessive plant reads and displays an abnormal pattern compared to the other 11 saliva samples. Excluding this sample the other 11 saliva samples ranged from 26.7 - 47.2% of the reads per sample mapping to bacteria. 17 of the 233 blood collected samples had above 25% of reads mapping to bacteria. Comparing the total bacterial reads shows that most saliva samples have upwards of 1,000 bacterial read matches, averaging 24,041 matches per sample whereas for blood only have 4 of 233 samples above 1,000 reads and average 78 bacterial matches per sample. Unknown collection samples averaged 2,025 bacterial read matches per sample.

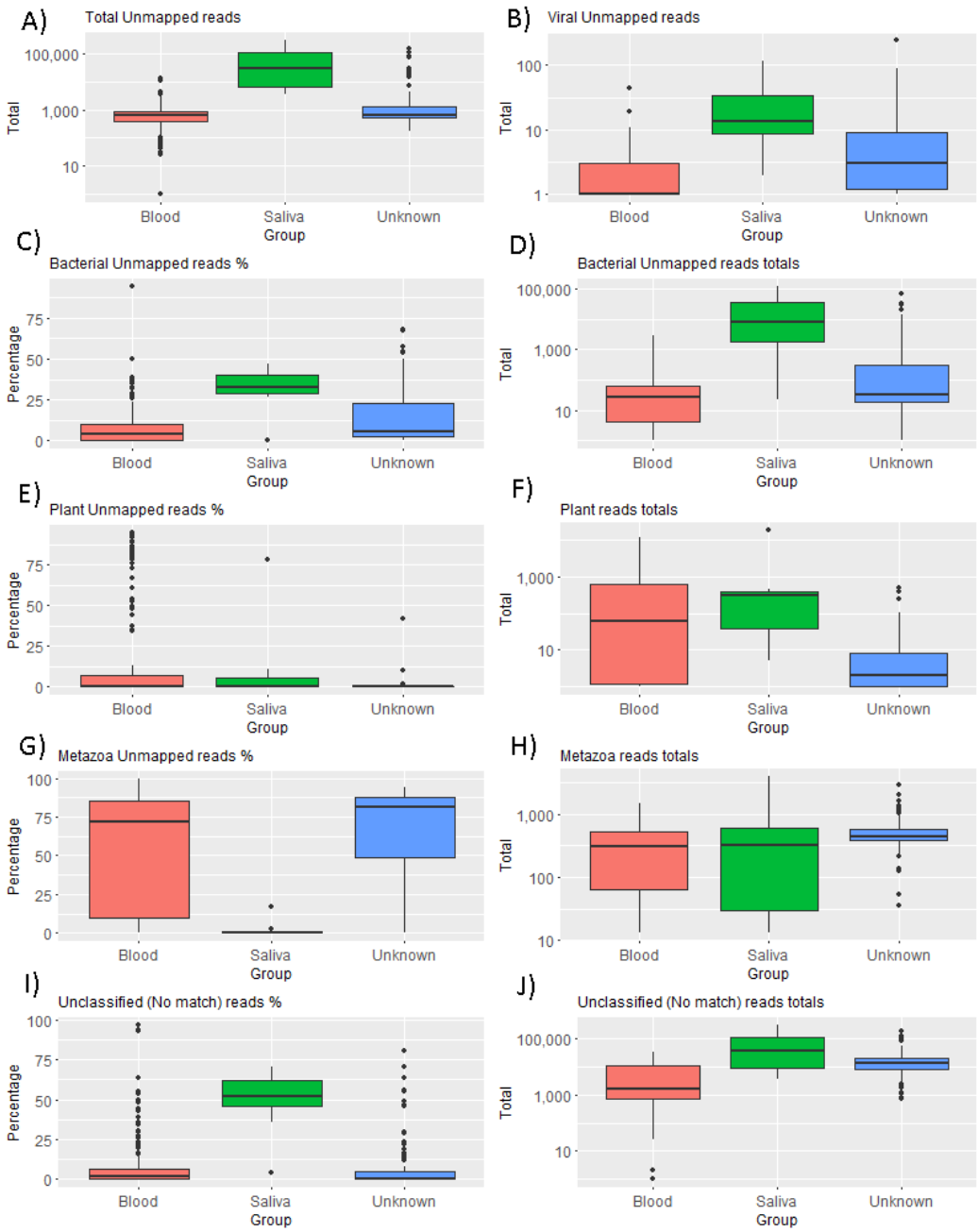

Figure 3.5: Comparison of unmapped read totals and percentages by collection methods of blood, saliva and unknown. **A)** Total unmapped reads were higher in saliva samples compared to blood **B)** Number of viral unmapped reads by collection method, most samples had few viral unmapped reads. **C)** Percentage matches to bacterial taxa per sample, higher percentages of bacterial reads were found in saliva samples. **D)** Number of bacterial taxa matches per sample, higher bacterial read totals were identified from saliva. **E)** Percentage matches to plant taxa per sample, few differences are seen between blood and saliva. **F)** Number of plant reads matched per sample. **G)** Percentage matches to metazoan unmapped reads. **H)** Number of metazoan read matches per sample shows similar numbers between blood and saliva. **I)** Percentage of unclassified reads were typically low for blood and higher at around 50% for saliva **J)** Number of unclassified was similar between collection methods but was slightly higher than blood samples.

Panels E, F & G, H show the percentage and total reads for plants and metazoan matches respectively. There were few differences between collection methods when comparing the read totals within panels F and H as the distributions overlap almost entirely between blood and saliva collections. Panel E for the percentage of matches to plants showed some differences between blood and saliva, with saliva samples between 0 and 11% when excluding a single sample with 79% plant matches. 183 blood samples have below 13% plant read matches, leaving 50 samples with higher plant percentages but the majority of samples still overlapped with saliva. Metazoan read matches display the same pattern as plants with near complete overlap of the distributions for total reads as shown in Panel H. Percentage matches as shown in Panel G shows that most saliva samples range from 0.2 to 2.3% with the same odd sample as in panel E having 17.2% metazoan matches. Blood samples however are much more extremely distributed with: 70 samples below 15%, 22 samples between 16 - 60 % and 141 samples above 60 %. This made 19.5 % of blood samples with below 15% metazoan read matches.

The final category visualised was unclassified reads, when no high quality match for the read could be found. In both the percentage and total reads plots, Panels I & J respectively, the saliva species distributions are higher with a median of 37,339 unclassified reads per sample compared with 1,552 reads for blood. Percentage matches display a similar pattern with 52.4% of the total reads being unclassified in saliva compared to 1.93% in blood. Outliers of the blood collection samples overlapped or were higher than the saliva distribution as shown in Panel I and Panel J looking at the read totals also has overlap of the boxes in the plot.

Individually when examining the total and percentage matches to each of the categories shown in Figure 3.5 differences can be seen between collection methods. However outliers from the blood collected samples often overlap with the saliva distributions complicating the analyses. To try and leverage all of the available data into a single analysis therefore and better pull apart the collection methods principal component analysis was performed as shown below in Figure 3.6.

Panel A shows PCA and clustering performed using the total read matches per sample for: bacteria, virus, fungi, archaea, plants, metazoa, human and unclassified matches, While in Panel B percentage values were used instead for the same categories. Panel A shows that principal component 1 and principal component 2 that explain 30.3% and 19.7% of the total variance, respectively adding to 50%. Panel A shows tight clustering of the blood collection samples, as indicated by the dashed red line, with a minority of samples visibly outlying. With Panel B principal component 1 and principal component 2 explain 26.7% and 16.6% of the total variance, respectively adding to 43.4%.

Panel B is based on the percentage values and does not cluster the blood collection samples as tightly as in Panel A and appears to split the data into multiple smaller clusters. A single saliva sample is placed near to a blood cluster in the bottom left corner which was the sample with abnormally high plant content.

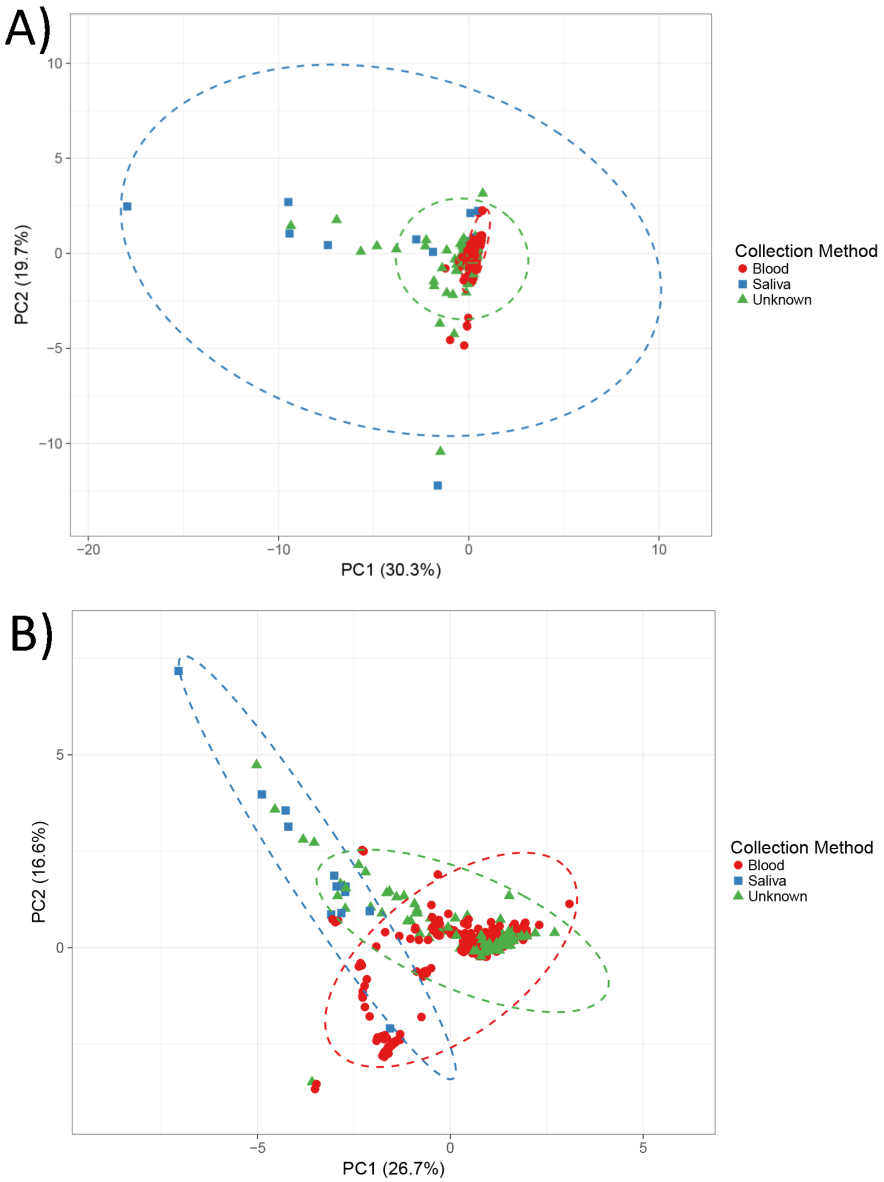

Figure 3.6: **A)** Clustering performed using total read matches to each of the databases used for all of the 358 exome samples. Blood samples formed a reasonably tight cluster with many unknown collection samples clustering with the blood samples. 21 unknown samples are distinguishable from the blood samples and are more likely saliva samples. X and Y axis show principal component 1 and principal component 2 that explain 30.3% and 19.7% of the total variance, respectively. Prediction ellipses are such that with probability 0.95. **B)** Clustering performed using percentage read matches. Blood samples do not cluster tightly instead forming two clusters and several samples scattered around either of the larger blood clusters. Saliva and blood samples are more clearly differentiable with 24 samples appearing to be saliva. X and Y axis show principal component 1 and principal component 2 that explain 26.7% and 16.6% of the total variance, respectively. Prediction ellipses are such that with probability 0.95, a new observation from the same group will fall inside the ellipse.

3.4.3 Unmapped reads of RNA-SEQ and WES from the same individual

For the 10 samples shown in Table 3.1 with RNA-seq and or WES performed unmapped reads were extracted and compared using the same pipeline as previously used for exome sequencing. 7 of the 10 samples were cases of IBD, which were also the samples with WES data was also available all from blood collections. The total reads in the BAM files, the unmapped reads for RNA-seq and for the corresponding WES samples are shown in Table 3.12. WES samples had a higher number of reads in the original BAM files but fewer unmapped reads than RNA-seq.

Individual	Type	Site	Total reads	Unmapped Reads
SOPR310	WES	Blood	48,769,525	131
SOPR311	WES	Blood	52,280,734	1099
SOPR313	WES	Blood	59,858,582	183
SOPR314	WES	Blood	56,416,935	190
SOPR317	WES	Blood	48,637,996	250
SOPR318	WES	Blood	55,894,788	10471
SOPR320	WES	Blood	52,024,278	131
SOPR310	RNA-seq	Illeal	5,391,979	96,420
SOPR311	RNA-seq	Illeal	4,375,818	30,713
SOPR314	RNA-seq	Illeal	4,170,622	36,946
SOPR317	RNA-seq	Illeal	4,852,103	71,738
SOPR318	RNA-seq	Illeal	4,880,614	85,827
SOPR322	RNA-seq	Illeal	5,877,048	91,692
SOPR315	RNA-seq	Illeal/Caecum	5,480,742	33,044
SOPR321	RNA-seq	Illeal/Caecum	4,562,559	49,100
SOPR310	RNA-seq	Rectal	5,784,351	114,667
SOPR311	RNA-seq	Rectal	6,541,305	200,430
SOPR313	RNA-seq	Rectal	4,491,428	52,171
SOPR315	RNA-seq	Rectal	4,833,543	52,083
SOPR317	RNA-seq	Rectal	6,389,926	126,438
SOPR318	RNA-seq	Rectal	9,524,197	288,415
SOPR320	RNA-seq	Rectal	6,126,678	77,725
SOPR321	RNA-seq	Rectal	5,904,713	62,713
SOPR322	RNA-seq	Rectal	4,522,074	33,804

Table 3.12: Total unmapped reads for seven WES and all 17 RNA-seq samples. WES samples had few reads with the exception being SOPR318 though 80% of the reads were to plant species and were likely contamination. RNA-seq samples contained between 30,713 and 288,415 unmapped reads.

Unmapped reads for each of the 17 RNA-seq and 10 exome samples were matched to each of the databases using BLAST with the pipeline as shown in Figure 3.1. Best read matches were then drawn in MEGAN6 and summarised as shown in Figure 3.7 with Panel A showing the percentage of read matches per sample to each database for exomes and Panel B the matches for reads from RNA-seq samples.

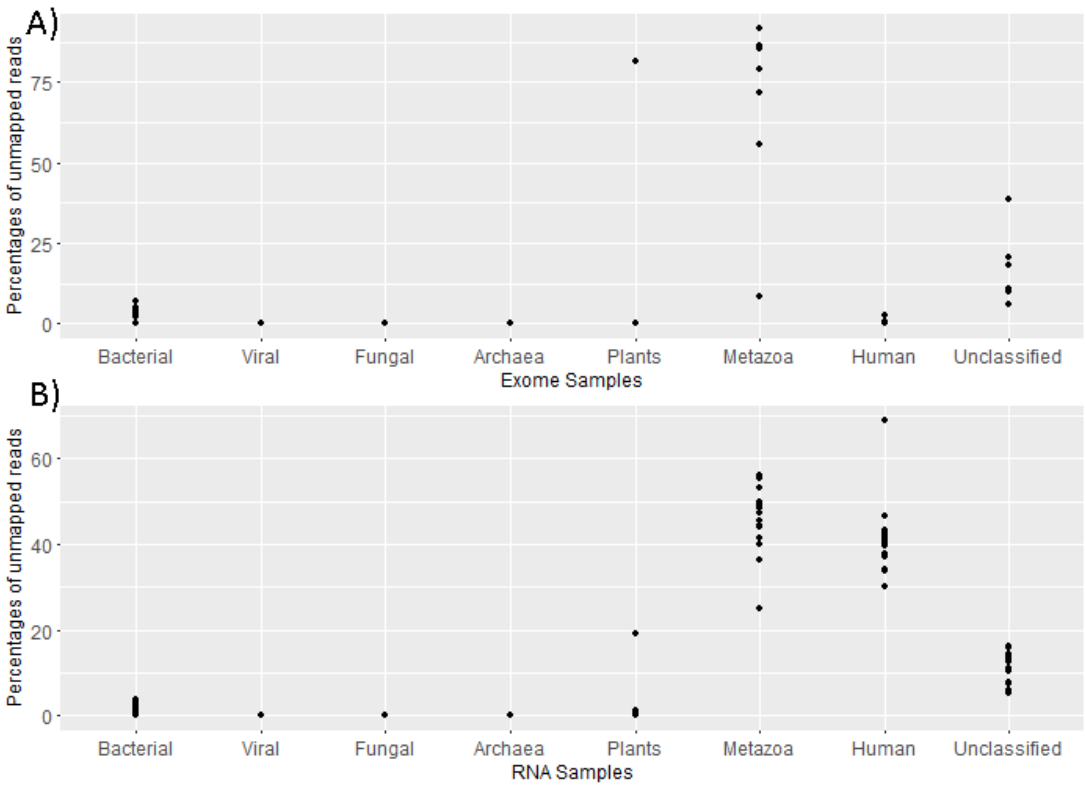

Figure 3.7: Exome and RNA-seq sample unmapped reads by database. **A)** Percentage matches to databases for WES samples showed most notable matches were to metazoa or unclassified. A single sample was sequenced in the batch with plant contamination and explains the 80% plant matches. All samples had low bacteria matches likely due to the blood collection method. **B)** Percentage matches to databases for RNA-seq samples which mirrored the trends seen in exome samples but have a much higher percentage human match due to the use of hg19 for initial alignment and subsequent matches then to the hg38 reference used by BLAST.

Figure 3.7 shows that both the WES and RNA-seq have low bacterial mapping percentages up to 7% but always remain a minority of the unmapped reads per sample. In WES samples only one sample (SOPR318) was identified with any plant reads, this sample was unusual and was sequenced in a different batch to the other six samples. By contrast human matches for unmapped reads in RNA-seq data is much

higher at 30 - 70%. In both WES and RNA-seq data a high proportion of matches were either to other metazoan species or were unclassifiable.

Exome matches

Examination of the species from the 10 WES samples with above 10 read matches was performed and identified 19 species as shown in Table 3.13.

Species	Common Name	SOPR-310	311	313	314	317	318	320	Total
Arabidopsis thaliana	Plant model organism	0	0	0	0	0	4270	0	4270
Gossypium hirsutum	Upland cotton (South,Central America)	0	0	0	0	0	3416	0	3416
Pan paniscus	Bonobo	78	474	82	116	96	50	79	975
Mus musculus	Mouse	0	0	0	0	0	782	0	782
Solanum lycopersicum	Tomato	0	0	0	0	0	319	0	319
Gorilla gorilla	Gorilla	14	171	27	27	23	16	16	294
Pan troglodytes	Chimpanzee	13	174	20	9	12	10	17	255
Solanum pennellii	Tomato	0	0	0	0	0	126	0	126
Gossypium arboreum	Tree cotton (Asian)	0	0	0	0	0	117	0	117
Theobroma cacao	Cocoa tree	0	0	0	0	0	74	0	74
Gossypium raimondii	Cotton plant (Peru)	0	0	0	0	0	71	0	71
Capsicum annuum	Chilli Pepper	0	0	0	0	0	55	0	55
Pongo abelii	Sumatran orangutan	5	27	2	9	6	1	4	54
Nicotiana tomentosiformis	Wild Tobacco	0	0	0	0	0	41	0	41
Homo sapiens	Human	1	28	1	0	1	0	1	32
Brassica napus	Rapeseed	0	0	0	0	0	15	0	15
Nomascus leucogenys	Northern white-cheeked gibbon	1	8	0	2	0	3	1	15
Rhinopithecus bieti	Black snub-nosed monkey	0	5	0	1	1	3	1	11
Raphanus sativus	Radish	0	0	0	0	0	10	0	10

Table 3.13: Exome species with 10 or more unmapped reads. 19 species were identified, none of which were bacterial. All plant species reads originated from the sample SOPR318. The top non-plant species were all metazoan with primates and mouse highest amongst the matches.

The top two matched species *Arabidopsis thaliana* & *Gossypium hirsutum* are plant species and were only detected in the sample SOPR318. Amongst the 19 species there were 11 plant species detected with above 10 reads and were only detected from the sample SOPR318. Of the eight other species detected with above 10 reads all were best matched to primates mouse or human.

PCA was performed based on species above 10 reads in exome samples as shown in 3.8. Most exome samples are clustered together whilst two samples are separated. SOPR318 is unique and had the high number of plant matched reads. Also sample SOPR311 WES had the second highest number of unmapped reads most of which were found by BLAST to map back to the human reference at 28 compared to the next highest of sample with one as shown in Table 3.13. From PCA analysis PC1 explained 68.5% of variance whilst PC2 explained 27.4% a total of 95.9% of all variance.

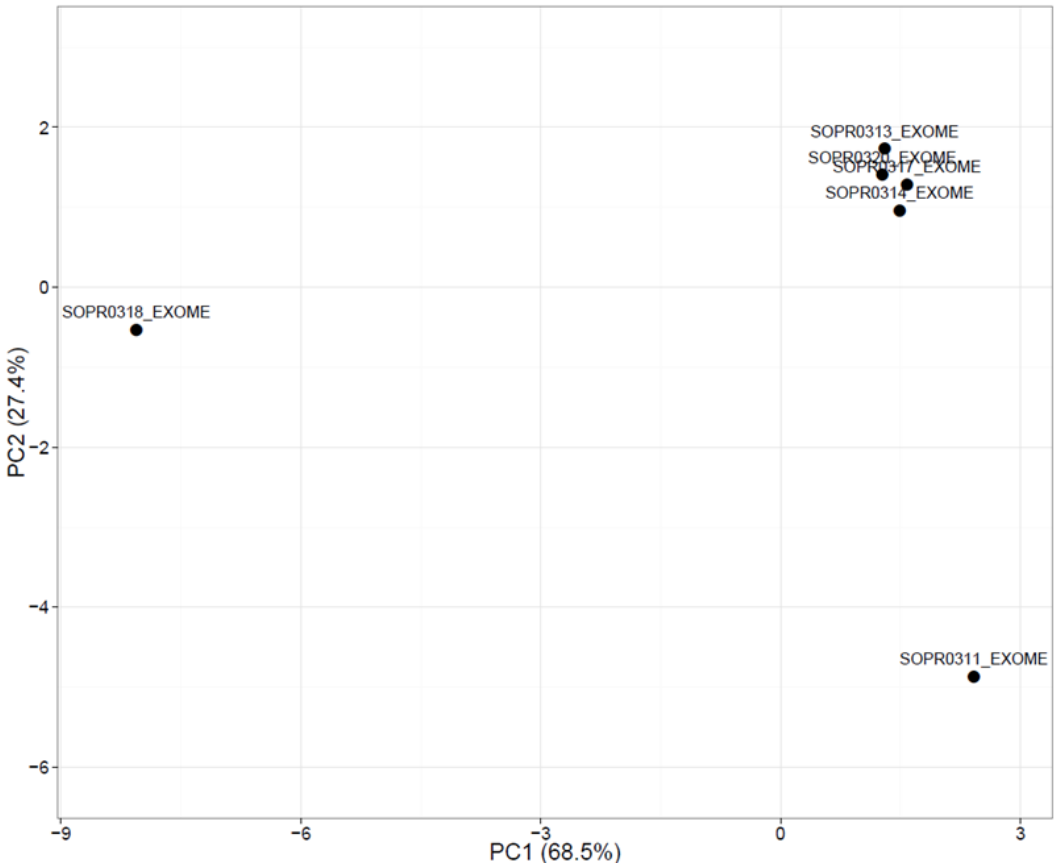

Figure 3.8: Clustering of the seven WES samples from individuals which also had RNA-seq performed. Four of the seven WES samples cluster due to the low species matches. SOPR0318 had excessive plant matches and SOPR0311 had higher number of unmapped reads extracted and the most reads mapping back to the human reference at 28 compared to the next highest of one. PC1 explained 68.5% of variance whilst PC2 explained 27.4% a total of 95.9% of all variance.

RNA-seq matches

Analysis of the 17 RNA-seq samples contained 182 species with 10 or more read matches. To reduce the number to a manageable list a filter of 1000 read matches or above was applied which left 32 species as shown in Table 3.14. The top matched species was *homo sapiens* with 691,282 reads across the 17 samples followed by *Pan troglodytes* (chimpanzee) with a total of 239,977. The next nine species range from 11,682 - 78,403 and are all primates. The highest matched non-mammalian species was *Gossypium hirsutum* (Upland cotton) though it had a total of 11,211 reads, 11,054 (98.6%) of which came from the sample SOPR0321-Rectal.

Species	0310Ileal	0310Rectal	0311Ileal	0311Rectal	0313Rectal	0314Ileal	0315Ileal	0315Rectal	0317Ileal	0317Rectal	0318Ileal	0318Rectal	0320Rectal	0321Ileal	0321Rectal	0322Ileal	0322Rectal	all
Homo sapiens	40989	47129	10335	86906	19684	12652	13387	19243	30003	58957	34846	198130	30759	19776	18897	37092	12497	691282
Pan troglodytes	17326	23974	6671	33367	11367	8759	6662	11936	15276	19891	18955	0	14356	11740	12276	20133	7288	239977
Pan paniscus	3324	4476	1302	8876	2510	1899	1489	2496	2683	3497	4290	29020	3316	1939	2064	4024	1198	78403
Gorilla gorilla	4298	4523	627	14416	2033	740	1168	1379	2134	8305	2055	10156	3250	1314	1568	2250	917	61133
Aotus nancymaae	3630	7038	1464	4665	2384	1951	1171	2070	2650	2459	3342	0	2293	1195	1578	3838	1425	43153
Macaca mulatta	1053	2008	405	1716	741	856	621	922	793	1146	1131	6238	1081	871	912	1097	596	22187
Colobus angolensis	1285	2024	555	2383	1013	836	861	1287	1171	1638	1564	0	1328	1190	1144	1276	820	20375
Callithrix jacchus	1189	2062	473	1579	627	652	389	680	1021	0	1075	3035	643	461	480	1275	497	16138
Pongo abelii	1398	1665	190	1647	547	444	308	392	632	1007	785	3106	579	443	376	801	253	14573
Nomascus leucogenys	769	859	457	1295	386	437	255	447	528	598	1108	2119	460	470	478	1329	303	12298
Papio anubis	709	897	234	1322	350	338	285	433	571	856	761	2253	570	475	500	852	276	11682
Gossypium hirsutum	0	0	11	0	0	3	0	0	0	0	0	128	6	4	11054	0	5	11211
Rhinopithecus bieti	632	1045	383	726	358	770	162	313	483	496	712	1700	451	396	435	695	202	9959
Mycoplasma pneumoniae	888	822	824	669	916	597	340	824	0	375	818	644	536	655	0	515	328	9751
Equus przewalskii	0	0	1457	26	4	0	2	6	17	1416	7	3359	1932	8	6	17	1410	9667
Chlorocebus sabaeus	426	699	226	850	327	505	165	284	346	0	645	955	506	269	283	437	180	7103
Canis lupus	321	1140	132	590	328	607	119	457	244	438	273	0	190	163	481	1264	224	6971
Macaca fascicularis	301	379	149	733	173	153	109	213	256	311	380	1219	247	177	193	414	146	5553
Cercocebus atys	232	371	101	599	208	194	193	240	263	386	324	706	216	199	243	282	148	4905
Saimiri boliviensis	206	309	81	359	163	130	76	182	162	0	229	762	176	139	182	233	99	3488
Equus caballus	9	10	582	18	7	4	4	0	7	444	22	990	678	0	0	15	464	3254
Rhinopithecus roxellana	172	232	63	390	87	33	65	104	172	263	176	513	139	134	127	183	70	2923
Macaca nemestrina	306	260	36	253	115	67	52	65	91	150	166	466	95	84	82	154	58	2500
Myotis brandtii	219	458	75	395	89	32	56	123	123	177	175	0	129	84	85	154	36	2410
Mandrillus leucophaeus	178	158	19	239	97	24	41	70	95	179	157	588	118	79	73	97	26	2238
Equus asinus	10	12	501	24	4	0	11	21	10	471	17	0	629	17	10	12	424	2173
Myotis davidii	89	151	38	268	22	30	33	59	128	210	105	448	56	38	40	134	20	1869
Tupaia chinensis	63	125	65	122	100	151	39	98	83	0	113	136	49	142	116	84	79	1565
Elephantulus edwardii	66	101	53	122	53	57	30	58	79	95	137	139	57	54	64	116	46	1327
Morus notabilis	48	247	36	76	67	139	18	0	23	59	50	43	29	13	85	280	50	1263
Microcebus murinus	86	114	28	161	48	31	28	61	82	0	72	263	37	58	54	60	45	1228
Myotis lucifugus	65	76	11	167	29	17	21	31	55	0	88	345	33	27	25	74	26	1090

Table 3.14: RNA-seq species with above 500 read matches. 57 species were found with above 500 matches. The top matched species was *homo sapiens* with 691,282 reads across the 17 samples followed by *Pan troglodytes* (chimpanzee) with a total of 239,977.

As RNA samples were mostly from IBD patients the bacterial species were those most of interest to ascertain differences between the sites and if differences between the disease sub-types could be detected. Bacterial species with above 10 reads were extracted as shown in Table 3.15.

Species	0310-I	0310-R	0311-I	0311-R	0313-R	0314-I	0315-I	0315-R	0317-I	0317-R	0318-I	0318-R	0320-R	0321-I	0321-R	0322-I	0322-R	Sum
Mycoplasma pneumoniae	888	822	824	669	916	597	340	824	0	375	818	644	536	655	0	515	328	9751
Propionibacterium acnes	338	16	19	0	16	11	8	22	11	13	21	34	72	79	31	26	7	724
Cronobacter sakazakii	0	0	0	0	0	0	0	0	0	182	0	345	0	0	0	0	0	527
Acinetobacter baumannii	280	0	0	0	0	0	3	10	0	0	0	0	0	4	0	0	0	297
Brevundimonas subvibrioides	191	0	0	0	14	5	5	14	6	0	0	0	8	8	0	0	3	254
Bacteroides vulgatus	21	35	0	0	28	0	11	51	20	37	0	0	7	0	11	0	0	221
Escherichia coli	0	0	57	20	30	0	0	13	0	0	0	0	0	0	70	0	0	190
Micrococcus luteus	12	0	0	0	0	0	2	85	0	0	0	0	11	16	0	9	0	135
Acidovorax sp. KKS102	21	12	0	0	25	12	5	16	0	0	10	0	0	6	0	15	0	122
Neisseria meningitidis	15	14	2	0	21	6	2	17	0	0	8	0	0	9	0	0	28	122
Listeria monocytogenes	0	0	0	0	0	0	0	0	0	46	0	72	0	0	0	0	0	118

Table 3.15: RNA-seq bacterial species with above 10 read matches. Illeal samples are denoted by an "-I " and Rectal by "-R" following the abbreviated sample names. 11 species were identified with *Mycoplasma pneumoniae* was top matched with 9,751 reads spread fairly evenly over 15 of the 17 samples. Also identified were two samples with high read matches to the species *cronobacter sakazakii*.

Mycoplasma pneumoniae was the top matched species with 9,751. The second highest match was to the skin bacteria *Propionibacterium acnes* with 724 matches. *Cronobacter sakazakii* was the third most matched bacteria, but was only found in two samples and was previously only found in three exome samples. Amongst the other bacterial species of interest were *Bacteroides vulgatus* was detected with 221 reads. Finally *Listeria monocytogenes* - a Gram-positive, particularly pathogenic bacteria totalling 118 reads though only from two rectal samples.

As shown in Figure 3.7 and Table 3.13 there were few bacterial read matches making comparisons between the exome and RNA-seq data infeasible. Clustering was performed for RNA-seq rectal and illeal samples to investigate differences in the species detected between sites and then by the sub-type of IBD as shown in Figure 3.9. As shown by both panels there was overlap of the clusters. In Panel A total read matches to each species were analysed with PCA and shaded by the collection site. Only three rectal samples can be seen outside of the red oval including the other 14 samples. Analysis of the species matches shown in Table 3.14 shows sample 0318-Rectal has a unique mapping pattern for metazoan species with no matches to *Pan troglodytes* and *Aotus nancymaae* whereas other samples all have upwards of 1,000 read matches. Sample 0315-Rectal has few distinct species matches which appear unusual compared to other samples. Sample 0311-Rectal is close to the red oval and has a noticeably high number of *Gorilla gorilla* read matches at 14,416 and *Pan troglodytes* at 33,367. Panel B uses the same values as Figure A but shades by the disease sub-types which show overlap. Principal component-1 only explained 23.3% of variance and principal component-2 14.7% giving a total of 38% of explained variance.

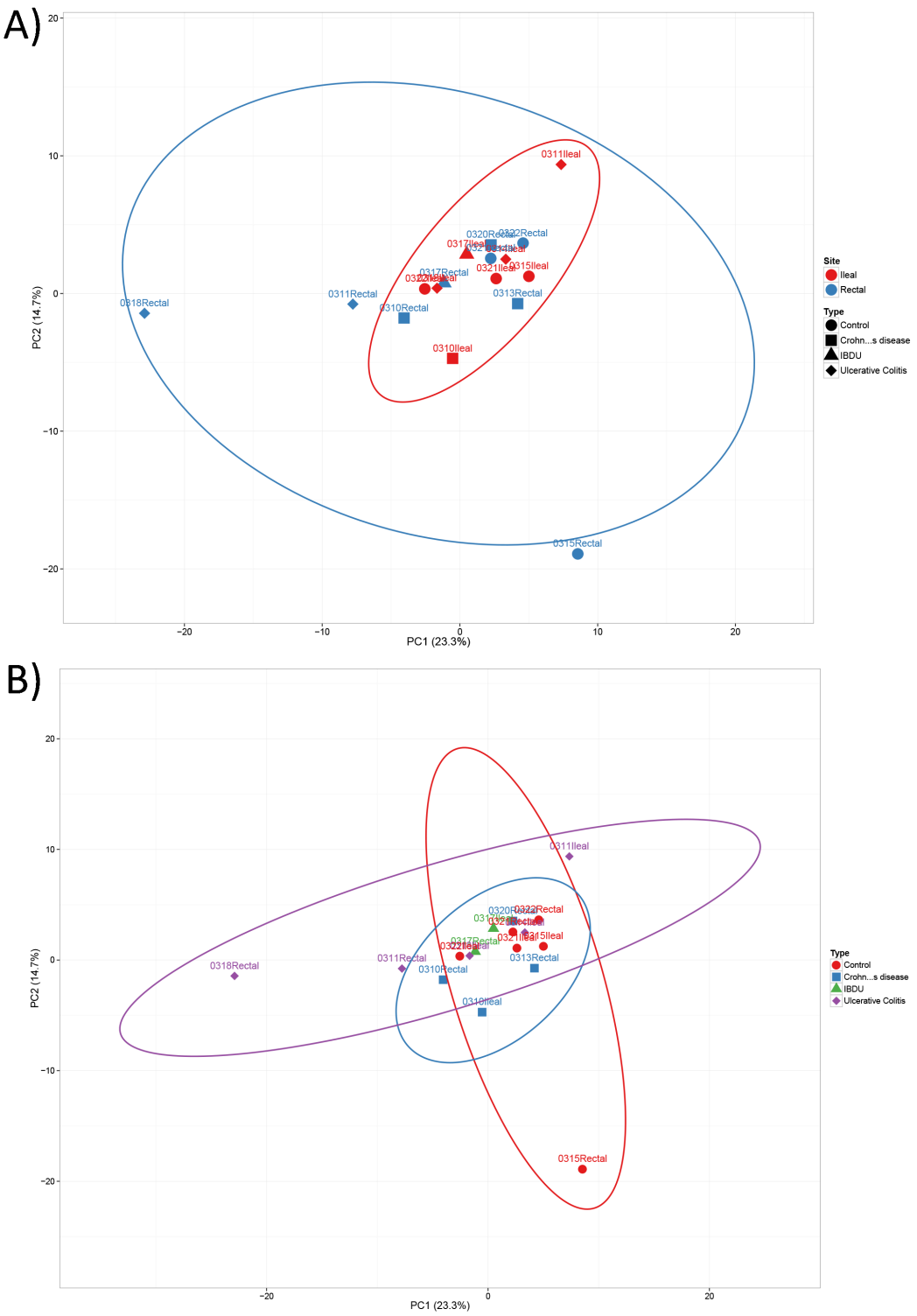

Figure 3.9: Clustering of RNA-seq samples by collection site and disease sub-type using species match totals for both panels A and B. Principal component 1 only explained 23.3% of variance and principal component 2 explained 14.7% giving a total of 38% of explained variance. **A)** Clustering by disease site, either rectal or ileal, failed to cluster the samples into two groups when using the species matches per sample. **B)** Similarly attempting to cluster samples by IBD sub-type fails to separate samples and there is visible large overlaps between the disease sub-types.

3.4.4 Comparison of 16S rRNA sequencing with unmapped WES reads mapped to bacteria

For six samples 16S rRNA data was compared with unmapped reads from blood collected exome samples. Unmapped reads extracted for each of the WES samples were under 3% of total reads as shown in Table 3.16. Sample SOPR0240 reported a high mapping percentage with 99.96% of reads mapping to the reference sequence, once unmapped reads were filtered a comparatively small 450 reads were left, shown in Table 3.16. SOPR0244 reported a reduced number of unmapped reads with approximately 78,000 compared to approximately 300,000 for the other four samples. 300,000 reads is approximately 0.6% of the total reads of the exome samples.

Sample	Mapped reads	Filtered unmapped reads	PC unmapped
SOPR0240	43,114,069	450	0.001
SOPR0241	43,462,787	291,675	0.671
SOPR0243	53,205,274	326,210	0.613
SOPR0244	51,325,796	78,447	0.153
SOPR0245	48,065,992	308,524	0.642
SOPR0246	49,317,353	315,338	0.639

Table 3.16: Unmapped read totals from exome sequencing for samples with 16S sequencing also.

Only the bacterial species were compared with between the WES and 16S data provided by Dr James Ashton. However between the samples only 1,963 reads mapped to bacterial phylum and displayed no relation to the genera to the bacteria listed in Table 3.17.

Sample	High Genera at time of Diagnosis
SOPR0240	*Prevotella(H)*
SOPR0241	*Prevotella(H), Faecalibacterium(H), Bacteroides, Ruminococcaceae, Lachnospiraceae*
SOPR0243	*Lachnospiraceae(H), Bacteroides, Blautia, Ruminococcaceae*
SOPR0244	*Bacteroides (H), Paraprevotella, Faecalibacterium*
SOPR0245	*Streptococcus (H), Sutterella(H)*
SOPR0246	*Bifidobacterium(H), Lachnospiraceae, Blautia, Ruminococcaceae*

Table 3.17: Summary results of 16S rRNA heatmap. For the 16S data analysed using Qiime[146] several genera were identified which were more prevalent in the data, genera which were particularly high are denoted by a (H). Qiime summary results were supplied by Dr James Ashton.

3.4.5 Investigating *cronobacter sakazakii* read matches

Cronobacter sakazakii was detected at relatively high levels in five samples. Due to the potential for this bacteria to cause necrotising enterocolitis and the phenotypic overlap with the inflammation of the gut caused by IBD the bacteria was investigated further. Firstly all the samples with any read matches to the *Cronobacter* genus were extracted and summarised as shown in Table 3.18.

Sample	*Cronobacter (G)*	*Cronobacter sakazakii (S)*	*C. Sak. SP291 or Plasmid (Sub-S)*	*Cronobacter turicensis (S)*
DW001	1	1	0	0
PR0018	1	1	1	0
PR0061	2	2	2	0
PR0076	1	1	1	0
PR0083	1	1	1	0
PR0097	1	1	1	0
PR0104	2710	2701	1847	9
PR0110	1	1	1	0
PR0204	1467	1460	1016	7
PR0260	1461	1455	1023	6
PT001	2	2	1	0
PT002	1	1	1	0
RNAR1	7	7	0	0
SOPR0255	5	0	0	0
0317Rectal rna	182	182	104	0
0318Rectal rna	350	345	186	0
Total	6193	6161	4185	22

Table 3.18: *Cronobacter* read match samples and totals. While 17 samples had any reads mapping to the genus only three exomes and two RNA-seq had above 10 reads mapping to the genus. Most matches when analysing at the species level were to the species *C. sakazakii SP291* or its plasmid at the sub-species level, though species level and below with small read totals of short-reads are likely to be inaccurate.

Three exome samples: PR0104, PR0204 and PR0260 and two RNA-seq samples 0317-rectal, 0318-rectal were the only samples with above 10 reads to the *Cronobacter* genus. The majority of reads for each of these samples all mapped to the *Cronobacter sakazakii* species, in particular the sub-species *C. sakazakii SP291* and its sequenced plasmid.

To assess the reads matching to *Cronobacter sakazakii* they were extracted and concatenated into a single fastq file and analysed using FastQC (v0.11.3), as reads

were filtered before analysing with BLAST so all reads had a quality greater than 20. Per base sequence content was highlighted by FastQC with elevated per base G-C content over reads shown in Figure 3.10.

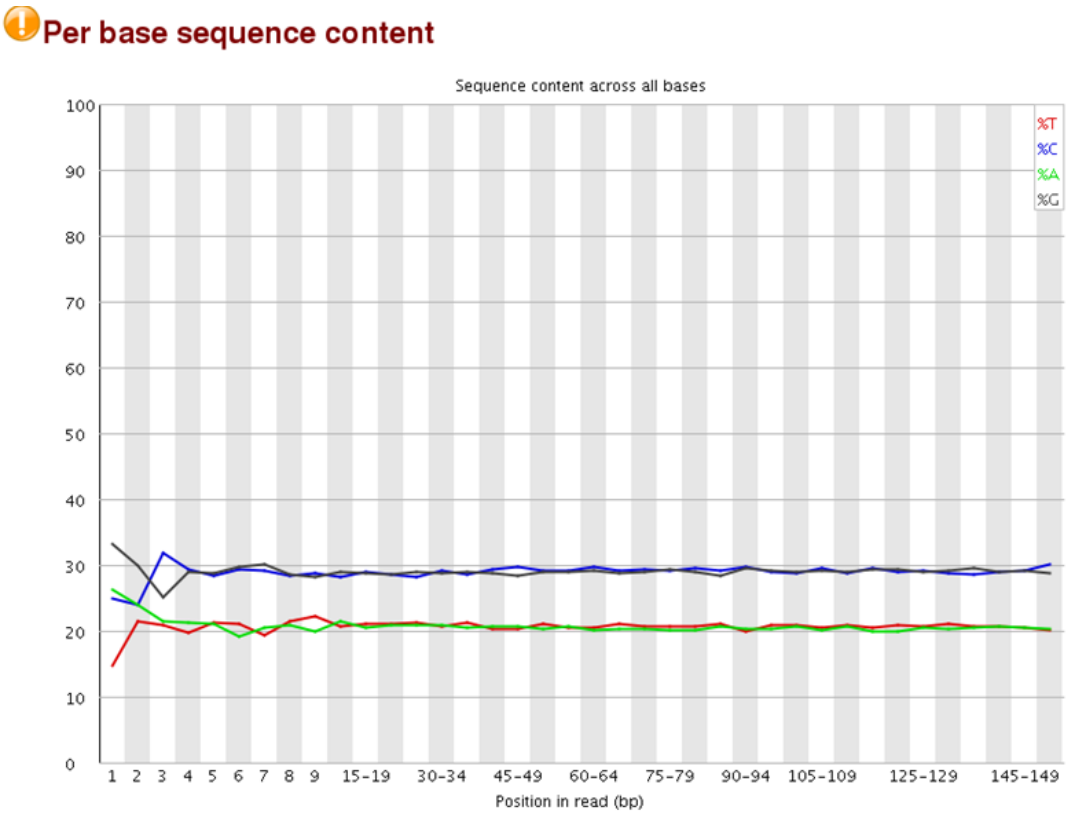

Figure 3.10: FastQC analysis of the *Cronobacter* genus reads shows an average G-C content of 58%, in line with the estimated genomic G-C content of *Cronobacter sakazakii* at 56 %[147].

Analysis of over-represented sequences from FastQC identifies no over-represented sequences when all sample reads are pooled for analysis. Over-represented sequences may have been indicative of repeats or sequences which failed to map initially with BWA to the human genome.

Analysis using the program Tandem Repeat Finder was performed to look within the reads for repeated motifs. If reads contained multiple repeat sequences they may have been problematic to align to the human genome. Read sequences from each of the samples were joined in a fasta file before submission to Tandem Repeat Finder, repeats at low copy number were detected as shown in Table 3.19.

Repeat (bp)	Copies	% Alignment	% Indels	Score	%A	%C	%G	%T	Entropy (0-2)	Repeat Bases (bp)	% repeat bases
14	2.1	93	0	51	36	20	20	23	1.95	29	0.003
15	2.5	95	0	67	34	31	26	7	1.85	38	0.004
15	1.9	100	0	58	13	20	34	31	1.92	29	0.003
16	1.9	93	6	53	0	46	40	13	1.43	30	0.003
18	2	88	0	54	16	47	30	5	1.7	36	0.004
24	1.9	86	0	65	15	21	28	34	1.94	46	0.005
24	2.5	97	0	109	6	32	42	18	1.77	60	0.006
150	2	99	0	591	19	29	35	16	1.94	300	0.032
150	2	100	0	602	21	23	19	35	1.96	300	0.032
150	2	100	0	600	29	29	22	19	1.98	300	0.032
150	2	100	0	600	20	30	30	19	1.97	300	0.032
150	2	100	0	600	35	24	20	20	1.96	300	0.032
150	2	100	0	602	23	26	27	22	1.99	300	0.032
150	2	100	0	600	23	25	19	32	1.98	300	0.032
150	2	100	0	600	22	32	30	14	1.94	300	0.032

Table 3.19: Tandem repeat finder results for *Cronobacter sakazakii* matches. Reads were concatenated and searched for over-represented kmers in the joint sequences. All of the possible repeats were calculated to be a very small percentage of the overall bases at a maximum of 0.032% for the largest repeat.

15 repeat motifs were identified in the reads of lengths 14 - 150bp, the default alignment score of 50 was used with the program. The repeats found are present at low frequencies with the maximum recorded copy number 2.5 showing the individual reads are not highly repetitive. The repeats comprise a small percentage of the total bases of the aggregated reads of up to 0.006%.

Using the mapping locations of reads to the *Cronobacter* subspecies bedtools coverage was used to calculate the range of depths of coverage across the genome and the percentage of the genome covered at each depth as shown in Table 3.20. A maximum depth of three was obtained showing reads do not all map to a few loci at high depths.

Name	Depth(x)	PR0104	PR0204	PR0260	SOPR0317	SOPR318	Genome (bp)	Avg. bases at dp	% Genome at dp
Cronobacter turicensis z3032	0	4,383,124	4,384,164	4,383,717	4,384,462	4,384,462	4,384,462	4,383,986	99.99
	1	1,338	298	745	0	0		476	0.01
Cronobacter sakazakii SP291	0	4,190,134	4,262,454	4,191,010	4,318,875	4,306,804	4,333,091	4,253,855	98.17
	1	137,693	68,605	136,831	13,771	25,625		76,505	1.77
	2	5,142	2,032	4,742	373	662		2,590	0.06
	3	122	0	508	72	0		140	3.24E-03
Cronobacter sakazakii SP291 plasmid	0	114,560	116,347	113,678	117,542	117,394	118,135	115,904	98.11
	1	3,575	1,788	4,457	593	741		2,231	1.89

Table 3.20: Number of bases in the *Cronobacter sakazakii* species with depths of coverage of 0, 1, 2, 3x. Depths across the genomes were low and not piled up at high depth suggesting that the matches are not caused by several loci of homology to humans.

All *Cronobacter* read matches were compared to the best human reference matches using BLAST. Matches were using the bit-scores and length of alignments as shown in Figure 3.11. From both the bit-scores in Panel A and lengths of alignments shown in Panel B. All of the *Cronobacter* matches were better quality as indicated by higher bit scores and longer match lengths for the same read.

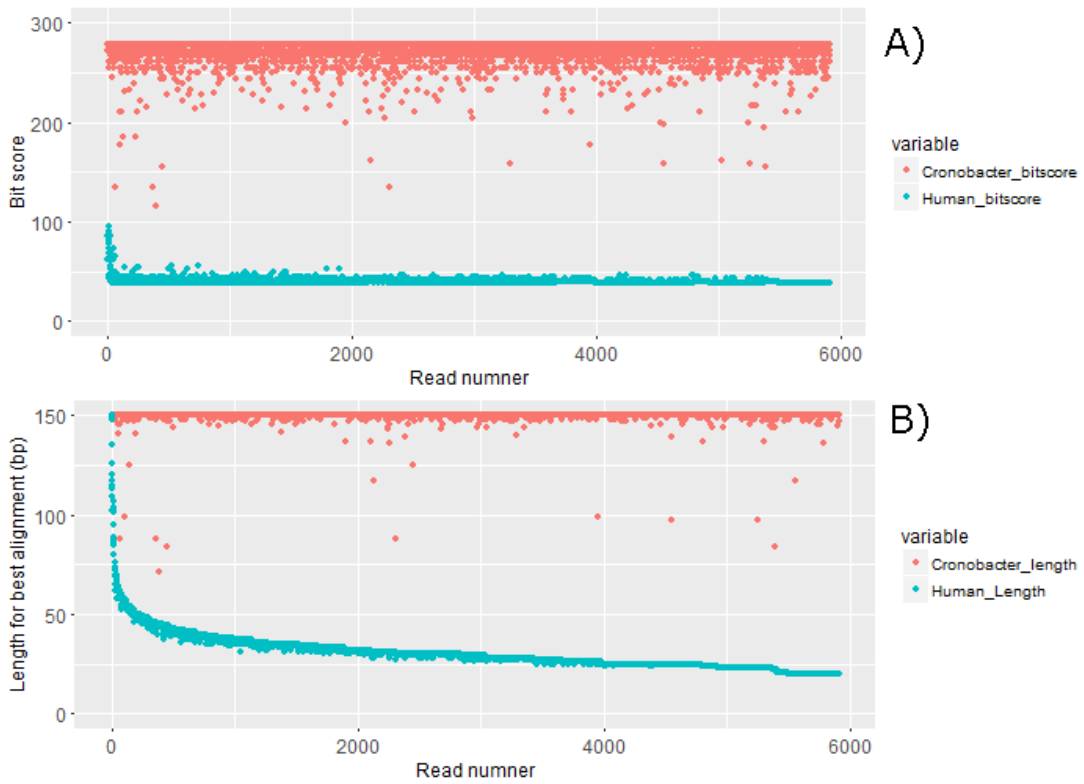

Figure 3.11: Comparison of quality of *Cronobacter* reads with the same reads mapped also to human. **A)** Comparison of bit scores for reads which were matched to *Cronobacter* and when they were mapped only to humans. Bit score was always higher for *Cronobacter* read matches. **B)** The length for alignments was equal or longer for the *Cronobacter* matches indicating a higher quality match. Where lengths were comparable the *Cronobacter* matches still had higher bit scores supporting the alignments as higher quality.

All of the reads mapping to the *Cronobacter* genus from the three WES and two RNA-seq samples were extracted and mapped to a single sub-species *Cronobacter sakazakii SP291* which had the most read matches. As shown in Table 3.18 not all the reads in the genus mapped to the highlighted sub-species as there are four *Cronobacter sakazakii* species: *ATCC BAA-894 / Plasmid , ES15, SP291/Plasmid, Malonaticus (CMCC 45402)*. To compare the species the average genomic similarity

was calculated using the program Mummer (v3.23)[148]. Mummer performs genome sequence alignments when using the nucmer command for DNA with the –maxmatch option to calculate maximum similarity as shown in Table 3.21 similarities are upwards of 91%.

Sub-species	**Percentage similarity to *CS-SP291***
Cronobacter sakazakii ATCC BAA-894	91.49
Cronobacter sakazakii ES15	92.25
Cronobacter sakazakii malonaticus	94.15

Table 3.21: *Cronobacter* sub-species identity shows that the other matched sub-species are upwards of 91% similar to the highest matched sub-species *CS SP291/Plasmid.*

All of the read matches all of the reads from the five samples were mapped to this sub-species and plotted using Artemis and DNA plotter as a circularised genome with reads from each of the samples overlaid showing the reads are evenly distributed as shown in Figure 3.12.

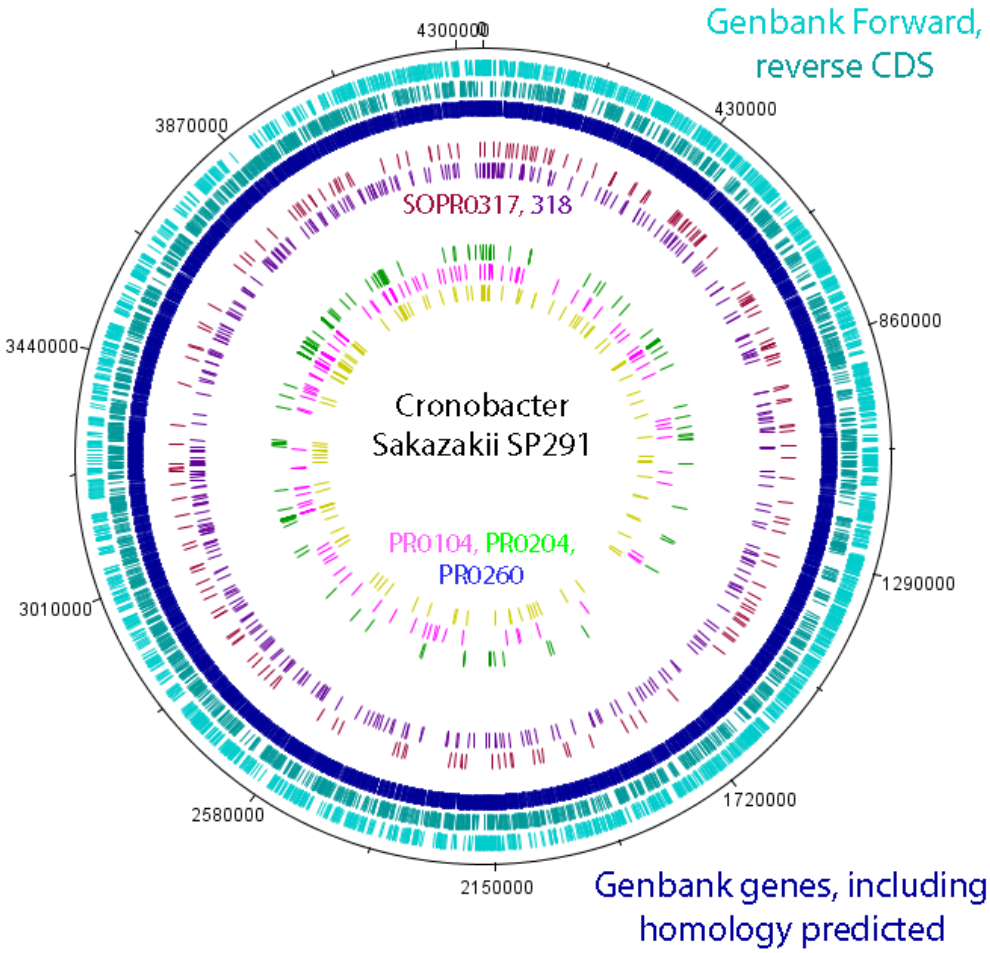

Figure 3.12: Visualisation of the *cronobacter* genus read matches against the single species of *CS-SP291*. Light blue tracks are the Forward and reverse CDS sequences, followed by genes defined by the Genbank file including many predicted genes on both the forward and reverse genome strands. In red and purple are RNA-seq read matches to the *CS-SP291* genome followed inside by the three exome samples in green, pink and yellow. Read matches as shown previously are widely spread across the genome.

Duplicates of the three exome samples with *Cronobacter* were sequenced in a new batch. None of these three new samples returned any *Cronobacter* genus reads.

3.4.6 Mapping unmapped reads back against the human genome

Mapping only to the human database removes the matches to other metazoan species, primarily primates to allow the determination of the best human mapping position. Using human mapped positions it was investigated to see if unmapped reads originate from specific loci across the genome and or overlap genes.

A total of 134 genes were overlapped by at least one re-mapped read. 36 genes had more than 100 re-mapped reads which overlapped as summarised in Table 3.22.

From the re-map most of the gene with the highest re-mapped reads was calcineurin binding protein 1 (*CABIN1*) with 69,446 reads in 305 samples, with the highest number of 1036 reads per sample, approximately 40 times more matches than the second most matched gene. Investigations were performed to examine the distribution of reads across the gene as shown in Figure 3.13.

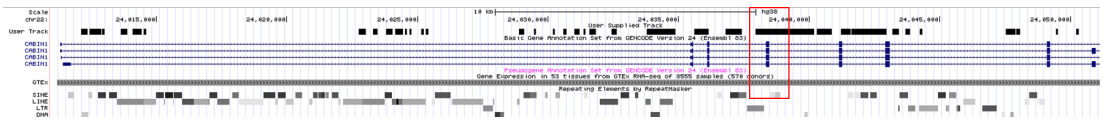

Figure 3.13: *CABIN1* Re-mapped reads visualised as the top black track over the gene, only exon four, highlighted by the red box had re-mapped reads which overlapped an exon but average coverage was not low enough to suggest a severe mapping or variant calling issue.

Only exon 4 of the gene was overlapped by the re-mapped reads. Coverage was calculated for all samples for the *CABIN1* gene per exon as shown in Appendix B, Table 8.6. Exon four was lower than the average coverage of exons at 97.9x but was not the lowest of the exons. Exon one was the worst covered with 1.7x depth likely due to the target capture kit not capturing the exon. The average exon depth of coverage was 191x across all 37 *CABIN1* exons.

Gene	Description	Chr	Start	Stop	Total reads	No. Samples
CABIN1	calcineurin binding protein 1	22	24011192	24178628	69446	305
MEGF11	multiple EGF like domains 11	15	65895079	66253747	1709	242
PTP4A3	protein tyrosine phosphatase type IVA, member 3	8	141391993	141432454	1112	216
FCGBP	Fc fragment of IgG binding protein	19	39863323	39934626	226	45
SPC25	SPC25, NDC80 kinetochore complex component	2	168834132	168913371	402	101
CTIF	cap binding complex dependent translation initiation factor	18	48539046	48863217	361	79
MUC6	mucin 6, oligomeric mucus/gel-forming	11	1012821	1036706	246	44
TJP1	tight junction protein 1	15	29699367	29968865	320	88
SLC39A11	solute carrier family 39 member 11	17	72645949	73092712	347	89
ATP8A2	ATPase phospholipid transporting 8A2	13	25372071	26025851	287	80
NCF4	neutrophil cytosolic factor 4	22	36860988	36878015	293	72
BAAT	bile acid-CoA:amino acid N-acyltransferase	9	101360417	101383519	354	71
KIF26B	kinesin family member 26B	1	245154985	245709431	247	59
ARHGEF26	Rho guanine nucleotide exchange factor 26	3	154121003	154257827	279	70
LA16c-325D7.2	Uncharacterised	16	2866348	2867618	242	61
RP11-426J5.2	Uncharacterised	18	48673575	48688419	232	45
RP11-290O12.2	Uncharacterised	4	154142122	154298819	229	61
CATSPERB	cation channel sperm associated auxiliary subunit beta	14	91580696	91780707	207	46
MCTP1	multiple C2 and transmembrane domain containing 1	5	94703741	95284575	222	51
RP4-591L5.1	Uncharacterised	1	30409560	30411638	188	47
PTPRK	protein tyrosine phosphatase, receptor type K	6	127968779	128520674	213	55
TRAPPC12	trafficking protein particle complex 12	2	3379675	3485094	180	46
SERPINA2	serpin family A member 2 (gene/pseudogene)	14	94364313	94366698	164	40
RP11-638L3.1	Uncharacterised	18	67516546	67899619	191	51
RP11-35J10.5	Uncharacterised	11	7704628	7882947	135	34
PHACTR1	phosphatase and actin regulator 1	6	12716805	13290484	184	37
LINC01355	long intergenic non-protein coding RNA 1355	1	23281309	23286752	150	30
LINC01194	long intergenic non-protein coding RNA 1194	5	12574857	12804363	142	40
RBFOX2	RNA binding protein, fox-1 homolog 2	22	35738736	36028425	134	25
AKT3	AKT serine/threonine kinase 3	1	243488233	243851079	129	26
MYO16	myosin XVI	13	108596152	109208007	109	22
SLC35F3	solute carrier family 35 member F3	1	233904933	234324516	110	24
TBCE	tubulin folding cofactor E	1	235367360	235448968	114	28
PTPRD	protein tyrosine phosphatase, receptor type D	9	8314246	10612723	106	23
MUC16	mucin 16, cell surface associated	19	8848844	8981342	121	23
CPEB2-AS1	CPEB2 antisense RNA 1 (head to head)	4	14909961	15002045	106	27

Table 3.22: Re-mapping reads to human identifies 134 genes overlapped by at least one of the unmapped reads. The top matched gene was *CABIN1* with 69,446 reads spread over 305. This was far more than the second most matched species with only 1,709 matches.

Mapping of RNA-seq unmapped reads as shown in Figure 3.7 had a higher number of reads mapping to human genome, this is likely more a technical artefact due to the data supplied being aligned to the hg19 reference and the BLAST matches from samples processed with the hg38 reference. Matches were found across all chromosomes as shown in Figure 3.14. In particular though chromosomes 1, 2, 6, 14 and 22 had higher number of read matches though matches were spread across the chromosomes and not to single loci on the chromosomes.

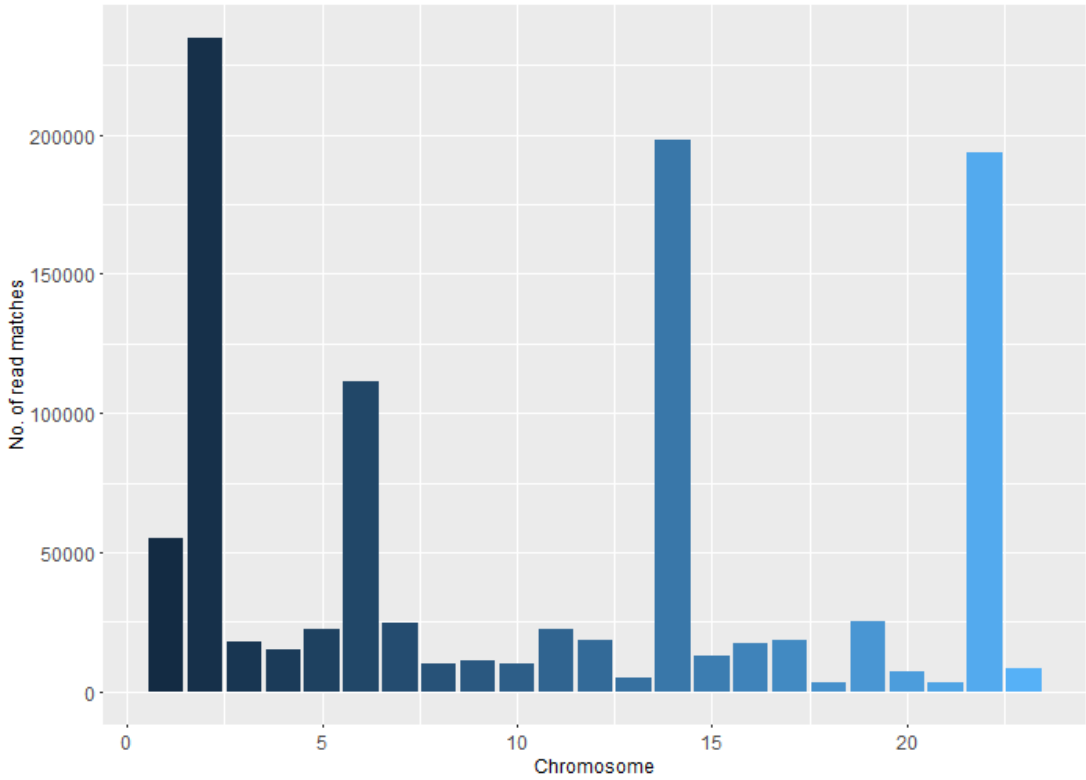

Figure 3.14: RNA-seq re-mapped read locations were distributed across the chromosomes though in particular chromosomes 1,2,6,14 and 22 had higher number of read matches.

3.4.7 Unclassifiable reads

In total 3,360,182 reads were un-mappable to any taxa over all of the analysed unmapped reads. FastQC analysis of the concatenated reads to look for unusual patterns. Base quality was already known to be above 20x for the majority each read as found also from FastQC. Per base sequence content was also normal at between 20 - 30% for all four nucleotides across the read lengths, no over-represented or adapter sequences were also identified. However FastQC identified high sequence duplication levels as shown in Figure 3.15 this occurs when non-unique sequences make up more than 50% of the total read sequences. In the plot the blue line represents the full sequence set whilst the red line shows de-duplicated sequences. The proportions shown are the proportions of the de-duplicated set which come from different duplication levels in the original data.

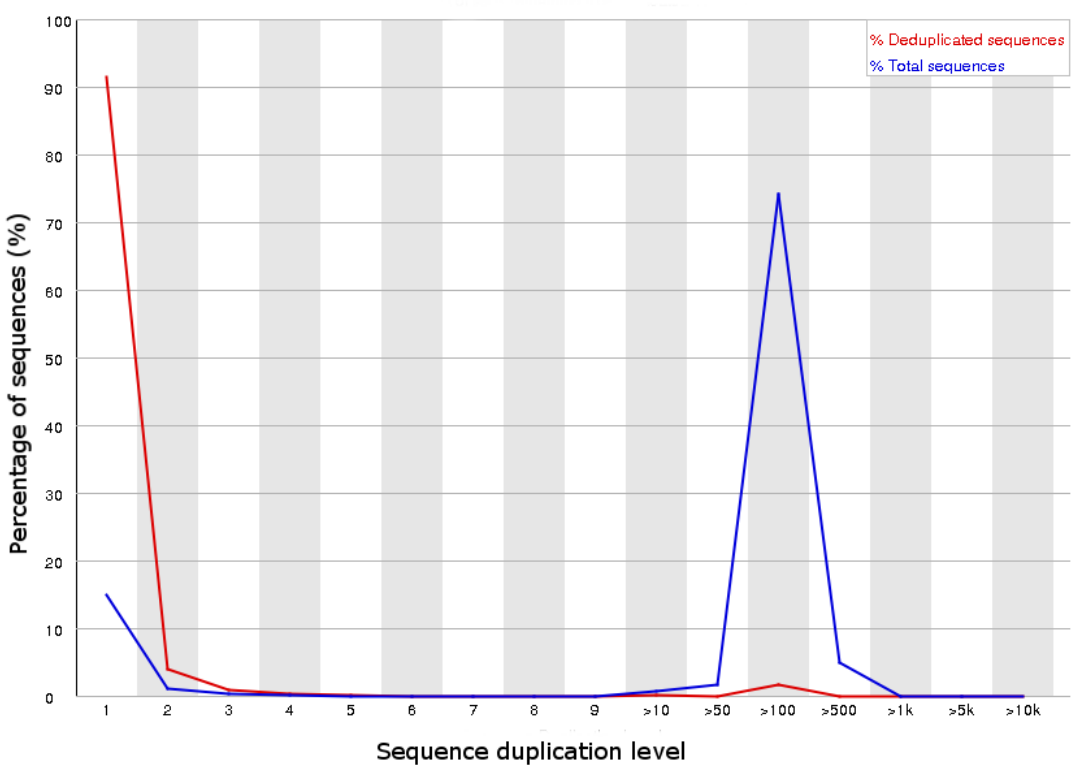

Figure 3.15: Sequence duplication levels of all unclassified reads. This module was flagged as a fail by FastQC as non-unique sequences make up more than 50% of the total reads. 16.48% of sequences were remaining after de-duplication.

3.5 Discussion

Cross-species contamination of NGS samples have been reported in a variety of studies as technical artefacts from sequencing steps such as *phiX174* contamination[87] to possible environmental contamination at sequencing centres[85] to NGS samples actually containing microbial sequencing in unmapped reads[89]. In this study we sought to test if unmapped reads could be extracted from 358 WES and 17 RNA-seq samples and assess if unmapped reads could either provide microbiome information and identify differences between IBD disease samples or how collection methods and the site of collection affected the profile of unmapped reads.

Extraction of unmapped reads typically only obtained a small percentage of the total reads in BAM files as shown from Figure 3.2 and Table 8.4 ranging from close to 1.58 e-06% for a sample with only one unmapped read up to a sample with 0.56% (303,033 reads) unmapped. These totals are lower than previous studies which found the unmapped reads were of the order of millions[85]. Differences are likely due to this study excluding phiX174 reads from totals, filtering to select only pair unmapped reads and requiring 90% each of reads to be above a PHRED quality of 20. By far the largest matched species from samples was phiX174 which totalled 2,555,986 million across the all exome samples. Usually the traces of the phiX174 virus should be removed from samples and so they were excluded from sample results shown in this chapter and all values have been reduced to reflect this. In both this project and previous studies using unmapped reads the mapping reads still made the majority of the files, in this study mapping reads were upwards of 99.4%.

The low number of unmapped reads passing the filtering is also due to the blood collection method known to have been used for 233 of the WES samples. Blood samples typically had fewer unmapped reads than saliva as shown by Panel A of Figure 3.5. Immunologically this makes sense as colonising flora is present in the oral cavity and upper respiratory tract and blood would be expected to be sterile under normal conditions and free of bacterial pathogens. With WES sequencing it was thought that the hybridisation step should have removed the un-bound and therefore non-target DNA but previous studies have also found that DNA from other species is

still detectable from exome samples[89,149,150].

Due to the majority of samples investigated being blood captures and most of the unmapped reads being excluded as artificial contamination of *phiX174* filtered totals were low. The low read totals limit the interpretation of results as there is insufficient power to make robust conclusions and evaluations of observed trends. In recent years Illumina has also introduced double bar-coding to allow identification of tag-hopping between plexed samples and exclude these reads also reducing the likelihood of cross-species contamination during sequencing. Therefore due to the availability of primarily blood samples which are usually sterile and with few unmapped reads other than technical artefacts and relatively few saliva samples from which more bacterial species were identified. Combined with improvements in sequencing methods all of which meant that unmapped reads were statistically too few to draw meaningful and reliable conclusion about the detection of cross species contamination.

Also as a consequence of using short read technologies and the low numbers of unmapped reads being returned per sample after filtering it is unlikely that mapping results down to the species level is likely to be reliable and results should be taken at a broader classification level such as to the genus level. Species within the same genus are similar, hence from short reads the mapping to a single species is not likely to be reliable without greater number of reads which were either longer and or assembled into large contigs to differentiate down to the species level.

Comparisons of the unmapped read totals extracted from the 245 paediatric IBD cases and 113 non-IBD patient control samples revealed relatively few differences. Figure 3.2 showed IBD cases and control sample ranges substantially overlapped, with control samples averaging 5,967 reads compared to 4,151 for cases. Only four databases saw high percentage matches between cases and controls with most differences caused due to plant sequence contamination seen only in the IBD samples in a single batch. As IBD samples in this batch have high plant contamination the percentage metazoan matches are lower for a number of samples, combined with the larger batch size of IBD samples, with more samples also with lower metazoan matches leading to the boxplot covering a larger range. PCA analysis as shown in

Figure 3.4 also returned overlapping clusters when analysing the total read matches and percentages to each of the databases. Therefore it does not appear that the unmapped reads from exomes could distinguish between the unmapped read of IBD case or non-IBD samples.

As highlighted when reviewing the top matched bacterial species they were predominantly those found in the upper respiratory tract or oral cavity. Particularly for species only found in in the oral cavity the majority of the read matches were from saliva samples. As shown in Panel A of Figure 3.5 there was a clear increase in the total unmapped reads for the saliva samples compared to blood samples.

As there are identifiable features that can distinguish the saliva samples from blood samples clustering analysis was performed using the total matches to each of the databases and then the percentages. Species level clustering was not performed as the range of bacteria present in samples meant it would likely result in multiple smaller clusters which would increase noise and increase the difficulty of interpretation. Using the total number of species, Panel A of Figure 3.4 produced a tight central cluster for 239 of the 245 blood collected samples which also overlap a large number of the unknown samples as the majority of those would also be expected to have been blood collected. Outside of the clustered blood samples are a number of blood samples which clustered less tightly along with several saliva samples and the majority of unknown collection samples. Due to the total number of unmapped reads being variable between samples of the same collection methods the samples cluster very heterogeneously and a clear separation or boundary is less obvious.

Analysis of unmapped reads from RNA-seq identified many more reads than from the equivalent exomes as shown by Table 3.1. This is due in part to the samples being from gut biopsies compared to blood but also partly due to the files provided being aligned against the human hg19 reference genome. As indicated when the RNA-seq unmapped reads were re-mapped back to the human genome the reads were distributed across the chromosomes though chromosomes 1, 2, 5, 14, 22 had higher numbers likely due to the reduced gaps in the reference and better capture of diverse sequence regions. Hence the human percentage and total matches are higher as

technical differences between the genome reference versions are highlighted. Relatively few bacterial matches were returned likely due to the poly-A targeting. Few bacterial RNAs have poly-A tails and so most bacteria would not have been selected for sequencing[140]. WES samples all had few unmapped reads bar SOPR0318 which had 81.74% of reads mapped to plant species as it was sequenced in a different batch to the other exomes. As none of the exomes had any bacterial species detected above 10 reads over the 7 exomes, likely due to all samples being blood collection, there was no comparison possible for the IBD patients in terms of bacteria from the rectal or ileal site and the blood based WES. The only species of note detected from the WES samples were bonobo, gorilla and chimpanzee[151–153].

Other than matches to other primates or back to the human genome the unmapped RNA-seq data predominantly matches to bacterial species. The only exception was in samples SOPR0321 rectal which had a high match count to *Gossypium hirsutum* (Upland cotton) and is suspected to be environmental contamination. The top matched bacteria was *Mycoplasma pneumoniae* with 9,751 read matches. The reads were fairly evenly distributed across the samples from 328 - 916. As this bacteria was not seen in the other exome samples also sequenced in the same batches as these RNA-seq samples it suggests the bacteria in these patients is more unique to the ileal and rectal sites in these patients. *Cronobacter sakazakii* was the third most matched bacteria, but was only found in two samples and was previously only found in three exome samples. Amongst the other bacterial species of interest were *Bacteroides vulgatus* was detected with 221 reads and is commonly detected in the gut as a commensal species. Using the unmapped read matches at a species level it was hoped that differences between the sites and IBD disease sub-types would be distinguishable. However as shown from Figure 3.9 by both collection site and disease subtypes there are insufficient differences to differentiate the categories.

Species matches were analysed across all 358 exome samples to then identify the most common potential cross-contaminants but also to investigate odd samples such as those displaying high plant or bacteria percentage matches. Bacterial read matches were to 132 species of bacteria with greater than 100 reads, 56 of which had above 500 reads which were investigated further by splitting the species by phylum as shown

in Tables 3.4 - 3.7. The common theme grouping the top matched bacterial species was the location in the human body where they are found, predominantly in the oral cavity or upper respiratory tract. High matching or species of particular interest are discussed here.

The phylum Bacteroides contains Gram-negative, obligate anaerobic bacteria commonly found to comprise a large proportion of the normal gut flora. Nine species were identified with above 500 reads, five of the species are from the genus *Prevotella*. Most species of *Prevotella* found were associated with the oral cavity and diseases such as gingivitis[154]. The Firmicutes phylum contains both aerobic and anaerobic Gram-positive bacteria. Due to the bacteria under this phylum tolerating variable levels of oxygen species are often found in a variety of environments. 18 species were identified including 16 species in the genus *Streptococcus*. Using the highest matched species *Streptococcus parasanguinis* which had 26,962 reads, 21,138 of (78.4%) which originated from 16 saliva or suspected saliva samples. *Streptococcus* species are usually colonisers of the upper respiratory tract and are amongst the most common causes of "Strep" or sore throat as an opportunistic pathogen[155–160].

The phylum of Proteobacteria contained many of the most pathogenic species and are found on a variety of mucosal surfaces. 22 species were detected above 500 reads. Species in the genera of: *Aggregatibacter, Escherichia, Haemophilus, Neisseria, Klebsiella* were all noted to contain species present in healthy individuals[161–164]. While other genera: *Burkholderia, Pseudomonas, Variovorax* report species that are suggested to be present environmentally and only opportunistically pathogenic[165–169]. Focusing on the most pathogenic species identified five species of interest. Firstly *Campylobacter concisus* is a an oral species associated with dental disease but also has limited evidence suggesting that the bacteria is higher IBD cases[170]. 82% of the reads for the species originated from saliva samples, comparing the reads matched in control samples to IBD samples revealed 2,642 reads came from 13 control samples and 6,768 from 23 cases suggesting a similar average though the highest control sample had 1,000 reads and 2 IBD samples had 2,572 and 2,406 read matches. This suggested that *Campylobacter concisus* may be found at a higher level in some of the IBD samples but more saliva samples would be needed to confirm the trend and the

importance of the elevated levels.

The bacteria *Ralstonia pickettii* was also implicated in Crohn's disease patients[171] but in the unmapped reads here 1121 out of 1202 reads were detected from control samples and so argued against elevated levels in CD but due to the low match total further samples would be needed to validate. *Salmonella enterica* has approximately 2500 serovars and typically are orally acquired from contaminated food, amongst the diseases caused by *S. enterica* include: enteric fever (typhoid), enterocolitis/diarrhea, bacteremia and chronic asymptomatic carriage[172]. *Shigella boydii* is associated with acute diarrhoea usually spread by contaminated food or water[173]. Finally *Cronobacter sakazakii* was identified which among infants in associated with necrotising enterocolitis suggested to be spread via food sources such as contaminated infant formula[147,174].

Plant species were an unexpected find and were found at above 10% of sample read matches for 54 samples. Interestingly 57 samples had above 10 reads matching to the any of the top three matched plants database. 55 of the 57 samples were performed at the same sequencing centre spread across three batches. There are several possible theories that could explain how plant DNA could have entered samples but most can be eliminated. Firstly that the plant species were consumed and DNA was present in the oral cavity which survived degradation and was captured by saliva collection. This theory can be discounted as only 5 of the 57 samples were saliva captured also the top matched species were those which were not likely to be consumed: *Gossypium hirsutum* (Cotton) - 28,951 reads, *Arabidopsis thaliana* (Model plant organism) - 23,058 reads and *Citrus sinensis* (Sour orange) - 14,291. Effectively this theory can be discounted due to the combination of factors required: low proportion of saliva captured samples amongst the 57, plant DNA avoiding degradation, the uncommon species detected not being commonly consumed as food and the free plant DNA being unlikely to be captured by the hybridisation step. A counter-theory has been suggested that plant genes may be able to pass from the digestive tract across into the bloodstream by an unknown mechanism[113]. However this theory and results have not been validated and no details of the method of transfer have been proposed to generate cell-free plant DNA and how it could avoid degradation. The most plausible

theory for the origins of the plant DNA are that they are environmental contaminants present at sequencing centres. This theory has the most evidence as sequence centre specific contamination has previously been shown by the study of Tae *et al.*[85]. As 55 of the 57 samples all were sequenced at the same sequencing centre which supports this theory. It is possible that the source of contamination has come from other samples being sequenced at the same time as the top matched species are those which are widely sequenced and studied[175–178].

The metazoa database excluded *homo sapiens* but included many primates and mammalian species such as *mus musculus* (Mouse)[151,152,179]. When considering only the best match per read most samples have a higher percentage match to metazoa, particularly to primate species with high homology to human such as *Pan troglodytes* (Chimpanzee)- 80,958 reads, *Gorilla gorilla* - 21,574 reads and *Pan paniscus* (Bonobo) - 10,128 reads. Ensembl provides pairwise and multiple whole genome alignments from which large-scale synteny, conservation scores and constrained elements are obtained[153]. One of the methods by which pairwise whole genome alignments are performed was the Enredo-Pecan-Ortheus (EPO) pipeline. From the comparisons performed with the human genome against other primates coverage was calculated: Chimpanzee 91%, Gorilla 90%, Bonobo 90%, Orangutan 86%, Gibbon 83% of the human genome covered. Due to the high levels of overlap sequences between the primates it is highly likely that short read sequences would be able to map to multiple species. Usually reads are only mapped to *homo sapiens* using BWA as it avoids the problems caused by highly homologous species and explains why when considering only the best matches that the human only database has a low percentage of matches.

While archeal reads may not have been expected to have been detected, both viral and fungal reads might have been expected to be detected. Viruses can be found in blood and have been documented to integrate into host genomes. One reason why they may not have been detected is due the requirement for customised reference genomes to detect the viruses due to the high mutation rates, combined with short reads making detection difficult[90]. Only two viral species, *Enterobacteria phage HK629* - 340 reads & *Enterobacteria phage lambda* -195 reads, were detected at above

100 reads and had comparatively low read matches per sample and so were considered further. Fungal DNA would not likely be present from blood samples but only from saliva samples.

The bacteria *Cronobacter sakazakii* was identified in a total of five samples (three exomes and two RNA) at relatively high levels of several hundred reads. Due to its association with causing necrotising enterocolitis in infants it was identified as a potential bacteria of interest[174,180]. Elevated G-C content was detected at 58% across all of the reads from the five samples which mapped to the genus *Cronobacter*. In approximate agreement with the genome wide figure of 56 % G-C obtained by Kim et al.[147]. Analyses of the reads for repeats and the quality of matches failed to identify repeated reads or repetitive sequences. Using match locations to bacteria also found a maximum read depth of 3x. If the reads all mapped to a limited number of positions at high depth then it would have suggested a region of homology to another species. The sub-species *C. sakazakii SP291* was identified as the sub-species with the most matches from the initial mapping and analysis and was upwards of 91% similar to other sub-species. Hence all the 6,193 reads were able to then be mapped to the single sub-species and plotted visually which confirmed the reads are distributed across the genome as shown in Figure 3.12. An estimated 88.27% of the genome is coding for *C. sakazakii* and so a spread coverage across the genome would be expected. Therefore all the evidence supports *Cronobacter* species actually being present in the samples. There were higher numbers of read matches in the RNA-seq samples despite the original exome BAMs containing approximately 5 times as many reads. The reduction in unmapped reads could be explained by the additional hybridisation step in WES. Which likely leads to the loss of an proportion of any *C. sakazakii* DNA. All the samples with *C. sakazakii* reads originated from the same batch, suggesting that this could be a contamination event. *C. sakazakii* is highly survivable as it has been shown to survive in dried milk formula powder for long periods of time. From the recent genomic and laboratory analysis *C. sakazakii* was shown to produce biofilms which aid its survival[147]. Given the documented ability to survive in extreme conditions it would be feasible for some of the DNA to survive into the sequencing stages regardless of the origin. The subsequent re-sequencing of the three exome samples which did not detect the *Cronobacter sakazakii* bacteria makes

the cross-species contamination more likely environmentally inherited at the sequencing centre. This finding limits the use of the bacterial read match but highlights the variability of contaminants present in sequencing centres.

The final investigation performed looked at mapping the unmapped reads from the exome samples back to the human genome to try and identify is there were any genes which had identifiable poor mapping. The top matched gene was calcineurin binding protein 1 (*CABIN1*) with 69,446 reads in 305 samples with the re-mapped reads scattered around the first few exons but only overlapping exon four. Despite the average coverage depth being lower over exon four other exons in the gene were at lower levels and did not have any re-mapped reads arguing against the importance of these re-mapped reads.

Unclassified reads were grouped and analysed by FastQC to try and identify if reads had features which could be used to identify their origins. No over-represented or adapter sequences were detected suggesting the reads were unique and not problematic read sequences such as repeats. G-C content was normal across reads also at between 20 - 30%. However the reads were reported to be highly duplicated by the sequence duplication module of FastQC. This module however does not consider the entire read length when looking for duplicates and only compares 50 bases likely increasing the number of reported duplications but indicating the duplication levels are an issue[32]. If the reads which are unclassified are from a similar species which are not included in the selected database and are present in many species this could be be why the unclassified reads are high across the dataset. Alternatively they could be sequences which are difficult to map to the human genome due to variants such as insertions or repeats. Using a *de-novo* assembly approach with the unmapped reads it may be possible to assemble all or part of the genome of a potential contaminating species. If contigs were able to be created they could alternatively be able to be re-mapped to the human genome and improve the initial mapping by using the additional contigs.

3.6 Conclusions

It has been shown that unmapped reads can be extracted from samples and the species present detected. As would be expected from blood samples there are few unmapped reads from other species due to the sterile conditions in normal blood samples. Most of the unmapped reads identified from samples are technical artefacts of *phiX174* contamination which combined with blood samples meant most analyses in this chapter were underpowered to generate robust and statistically reliable results. Comparisons of IBD case and control samples could not be distinguished from unmapped reads alone. While there is evidence for differences in the gut microbiomes of IBD and non-IBD patients[117] the collection and analysis of blood based WES did not appear to detect these differences. RNA-seq samples taken from gut biopsies failed to also identify the changes in the gut microbiome between disease sub-types of IBD. Most differences in unmapped reads were between the collection methods of blood and saliva samples. Saliva samples have more unmapped reads, in particular more bacterial taxa associated with the oral cavity or respiratory tract. Investigations by batch also helped to identify potential environmental contamination as nearly all the samples of a single batch had a significant proportion of plant matches to commonly sequenced species such as the model plant *Arabidopsis thaliana*. Several bacteria such as *Cronobacter sakazakii* were identified in several samples at unusually high levels. Further examination of the matches suggest they are likely to be from *Cronobacter* bacteria but the origins of the bacteria remain unclear and appear to be environmental contamination. Finally re-mapping reads back to the human genome failed to identify loci with coverage issues. There remain a large number of unmapped reads per sample and are believed to be either problematic reads which cannot be well mapped to the human genome or are from a contaminating species not included in the databases used in this project.

Chapter 4

Comparing variant calling using a whole exome and genome sequencing trio

4.1 Introduction

Second generation sequencing became commercially available around 2007 which reduced the time and cost to sequence a genome due to the massively parallel nature of the technology. Despite the fall in the time required the cost per genome was still prohibitively high with costs upwards of $1 million per genome in the year of launch as shown in Figure 1.3. Though the cost per genome rapidly decreased to under $10,000 by 2011 it was still prohibitively high for most lab groups. In the period 2011-2015 sequencing costs per genome remained steady at around $7,000 - 9,000[73].

Cost is not the only factor when considering to use WGS as it also introduces additional storage and processing challenges. Firstly, WGS requires additional data storage compared to targeted approaches. WES only targets the 2-3% of the genome which codes for exons of genes. Hence computational storage of WGS performed at the same depth of coverage as WES would require between 33 and 50 times more storage space per sample. Considering an average WES sample may produce a BAM file around 4 GB an equivalent depth of coverage genome could easily be to 132 - 200GB per sample. Secondly the time to process WGS samples increases due to the

increase in data from sequencing of the non-exonic regions. To reduce this bottleneck jobs require software to make use of parallelisation or input data to be split and run in multiple jobs to complete in a reasonable time.

Targeted approaches such as WES are intended to only capture gene exons. By primarily sequencing exons which code for proteins WES will capture most of the regions and variants known to be implicated in diseases[181]. This reduces the amount of the genome sequenced to 2-3%, making the target size around 51MB[182], proportionally reducing the: cost, storage and computational burden. However with the introduction of a higher throughput sequencers such as Illuminas Hi-Ten X sequencing system around 2015 the cost of WGS has fallen to close to $1000 per genome compared with around $151 per exome[73,182]. As the costs of WES and WGS converge WGS may become the more preferred method due to the additional information provided. Recently the introduction of the newest Illumina sequencer the Novaseq has further enabled a reduction in sequencing costs with WES samples effectively now the same cost as many targeted panels sequenced on the X10.

4.1.1 Non-coding variant annotation

WGS samples compared to targeted methods such as WES give a large increase in the number of intergenic and intronic variants called. Annotation tools prior to 2011 and the release of ENCODE[22] data focused on coding variants as with tools such as SIFT[60] which predict the effect of amino acid changes. In non-coding regions of the genome these tools which are solely amino acid based will not return any prediction. Often only AAFs from genome wide studies such as the 1000 genomes project or sequence conservation with other organisms would have been the only available annotations for non-coding variants. At the same time data from the ESP 6500 and ExAC with 65,000[63] WES samples were available for exome variants providing additional sources for AAFs. Due to the limited evidence and methods to classify non-coding variants combined with the additional sequencing cost of WGS methods such as WES were often preferred.

ENCODE data has also enabled the development of tools to combine multiple data sources and predict the consequences of variants. Due to the complexity of:

integrating multiple sources of evidence, weighting evidences and creating or refining algorithms to analyse the data prediction tools have improved with each generation. Initial tools able to classify intronic variants were released between 2012 and 2015 such as: RegulomeDB[183], GWAVA[184], CADD (v1.0)[66] and FATHMM-MKL[69]. Each of these tools used differing combinations of evidence and often displayed a lack of agreement between tools for variants[62,66,69].

RegulomeDB focused on using data to identify potential transcription factor binding sites and score variants based on the number of ENCODE features detected at that position but did not predict which would be more damaging or important to function. GWAVA took a more sophisticated approach by using ENCODE data combined with a classifier algorithm based on a random forest model trained on three variant datasets: randomly distributed SNPs across the genome, within 2kb of an annotated transcription start site or all variants in the 1000 genomes dataset within 1kb of an HGMD variant. GWAVA found the randomly-trained model gave the highest AUC value of 0.97[184].

CADD (Combined Annotation Dependent Depletion) aggregated data from multiple ENCODE experiments with annotations from VEP and UCSC for measures including sequence conservation (GERP, phastCons, phyloP), deleterious measures for amino acid changes (Grantham, SIFT, PolyPhen) and transcript information such as intron-exon boundaries[66]. In total 63 annotation sources were used, all of the available annotations were supplied into a support vector machine algorithm which was trained on 10 sets of data containing 13,141,299 SNPs and 1,554,039 INDELs. The final model generated scores for all 8.6 billion possible SNPs for the GRCh37 reference and accounted for the number of annotations available for each varaint prediction. CADD scores (c-scores) were converted to a PHRED like scale with values above 20 indicating variants were in the top 1% of most pathogenic. CADD has had five version updates since initial release, the latest in February 2019 (v1.5), but has been suggested by tools: EIGEN[185], DANN[67] and GAVIN[68] that CADD over-fits using its training dataset and could be improved by using alternate or revised algorithms. Revisions have been made in CADD from versions 1.4 (July 2018) including updating to hg38 but the release was too late for inclusion in this study.

FATHMM was originally a program like SIFT which focused on the prediction of functional consequences of amino acid changes using hidden Markov models[62]. The program was revised to use multiple evidence kernels in 2015 (FATHMM-MKL) with a semi-supervised machine learning algorithm, a mix of labelled and unlabelled data[69]. Each of the kernels converts the input data to a matrix which is combined into a single kernel matrix with individual kernel weights for each annotation type. A support vector machine algorithm then classifies the input query based on a training data set using the kernel weights. FATHMM-MKL uses 10 data sources: 46 & 100 Way Sequence Conservation, GC content measure, ChIP-Seq, TFBS PeakSeq, DNase-Seq, FAIRE, Genome Segmentation, DNAse-I footprinting. FATHMM-MKL was benchmarked against the other tools CADD and GWAVA using a dataset of 87,518 coding examples consisting of 67,305 positive (HGMD) and 20,213 negative examples. FATHMM-MKL and CADD were found to both achieve an AUC of 0.93 for coding variants. For non-coding variants FATHMM-MKL achieved an AUC of 0.91 compared with 0.87 for CADD and 0.69 for GWAVA. These results show that the performance of CADD and FATHMM-MKL is similar for coding variants but that FATHMM-MKL has a small performance edge in the prediction of non-coding variants with functional consequences. FATHMM-MKL scores variants on a scale of 0-1 with above 0.5 being called as pathogenic, though the confidence of pathogenicity is described as being higher the closer the score is to either extremity of 0 or 1[69]. As with CADD the program adjusts the scoring algorithm for the available annotations per variant predicted.

An updated version of FATHMM-MKL was released in 2017 called FATHMM-XF (eXtended Features), which made use of 27 datasets for eight feature groups (conservation, histone modifications, open chromatin, transcription factor binding sites, gene expression, methylation, digital genomic footprinting, interaction networks) for both coding and non-coding variants[70]. A comparison was also performed using FATHMM-XF against CADD(v1.3)[66], GAVIN[68] and DANN[67] using 31,099 non-coding variants and 62,884 coding ClinVar variants. Accuracies of prediction for each tool were calculated for coding and non-coding variants: FATHMM-XF (0.89, 0.88), GAVIN (0.89, 0.87), CADD (0.63, 0.64) and DANN (0.60,

0.61). FATHMM-XF gave similar results to GAVIN with FATHMM-XF achieving 0.01 higher accuracy than GAVIN. CADD annotations performed worse than previous literature had suggested supporting the claims that the tool has issues in either the SVM algorithm or in over-fitting the training set[67,68,186].

GAVIN is based upon CADD scores (v1.3) and considers that the use of a single threshold value for variant pathogenicity as inappropriate for many genes and that instead calibrations are needed to correct pathogenicity scores per gene[68]. To calculate re-calibrated thresholds GAVIN uses ExAC AAFs and snpEFF[59] impact scores per gene for variants listed in ClinVar (Nov2015)[65] described as pathogenic or likely pathogenic. For all other variants not listed in ClinVar they were used as a benign set to allow mean pathogenicity and benign thresholds to be calculated per gene. Currently the public version of GAVIN (r0.3) is only available with re-calibrations for includes 3,498 genes and primarily focuses on coding variants. In future versions of GAVIN the introduction of gnomAD genome wide allele frequencies should enable full annotations of non-coding regions including deep intronic variants. With improvements in algorithms and availability of data for non-coding regions, it is increasingly possible to identify functional variants in non-exonic regions which could be pathogenic.

4.1.2 Limitations of WES compared to WGS

All variants called in WES should also be present in WGS, resulting in WGS at least matching the diagnostic rates of WES theoretically. Studies comparing WES and WGS have supported higher diagnostic yields for WGS[181,187,188]. Comparisons between the sequencing technologies also identify the limitations of WES which lead to the lower diagnostic rates. Firstly coverage across genes is not uniform in WES due to overlapping targets, for instance when two exons are in close proximity and ends overlap this leads to potentially twice the coverage in the overlap. Biases also arise in coverage from the PCR amplification as GC-rich regions have higher annealing temperature due to the increased stability of GC rich regions[189]. WES also has preferential capture of reference alleles compared to WGS which leads to bias against calling heterozygous SNPs[181].

Comparisons between samples sequenced both by WES and WGS showed that WES samples displayed a right skewed depth of coverage distribution, with a median of 50x depth but a mode of 18x depth[188]. A comparison of samples sequenced using commercially available WES capture kits and WGS over only the common targets regions between WES kits showed that when comparing the fraction of the exome targeted regions covered at above 20x depth was higher for WGS samples sequenced at only 44x and 56x average depth of coverage than with equivalent WES samples. To obtain a similar fraction of the targeted exome at 20x coverage as WGS the WES samples needed an average depth of coverage of 160x for Agilent V4 and around 80-100x for Agilent V5 and NimbleGen V3 kits[189]. Due to the outlined biases WES kits are more susceptible to fail to sufficiently capture all the exons targeted in the capture kit compared to WGS where there is no capture step[187].

WES is reliant on knowledge of coding genes and transcripts to effectively target them for capture. Our knowledge of coding and functionally important regions of the genome is constantly improving as highlighted by comparison of GENCODE versions 27 (Jan 2017) to 28 (Nov 2017) where the number of known genes increased from 58,288 by 93 to 58,381[24]. The number of protein coding transcripts also rose from 80,930 to 82,335 up by 1,405. As successive versions of databases such as GENCODE and RefSeq include new or re-classified genes and transcripts exon capture kits require revisions to update targeted regions from previous generations of capture kits which may not have captured all coding genes and exons[189].

In a study by Lionel *et al.*[187] a total of 103 patients were compared for diagnostic yield between targeted methods and WGS. Across patients a median of three tests were performed sequencing a median 19 genes, the most common test performed was chromosomal micro-array genotyping for 44 of the 103 patients which yielded no diagnostic variants. Median sequencing costs were calculated as $5,173, far above the current cost of WGS. 70 samples were both WES and WGS sequenced, which resulted in 18 additional diagnoses for WGS compared to WES. Several of the additional diagnoses were due to WES not targeting the genes or variants which were deep-intronic and so not captured. The genes identified included: *PIGG*, *RNU4A-TAC*, *TRIO* and *UNC13A* which have relatively recently been identified as

disease causing from 2015/2016[187]. Two CNVs were also identified in WGS which were not detectable in WES due to the coverage patterns[187].

Structural variants (SVs) and copy number variants (CNVs) are significantly more complicated to detect from WES and targeted approaches. Most of the problems with detecting variants arise from the coverage biases previously mentioned. Approaches which use read depths such as CNVkit[80] need to apply corrections to calibrate samples to adjust for lower coverage caused by GC-rich regions. Secondly for repetitive sequence regions where mapping is likely to be variable it leads to increased false positive calls. Finally in targeted sequencing the density or proximity of targets affects the sequence distributions. Two targets in close proximity may cause reads to overlap leading to shoulders and the appearance of increased depth, leading to erroneous gain of copy calls. Conversely two exons with a small gap may appear with a drop in coverage where the ends of the two sequence read distributions meet resulting in erroneous deletion calls at the edges of exons. Often SVs are also large events with breakpoints which lie outside of the exons, making full length copy or SVs difficult to accurately detect from targeted approaches. With WGS the entire genome is covered and with more uniform coverage, empowering CNV and SV calling[187,188].

The obvious limitation of WES and targeted approaches is that they do not encompass as much variation as WGS, consequently for complicated cases where no diagnostic variants can be identified at the time of first analysis there is the possibility a novel gene could be subsequently identified from other studies. If the novel gene was not included in the WES capture kit used then this case would remain un-solved[190]. This limits the future improvements to diagnostic rates by re-analysis of data compared to WGS. By using WGS the additional non-coding, mitochondrial and CNVs variants can all be detected, which minimises the possibility that additional sequencing may be required later to confirm a new suspect gene or intronic region. By reducing future re-sequencing WGS is more cost-effective and saves DNA samples which may be limited for validation of suspected variants such as using additional Sanger sequencing[187].

4.1.3 Variant calling with GATK HaplotypeCaller

As the cost of NGS decreases large studies with many samples are now feasible such as the 100,000 genomes project[18]. This enables joint variant calling to improve detection of rare variants and at sites where calls are ambiguous in some samples but seen clearly in others, enabling quality thresholds to be altered to empower calling. Variant calling pipelines such as SAMtools[46] perform joint calling of samples by supplying multiple BAMs into a single variant call step. Reading the data from multiple BAMs and then performing calculations while holding data in memory is extremely computationally intense. Also should a cohort be expanded with additional samples it then becomes necessary to then re-call all samples using the same computationally expensive step.

In this study GATK HaplotypeCaller (HC) was the chosen to call variants due to its balance of calling accuracy, sensitivity and scalability of the pipeline to joint call large cohorts[47]. GATK HC performed very similar to the class leading SNP caller SAMtools[46,48] but achieved better performance with calling of INDELs[54,191,192]. Overall this made GATK HC the better caller combined with its native support for the gVCF format which deals better with larger sequencing such as WGS. The scalability of the HC pipeline makes it a more useful caller to take advantage of larger datasets as demonstrated by the ExAC consortium which used the program to call over 90,000 exomes[48]. A comparison of the time taken to call 1,500 exomes showed that the CPU-hours taken was approximately halved at 19,000 for HC compared to upwards of 37,000 to 46,000 for other callers.

HC begins calling by defining the active regions for samples by calculation of an 'active probability' across the genome for each sample. Sites are assigned a probability based on genotype likelihoods and a heterozygosity prior created from a pileup of the BAM file. If a region is deemed to produce a signal above a default threshold of 0.002 it is defined as active and proceeds to generate a de Bruijn-like assembly graph for the region using all reads. In each of the paths across the graph edges are assigned weights based on the number reads matching between the nodes and then pruned based a Hidden Markov Model (Pair-HMM) calculating the most likely haplotype across the entire active region. Using the probabilities obtained from

the HMM model for each haplotype that contain the variant allele at that position the probabilities can be supplied to a Bayesian model which calculates the genotype probabilities for each site[48].

For each sample HC produces a gVCF file which omits a reference confidence model and collapses blocks of the genome where there are no non-reference allele variants, calling them as homozygous reference blocks. For the sites which are not homozygous reference the likelihoods at each site that a variant is homozygous alternate or heterozygous is stored. HC is run per sample so each sample will produce a gVCF of likelihoods and avoiding reading of many BAM file simultaneously. gVCFs are then merged using the combineGVCFs tool for large cohorts which generates a single gVCF with all intervals covered in the genome either as collapsed block or sites with variant likelihoods, this method also allows additional samples to be added retrospectively at a later date using relatively cheap processing steps. The joint gVCF file then contains all the variant likelihoods for each sample at each site which allows application of the confidence model to genotype the sites, using the likelihood information from all samples to assign genotypes per sample but also to help resolve poorly covered regions in some samples where there is limited evidence supporting a call or lack of a variant call.

4.1.4 ACMG genes

American College of Medical Genetics (ACMG) genes are a list of 56 curated genes produced by the American College of Medical Genetics and Genomics. Each of the genes in the list have been determined to be actionable by a panel of the ACMG and are recommended to be reported to clinicians. Genes included on this list include: *TP53* or *BRCA2* which are strongly associated with an individuals risk for forms of cancer[193].

4.2 Aims

This project will assess the quality, coverage and variant calling in WES and WGS which was performed on the same family trio. By comparing the variant calls of the WES and WGS the project will assess additional benefits and potential limitations of each method for identifying causal variants both now and by future re-analysis.

4.3 Materials & methods

4.3.1 Samples & sequencing

Blood samples were taken from a family trio consisting of both unaffected parents (SD001 - Father & SD002 - Mother) and a child (SD003) with a rare disease of Sedaghatian-type Spondylometaphyseal Dysplasia (SSMD). Blood samples were taken by clinicians during the presentation in Southampton general hospital. From blood samples DNA was extracted at the Wessex regional genetics laboratory in Salisbury and then sent for WES and subsequently WGS.

Exome sequencing

Exome sequencing was performed by using the Agilent SureSelect Human All Exon (v5) capture kit spanning 50.4 Mb, targeting 357,999 exons over 21,522 genes performed at the Wellcome Trust Centre for Human Genetics (WTCHG) in Oxford. Sequencing was Illumina paired-end on a Hi-Seq 2000 with read lengths of 100 base pairs.

Samples SD001 (Father) and SD002 (Mother) were both run on a single lane, generating paired fastq files. The affected sample SD003 (Child) was subject to additional sequencing as pre-agreed minimal sequencing was not achieved from the first sequencing run, hence the sample was comprised of four fastq files.

Genome sequencing

Genome sequencing was performed at the Beijing Genomics Institute (BGI) in Hong Kong using an Illumina Hiseq X10, paired end sequencing was performed to try and achieve mean depth of coverage of 30x across the genome. Reads from the X10 were 150 base pairs in length. All three samples were sequenced on three lanes yielding six fastq files per sample.

4.3.2 Quality control

FastQC

Checking of sequencing quality was performed using FastQC (v0.11.3)[32]. Results from FastQC report on 12 different measures of data quality. Each of the FastQC reports were then aggregated using the MultiQC tool.

SNP fingerprinting

Verification of sample identities was performed using a SNP-fingerprinting method[99]. Alleles at 25 specified positions were extracted from VCF files which contained the reference, alternate alleles and the genotype. Results were compared against sample KASPR genotyping performed at LGC to verify that the samples were correctly labelled.

To confirm sample sex and familial relationships are as expected pairwise variant comparisons were performed to calculate the number of shared variants but also the percentage similarity between samples using a custom bash script before plotting in R as a heatmap. The percentage of chromosome X heterozygotes also allowed determination of sample sex.

Alignment & reference genome

Alignment was performed with the reference sequence GRCh38.p11 reference sequence using Burrows Wheeler Aligner (BWA) v0.7.12[43].

SAM files generated from BWA were converted to binary BAM files using Picard tools[49]. Picard tools was also used to mark duplicate reads. Marked duplicates are flagged in BAM files such that they are excluded by the read filters in GATK HaplotypeCaller from variant calling but will remain in BAM files.

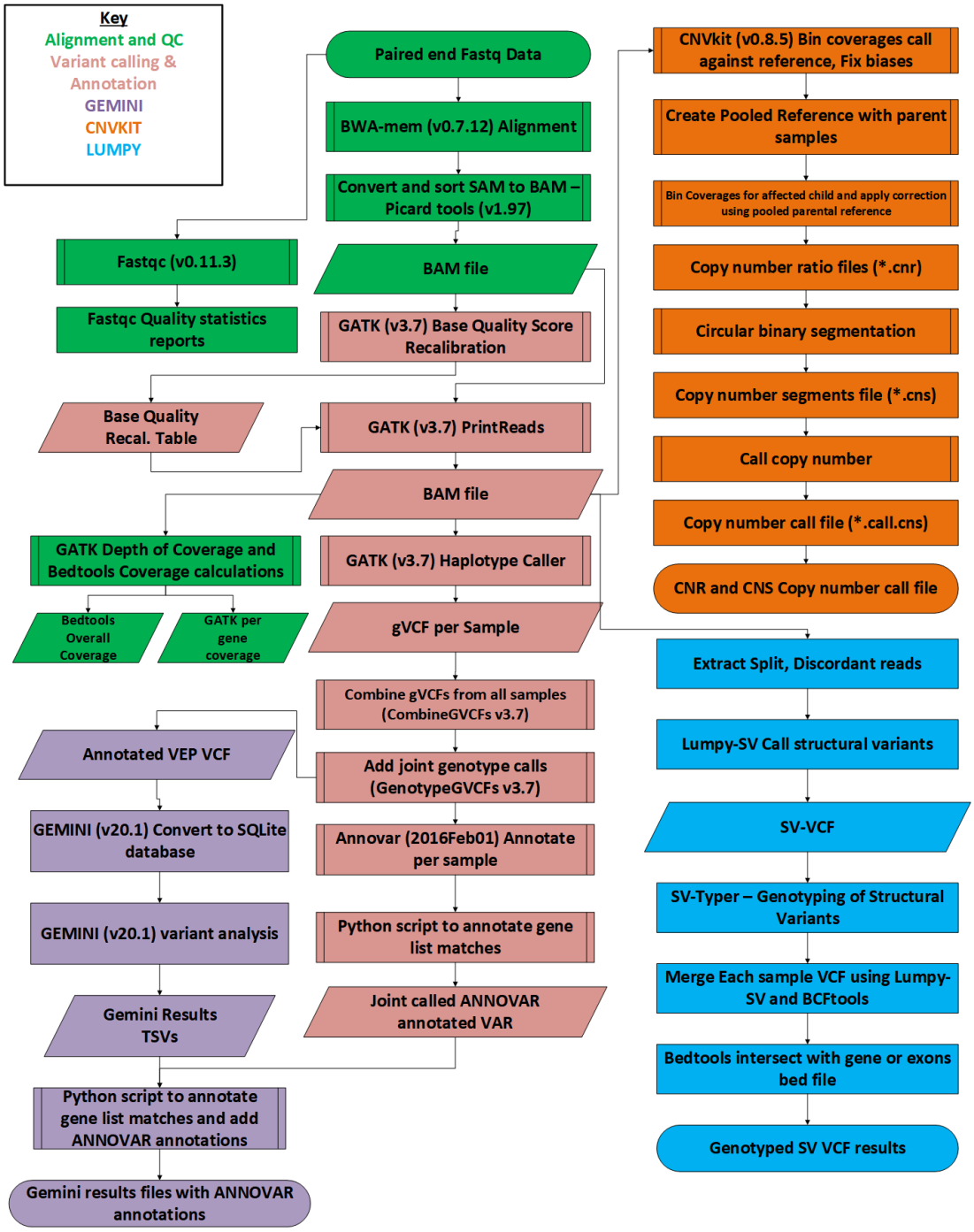

Figure 4.1: Flowchart of the BWA-GATK variant calling pipeline used for the trio WES and WGS samples, including GEMINI and LUMPY analysis steps. Shown in blue are the alignment and variant calling steps performed by BWA (v0.7.12), Picard tools(v1.97) and Genome Analysis ToolKit (GATK v3.7). Once variants are called by HaplotypeCaller the VCF is supplied to GEMINI (v0.20.1) to classify variants by whether they are non-Mendelian, compound heterozygotes, autosomal recessive or dominant variants. Using the BAM file structural variants are called using the program LUMPY, with genotype calls added by SV-Typer. Copy number variants were called also using CNVkit.

4.3.3 Sample coverage

Coverage statistics for samples were calculated using the GATK (v3.7) base quality recalibrated BAM files. Two methods were used to calculate coverage as shown in Figure 4.1. Firstly, SAMtools (v1.2.1)[46] was used to extract only mapped reads with mapping quality above 20 and filter by bit-flag (*1796*) to remove reads marked as duplicates, unmapped or not in primary alignments or those that fail vendor quality checks. Reads which passed filtering were piped into bedtools (v2.21.0)[194]. For WES samples the bedtools command "coverage" calculates coverage based on a set of intervals, in this case a BED file for the capture kit (Agilent SureSelect Human All Exon v5). For WGS samples the bedtools command "genomecov" was used to calculate the coverage over all bases. Coverage was output in a BED file which as with exome coverage were plotted graphically using R. The "-hist" option was also used to output a BED file per method containing the number of bases and percentage of the kit at particular depth which were plotted as a decreasing cumulative plot for each sample.

Coverage per gene was also calculated using GATKs Depth Of Coverage (DOC) tool which applies read filters equivalent to those used in HC and also those applied by SAMtools for bedtools calculations. DOC also used a RefSeq gene list[195] containing names and location which was obtained using the UCSC genome browser[196] for the hg38 genome in November 2017. DOC was run for WES using the same targets BED file as with bedtools generating seven files (four summary and three per base) for the targeted regions describing coverage over: RefSeq genes, sample averages, per chromosome, cumulative counts and per base for genes, chromosomes and entire samples. Seven files were also produced for WGS which performed the same calculations but using all bases in the genome. Calling thresholds were then set to 1, 5, 10, 20, 30, 50 and 100x depth to calculate the percentage of each WES and WGS sample covered at each of the specified depths.

To compare coverage between WES and WGS over exonic regions the WGS data was restricted to the WES target regions. The WES data and exome locations extracted from WGS files was then directly comparable to assess potential biases and poorly covered regions in either method. A comparison was also performed looking at ACMG

genes, comparing coverage over the coding exons which provides a fair comparison of coverage and variant calling of WES and WGS.

4.3.4 Variant calling

GATK (v3.7) was used to call variants for the HG38 aligned files. As shown in Figure 4.1 base quality score recalibration was performed to correct for systemic biases. Variant calling was performed using GATK HaplotypeCaller (v3.7) to generate a gVCF per sample. Each of the gVCFs were then combined into a single gVCF using the combineGVCFs tool. Joint genotyping of variants using GenotypeGVCFs was then performed to yield a single VCF for the trio.

4.3.5 Annotation of VCF variants

VCF files for the WGS and WES trios were created from each of the GenotypeGVCF steps. VCF files were annotated using ANNOVAR (FEB 2016)[56]. VCF files were filtered on a total depth of four across all samples to exclude variants of low depth across the joint calls. Variants passing filtering were converted into the ANNOVAR input file using the "convert2annovar" perl script and supplied into the "table_annovar.pl" perl script. A summary of the databases used for annotation is shown below in Table 4.1.

Additional annotations were also added to ANNOVAR VCF files for identification of candidate genes using a custom python script which would match gene names from genes lists. In total there were four gene lists used to flag variants of interest and one to flag potential false positive variants[197]. ACMG genes were also flagged along with two candidate gene lists for SSMD. VCF files were also annotated using VEP v88 with RefSeq annotations to obtain summary statistics per sample for the variants called and consequences.

Database	Description
refGene	FASTA sequences for all annotated transcripts in RefSeq Gene
knownGene	FASTA sequences for all annotated transcripts in UCSC Known Gene
avsnp144	dbSNP144 with allelic splitting and left-normalization
avsnp147	dbSNP147 with allelic splitting and left-normalization
gnomad genome	gnomAD genome alternative allele frequencies collection
1000g2015	Alternative allele frequency data in 1000 Genomes Project
esp6500siv2	Alternative allele frequency for the NHLBI-ESP project with 6500 exomes
exac03	Alternative allele frequency for the ExAC 65000 exomes
cosmic70	Cosmic 70 database IDs
dbnsfp31a interpro	*
dbscsnv11	dbscSNV version 1.1 for splice site prediction by AdaBoost and Random Forest
clinvar 2016	CLINVAR database with Variant Clinical Significance
hrcr1	40 million variants from 32K samples in haplotype reference consortium
kaviar	170 million Known VARiants from 13K genomes and 64K exomes in 34 projects
nci60	NCI-60 human tumor cell line panel exome sequencing allele frequency data
spidex	Machine-learning prediction on how genetic variants affect RNA splicing
FATHMM-XF	Pathogenicity scores from FATHMM-XF

Table 4.1: List of databases used with ANNOVAR to annotate variants called in the GATK VCF files. * DBNSFP has a combination of annotation. For pathogenicity predictions for selected variants were provided for: SIFT, PolyPhen2, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN and fitCons. Sequence conservation scores were also provided using the tools PhyloP, SiPhy while conserved sequence domains were identified using InterPro domains.

4.3.6 Annotation of non-coding variants

Non-coding variants were called for WGS but also in WES where targets overlap with intronic, intergenic regions or due to some off-target reads being retained in sequencing post hybridisation. Currently non-coding variants can be annotated with alternate allele frequency information from WGS based databases such as the 1KG project[27] and gnomAD[63]. However not all variants will be contained in these databases as it either may be rare or data for the region may be missing as indicated by a frequency of 0 or '.' . Annotations were generated using hg19/GRCh37, so annotations were lifted-over to hg38 using Crossmap[198]. Lift-over can also cause loss of regions where assemblies have changed and cause some incompatibility where variants are reported for the opposite strands in the reference builds.

Pathogenicity scores can be added to non-coding variants however all sources are created using hg19/GRCh37 data, again requiring either lift-over of annotation sources to hg38 or results files back to hg19/GRCh37. DBNSFP version 31a for hg38[199] was used with ANNOVAR to annotate variants with pathogenicity scores. DBNSFP provides scores for some non-coding variants from CADD, DANN and FATHMM-MKL, but is only intended to be used with non-synonymous SNPs hence most non-coding variants are only incidentally annotated and generally lack annotations. To improve ANNOVAR annotations the full hg19 database for FATHMM-XF[70] was downloaded and lifted over to hg38 and annotations added to the VCFs.

To annotate VCFs with CADD the VCF files were lifted over to hg19 due to the large size of the CADD (v1.3) databases of 200GB compared to a relatively small VCF[66]. CADD annotation were added using a local installation of snpEFF (V4.1)[59]. Missing scores were calculated *de-novo* from the CADD website: http://cadd.gs.washington.edu/score.

4.3.7 GEMINI variant analysis

To analyse variants across the trio the program GEnome MINIng (GEMINI) v0.20.1 was utilised[200]. GEMINI takes an input trio VCF file which has been normalised using VT-tools to create unique records for multi-allelic variants, one record for each

REF/ALT combination. VCFs are also left-aligned and represented using the most parsimonious alleles[201].

Variant Effect Predictor (VEP) (v88)[57] was used to add GEMINI compatible annotations and retain the VCF format. Each trio VCF was converted into an SQL-lite database which GEMINI then queries to pull out variants matching the program mode. Additional annotations from ANNOVAR were added to the outputs of each GEMINI program such as for FATHMM-XF as shown in Figure 4.1.

4.3.8 Structural and copy number variant pipeline

Larger chromosomal events than SNPs or INDELs (INDELs defined as upto 50bp) are possible, such events are known as structural variants. The exact definition of the starting size for structural variants can vary between programs. GATK HC will call variants up to 20bp as INDELs relatively reliably, beyond this size it is not reliable and dedicated calling programs are required.

LUMPY

To call large chromosomal variants such as duplications, deletions or inversions the structural variant caller LUMPY (v0.2.13) was used with WGS data[79].

LUMPY requires split and discordant reads to be extracted: split reads - when reads map to locations further apart than the read length, discordant reads - when reads do not map as expected for paired reads such as incorrect orientations or sizes compared against the reference genome. Both split and discordant reads are extracted from WGS BAM files using SAMtools (v1.2.1) to identify potential breakpoints. A flow diagram showing the steps performed to call structural variants is shown in Figure 4.1.

Results from LUMPY are in VCF 4.2 format, which contains the latest nomenclature for structural variants. Each VCF contains a brief description of which variants such as the location and structural variant type along with size and confidence intervals where appropriate. There are five possible structural variant descriptions in the VCF files: duplications, inversions, deletions and breakends or insertions.

Genotypes were added to the variants called by LUMPY using the Bayesian genotyping program SVTyper-SSO (https://github.com/hall-lab/svtyper) based on the algorithms used in SpeedSeq[202]. Each sample was combined using BCFtools (v1.2.1) to generate a trio structural variant VCF.

A BED file of gene locations rather than canonical exons was used for the intersection as intronic or UTR variants could affect splicing or gene regulation. Bedtools intersect[194] was also run with the *-wao* flag which will write all out, so entries in the VCF which do not intersect a gene will still be reported but can easily be filtered out using the gene names column added by the intersection.

CNVkit

CNVkit (v0.8.5) was used to call CNVs using read depth, specifically for WES and targeted sequencing but can also be used with genome and amplicon based sequencing methods[80].

For exome sequencing CNVkit required a target regions BED file, in this case for the Agilent SureSelect Human All Exon (v5) capture kit. This BED file defines the on-target regions using the "*target*" command. This BED file was annotated using a Reflat text file of gene locations and names. The BED file was also used for "*antitarget*" command which then defines the off-target regions along with a list of genomic intervals unable to be mapped in the reference genome. Un-mappable regions such as centromeres, telomeres, and highly repetitive regions can be in a pre-calculated file from UCSC or calculated using the fasta sequence using the "*access*" command. Bin size was selected using the "*split*" command which defines bins with a minimum size of 200bp and target size of 267bp. The average size of a human exon is 200 bp, by choosing a bin size above 200bp this reduces noise from small bins with few reads.

The "*coverage*" command was then used along with the BAM file, target and antitarget region files to bin read depths for the samples. Binning over the on and off target regions each generated a coverage '.cnn' region file. A single pooled reference

'.cnn' file was then be created using the '*reference*' command with both of the parent samples using the target and antitarget coverage '.cnn' files with a fasta file for the hg38 human genome.

Copy number ratios for the samples were then calculated using the pooled reference file with the '*fix*' command. This command also applied corrections for GC-sequence bias, repeat sequences and edge correction where exons in close proximity lead to a multi-modal coverage distribution with losses of copy potentially called in the dips between exons.

Each copy number ratio file required segmentation using the "*segment*" command to group similar depth ratios across the genome. There are three segmentation algorithms supported in v0.8.5, the default circular binary which is a python implementation of DNAcopy[203] which reportedly performed best in CNVkit development. Other available algorithms were fused lasso and haar seg which were not used as testing with other projects showed no significant improvements[80]. Segmentation will yield one copy number segment (CNS) file per sample which can then have thresholds applied to make the copy number prediction based on the log2 ratio values per segment using the default thresholds with the *"call"* command as shown below in Table 4.2 .

log2 upto	Copy number	Interpretation
-1.1	0	Homozygous deletion
-0.4	1	Heterozygous deletion
0.3	2	Neutral copy number
0.7	3	Single Gain of copy

Table 4.2: Thresholds used for CNV segement calling with CNVkit with loss of copy below a copy number of 2 and gains above 2.

WGS data was not intended for CNVkit as there is no off-target region only the un-mappable positions. To accommodate genome data the target regions were defined as the entire genome excluding only the un-mappable regions. Also for genome data the "*fix*" command does not use the edge correction as coverage should be uniform

across exonic and intronic regions in theory.

4.3.9 Comparison of exome and genome variants

Variants contained in each of the trio files were summarised by VEP when annotating the VCFs[57]. To identify variants called exclusively in WES or WGS each of the VCFs per sample were compared using the BCFtools command "*isec*" to calculate the variant intersection. The '*none*' option was also used to only count records with identical reference and alternate alleles. Comparisons per sample were initially performed unfiltered and were subsequently using a depth filter to exclude variants below a depth of 10. BCFtools returned a VCF of intersected variants and two further VCFs of the variants unique to WES and WGS.

Variants unique to either WES and WGS were then filtered to identify those variants more likely to be false positives by removing heterozygous variants below alternate allele frequencies below 0.2 before assessing remaining variants using the annotations from VEP and ANNOVAR as previously described. Gene symbols for each of the variants passing filtering were then supplied into the tool DAVID (v6.8) and analysed using the 'Functional Annotation Clustering' tool[204]. DAVID is a bioinformatic tool used to provide functional analysis of gene lists. The tool was first set to *Homo sapiens* to restrict functional matches to the correct species. Clustering was run using the pre-defined DAVID classification stringency of 'Medium' (Similarity overlap = 3, Similarity threshold =0.5, Initial group membership = 3, Final Group membership = 3 , Multiple Linkage Threshold = 0.5). From each of the clusters InterPro terms were reviewed to assess the functions of each cluster identified and hence the type and functions of genes only called in either WES or WGS.

4.4 Results

4.4.1 Quality control

FastQC

WES samples showed good quality as shown in Figure 4.2, mean quality of fastq files were all above a PHRED score of 30 across the read positions. None of the fastq files were reported to contain over-represented sequences.

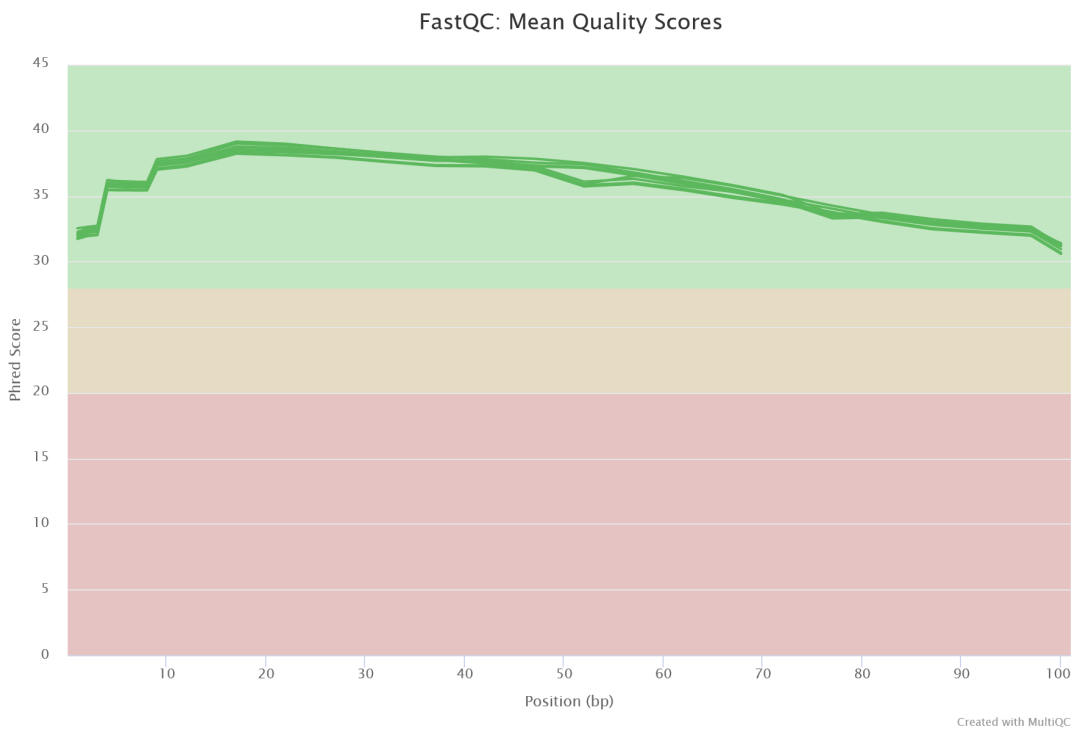

Figure 4.2: MultiQC visualisation of per base quality of WES data. All fastq files are shown with mean quality above 30 across the entire read length.

Sample SD003 was subject to additional sequencing so was comprised of four fastq files. As shown in Table 4.3 the read total for lane one of WES sample SD003 was four million reads fewer than SD002, an approximate 19% reduction. To make up the total reads the sample was sequenced a second time to give a combined total of 35.6 million reads.

Sample	Sequencing	Lane 1	Lane 2	Lane 3	Total Sequences
SD001	Exome	25,259,782	NA	NA	25,259,782
SD002	Exome	21,983,447	NA	NA	21,983,447
SD003	Exome	17,846,534	17,792,799	NA	35,639,333
SD001	Genome	126,375,452	127,753,166	126,344,377	380,472,995
SD002	Genome	141,951,093	143,781,207	142,282,827	428,015,127
SD003	Gemome	121,194,041	122,574,893	120,635,843	364,404,777

Table 4.3: Total reads (sequences) generated for WGS and WES samples divided by lane. WES samples SD001 and SD002 were sequenced with sufficient reads. As seen in lane one for WES sample SD003 the total reads were four million reads fewer, a 19% reduction in data. The WES SD003 sample was sequenced a second time, with the combination of lanes one and two then gave the sample around 35.6 million reads total. For WGS samples SD002 returned the most reads with 48 million more than SD001 and 64 million than SD003.

All three WGS samples were sequenced using three lanes yielding six fastq files. Per base sequence quality for all samples is shown in Figure 4.3. Mean quality can be seen to be above a PHRED score of 30 for positions up to base 130 after which mean qualities fall below 30 for some files down to a low of 28 as sequencing progresses. This trend is commonly observed with Illumina sequencing as the chemistry degrades and loses performance as sequencing proceeds, particularly with longer reads.

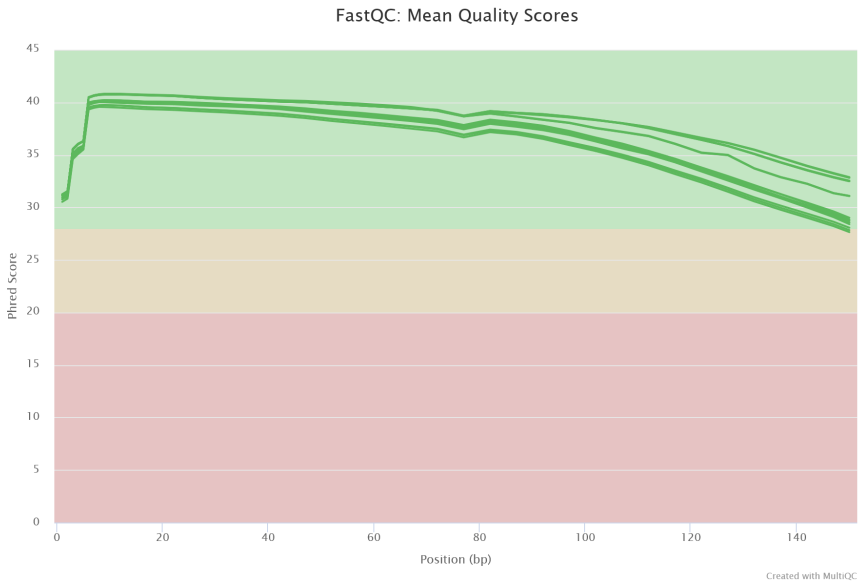

Figure 4.3: MultiQC visualisation of per base quality of WGS data. Fastq files are shown with mean quality above 30 up to positions 130 for most fastq files but some file decrease in quality down to 28 by positions close to 150bp.

In samples SD001 and SD003 adapter sequences were identified at levels above 0.1% and so were flagged by FastQC. As seen in Table 4.4 the percentage of each adapter is low with the highest reported being 0.3%. It was decided therefore based on mean per base quality, low percentages of adapters and GATK best practice guidelines not to perform trimming and or filtering of reads.

Sample	Lane	Forward or Reverse	Sequence	%
SD001	1	Forward	TruSeq Adapter, Index 7 (96percent over 33bp)	0.26
SD001	1	Reverse	Illumina Single End PCR Primer 1 (96percent over 31bp)	0.29
SD001	2	Forward	TruSeq Adapter, Index 7 (96percent over 33bp)	0.26
SD001	2	Reverse	Illumina Single End PCR Primer 1 (96percent over 31bp)	0.3
SD001	3	Forward	TruSeq Adapter, Index 7 (96percent over 33bp)	0.26
SD001	3	Reverse	Illumina Single End PCR Primer 1 (96percent over 31bp)	0.29
SD003	1	Forward	TruSeq Adapter, Index 3 (97percent over 34bp)	0.14
SD003	1	Reverse	Illumina Single End PCR Primer 1 (96percent over 31bp)	0.17
SD003	2	Forward	TruSeq Adapter, Index 3 (97percent over 34bp)	0.14
SD003	2	Reverse	Illumina Single End PCR Primer 1 (96percent over 31bp)	0.17
SD003	3	Forward	TruSeq Adapter, Index 3 (97percent over 34bp)	0.14
SD003	3	Reverse	Illumina Single End PCR Primer 1 (96percent over 31bp)	0.17

Table 4.4: Adapter sequences were only detected in WGS samples at over-represented levels. Levels of adapter sequences in lanes were still at low levels with at most being 0.3%.

Alignment

Alignment of reads to hg38 resulted in over 99% of reads mapping in all WES and WGS samples as shown below in Table 4.5.

Sample	**SD001**			**SD002**			**SD003**		
Reads	**Total**	**Mapped**	**Mapped (%)**	**Total**	**Mapped**	**Mapped (%)**	**Total**	**Mapped**	**Mapped (%)**
Exome	50,711,354	50,533,404	99.65	44,129,566	43,978,057	99.66	71,547,167	71,231,490	99.56
Genome	519,451,403	517,423,990	99.61	583,554,661	582,043,893	99.74	498,137,509	496,743,331	99.72

Table 4.5: Read alignment mapping summary for exome and genome sequencing for all members of the family trio. All samples reported above 99% of reads mapping to the human hg38 reference genome.

Exome sample coverage

Samples SD001 and SD002 have modal coverage of 36x and 32x respectively over all targeted regions as ca lculated with bedtools and shown in Figure 4.4. Modal depth of coverage is lower than sample SD003 at 52x due to the additional sequencing performed for the sample.

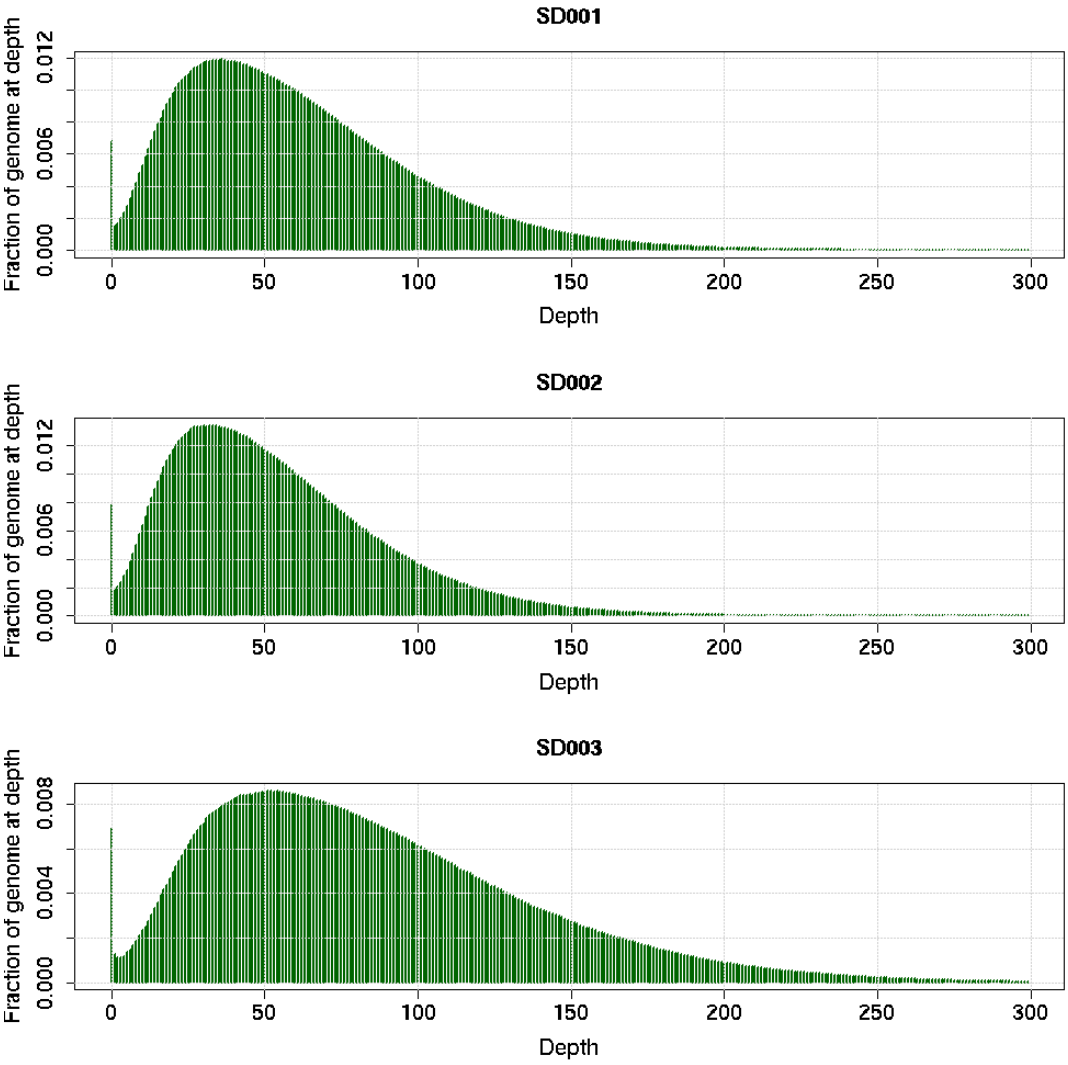

Figure 4.4: Coverage histograms for samples SD001(Top), SD002, SD003. Plots show the fraction of samples at each depth level upto 300. For SD001 and SD002 modal depth of coverage was 36 and 32 while SD003 was 52 due to additional sequencing.

A spike in the coverage fraction at 0x depth of coverage is visible in all samples. The fraction of the exome lacking any coverage was 0.0068, 0.0078 and 0.0069 for SD001, SD002 and SD003 respectively. The lack of coverage is consistent across the three

samples. This could be explained by regions not mapping well to the hg38 reference or specific sequences which were not captured efficiently by hybridisation steps during exon capture resulting in the lack of coverage. Alternatively if deletions are inherited by the child from the parents regions would be lacking or missing depending upon zygosity of the deletion.

Coverage was also visualised as an declining cumulative frequency plot as shown in Figure 4.5. Each series on the plot represents the fraction of the sample which was at or above that depth level. In samples SD001 - 55% and SD002 - 48% of positions were of depth 50x or above. Sample SD003 had close to 72% of positions equal to or above 50x depth, the curve of SD003 was comparatively shifted to the right due to the additional sequencing performed.

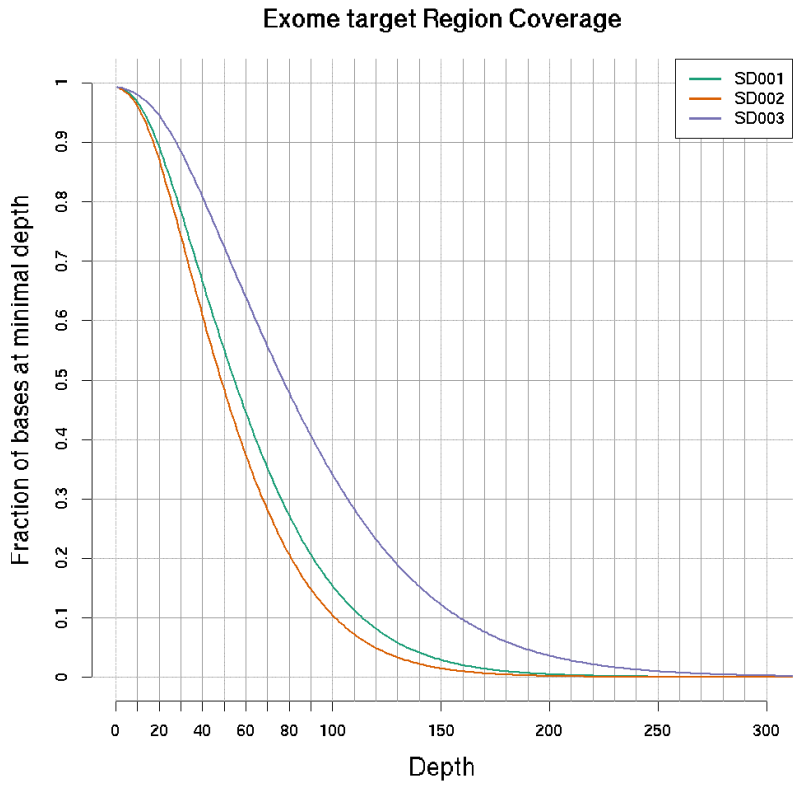

Figure 4.5: Plot of decreasing cumulative frequency for WES samples. At each depth level the fraction of the sample at that depth of coverage level or greater is shown. SD003 is shifted to the right of other samples due to the extra sequencing. All samples were covered at above 75% of the exome at a minimum depth of coverage of 30x.

Genome sequencing coverage

The expected sequencing level for genome samples was a mean of 30x depth for each sample across the genome. Figure 4.6 shows the modal value for depth of coverage lies approximately at 24x for SD001, 23x for SD003 and SD002 covered at a slightly higher 26x.

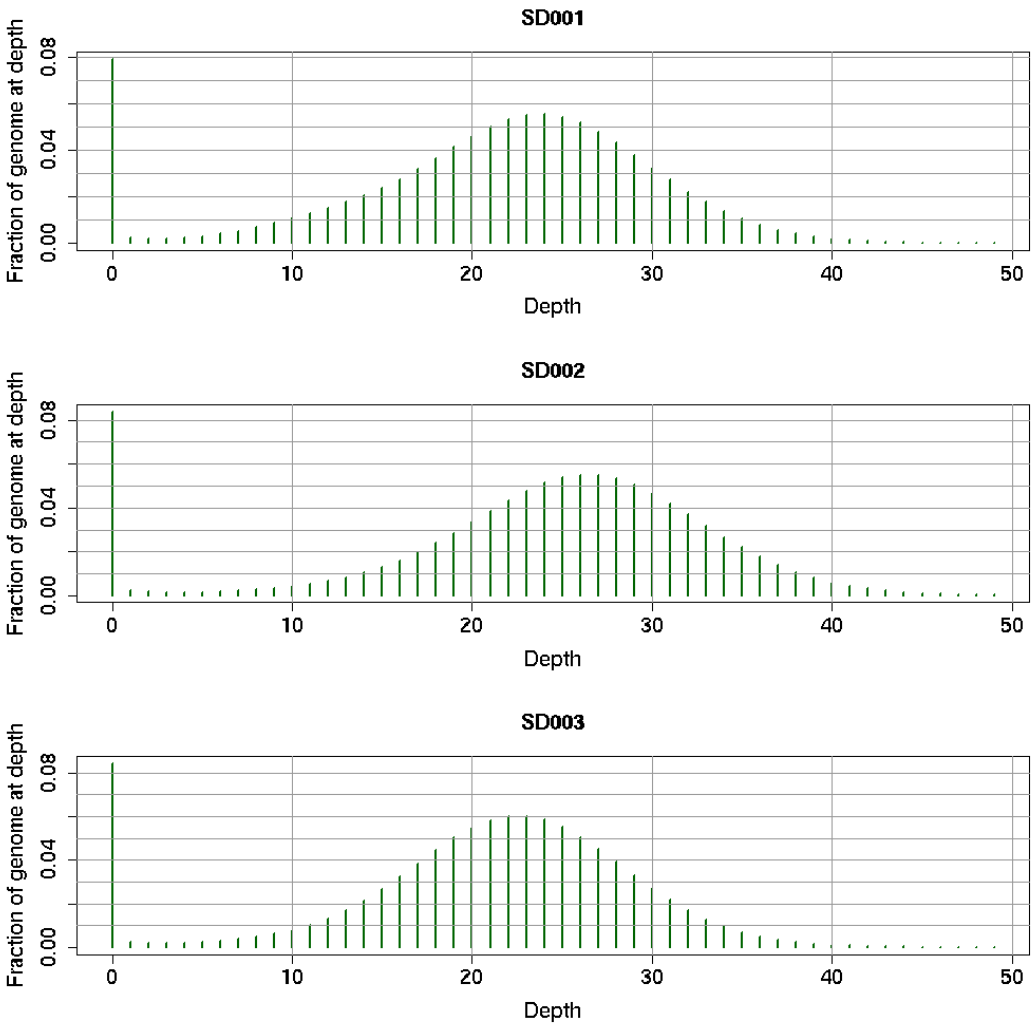

Figure 4.6: Coverage across WGS samples as a fraction of total sequencing for samples SD001, SD002 and SD003.

Unlike WES there is no hybridisation step, therefore regions with no coverage are likely due to mapping issues to hg38 or deletions in samples. WGS coverage calculations also do not account for regions which can be difficult to map to around such as: centromeres, repetitive and pseudoautomal regions.

Plotting coverage as a decreasing cumulative shown in Figure 4.7 reveals the cumulative decrease in depth of coverage is greatest between 20x and 30x. Between 10 and 20% of each sample is coverage at a depth of 30x or greater.

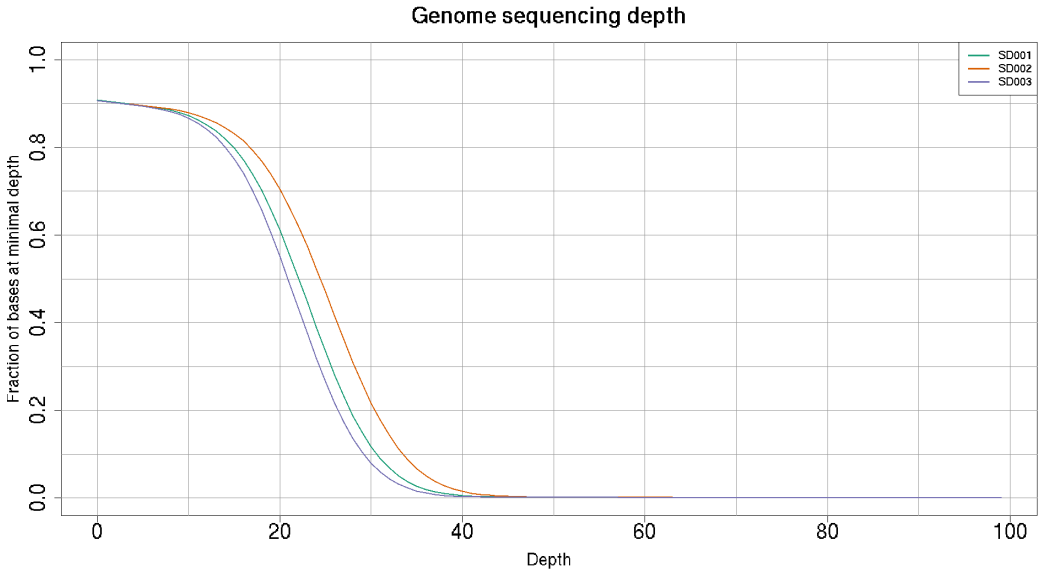

Figure 4.7: Coverage across genomic sequencing samples as a decreasing cumulative total of sequencing. Mean coverage for samples is 22x for SD003, 23x for SD001 and 25x for SD002.

4.4.2 Variant calling

Checking sex and familial relationship

Figure 4.8 shows the percentage similarity of VCF files using only coding variants from exome samples. The child (SD003) should be related to both parents, hence the similarity between the child and parents is higher than between the parents. From the matrix the similarity between the child and both parents is 75% and 79%, whilst (SD001 & SD002), both show similarity of 61.8% and 63%.

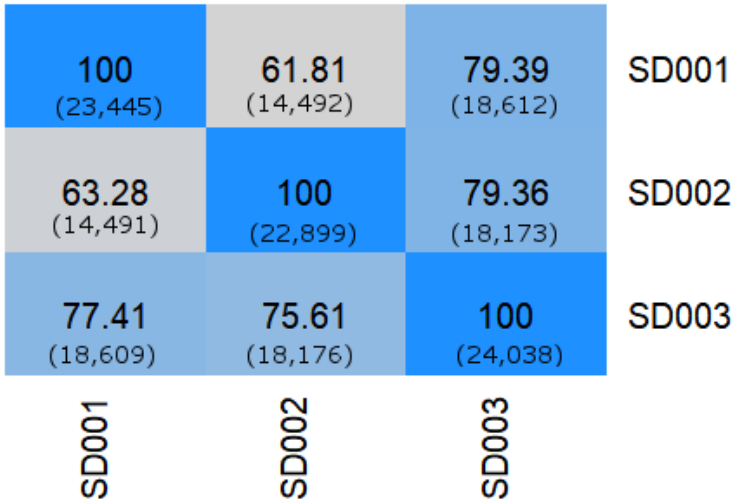

Figure 4.8: Heatmap of similarity for WES samples. When comparing the coding variants contained within each of the VCF files a measure of the percentage matches of variants is calculated. For the child sample SD003 the relatedness should be higher than that of the two parent samples SD001(Father) and SD002(Mother). Comparisons were performed in both directions in the matrix.

The WGS sample matrix of similarity is shown below in Figure 4.9. Similarity between parents and child is around 56% compared with 42% for between parents. The difference between parental and parent-child similarity seems similar to the 11% difference measured in WES samples though with an overall higher similarity percentage.

Testing the samples by measuring chromosome X heterozygote variants as described in Chapter 2 provided a quick check for the gender of the samples which can be matched to the phenotypic data. QC measures described in Chapter 2 were tested on the samples as shown on the following page in Table 4.6. From these results they

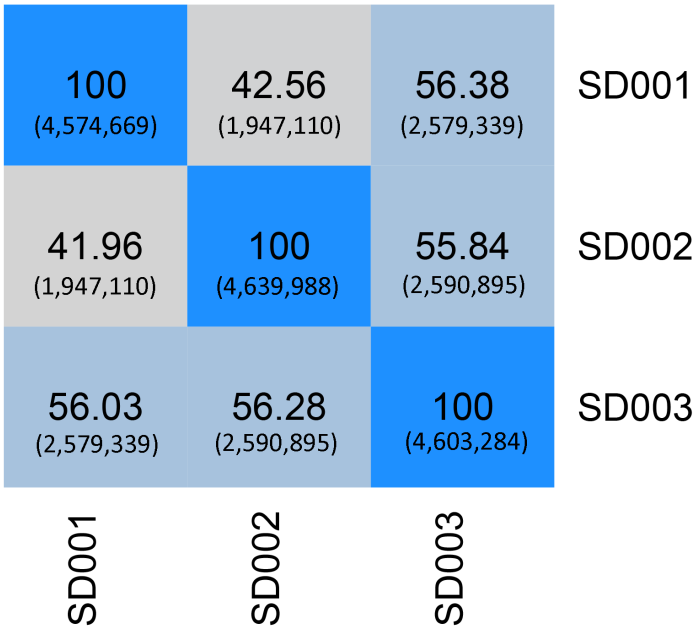

Figure 4.9: Heatmap of similarity for WGS samples. Comparing the variants contained within each of the VCF files a measure of the percentage matches of variants is calculated. For the child sample SD003 the relatedness should be higher than that of the two parent samples SD001(Father) and SD002(Mother). Comparisons were performed in both directions in the matrix.

match the expected sex of the samples, SD001 (Father) is assigned as a male, SD002 (Mother) and SD003 (Daughter) are both predicted as females as they have above 52% chromosome X heterozygote variants. VerifyBamID was run on all BAM files and found below 0.2% contamination for WES samples and below 0.1% for WGS samples.

SAMPLE	Total	PC AAF HOM	PC AAF HET	Total Hets	PC Xhets	Sex	Het:Hom Ratio	Dev Metric	Dev A	Dev B
SD001 EXOME	84672	99.8	45.2	31603	16.7	Male	1.7	5.0	0.2	4.8
SD002 EXOME	79504	99.8	45.7	28700	57.5	Female	1.8	4.5	0.2	4.3
SD003 EXOME	100168	99.7	44.9	35500	60.7	Female	1.8	5.4	0.3	5.1
SD001 WGS	4133509	99.8	46.5	1578574	10.1	Male	1.6	3.7	0.2	3.5
SD002 WGS	4271522	99.8	46.8	1561181	61.7	Female	1.7	3.5	0.2	3.2
SD003 WGS	4161478	99.8	46.5	1538281	57.3	Female	1.7	3.7	0.2	3.5

Table 4.6: Results from selected QC measures described in chapter 2. The number of variants retained after filtering (phred > 20 and depth > 10) are shown in the table. The percentage of X heterozygotes for SD001 is low at below 20% to be classified as a male. Samples SD002 and SD003 are both correctly predicted as female.

SNP fingerprinting

SNP fingerprinting compared the independently genotyped data for 24 SNPs with the alleles called from the WES and WGS data. A summary grid of the data is shown in Figure 4.11. There was only one mismatch between the genotyping and variant alleles amongst the six samples suggesting that samples were correctly labelled. Variant rs2298628 was the only reported disagreement with the genotyping in the exome sample SD003. The position was visualised and found to be of low depth for an exome variant totalling only 11 reads split as eight reference C alleles and three alternate T alleles and was called then as reference homozygote, a read pile-up of the variant is shown in Figure 4.10.

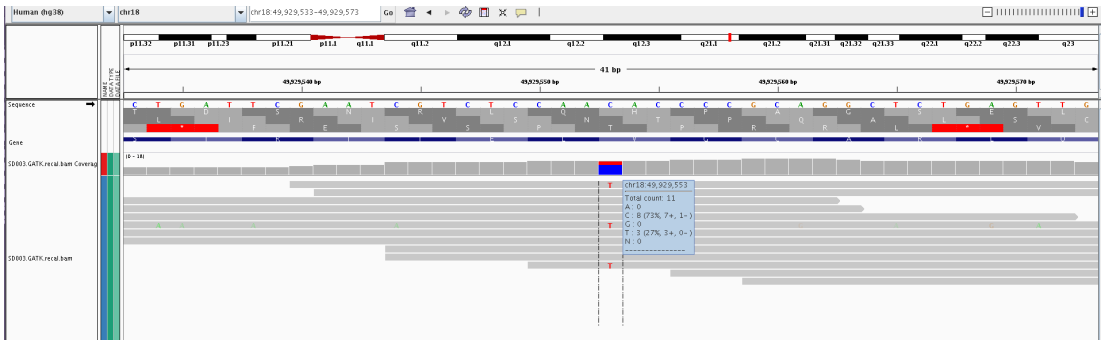

Figure 4.10: Visualisation of variant rs2298628 in exome sample SD003 with low depth of 11 total alleles split as eight reference C alleles and three alternate T alleles leading to a homozygous reference call and disagreement with genotyping.

Sample	rs2228611	rs497692	rs1410592	rs2229546	rs10203363	rs2819561	rs4688963	rs309557	rs2942	rs17548783	rs4735258	rs1381532	rs10883099	rs4617548	rs7300444	rs2297995	rs4577050	rs2070203	rs1037256	rs2298628	rs10373	rs4148973	rs4675	rs9962023	Concordance
SD001 WGS	?	Y	Y	Y	Y	Y	Y	Y	?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	23/23
SD001 Exome	?	Y	Y	Y	Y	Y	Y	Y	?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	23/23
SD002 WGS	?	Y	Y	Y	?	Y	Y	Y	?	Y	Y	Y	?	Y	Y	?	Y	Y	Y	?	Y	Y	Y	Y	19/19
SD002 Exome	?	Y	Y	Y	?	Y	Y	Y	?	Y	Y	Y	?	Y	Y	?	Y	Y	Y	?	Y	Y	Y	Y	19/19
SD003 WGS	?	?	?	Y	Y	?	Y	Y	Y	Y	?	Y	Y	Y	Y	?	Y	?	?	y	Y	?	?	Y	14/14
SD003 Exome	?	?	?	Y	Y	?	Y	Y	Y	Y	?	Y	Y	Y	Y	?	Y	?	?	N	Y	?	?	Y	13/14

Figure 4.11: SNP fingerprint grid for skeletal dysplasia samples comparing calls from genotyping with those from exome and genome sequencing. In the grid the alleles which match the genotyping data are shown in green, mismatches in red and non-matches due to missing data are shown in white. Alleles for 25 SNPs were extracted from VCF files, to determine if SNPs were heterozygous or homozygous alleles were referenced with the genotypes contained in the "Info" field (0/1 - Heteozygote, 1/1 - Homozygote). Sites where genotyping and or sequence data are not present are not shaded and contain question marks indicating missing data. One mismatch was identified for sample SD003 exome for variant rs2298628 highlighted in red. All sites not missing data or are at low depth, as in rs2298628 for WES SD003, match the results from genotyping.

Variants per sample

Using the RefSeq classifications from VEP annotated VCF files statistics were compiled to compare the totals of variants called per sample for both WES and WGS samples. Classifications of variants by type are shown in Figure 4.12. Most obvious from all panels of Figure 4.12 is that WGS samples contain many more variants than WES, as would be expected. Comparisons of total variants called show a joint call total of 582,406 for WES samples compared with 250,700 , 227,975 and 321,728 for SD001, SD002 and SD003 respectively. More variants may have been called in SD003 due to the additional sequencing performed leading to increased depth of coverage as shown in Figures 4.4 and 4.5. WGS samples returned a joint total of 6,461,339 variants and 4,541,959 , 4,604,340 , 4,567,810 for SD001, SD002 and SD003 respectively. The total number of variants in WGS is relatively constant only varying by 62,381 compared to WES which varies by 93,753.

WES samples report 232,617 , 211,830 and 297,541 SNPs for SD001, SD002 and SD003 respectively, a range of 85,711 variants. WGS samples are more consistent with 3,903,969 , 3,943,325 and 3,930,347 SNPs reported for SD001, SD002 and SD003 respectively, a smaller range of only 39,356 SNPs. Comparisons of joint called files show that 9.98 times more SNPs are called in WGS than WES. Approximately ten times more SNPs were detected in the joint called WGS at 5.4 million compared to 542,933 in the joint WES dataset.

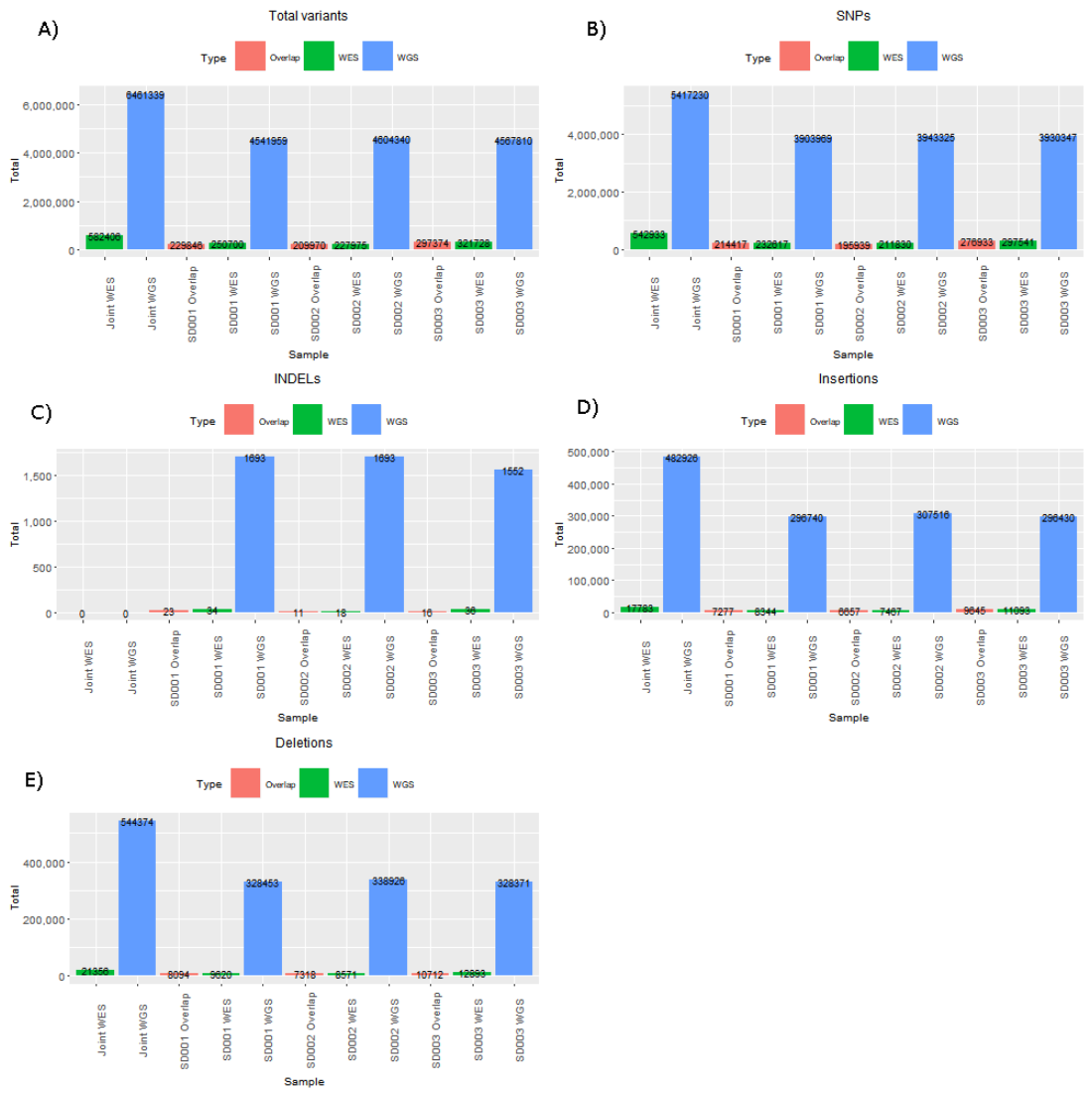

Figure 4.12: Comparison of variant calls by type for WES and WGS trios using statistics calculated using VEP (v88). Overlaps in calls between WES and WGS samples are also shown. **A)** Total variants called per sample was around 4.5-4.6 million per WGS sample and between 227,000 - 322,000 for WES samples. Comparison of trio calls shows WES identified 582,406 variants and WGS 6,461,339 variants. **B)** The majority of the variants in samples were SNPs with approximately 92% of WES and 85% of WGS variants being SNPs. **C)** INDELs were counted per sample and the same number of variants are identified in SD001 and SD002 in WGS with 1,693 but 34 and 18 are called for samples in WES. SD003 reported 1552 INDELs in WGS and 36 in WES. **D)** Between 7,467 and 11,093 insertions are detected in WES samples compared with 296,740 to 307,516 in WGS samples. **E)** WES samples report between 8,571 to 12,893 per sample compared to 328,371 to 328,435 for WGS samples.

VEP sub-divides insertions and deletions which would traditionally be grouped under the INDEL category and uses sequence ontology definitions. INDELs are only counted if an event contains both an insertion and deletion greater than 2 bases. Few INDELs are called in WES at 34, 18 and 36 for the three samples compared with 1,693 for both SD001, SD002 and 1552 for SD003 in WGS.

Insertions are classified as the addition of one or more bases between two adjacent bases. Conversely deletions are simply the excision of one or more nucleotides. For both insertions and deletions far fewer variants are detected in WES compared to WGS samples with around 25-41 times more variants per sample. Between 7,467 and 11,093 insertions are detected in WES samples compared with 296,740 to 307,516 in WGS samples. There is a proportionately larger difference between WES and WGS joint calls with 482,926 variant in WGS compared to 17,783 , approximately 27 times more variants. WES reports between 8,571 to 12,893 per sample compared to 328,371 to 328,435 for WGS. As with insertions in joint calls approximately 25 times more variants are called in WGS than by WES.

Examination of SNPs using the most severe consequences per variant were performed to compare the detection capability of WES and WGS as shown in Figure 4.13. Comparisons of missense variants reveal an average of 1,018 more variants were called per sample in WGS. An additional 1,634 variants in the joint called WGS file were identified compared to WES. A similar pattern is also observed for synonymous variants with an average of 523 additional synonymous variants called in WGS, with a further 803 called in the joint called compared to the WES joint call. For stop-loss and start-loss variants there are consistently between four and seven more variants called in WGS data per sample. Stop-gain variants exhibit a larger difference between WGS and WES with between 20 and 27 variants additionally called per sample in WGS data. Comparing calling of splice region variants revels noticeably more variants called in WGS data for splice region, donor and acceptor variants. An average of 1,800 additional splice region variants were called in WGS data per sample, focusing on splice donor and acceptor variants though shows the average difference per sample is 218 and 120 variants respectively.

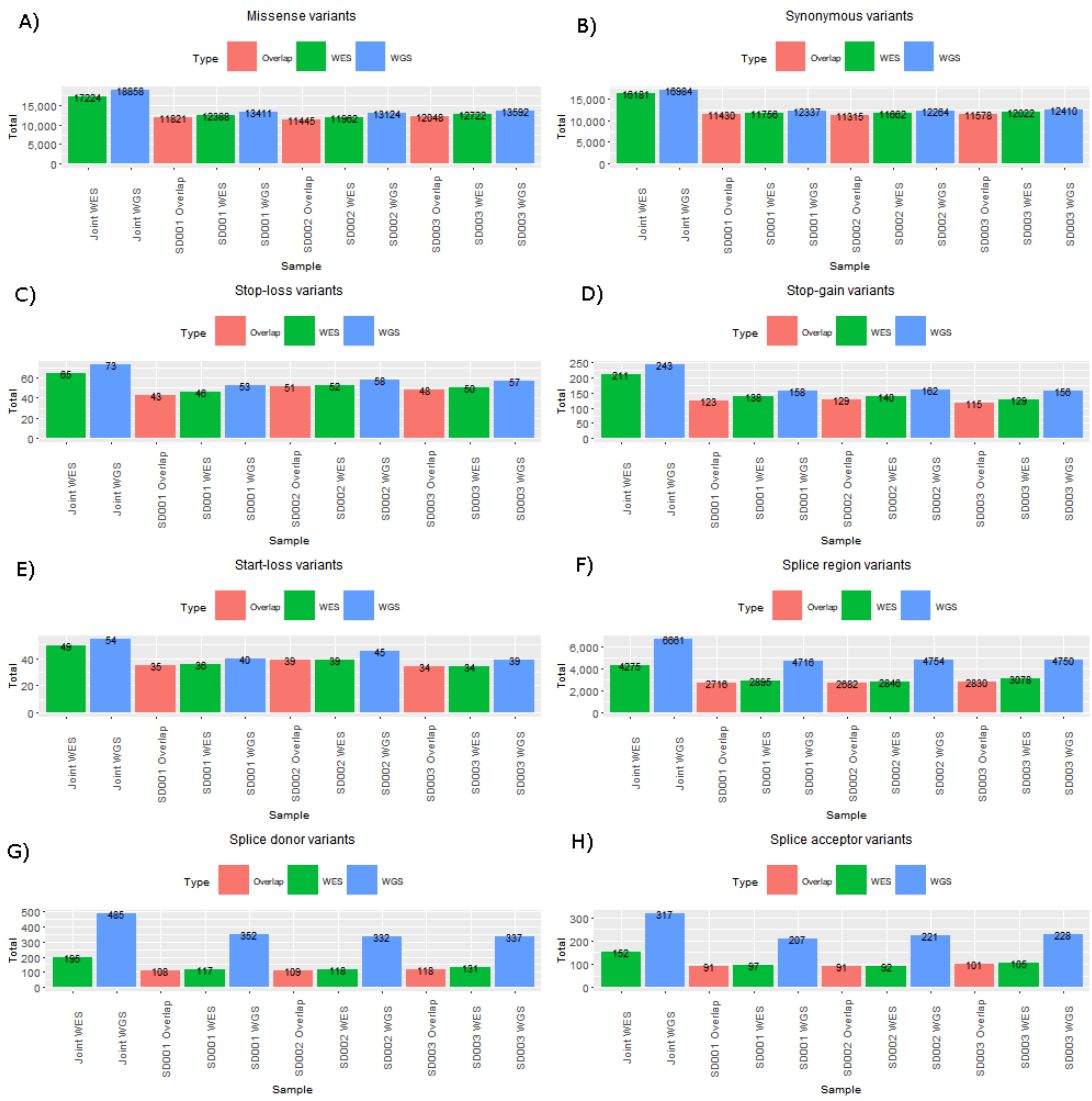

Figure 4.13: Comparison of coding and splicing variant calls by most severe consequence for WES and WGS trios using VEP. Overlaps in calls between WES and WGS samples are also shown. **A)** Around 12,000 missense variants are called in WES per sample, typically WGS called an additional 1000 missense variants per sample. For joint calls WGS called an additional 1,634 variants. **B)** Similar numbers of synonymous variants were identified between WES samples at between 11,662 and 12,022. Around 400-600 additional synonymous variants per sample were in WGS data. **C)** Similar numbers of stop-loss variants are called between WES and WGS with around 50 in WES and only 6 or 7 additional variants detected per sample for WGS. **D)** More stop-gain variants are detected in WGS than WES with 129-140 in WES and an additional 20-27 in WGS. **E)** Comparable numbers of start-loss variants are called in WES between 34 and 39 with only between 4-6 additional WGS variant calls. **F)** Around 2,000 more splice region variants are captured in WGS compared to WES **G)** Approximately 200-240 splice donor variants per sample are only detected in WGS calls. **H)** Approximately 100-130 splice donor variants per sample are only detected in WGS calls.

4.4.3 Non-coding variants

Many times more non-coding variants were called in WGS than WES as shown in Figure 4.14. In WGS samples around 2.1 million intronic and 1.7 million intergenic SNPs are called per sample compared to around of 120,000 and 36,000 in WES samples respectively. Comparison of regulatory regions as annotated by the ensembl regulatory build shows that there are around 130,000 variants per WGS sample compared to only 4,500-7,000 in WES.

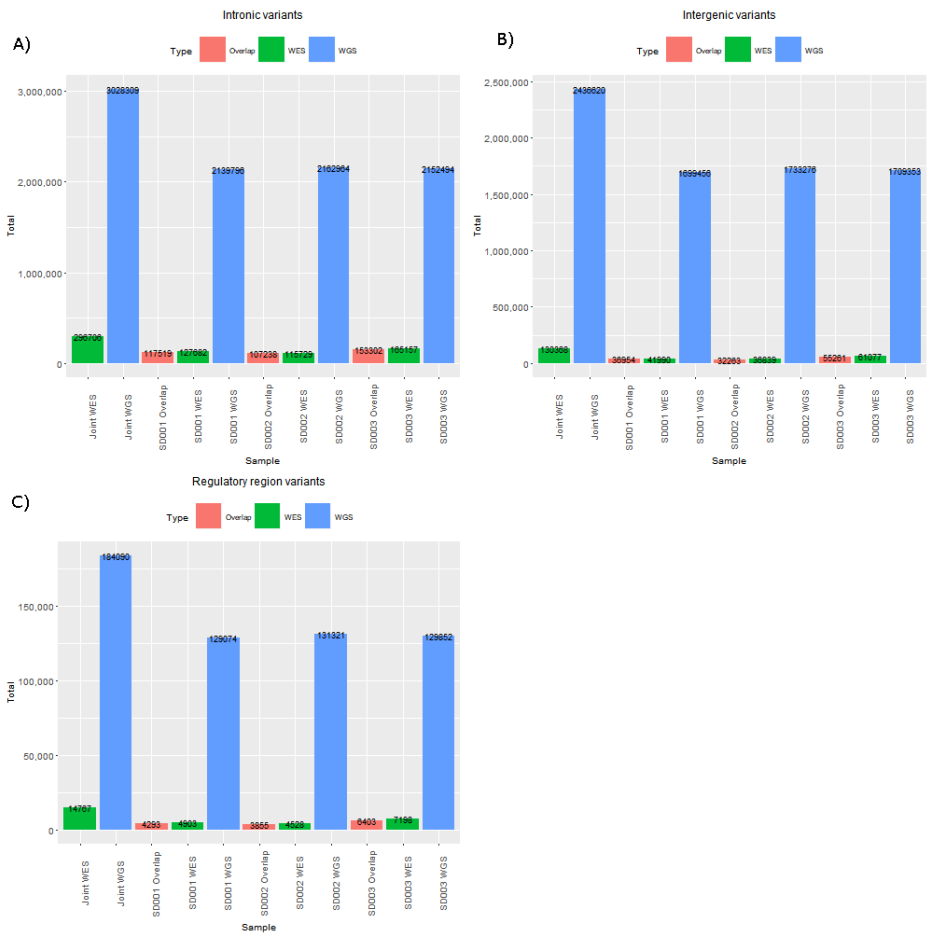

Figure 4.14: Comparison of non-coding variants called in WES and WGS data by sample. Overlaps in calls between WES and WGS samples are also shown. A) Intronic variants are the most commonly detected non-coding variant in both WES and WGS data. In WES samples between 115,729 and 165,157 variants were called compared with around 2.1 million in WGS. **B)** Intergenic variants were the second most prevalent event detected with a range of 36,839 to 61,077 in WES compared to 1.7 million variants per WGS sample. **C)** Regulatory region variants as calculated using the ensembl regulatory build only identified 4,903 , 4,528 and 7,198 variants in WES data per sample compared to approximately 130,000 variants in WGS per sample.

CADD

Using a version of the trio WGS calls lifted-over to HG19 all variants were annotated with CADD scores using snpEFF. Variants were then filtered using the a pathogenicity threshold of CADD-phred score above or equal to 15. A total of 5,118,438 variants were annotated across the WGS trio file. 66,497 variants had a CADD-phred score above or equal to 15. Variants were also annotated with ANNOVAR which allowed additional filters: 1000 genomes all populations allele frequency of below 0.05 or missing and selecting non-coding variants (no exonic or splicing variants) using Known Gene reduced the totals down to 8,735. Variants were then sub-divided using genotypes to plot per sample as shown in Figure 4.15. For each sample a similar number of variants were above a CADD-phred score of 15 at: 4,682, 4,783 and 4,776 for SD001, SD002 and SD003 respectively. Using a CADD-phred of 20 selects the top 1% most pathogenic variants which totalled 711, 726, 732 for SD001, SD002 and SD003 respectively.

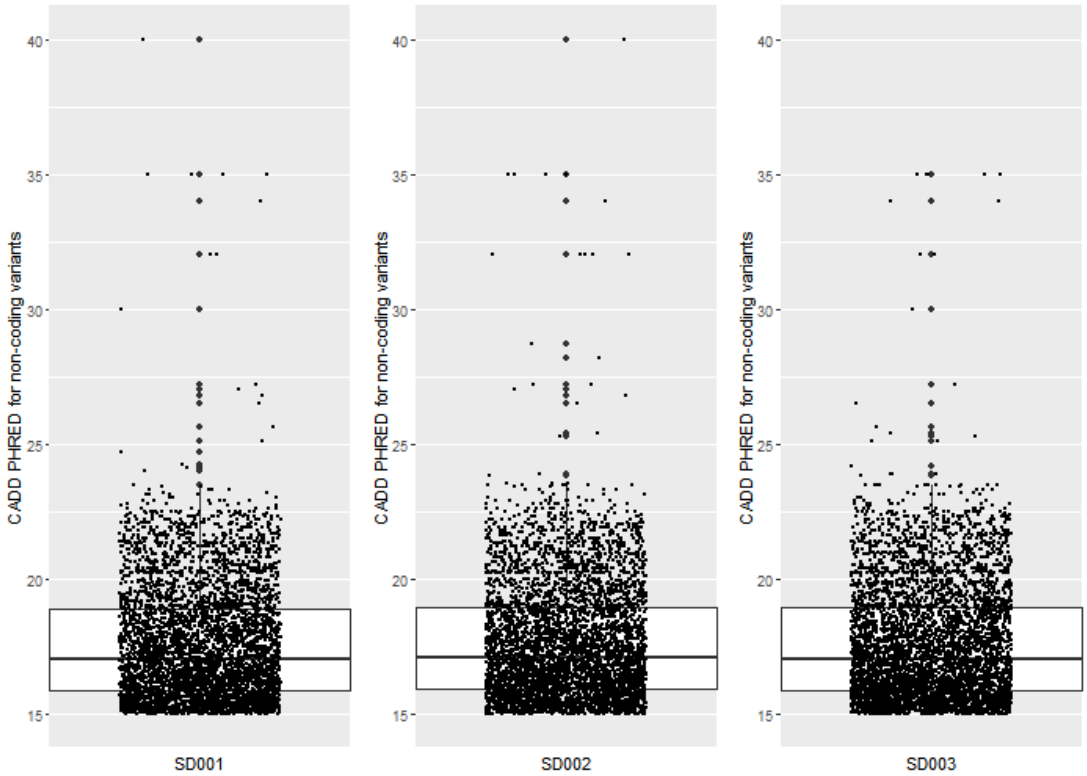

Figure 4.15: Distribution of non-coding variants annoated as pathogenic by CADD per sample. For each sample a similar number of variants were above a CADD-phred score of 15 at: 4,682 , 4,783 and 4,776 for SD001, SD002 and SD003 respectively.

FATHMM-XF

Using lifted over HG38 annotations for FATHMM-XF a total of 3,640,746, 3,636,958 and 3,641,644 SNPs were annotated for samples SD001, SD002 and SD003 respectively in the trio WGS VCF. Using the FATHMM-XF cut-off for pathogenicity as 0.5 variants with either a coding or non-coding score above the threshold were extracted, selecting 29,820 variants. Application of subsequent filters to exclude coding variants using Known Gene annotation and using genome-wide alternate allele frequency measure gnomAD all with frequency of below 0.05 or missing reduced this list down to 3,796 variants. Variants were then divided by sample giving 1,920 , 1,985 and 1,985 again for SD001, SD002 and SD003 respectively as shown in the Figure 4.16. Application of the high confidence pathogenicity cut-off of above 0.96 identified 27, 40 and 35 variants in SD001, SD002 and SD003 respectively.

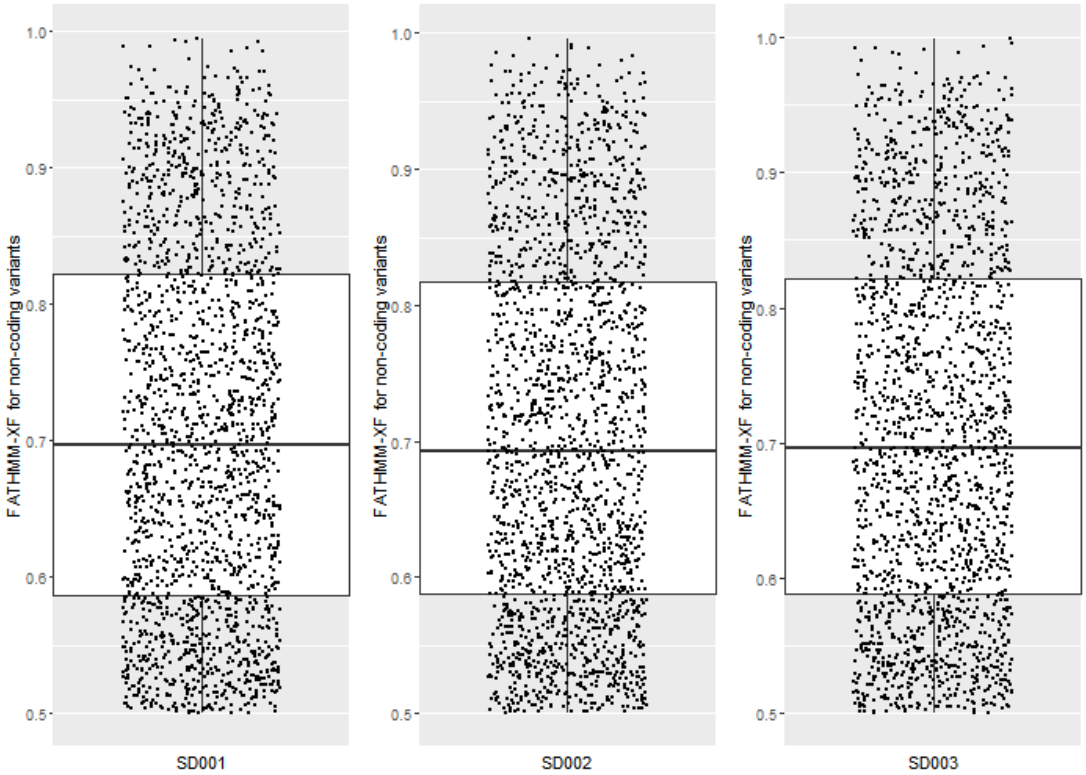

Figure 4.16: FATHMM-XF identified 3,796 non-coding variants with a score above 0.5 and which were rare with gnomAD below 0.05 or missing data. 1920, 1985 and 1985 variants above the 0.5 threshold were identified for SD001, SD002 and SD003 respectively.

4.4.4 GEMINI variant calls

Using trio called VCF files GEMINI created SQL-lite databases from which analyses to group variants by inheritance mode could be performed. Joint called VCFs contained genotype information for each member of the trio for every site where a variant was called. GEMINI was run with three analysis modes to prioritise possible causal variants which are discussed further in the Results section of Chapter 5. The totals for each mode are shown in Figure 4.17 for WES and WGS trio files divided in each plot by either coding or non-coding variants.

The first analysis tool run was to identify errors in Mendelian inheritance. There are four possible violations for when a single new allele is seen compared to either parent termed 'plausible *de novo*', when two new alleles are observed it is termed an 'implausible *de-novo*' variant. The third violation possible is a loss of heterozygosity when the child had only inherited alleles from one parent. Finally uniparental-disomy is when a child inherits the alleles from one parent. Examples of each mode are summarised below:

1. **Loss of Heterozygosity** - Child inherits one allele from one parent.
   $G/A + G/G \rightarrow A/A$

2. **Implausible *de-novo* mutations** - Child has two new alleles compared to parents.
   $G/G + G/G \rightarrow A/A$

3. **Plausible *de-novo* mutations** - Child has one new allele compared to parents.
   $T/T + T/T \rightarrow T/C$

4. **Uniparental Disomy** - Child's alleles are same as one parent.
   $A/A + G/G \rightarrow A/A$

For all four plots in Figure 4.17 there were many times more non-coding than coding variants identified from WGS than WES. In Panels A and B the comparisons of

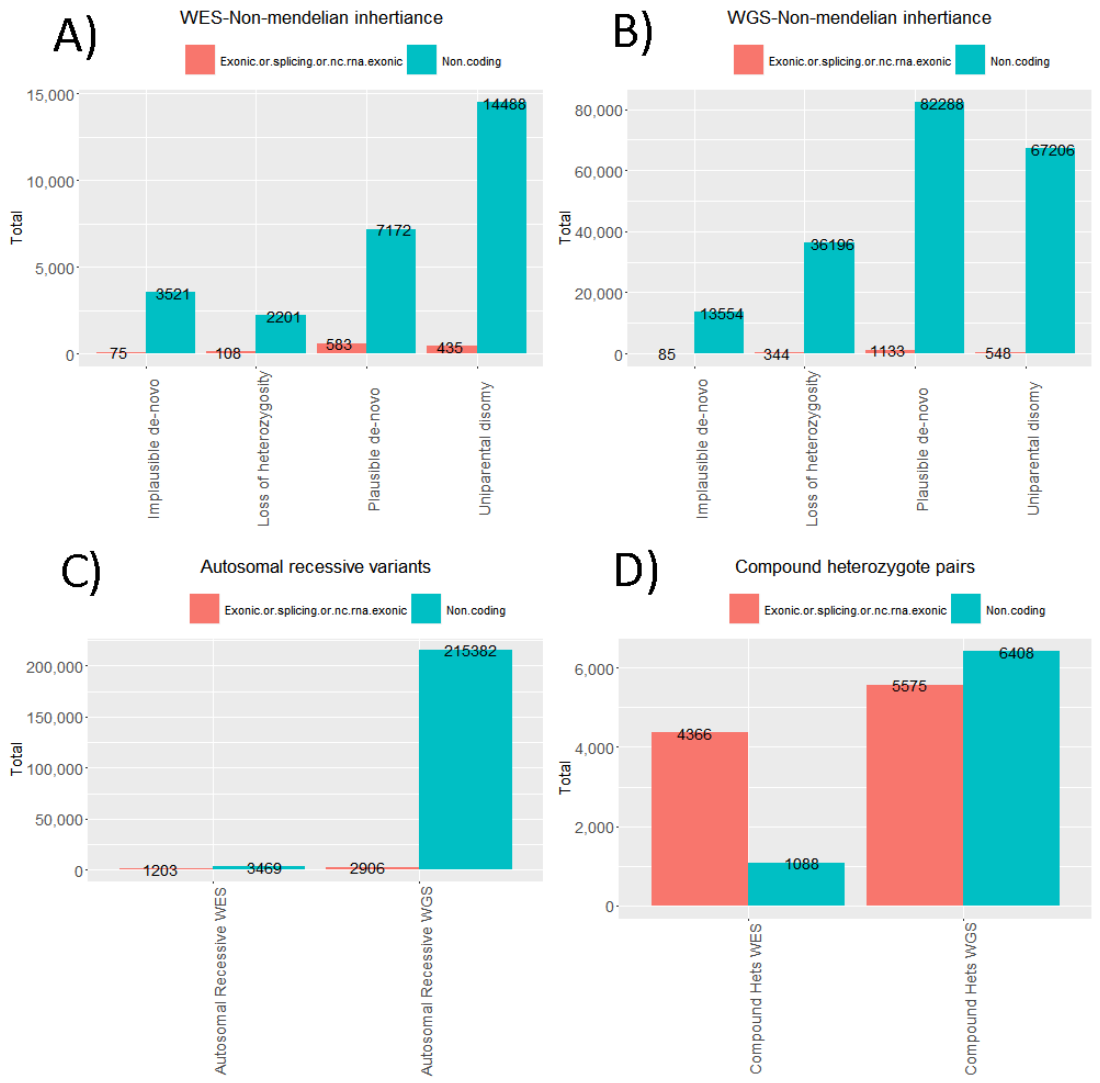

Figure 4.17: GEMINI variant call totals by tool. **A)** Non-Mendelian variants for WES split by the four variants types detected by the tool. Most variants detected were non-coding and uniparental-disomy events. **B)** Non-Mendelian variants for WGS split by the four variants types detected by the tool. As with Panel A most of the variants detected were non-coding. The number of coding implausible-*de novo* variants were similar at 75 in WES and 85 in WGS. Around three times as many coding LOH events are detected in WGS compared to WES at 344 compared to 108. Double the number of coding plausible-*de novo* variants were called in WGS compared to WES at 1,133. 113 additional coding uniparental disomy events are called in the WGS data giving a total of 548. **C)** Autosomal recessive variants detected in WES were 1,203 coding and 3,469 non-coding. WGS had 2,906 coding variants and 215,392 non-coding variants. **D)** Compound heterozygote variant pairs for WES total 4,366 which have both variants in the pair coding compared to 5,575 coding pairs in the genome. Only 1,088 variant pairs contain a non-coding variant in the pair, WGS however had 6,408 pairs with a non-coding variant, likely due to regions missed in WES such as UTRs, splice regions and introns.

coding variants show a much closer pattern with a similar number of implausible *de-novo* variants called at 75 in WES and 85 in WGS and for uniparental-disomy 435 compared with 548 in WGS. A larger discrepancy can be seen when comparing the totals of coding - loss of heterozygosity variants with 108 in WES compared with 344 in WGS. The largest discrepancy in the number of coding variants is in the number of plausible *de-novo* variants with one allele change relative to parents which has only 583 variants identified compared to 1,133 in WGS. The largest difference in non-coding variant is also for plausible *de-novo* violations with around 11.5 times more variants called in the WGS data.

The second GEMINI analysis mode used was to check for autosomal recessive variants, shown in Panel C of Figure 4.17, where the affected child is homozygous alternate while both unaffected parents are heterozygous for the same site. An additional 1,703 coding variants are called in WGS data and approximately 62 times more non-coding variants. The final analysis module used was for compounded heterozygote mutations, shown in Panel D, where multiple inactivating heterozygote mutations are present in the same gene leading to loss of function. In WES 4,366 coding variant pairs were identified compared to 5,575 coding pairs in WGS. Only 1,088 variant pairs contain a non-coding variant in WES while WGS had 6,408 pairs with a non-coding variant. There was smaller variance in the number of coding variants with only an additional 1,209 variants called and approximately 5.8 times more intronic variants in WGS compared to WES.

4.4.5 Structural & copy number variants

Large scale variants were called using CNVkit for both WES and WGS samples and with LUMPY for WGS data only. CNVkit can detect alterations in copy number while LUMPY aims to detect structural variants such as deletions, inversions and insertions.

LUMPY

LUMPY was only built to analyse whole genome data so only WGS data was used. After joining the individual genotyped VCF files data from each of the three files were merged into a single file with genotypes. From the single file 108,250 variant positions were identified with genotypes for all three samples. For each sample the total number of variants identified with non-reference genotypes were consistent at 56,000 - 58,000 per sample as shown in Table 4.7.

Sample	./.	0/0	0/1	1/1	Total (0/1, 1/1)
SD001	906	50,469	33,856	23,019	56,875
SD002	956	48,410	36,361	22,523	58,884
SD003	904	49,665	35,141	22,540	57,681

Table 4.7: For each of the samples in the trio joint genotyping was performed, providing genotypes for all 108,250 variants identified between the three samples. Only 0/1 (Heterozygous) and 1/1 (homozygous alternate) indicate the presence of a variant. Each of the samples contained a similar number of variants between 56,875 and 58,884.

Examination of the types of structural variant per sample as shown below in Figure 4.18 reveals that the most commonly detected events were deletions. In each WGS sample there were approximately 50,000 deletions detected compared to the next most commonly detected event of breakends at around 7,000 per sample (3,500 pairs). Between 1,135 and 1,237 duplications event were also detected per sample. The least detected events were inversions where the orientation of a section of DNA is inverted. Patterns are remarkably similar between samples in all categories, particularly between SD002 and SD003.

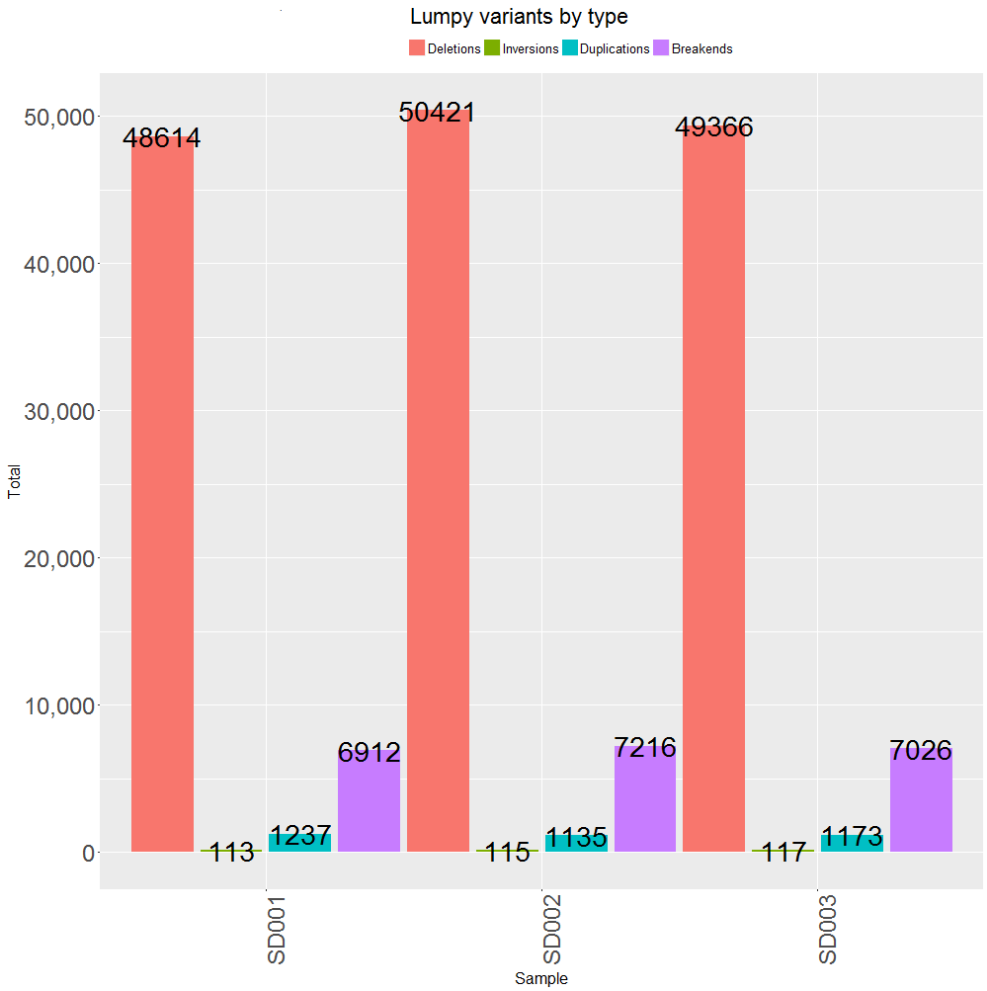

Figure 4.18: LUMPY variants called per WGS sample divided by variant type. For all categories the total number of variants called are very consistent across the three samples. The most commonly detected variant type was deletions with approximately 50,000 variants detected per sample. The second most prevalent type was breakends which describe intra or cross-chromosomal events such as translocations at around 7,000 per sample (3,500 pairs). Between 1,135 and 1,237 duplication events were also detected per sample. Relatively few inversions were detected with around 115 detected per sample.

Further investigations looked at the distribution of variants across chromosomes. The number of deletions detected were shown to be related to the the length of the chromosome. The number of deletions per chromosome peaks at around 2,100 for all three samples. For breakend events the number reported per chromosome does not seem to be related to the chromosome length and variants per chromosome were more evenly distributed at between 150 to 400 per chromosome. Duplications were also shared evenly across chromosomes though the total of the variants are lower. To adjust for the length of chromosomes the number of variants per megabase are shown

in Figure 4.19. The number of deletions, duplications and inversions per megabase were generally similar. Though the number of deletions per megabase is lower for chromosomes 9, 14, 15 and 16 across all three samples. The number of breakend events per megabase were also higher for chromosomes 20, 21 and 22 in all samples.

Figure 4.19: LUMPY variants called from WGS samples split by type and plotted per MB for each chromosome. The number of deletions, duplications and inversions per megabase are generally similar. Though the number of deletions per megabase is lower for chromosomes 9, 14, 15 and 16 across all three samples. The number of breakend events per megabase is also higher for chromosomes 20, 21 and 22 in all samples. Due to the sparsity of inversions few were called and are depicted in more detail in Figure 4.20

Due to the low number of inversions the per chromosome totals were plotted below in Figure 4.20. Between all samples inversions are most often called on chromosomes 1, 5, 7, 10, 11, 12, 17 and in SD002 and SD003 also on chromosome 16. The maximum number of inversion variants called per sample on a chromosome is 14 on chromosome 7 for sample SD001.

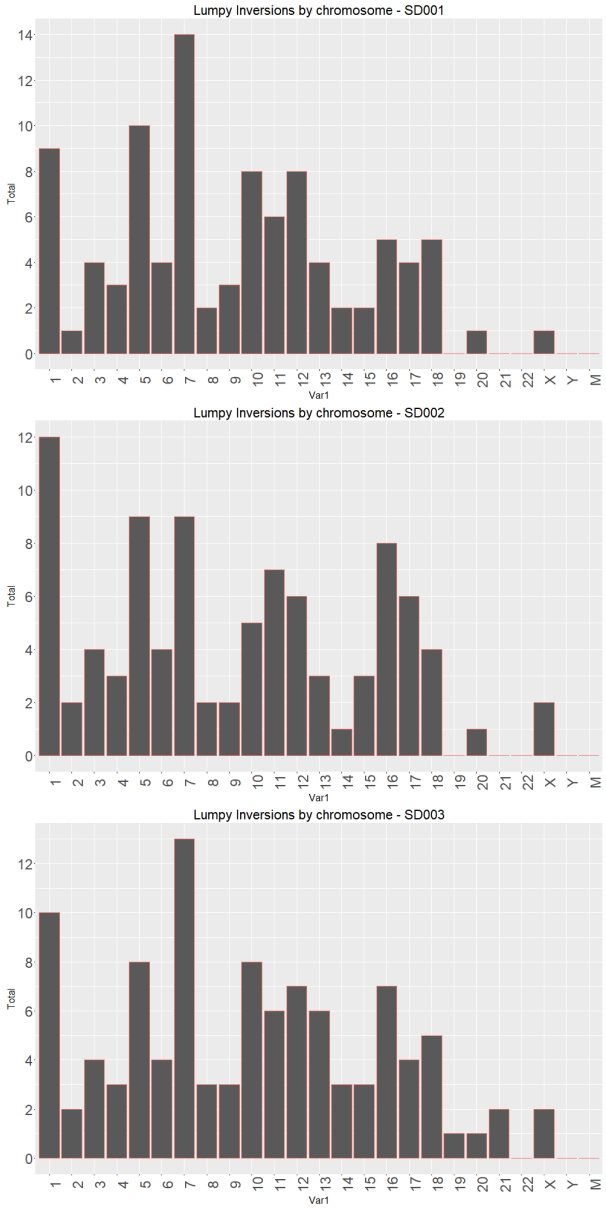

Figure 4.20: Total inversions called per chromosome by LUMPY for each sample. In all three samples most inversions are on chromosomes 1, 5, 7, 10, 11, 12, 17. Some similarities can be seen between SD001 or SD002 with SD003 and so are likely inherited inversions.

CNVkit

For CNVkit the total segment calls per sample are shown in Figure 4.21. Comparing the totals of gain, loss and neutral segments per sample was performed with exclusion of chromosome Y for female samples SD002, SD003. Segment calls in WES samples were consistent at 49 or 50 neutral segments per sample. The number of gains in WES samples were low with only 4, 3, 5 gains called in SD001, SD002 and SD003 respectively, a similar pattern with losses was also observed at 9, 7 and 12 per sample.

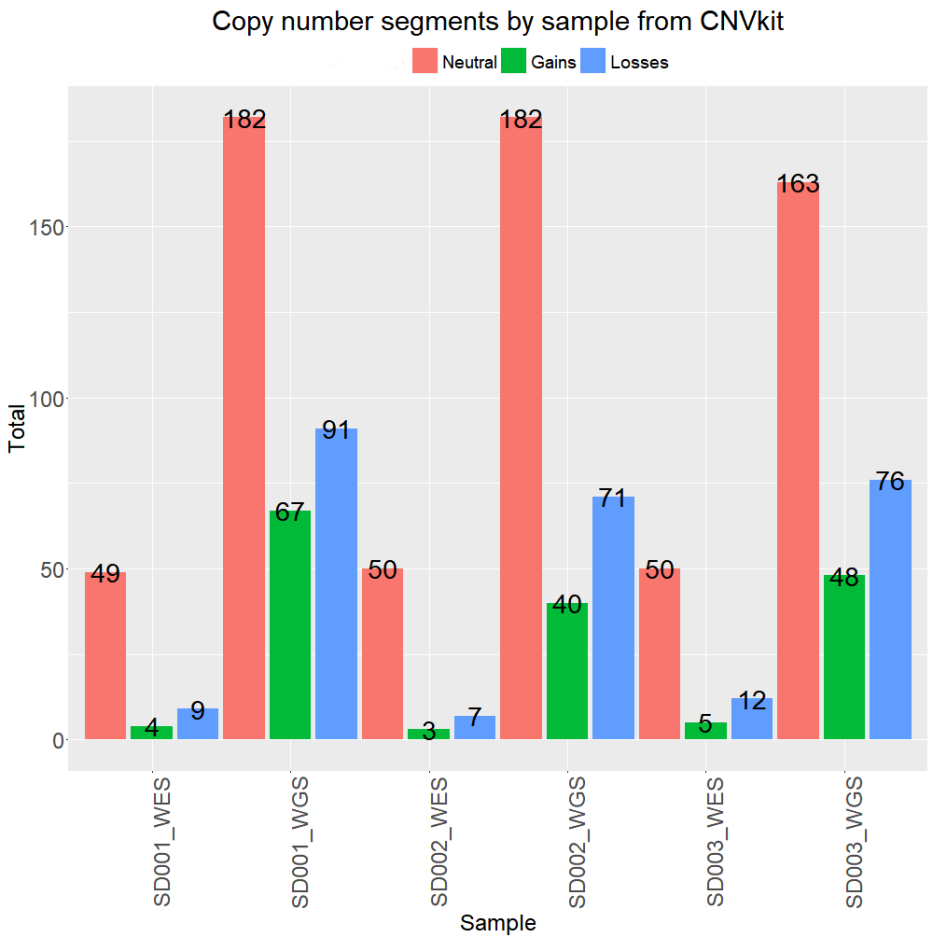

Figure 4.21: Total segment calls per sample divided by type with neutral segments of copy number 2, losses below 2 and gains above 2. In WES samples there were either 49 or 50 neutral segments per sample, compared with 182 neutral segments for WGS SD001, SD002. Fewer neutral segments were comparatively called in SD003 at 163. In WES samples relatively few gains or losses were identified with 4, 3, 5 gains called in SD001, SD002, SD003 respectively while in WGS 67, 40 and 48 were identified. For loss of copy segments 9, 7 and 12 were called in WES contrasted with 91, 71 and 76 in WGS.

WGS samples also displayed a majority of segments called as neutral with 182 in both SD001, SD002 and 163 in SD003. Despite the small reduction in the number of

neutral segment calls there were not more loss and gain segments in SD003 than SD001. SD001 had 20 and 15 more loss segments than SD002 or SD003 respectively. Also SD001 had 27, 19 more gains again than SD002, SD003 respectively. Dividing segment calls into gains, losses and neutral segments per chromosome shows that there were 18 losses and 14 gains of copy segments calls on chromosome Y for the male sample SD001. It is probable that calling of chromosome Y using a pooled reference of a single male and female has led to biases in segmental calling only for chromosome Y.

Examination of the distribution of variants per chromosome is shown in Figure 4.22. For all samples the number of segment calls does not correlate with length of the chromosomes. There were relatively high numbers of neutral segments on chromosomes 1, 15, 16, 17, 2, 5 and 9. This disagrees with patterns in WES where most chromosomes were neutral with only 2 segment calls except on chromosome 7. In WGS data chromosome 9 has a high number of gains called at 17.

For SD002 in WGS data there were a high number of neutral segment calls on chromosomes: 1, 15, 19, 20, 22 and 9. Gains and losses were highest on chromosomes 14, 15 and 9 with all other chromosomes called with fewer than 5. In WES data neutral segments were evenly distributed with two neutral calls per chromosome for all bar 11, 17, 21 and 7. Few gain or loss segments were called in WES with the events that were called distributed as one per chromosome on 11, 15, 17, 21, 7 and only 2 segments called on chromosome X.

Sample SD003 for the WGS data shows chromosome 9 had the most neutral segments and one of the highest counts of gains or losses. Chromosomes 1, 15 and 3 also have a high number of neutral segments with chromosome 15 also reporting the highest number of losses in both WGS and WES.

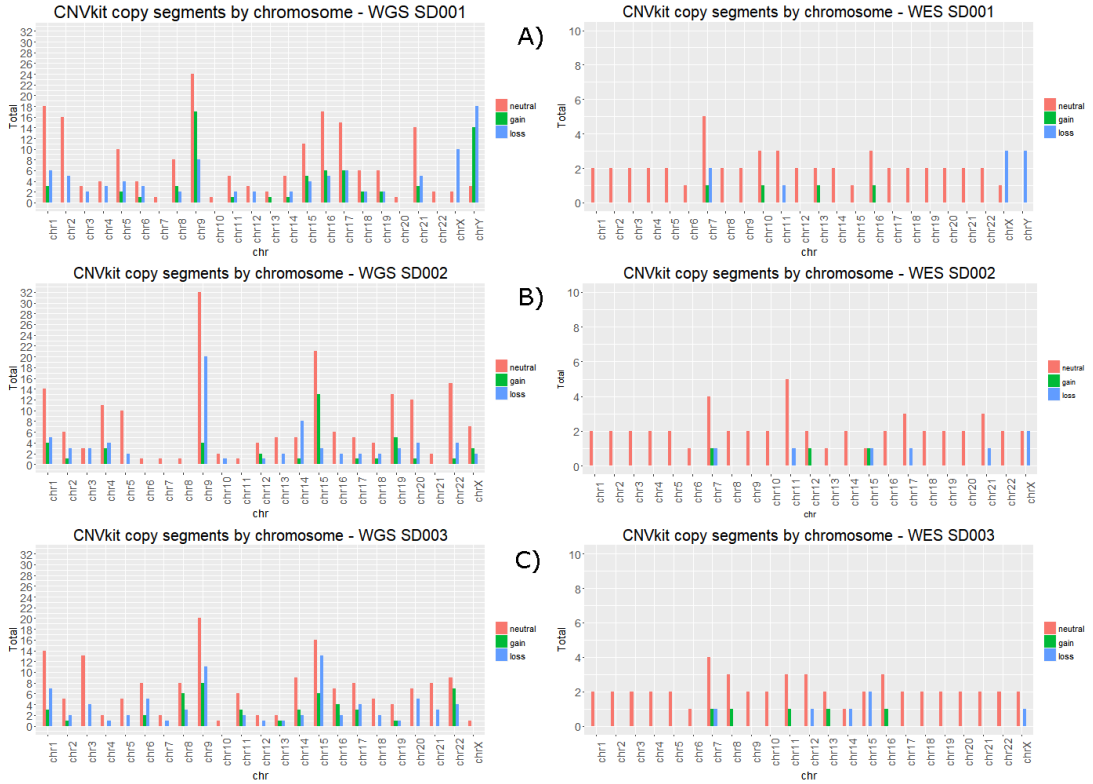

Figure 4.22: CNVkit segment calls split by chromosome and type per sample with WGS samples on the left and WES on the right. **A)** For SD001 in WGS data there were relatively high numbers of neutral segments on chromosomes: 1, 15, 16, 17, 2, 5 and 9. From WGS data chromosome 9 displays a high number of gains totalling 17. Gains and losses were then more evenly distributed across the other chromosomes at below 6. WES data called few CNVs other than neutral segments except on chromosome 7. **B)** For SD002 in WGS data there were a high number of neutral segment calls on chromosomes: 1, 15, 19, 20, 22 and 9. Gains and losses were highest on chromosome 14, 15 and 9 with all other chromosomes called with fewer than 5. In WES few gain or loss segments were called. **C)** SD003 has the most similar number of loss, gain and neutral segments per chromosome. In all the WGS samples over the autosome chromosome 9 consistently called the most neutral segments and one of the highest number of gains or losses. Chromosomes 1, 15 and 3 also had a high number of neutral segments and chromosome 15 also has the highest number of losses in WGS and WES.

Between the WES and WGS calls from CNVkit variants were compared to see if all the WES variants were also described in WGS. A total of 11 WES non-neutral CNVs were also identified in WGS, three examples of which are shown in Figure 4.23. One CNV is shown from each of the samples with the WGS CNV plotted on the left and WES on the right. Generally the CNVs show good agreement and with multiple points used to make segment calls as shown by the grey dots in each plot.

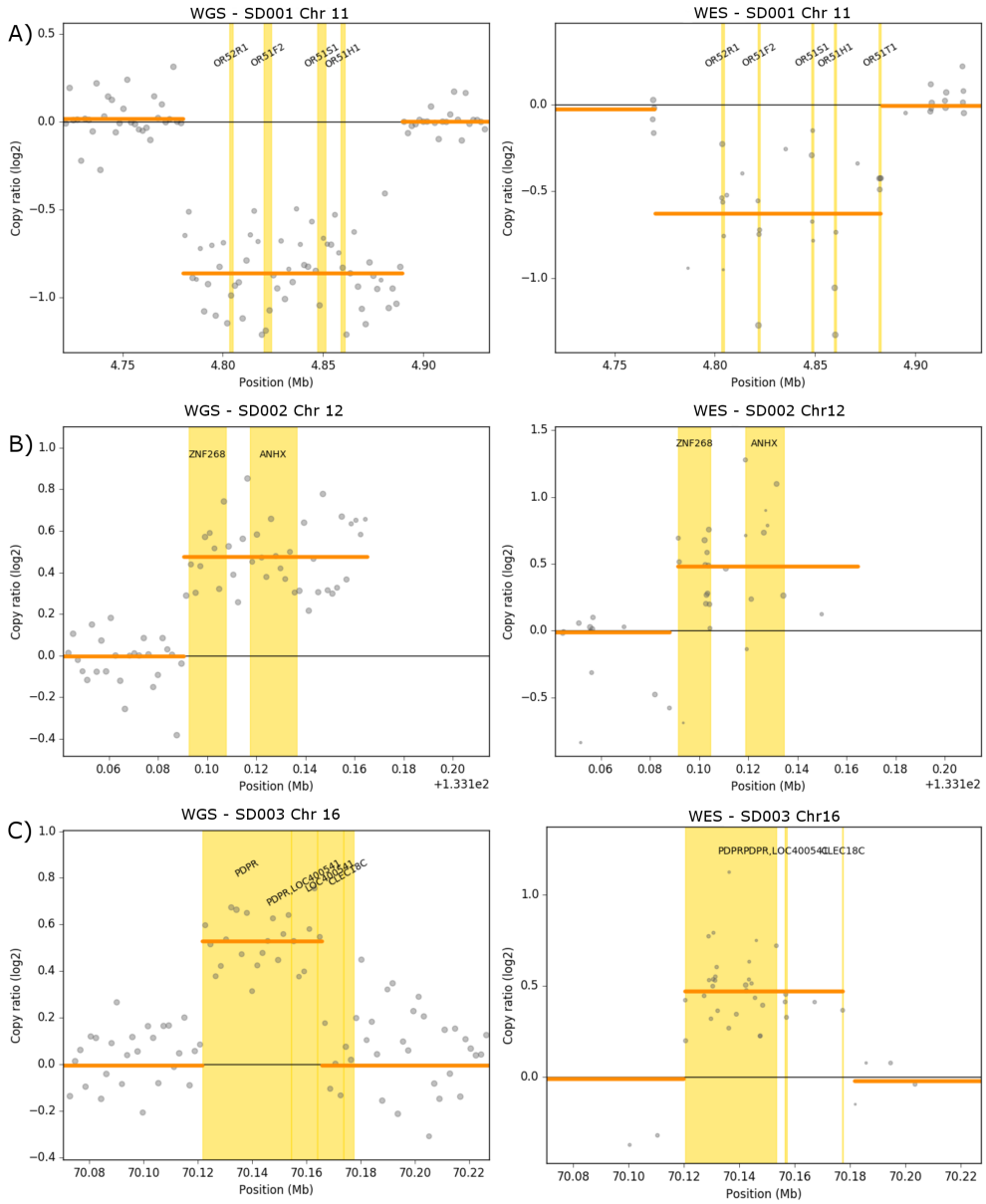

Figure 4.23: CNVkit agreeing CNVs between WGS (Left) and WES (Right). All three CNVs show good agreement between WGS and WES. There are more copy number ratio points called in WGS than in WES samples.

Whilst approximately one third of the non-neutral CNVs called agreed between WES and WGS the other two thirds did not. Sampling three of these disagreements as shown in Figure 4.24 shows that for Panels A and B there are a lack of grey points over the regions where the CNVs are called by WES. Panel C shows in the region of the deletion called in WES grey points with negative copy ratios suggesting evidence for a deletion in WGS.

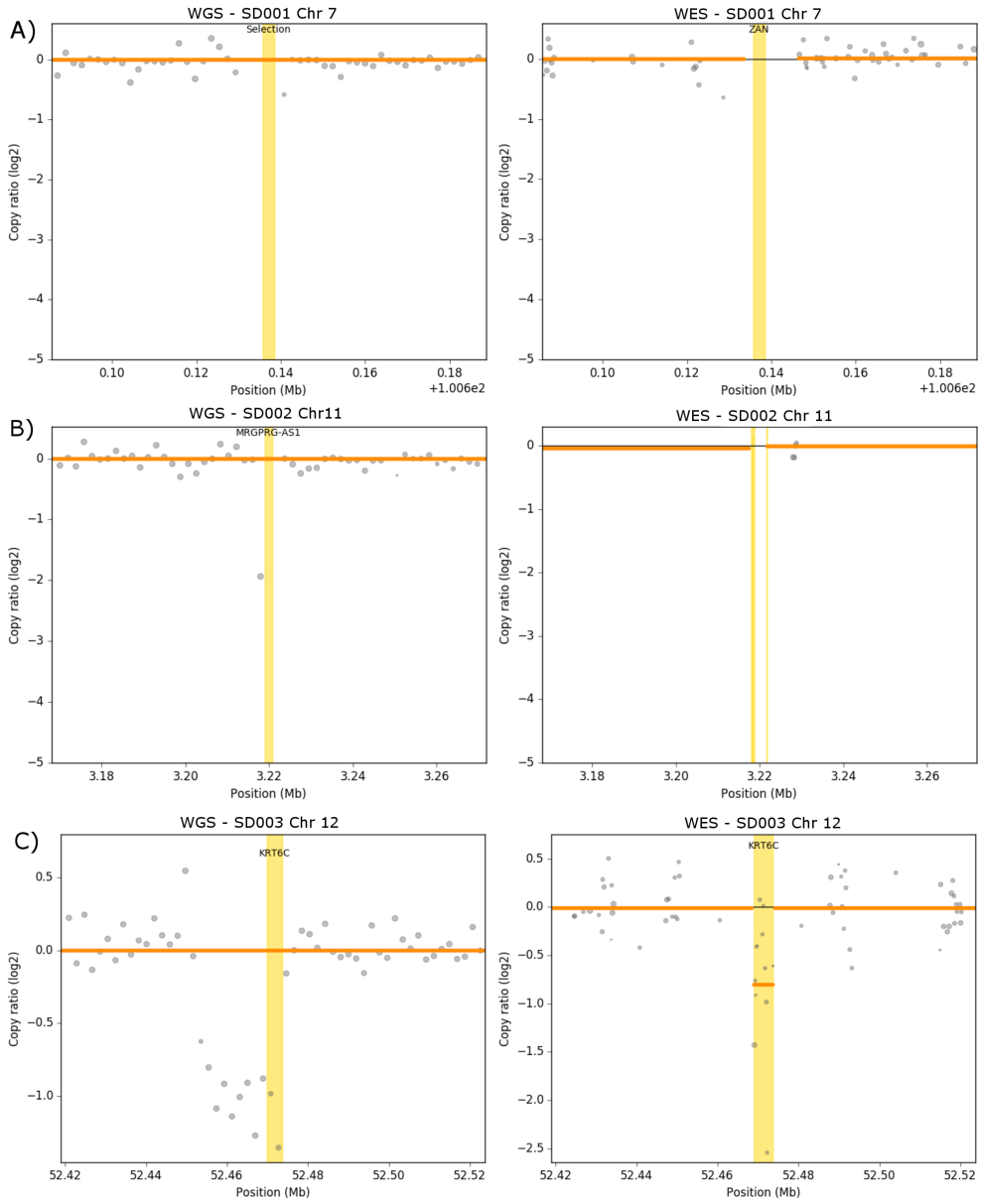

Figure 4.24: CNVkit disagreeing CNVs between WGS (Left) and WES (Right). Panels A and B show a lack of grey copy ratio points over the regions where the CNVs are called by WES. Panel C shows in the region of the deletion called in WES grey points with negative copy ratios suggesting evidence for a deletion in WGS.

4.4.6 Comparing whole exome with whole genome sequencing

Previously in Section 4.4.2 the total number of variants called per sample were investigated for WES and WGS samples. As the genome theoretically covers all non-coding regions the files are much larger, capturing more coding and many times more non-coding variants. An initial comparison was performed comparing how many of the variants which were called in the WES were also called in the WGS are shown in Figure 4.25.

For samples SD001, SD002 and SD003: 91.7, 92.1 and 92.4% of WES variants respectively were found in WGS for the same sample. 8-9% of variants were only called in WES samples.

Investigations of the 8-9% of WES only variants were performed to identify if the variants were of low depth or B-allele fraction and hence potentially false positive calls. A variant depth filter of 10 or above was applied which reduced the number of exome only variants from 20,854, 18,005 and 24,354 down to a total of 3,870 , 3,157 and 5,321 (down by 81.4%, 82.5% and 78.15%) for SD001, SD002 and SD003 respectively.

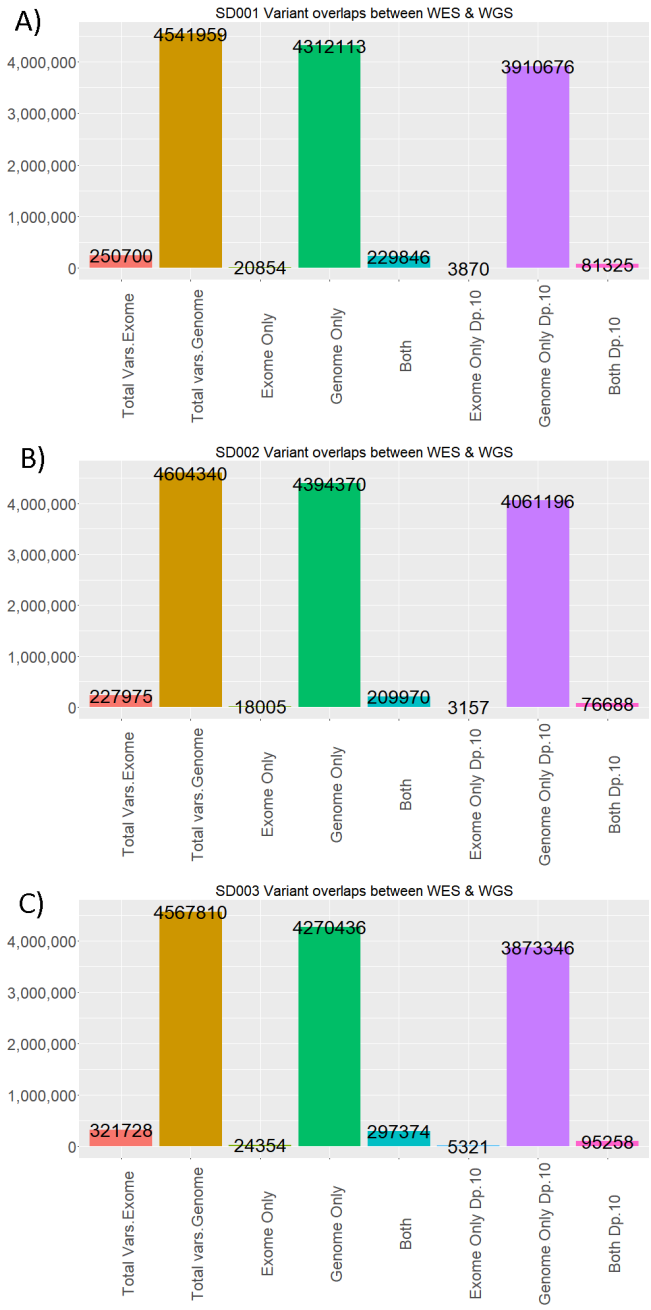

Figure 4.25: Overlap of variants called in WES (Exome) and WGS (Genome) samples SD001, SD002 and SD003. In all samples around 4.5-4.6 million variants were identified from WGS compared to between 227,000 to 321,000 in WES. **A)** SD001 had 20,854 variants only in WES (8.3%). 229,846 variants are found in both WES and WGS (91.7%). Application of a depth filter to remove variants with depth below 10 reduces the variants to 3,870 only in WES, removing 81.4% of unique WES variants. **B)** SD002 had 18,005 WES only variants (7.9%) and 209,970 (92.1%) variants in both WES and WGS. Application of the depth filter above 10 reduces WES only variants to 3,157 , excluding 82.5% of unique variants. **C)** SD003 had 24,354 WES only variants (7.56%) and 297,374 variants shared by WES and WGS (92.4%). Application of the depth filter above 10 reduces WES only variants to 5,321, excluding 78.15 % of unique variants.

A subsequent filter was applied to the WES only variants to remove variants called with a minor read ratio or B-allele frequency below 0.2. This is to remove variants with depths above 10 but also few non-reference alleles and more likely to be a false positive. Exome only variants for samples SD001, SD002 and SD003 were then reduced down to: 2,925 , 2,360 and 4,100 respectively. This makes the proportion of WES only variants of "high quality" between 1.0 - 1.3% of the total of called exome variants.

For the remaining "high quality" variants unique to WES samples annotation was performed to identify the types of variants. Results showed that an average of 82.5% of the variants remaining are non-coding encompassing either intronic, intergenic, down/upstream or UTR variants. That left 502, 446 and 672 coding variants in samples SD001, SD002 and SD003 respectively. A filter was applied to coding variants to identify if they were in genes which are commonly reported to contain false positives and are highly mutable[197]. Between 56 - 64% of remaining variants were identified in genes commonly containing false positives as shown below in Figure 4.26.

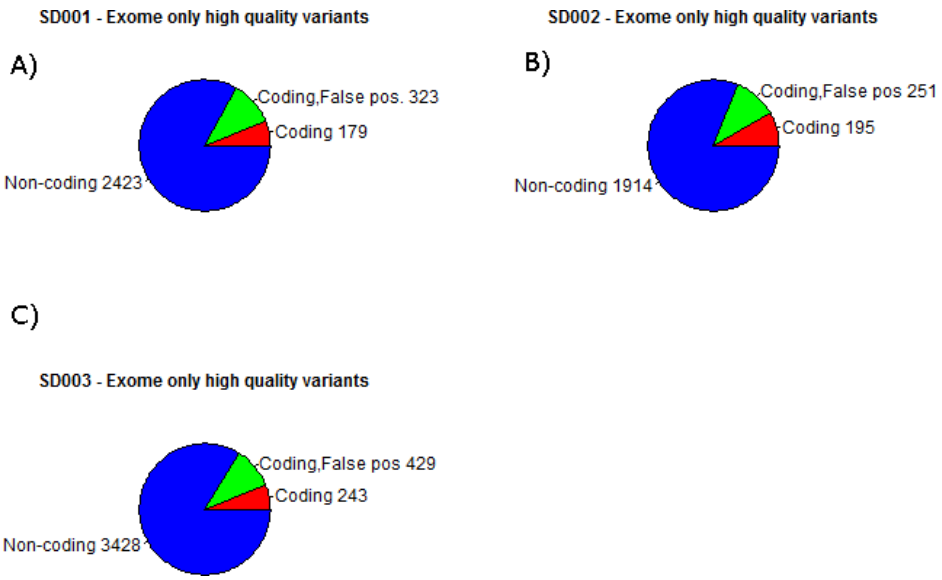

Figure 4.26: Classification of remaining high-quality Exome (WES) only variants per sample. **A)** SD001 returns 2,423 variants (82.8%) as non-coding with a total of 502 coding variants of which 179 are in genes not suspected to be false positives. **B)** SD002 reported 1914 variants (81.1%) as non-coding. 446 coding variants were called of which 251 of the variants are in genes commonly containing false positive variants and 195 coding variants in other genes. **C)** SD003 reported 3,428 non-coding variants (83.6%) with a total of 672 coding variants. 430 of the coding variants were in suspected false positive containing genes and 243 coding variants in other genes.

Aggregation of the coding variants called only in WES samples gave a total of 1,620 variants spread across 709 genes. Only 173 of the genes were reported in more than one sample containing a variant. Functional clustering was performed on the 173 genes to try and identify gene shared functions using DAVID[205]. Three of the genes were not recognised, all of which were "LOC" genes which are currently uncharacterised genes. Using InterPro domain data this covered the highest percentage of genes at 81.8% of the queried genes. Also as exonic coding genes were being investigated clustering using protein domains was deemed most appropriate. Five clusters were generated using the medium classification stringency from DAVID (v5.8) as summarised in Table 4.8 below which are most associated with immune related genes or transcription factors. Cluster-1 is associated with MHC class 1 and 2 molecules, both of which have the primary functions of antigen recognition, with variable recognition domains which are more likely to contain the mutations. Cluster-2 is also immune related also and associated with immunoglobulins involved also in antigen recognition. Cluster-3 groups a number of domains, von-Willebrand type-D domain is also found in many plasma proteins such as collagens and complement factors. Cysteine rich domains are found in commonly mutated proteins such as von-Willebrand factor, Zonadhesin and Mucins which regularly appearing as false positives in WES. Cluster-4 contains Epidermal growth factor like domains associated with tyrosine-kinase pathways. Cluster-5 contains zinc fingers which are common transcription factors and are often challenging to assign functional importance to owing to their involvement in multiple pathways.

Cluster	Enrichment Score	Interpro Term 1	Term 2	Term 3
1	4.62	MHC class I	MHC class II	Ig C1-set
2	4.41	Immunoglobulin	Alpha-1B-glycoprotein	
3	3.78	von Willebrand factor , type D domain	Trypsin Inhibitor-like	cysteine rich domain
4	1.85	Epidermal growth factor like		
5	0.31	Zinc finger, C2H2		

Table 4.8: Clustering of genes with coding variants in WES only. 5 clusters were identified, Cluster 1 is associated with MHC molecule. Cluster 2 is immune related also and associated with immunoglobulins. Cluster 3 groups a number of domains, primarily von Willebrand type D domain. Cluster 4 contains Epidermal growth factor like domains associated with tyrosine-kinase pathways. Cluster 5 contains zinc fingers which are common transcription factors.

Comparing coverage between whole exome and genome sequencing for exome targets

To directly compare WGS against WES, WGS data coverage calculations were restricted to the regions covered by the Agilent SureSelect human all exon V5 bed file. By restricting the calculations for WGS in effect this created an exome from WGS. Previous plots of coverage for genome sequencing were based upon genome wide intervals as shown in Figures 4.6 and 4.7 so the fraction of WGS samples at each coverage depth were calculated restricted to exome co-ordinates and plotted in Figure 4.27.

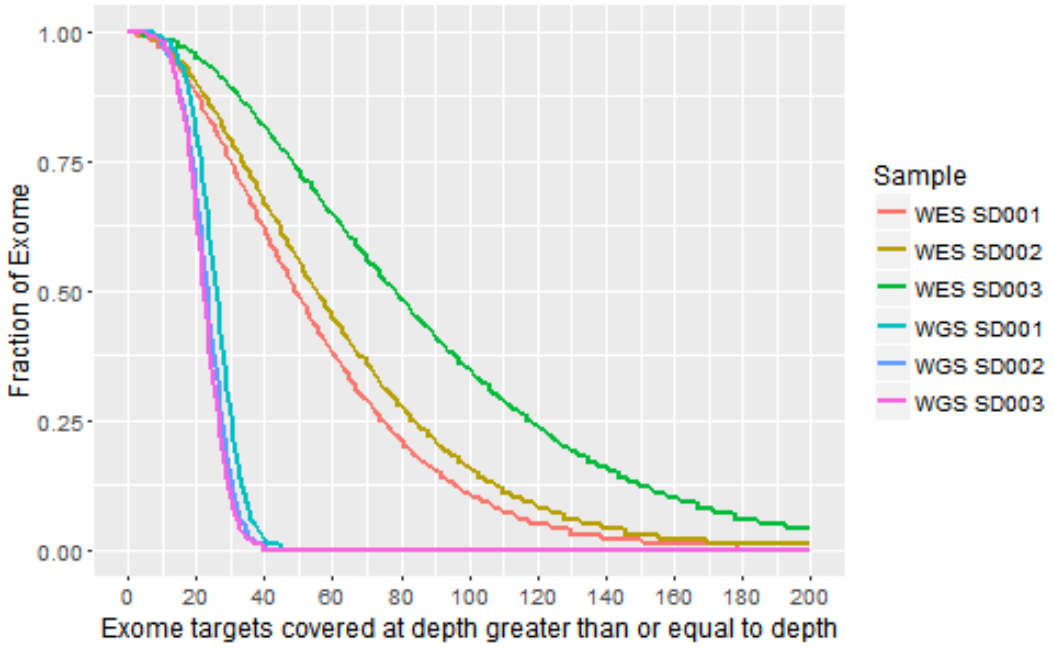

Figure 4.27: Fraction of exome covered at depths for WES and WGS. At higher depths of coverage WES differentiates from WGS samples which were only sequenced for a mean target of 30x. At lower coverage depths WGS samples have a higher fraction of the exome covered than WES samples

WES samples were covered at a higher depth than WGS samples, hence at higher depths of coverage WES and WGS samples begin to differentiate. However at lower depths below as shown in the top left of Figure 4.27 WGS samples cover a higher fraction of the exome than WES samples as shown in Figure 4.28 in more detail. Below 12x depth of coverage a greater fraction of the exome is actually covered by WGS than WES for SD001, SD002 and around 8x for SD003 due to the extra sequencing performed for SD003.

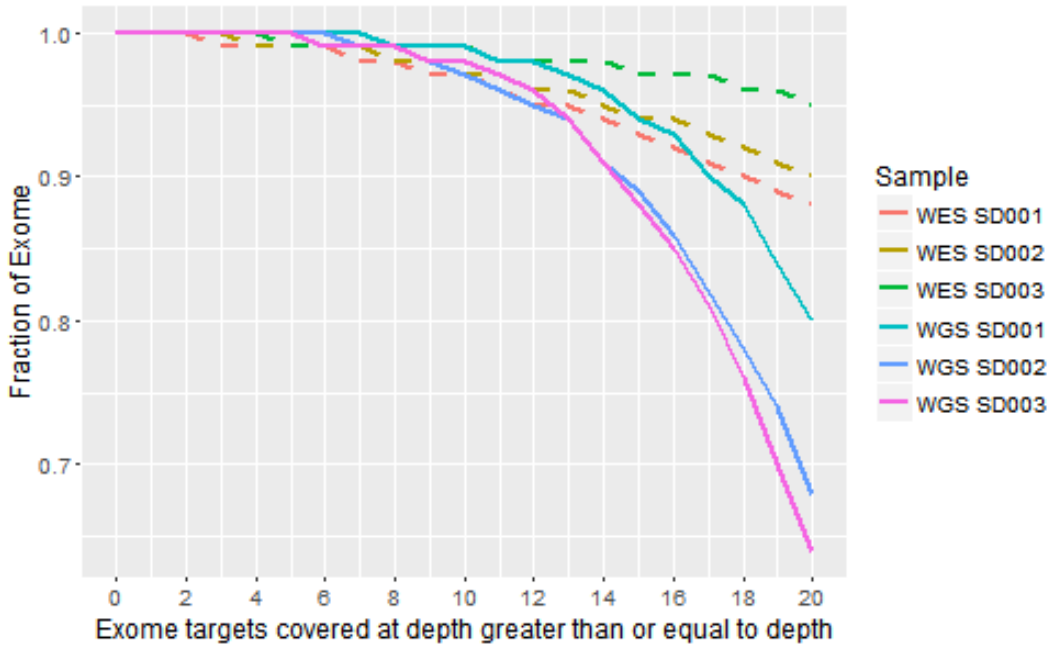

Figure 4.28: Fraction of the exome covered at depths for WES and WGS from 1 to 20x depth of coverage. WGS samples SD001 and SD002 show a higher fraction of coverage up to 12x depth of coverage. Despite the extra sequencing for the SD003 WES sample the WGS had a higher fraction covered up to 8x depth of coverage.

Also using the exome created from WGS samples the fraction of ACMG genes covered at a minimum of 10x depth of coverage was calculated as shown in Figure 4.29. As previously described in Section 4.1.4 the ACMG genes are a list of highly actionable genes such as *BRCA1* and *BRCA2* for which variants are strongly recommended to be reported back to clinicians. For the three WES samples they are covered at a higher median depth of coverage than the WGS boxplots however the WES samples have a larger inter-quartile range and more outlying genes with lower depths of coverage indicating the coverage is more variable than WGS.

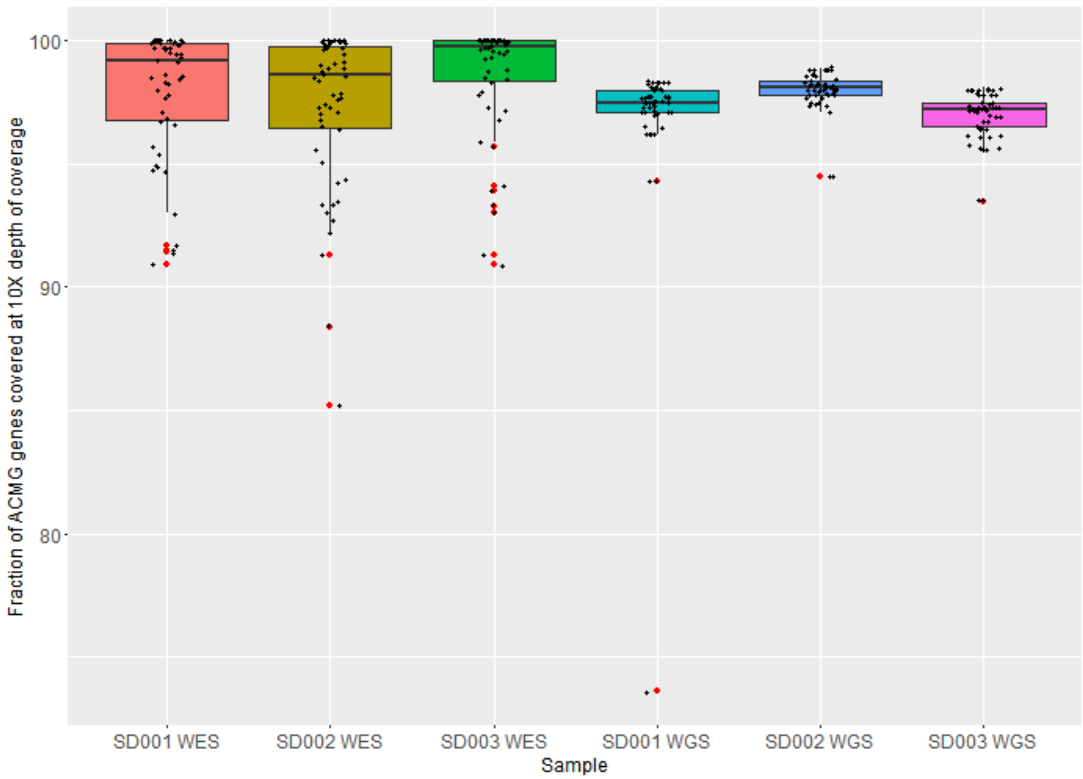

Figure 4.29: Coverage for the 55 ACMG gene for WES and WGS samples showing that genome samples are more consistent and have fewer outliers than WES. In WGS outliers below the distributions are *TP53* and *BRCA1*.

For WGS samples there are two genes which are identified in red as outliers for each of the samples which are *TP53* and *BRCA1* for all samples. To check the coverage for these genes they were visualised which confirmed the regions of below 10x depth are still covered as shown in Figures 4.30 and 4.31. SD001 also has a gene *GLA* covered at 10x across 73% which has some regions which are covered at lower depth.

Figure 4.30: WGS sample coverage for *BRCA1* on a scale of 0-50x depth of coverage. Coverage is generally good and evenly spread with only a few regions of no coverage in intronic positions such as around 43,100kb.

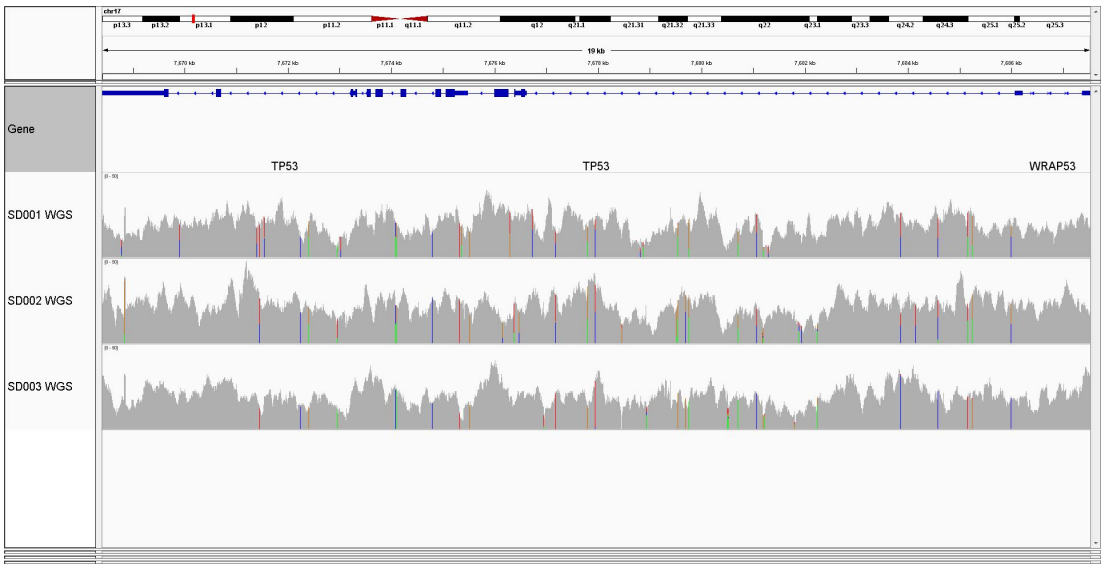

Figure 4.31: WGS sample coverage for *TP53* on a scale 0-50x depth of coverage. Coverage is generally good and evenly spread.

4.5 Discussion

This chapter aimed to determine the trade-off between performing targeted or WES compared to WGS given the convergence in sequencing costs. The obvious benefit of targeted approaches are the reductions in cost which can be substantial when multiplied for large cohorts. As the costs of NGS decreases the price difference between WES and WGS will narrow, leaving the question of what are the additional benefits of WGS and how does it currently compare against WES for the same samples.

By using WGS this future-proofs a dataset as all accessible regions will have been sequenced and would only require re-annotation to prioritise whereas in WES potentially variants of interest could be missing. The study by Lionel et al.[187] highlighted this exact pitfall where 103 patients were WES or target sequenced at high depths of average coverage upwards of 100x and compared with WGS at around 40x depth of coverage. The WGS samples yielded an additional 18 diagnoses compared to WES and targeted methods as they often did not include the genes in the target design and or variants were deep intronic so regardless of depth a WES based study would never be able to diagnose these patients[187].

WES sample SD003 initially failed to reach a specified minimum number of reads and required additional sequencing. Hence there was an additional lane of data compared to other WES samples. QC of WES and WGS samples showed that the mean quality of reads was above 30 for most of the fastq files however at positions upwards of 130 bases WGS fastq files were observed to dip down to a PHRED quality of 28. This is characteristic of Illumina sequencing due to imaging of fluorescence becoming technically more difficult as sequencing proceeds and clusters losing sync[4,7]. A mean phred score of 28 is still good as the error rate should be close to 0.2%. Adapters were also identified in WGS samples for SD001 and SD003 at levels of 0.3% per lane or less. However as the adapters were such a small proportion of the overall sequencing it was decided it was unnecessary to remove them. In GATK variant calling for positions with low base quality reads are not used and are by default ignored, removing potential issues of low phred quality reads. GATK also suggests that

adapters when present at low concentrations are not of importance and corrected by the program as recommended by developers[37,38].

Depth of coverage calculations showed the WES samples have a right skewed distribution. Samples had varying percentages of bases targeted at 50x depth of coverage from 55% and 48% for SD001, SD002 up to 72% for SD003. By comparison WGS coverage per sample displayed a more normally distributed range of coverages around 23-26x depth. WGS depth of coverage is typically lower as it was sequenced to a lower depth but should contain less bias than WES as it has no capture step. Capture steps introduce biases as GC-rich regions tend to be less well captured and sequenced[189]. Exons in close proximity will have overlapping targets leading to elevated coverage over the regions. Coverage will also be normally distributed over target regions in WES with the highest coverage in the middle of a target region[189]. WES samples had 75% of samples covered at 30x in SD001 and SD002 but for SD003 approximately 89% was covered at 30x for the defined target regions. It has been shown that in order to correctly call genotypes of variants upwards of 30x depth is required[107–109] suggesting that the WGS samples should be sequenced to a higher mean depth of coverage than 30 to ensure that variant genotypes are correctly called.

Checks of sample identity were performed by comparing the genotyping with WES and WGS results which confirmed the sample identity for all samples. A single mismatch was reported though for the WES sample SD003 was found and is believed to be caused by low depth. The variant rs2298628 was called with 11 alleles but only three were alternate from visualisation of the read pile-up in theory giving an AAF of 0.27, which could be called a heterozygote depending upon the threshold used by the variant caller. However none of the alternate alleles were counted by GATK HC. Alleles were not counted in variant calling post-recalibration if they fail the filtering thresholds for mapping, base quality or duplicates and so were excluded[47,188].

Analysis of the variant calls for WES and WGS samples highlight the number of additional variants called per sample with 220,000 to 320,000 in WES compared with 4.5 - 4.6 million for WGS. The majority of the additional variants called in WGS were SNPs, with between 3.6 and 3.7 million more SNPs called genome wide compared to

just in WES. Previous studies of WGS have also found a slightly lower number of SNPs per WGS sample at around 3.3 million when averaging over 40 genomes at 50x depth sequencing but using the hg19 reference genome[206].

Comparisons for each sample between WES and WGS for coding variants also consistently identified more variants in WGS. Whilst more synonymous and non-synonymous calls are expected in WGS, the number of stop and start codon loss or gain variants is concerning. Both of these variant types will have large consequences and most likely to alter function and stability of gene products. Compared to other genome studies higher variant totals per sample were obtained with an average 9,184 missense variants and 9,612 synonymous variants per genome previously found compared to the approximately 12,000 found in this study for both categories[206]. Differences in coding variants were most noticeable for splice region variants with around 2,000 further variants per sample called in WGS samples. Splice region variants will be around the intron-exon boundary and so WES kit design may bias against calling variants at the extremities of exon. WES has a normal coverage distribution over an exon, hence coverage will be lower at the boundaries[188,207]. Similar numbers of splice variants were also reported previously for WGS samples at an average of 2,296 on hg19, in the same range as the approximate 3,000 per sample found in this chapter[206]. WES kits may also not be designed or optimised to capture all of the alternate transcripts possible due to revisions of gene databases and so whilst it is not surprising the WGS calls more splice region variants the number per sample is concerning and does limit future re-analysis potential of WES[187,188]. Programs such as SPIDEX have more recently been developed to identify potential splicing variants occurring deep intronic regions that also would not be captured in WES[208].

Most of the variants called in WGS are non-coding with an average of 2.15 million intronic and 1.71 million intergenic variants, together comprising an average of 84.7% of variants. The large benefit of WGS sequencing is the capture of non-coding regions which can contain elements such as gene promoters and enhancers[24,72,209]. However to take advantage of the variants effective annotations are required, whilst FATHMM-XF[70], CADD[66] and GAVIN[68] were identified as the most promising tools

to identify possible pathogenic variants all were calculated on hg19. As variants were called on hg38 in this study this required additional fixes to lift-over either databases to hg38 or in the case of CADD where database files were approximately 200GB to lift-over the hg38 VCF to hg19. Not all variants were able to be annotated by these tools for either coding or non-coding. For coding variants only CADD annotated a limited number of more common INDELs whilst other tools do not annotate INDELs currently owing to the number of combinations possible which would make databases too large. Future improvements to annotation tools for pathogenicity could be able to perform *de novo* calculation of pathogenicity and therefore enable the annotation of all variants in a dataset along with the native support of hg38 will enable better prioritisation of both coding and non-coding variants including INDELs.

Most SNPs were able to be annotated but some were not, likely due to changes in the reference alleles being reported for the other strand between references. CADD with the application of a pathogenicity threshold of 15 for phred scaled scores and an allele frequency of 0.05 or missing from 1000genomes leaves a total of 8,735 non-coding variants. FATHMM-XF had clear filtering criteria with anything above 0.5 classified as pathogenic and above 0.96 as high confidence of being pathogenic. FATHMM-XF lifted to hg38 only annotates SNPs and managed to annotate around 3.64 million of the SNPs of which 29,820 were above 0.5. Selection of the non-coding variants with gnomAD alternate allele frequencies below 0.05 and a non-coding XF score of above or equal to 0.96 left 3,796 variants. When applying a filter for CADD phred score of 20 dividing the variants by sample there are between 711 and 732 variants in each of the samples. For FATHMM-XF using the high confidence threshold of 0.96 left between 27 and 40 variants. Both of these tools highlight the challenge of identifying causal, pathogenic variants as in the healthy individuals SD001 and SD002 there are many variants predicted pathogenic but appear to be tolerated. Similar findings have been reported in other studies which look at sequencing of healthy individuals where mutations annotated by ClinVar as pathogenic were found in exomes[210].

GEMINI integrated the genotype information to filter the variants and identify events such as *de-novo* alleles in the affected child it offers an effective method to help incorporate and filter the non-coding variants with coding variants. It has been shown

that the expected *de-novo* mutation rate between generations is estimated between $1x10^{-8}$ to $3x10^{-8}$ though the rate has been shown to vary by region, across families and by age. Therefore the expected numbers should be low, of the order of hundreds. The results from GEMINI show more *de-novo* variants are called than might be expected. This could be explained by the previously mentioned factors except for the high number of non-coding variants which could indicate errors in non-coding regions possibly due to depth as shown from coverage analysis.

Structural variants and copy number are more problematic to interpret and detect than SNPs or INDELs. Part of the problem is the variety of approaches undertaken from hidden Markov models, analysis of split and paired reads to read depths, often each with their own output format description. The introduction of the structural variant format guidelines in VCF specification version 4.2 offers one potential way variants could be uniformly reported allowing clearer interpretations. For targeted approaches the biases introduced from capture methods also bias CNV calling due to the uneven coverage which can lead to false positive and negative calls if not corrected[80,207].

LUMPY was the first structural variant caller used and also made use of trio calling the WGS samples detecting between 56,875 and 58,884 variants per sample. The majority of the variants called were deletions at between 85-90% of variants. This is in-line with studies looking at population or large datasets for CNVs or SVs which estimate about 1.9 million bases are in deletions larger than 100 bases[211]. Interestingly the number of deletions called per chromosome in samples appeared to be proportional to the length of the chromosome a trend that was not repeated for the other classes of structural variants. Visualisations of many of the variants called by LUMPY proved hard to see, while evidence used such as split reads and changes in depth which should accompany the event could not be seen suggesting many of the calls may be false positives. Other events such as duplications, breakends and inversions were much more evenly distributed across chromosomes and far fewer were called suggesting the deletion calling thresholds may be too sensitive. Several of the deletion calls are also improbably large at upwards of 10MB which are not detected by CNVkit.

CNVkit is usable on both WES and WGS data which allowed results from WES to be compared with WGS. There were more total segment calls in WGS data than WES due to the larger target region covered, most notably WGS samples reported more neutral segments than WES samples. Unlike LUMPY the program CNVkit did not call the number of deletions proportionately to the size of chromosomes instead loss segments appeared more randomly distributed. Sample SD001 calls excessive loss and deletion segments on chromosomes X and Y which may have been caused due to the reference used being pooled with both parents[80]. The rationale behind this decision was that pooled references are recommended by CNVkit and the sample where variants are most of interest is the affected child. By pooling both parents it provided a higher quality reference to call SD003. However for samples SD001 and SD002 that meant they were called against a reference averaged by themselves and a sample of the opposite sex which may have led to the excessive calling due to the imbalances in the reference.

Most of the variants called only in WES samples were hard to make a case for when visualised in IGV and appeared to be false positives possibly due to the sparsity of data points in WES samples in the regions where CNVs were called. In WGS samples deviations in copy number ratios in region identified by WES were often visible but were not called suggesting a lack of sensitivity in calling. In the use of CNVkit in Chapter 6 the bin size used to call CNVs was reduced down to 150bp from the default of around 267bp. The reason being the samples were of high coverage and so by using smaller bins there were still enough reads to group in the bins and give sufficient weight to the bins and reasonably accurately call CNVs. For the WES and WGS samples the depth of coverage is lower, so by using smaller bins there was insufficient data in the bins which led to lower weights and increased noise and problems with false positive calls . To solve this problems both samples would require additional depth of coverage to enable smaller bins and better CNV detection[80]. Other CNV and structural variant callers have been shown to give variable results and one common approach to account for this was to aggregate the calls for multiple programs and create a prioritisation system based on how many tools identified a variant[75]. However such an approach is both computationally challenging and requires

integration of multiple file formats. Programs such as SURVIVOR[212] have recently been created to try and perform some of these aggregations and comparisons. Therefore whilst structural variant calling may have improved with such tools as LUMPY[79] and CNVkit[80] issues still exist with regards to noise and prioritisation unless a target gene or region is known.

Comparisons of WES and WGS samples revealed that the vast majority of variants called in WES samples were also called in WGS samples. 91.7 - 92.4% of variants called in WES samples were also called in WGS samples. Adopting a similar approach to the study by Belkadi *et al.*[188] the 8 - 9% of WES only variants were investigated to identify why they were unique. Application of a depth filter of 10 or above reduced the variants from upwards of 20,000 to between 3,157 and 5,321. This showed that an average of 80.7% of the variants were those covered at low depth for the WES samples and likely to be false positive calls. Application of a B-allele fraction filter of 0.2 or above removed heterozygous variants with few alternate alleles and again more likely to be false positives. The filter on B-allele frequency reduced the variants down to between 2,360 to 4,100 per sample of which an average of 82.5% (2589) of variants were either intergenic or intronic which are not likely to have been targeted by the WES capture kit and were captured incidentally. A false positives gene list which are commonly seen in exome sequencing[197] and a list of highly mutable genes was applied to the remaining coding variants which showed over half of the remaining variants were in these genes. Finally this left 243 "high quality" variants unique to WES. This number was similar to the number obtained by Belkadi et al. who found 105-140 WES only SNPs per sample[188]. In their study they investigated some of the WES only variants further by attempting to validate variants by Sanger sequencing which showed 54.6% of the variants were false positive calls by WES. Also they suggested from a comparison of callers that differences were introduced between GATK Haplotype Caller and Unified Genotyper but could not identify the reason for the performance difference. The study concluded that only 26 of the WES only SNPs are likely real and missed by WGS. Most of the WES only coding variants were found in immune associated genes such as MHC or immunoglobulins which are naturally variable and the extra depth in WES samples may have helped in calling these variants.

A direct comparison was performed using WGS samples restricted to the WES target regions. This revealed that in direct comparison at lower depths of coverage up to 8-12x WGS samples actually covered a larger fraction of the targeted exome despite being sequenced genome wide at lower depths than the high depth WES restricted samples. Coverage at 10x depth was also shown to be more consistent for WGS samples over the 55 ACMG genes with only genes *BRCA1*, *TP53* consistently covered at 93 - 94%, higher than in WES. *GLA* was also covered relatively poorly in SD001 WGS which was caused by regions of low depth coverage but not regions of no coverage. This suggests the samples may have needed to be sequenced with a higher mean depth of coverage target. No additional coding ACMG gene variants were identified from the WES samples however when compared against WGS, supporting the notion WGS was just as effective.

4.6 Conclusions

Whole genome sequencing (WGS) consistently matched the performance of whole exome sequencing (WES) and calls many times more non-coding variants. A higher number of splicing variants are also identified which could be important for identification of causal genes. By performing WGS you gain all the same information as a WES sample but do not limit the sample for re-analysis by only targeting specific genes and transcripts, offering greater potential for future diagnoses. Currently the main issue preventing the large scale switch for most labs to WGS are the additional costs which will decrease with the future improvements in NGS.

Chapter 5

Prioritisation of trio called variants

5.1 Introduction

5.1.1 Using NGS to identify disease causing variants

One of the first successes of NGS used WES to identify the gene involved in the rare Mendelian disorder Miller syndrome[1]. The study included two affected siblings and both unaffected parents, all of which underwent WES with a mean coverage depth of 40x. Filtering of called variants was performed using dbSNP129 and exome data from HapMap. Variants were also filtered by type with non-synonymous, splicing and INDELs being prioritised. Models were then applied to the filtered variants for autosomal dominant and recessive inheritance which identified the gene Dihydroorotate dehydrogenase (*DHODH*) as containing causal variants[1]. This study provided a model on which familial sequencing can be performed to identify mutations present only in affected individuals and not in unaffected family members.

5.1.2 Sedaghatian-type SpondyloMetaphyseal Dysplasia

Skeletal dysplasias (osteochondrodysplasias) are an umbrella term used to describe a range of growth disorders of the bone and cartilage with symptoms varying from short stature, dwarfism to perinatal death. Over 450 different disorders fall under the skeletal dysplasia umbrella term[213]. The incidence rate of skeletal dysplasia cases is

approximately 1 in 5,000 births making the disease relatively rare[214].

Skeletal dysplasias are generally characterised by an abnormality of the skeleton, each form of skeletal dysplasia usually has a unique set of symptoms which can be used to identify a more precise form. Typically diagnosis of a form of skeletal dysplasia is performed by examination of familial histories with a set of skeletal radiographs for the affected patient. Forms of skeletal dysplasia can appear similar and not all symptoms may be presenting in an affected patient making incorrect diagnoses possible. Diagnosis difficulties can be compounded by the involvement of a number of other organs and systems including: pulmonary, neurologic, cardiac, renal and visual, all can be present with varying degrees of severity[214].

The rarity of cases of skeletal dysplasia and the number of forms presents problems with gathering large sample cohorts to detect genetic factors for specific forms. A summary of the analyses gathered on the various forms of skeletal dysplasia and the genetic causes was published in 2010 and 2015[213,215]. In these publications 456 forms of skeletal dysplasia were grouped into 40 categories based on forms with similar phenotypes.

In the case investigated in this chapter a child was diagnosed based on family records and radiography as a case of Sedaghatian-type SpondyloMetaphyseal Dysplasia (SSMD). This form presents with symptoms which can include: severe metaphyseal chondrodysplasia, mild limb shortening, platylspondyly, cardiac conduction defects, central nervous system abnormalities, delayed epiphyseal ossification, irregular iliac crests, pulmonary haemorrhage, agenesis of the corpus callosum, pronounced frontotemporal pachygyria, simplified gyral pattern, partial lissencephaly and severe cerebellar hypoplasia[216]. Few cases of SSMD have ever been documented, let alone had DNA samples collected due to the short life expectancy caused by disease complications. Glutathione Peroxidase 4 (*GPX4*) is currently the only gene that has currently been directly implicated with SSMD[216], it is possible that there are other genes causing SSMD or genes which cause similarly lethal forms skeletal dysplasia.

GPX4 is one of eight forms of glutathione peroxidases found in humans, it is the only

form with anti-lipid peroxidation functions. Loss of *GPX4* function leads to excessive lipid peroxidation and cell death[216]. *GPX4* is found on chromosome 19 band p13.3 (1,103,936 - 1,106,787 NC 000019.9) and was been identified in other cases as a cause of SSMD in 2014[216]. Three isoforms of *GPX4* are expressed, see Figure 5.1, cytosolic *GPX4* is ubiquitous in most mammalian cell types, mitochondrial *GPX4* is associated with development of sperm cells, the role of nuclear *GPX4* remains largely unresolved. Using knockout mice studies for the nuclear isoform mice remain viable but a minority show defects with the left atrium not forming[217]. Cytosolic *GPX4* can be detected in the nucleus despite the isoform being without the nuclear localisation sequence at the 5'UTR as with the nuclear isoform. Mitochondrial isoform *GPX4* contains a mitochondrial targeting peptide at the 5'UTR in contrast to the cytosolic form where the peptide is cleaved. Knockout studies of the mitochondrial isoform show male mice became infertile, supporting the predicted role of the isoform in sperm development. Loss of cytosolic *GPX4* resulted in embryonic lethality, addition of mitochondrial *GPX4* did not prevent lethality, only addition of the cytosolic isoform without the mitochondrial targeting peptide restored viability[217].

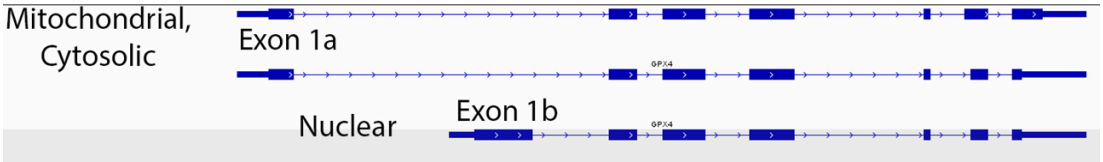

Figure 5.1: Isoforms of *GPX4*, the cytosolic isoform uses exon 1A and is ubiquitously expressed in mammalian cell types with knockout resulting in embryonic lethality which was not recovered with addition of mitochondrial and nuclear isoforms. The mitochondrial isoform is dominantly used in spermatogenic cells and nuclear *GPX4* using the alternate start exon 1B is expressed during development of spermatids.

Two binding sites have been identified upstream of the transcription initiation site of *GPX4* for the cytosolic and mitochondrial isoforms. The site furthest upstream is dominant in spermatogenic cells but the mechanism remains unresolved. The second site lies close to the 5'UTR of exon 1A and is dominant in most mammalian cells, which is under the control of general transcription factors Sp1 and NF-Y without a TATA-box[217,218]. Most evidence suggests that DJ-1 (encoded by gene *PARK7*-Parkinson protein 7) and G-Rich RNA Sequence Binding Factor 1 (*GRSF1*) are the post-transcriptional regulators for *GPX4*[217–219]. DJ-1 is believed to be redox

sensitive, acting to protect neurons against cell death and oxidative stress by binding to the 5'UTR of the *GPX4* RNA, when cysteine residues in DJ-1 are oxidised they lose affinity for RNA allowing *GPX4* release. *GRSF1* gene product is also believed to be redox sensitive, acting in a similar mechanism to DJ-1[217].

GPX4 was previously identified as the cause of SSMD in two family trios. Between the two cases two specific *GPX4* mutations were identified: $c.587 + 5G > A$ and c.588–8_588-4del[216]. In both trios variants were validated using Sanger sequencing and a mini gene system. A skin biopsy from the child of one family failed to grow, this fitted with the problems caused with fibroblast growth caused by *GPX4* mutations. Knockout experiments in mice show that *GPX4* loss causes loss of viability beyond gestation. Sequencing results showed exon 4-5-6 splicing in the control *GPX4* mini-gene construct, while the variant $c.587 + 5G > A$ resulted in the splicing out of part of exon 4. The variant c.588-8_588-4del in *GPX4* resulted in the loss of exon 5[216].

5.1.3 Short-Rib polydactyly syndromes

Diagnosis of SSMD can be complicated due to phenotypic overlap with Short-Rib Polydactyly Syndromes (SRPS) which can also be perinatal lethal. SRPS encompass a range of disorders commonly characterised by very short horizontal ribs, short limbs and variable degrees of polydactyly[220]. SRPS are a group of skeletal ciliopathies which include Asphyxiating Thoracic Dystrophy (ATD) also known as Jeune syndrome. Patients with Jeune syndrome typically display a narrow thorax with bell shaping in some cases caused by shortening of the ribs. Shortening of the ribs impairs pulmonary development leading to respiratory distress in the first two years of life. Cases of Jeune syndrome are rare with an estimated incidence of 1-5/500,000 live births[221,222]. As of March 2018 there are 19 phenotype entries listed on OMIM[223] for SRPS spread across 20 genes as shown below in Table 5.1.

Not all SRPS described are lethal with one study estimating that only 60% of JATD cases end in lethal respiratory distress such as Ellis-van Creveld syndrome, SRPS types 13 (*CEP120*), 14 (*KIAA0586*) and 17 (*TCTEX1D2*)[223,224]. Inheritance methods can complicate interpretation of Jeune syndrome with cases being described as autosomal recessive including compounded heterozygotes or digenic recessive[221,222].

Gene/Syndrome	Phenotype	Inheritance	MIM number
EVC2	Ellis-van Creveld syndrome	AR	225500
EVC	Ellis-van Creveld syndrome	AR	225500
SRTD1 -15q13.3 region	Short-rib thoracic dysplasia 1 with or without polydactyly	AR	208500
IFT80	Short-rib thoracic dysplasia 2 with or without polydactyly	AR	611263
DYNC2H1	Short-rib thoracic dysplasia 3 with or without polydactyly	AR - DR	613091
TTC21B	Short-rib thoracic dysplasia 4 with or without polydactyly	AR	613819
WDR19	Short-rib thoracic dysplasia 5 with or without polydactyly	AR	614376
NEK1	Short-rib thoracic dysplasia 6 with or without polydactyly	AR - DR	263520
WDR35	Short-rib thoracic dysplasia 7 with or without polydactyly	AR	614091
WDR60	Short-rib thoracic dysplasia 8 with or without polydactyly	AR	615503
IFT140	Short-rib thoracic dysplasia 9 with or without polydactyly	AR	266920
IFT172	Short-rib thoracic dysplasia 10 with or without polydactyly	AR	615630
WDR34	Short-rib thoracic dysplasia 11 with or without polydactyly	AR	615633
SRTD12 - Not mapped	Short-rib thoracic dysplasia 12	Unknown	269860
CEP120	Short-rib thoracic dysplasia 13 with or without polydactyly	AR	616300
KIAA0586	Short-rib thoracic dysplasia 14 with polydactyly	AR	616546
DYNC2LI1	Short-rib thoracic dysplasia 15 with polydactyly	AR	617088
IFT52	Short-rib thoracic dysplasia 16 with or without polydactyly	AR	617102
TCTEX1D2	Short-rib thoracic dysplasia 17 with or without polydactyly	AR	617405
IFT43	Short-rib thoracic dysplasia 18 with polydactyly	AR	617866

Table 5.1: Short-Rib Polydactyly Syndromes phenotypes and genes listed on OMIM as of March 2018 and the listed inheritance methods previously recorded. AR= Autosomal recessive , DR= Digenic recessive.

Cytoplasmic dynein 2 heavy chain 1 (*DYNC2H1*) was estimated to be the cause of approximately a third of Jeune syndrome patients of north-European patients and is the most commonly reported cause of Jeune syndrome[222,224]. *DYN2CH1* is located on chromosome 11 and comprises 90 exons spread over 370.4 kb making Sanger sequencing challenging in terms of time and cost for most clinical cases. *DYNC2H1* has only been implicated in SRPS with the introduction of next generation sequencing methods such as WES or WGS. *DYNC2H1* produces the IFT-A

cytoplasmic dynein-2 motor heavy chain which interacts with other gene products from *WDR60*[225], *WDR34* which produce intermediate chains and *DYNC2L1* producing light intermediate chains forming the IFT-A dynein-2 complex[220].

Other genes most associated with Jeuene syndrome are *IFT-80* (Intra-Flagellar Transport 80), *TTC21B* (Tetratricopeptide Repeat Protein 21B)/*IFT139*, *WDR 19* (WD-Containing repeat 19)/*IFT144* , *WDR35*[222,224,226]. All of these IFT and WDR genes including *DYNC2H1* affect ciliary intraflagellar transport, loss of function mutations therefore impair ciliogenesis preventing cell signalling required for normal human development. *NEK1* loss of function has also recently be suggested to affect ciliogenesis, the protein produced was localised to the basal body and within the cilium suggesting a role in IFT via digenic mutations or autosomal recessive inheritance[220,221,224].

Early analysis in 2003 of several patients diagnosed with SRPS also identified abnormalities around chromosome 15 band q13 (15q13) suggesting a link to SRPS type-1[117] but to date no study has been able to link a gene in the region to SRPS. Evidence also suggests regions in the 15q13 band are also associated with impairment of neurological development by genome instability[222,227]. SRPS type-12 as with type-1 has no gene specifically identified but also currently has no loci has been mapped for this type.

Due to the phenotypic overlap of lethal SRPS cases with SSMD misdiagnosis is possible from the phenotype alone making analysis of SRPS genes logical for suspected SSMD patients with negative results in *GPX4*.

Chromosome 15q13 micro-deletion

In the chromosome 15q13 region there are three well characterised breakpoints (BP3, BP4 and BP5) which can lead to inversions, predisposing this region to deletion[227]. Each breakpoint region contains some duplicated sequences, in particular between BP4 and BP5 though the sequences lie in opposite orientations, see Figure 5.2. BP3 - BP5 deletions are also possible which cause the disease Prader-Willi syndrome or Angelman syndrome[227]. This is the same locus that is also believed to be involved with SRPS type-1[117]. The locations of the breakpoints converted from hg17 to hg38 are shown below in Table 5.2.

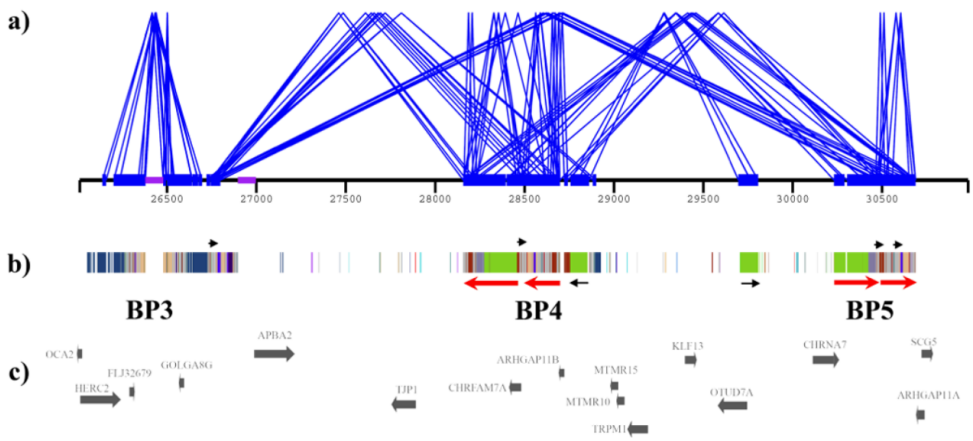

Figure 5.2: a) Homology map showing duplicated sequences between each of the breakpoint regions, with blue lines indicating similar sequences. b) Locations and structure of breakpoints. Highlighted by red arrows indicating the orientation are the segments which invert possibly causing the deletion of the BP4-5 critical region. c) Genes which are contained within and between each of the breakpoints. Adapted from the paper by Sharp et al.[227]

Both Prader-Willi and Angelman syndrome are caused by the loss or gain of copy in the 15q13 region leading to the altered expression of at least one of the 20 imprinted genes in this region. Imprinting occurs when only one of the parental copies of a gene are active. However 70% of cases of Prader-Willi syndrome have been shown to be caused by a large deletion in the the 15q11-q13 imprinted region, 25% of cases reported maternal uniparental disomy and only 5% as a imprinting defect leading to expression as though inherited from the mother. Similar patterns were also found for Angelman syndrome with approximately 74% of cases with region deletions, 11% with loss of the maternal copy of *UBE3A*, 8% with uniparental disomy and 7% with

imprinting[228].

Considering phenotypes and incidence of both Prader-Willi and Angelman syndromes there is only a limited match with SSMD. Both syndromes are too common to be in keeping with the rarity of the SSMD phenotype seen in the affected patient at an estimated frequency of between 1 in 15,000 - 30,000 live births[228]. Prader-Willi syndrome is characterized by: hypotonia, short stature with small hands and feet, hyperphagia leading to morbid obesity beginning during early childhood and intellectual disability[228]. Angelman syndrome is associated with: intellectual disability, seizures and an ataxic gait[228].

Breakpoint	HG17 - Start	HG17 - End	HG19 -Start	HG19- End	HG38 -Start	HG38- End
BP3(A)	26,103,472	26,350,000	28,429,877	28,676,405	28,184,731	28,431,259
BP3(B)	26,500,000	26,900,000	28,710,238	29,097,694	28,465,092	28,852,548
BP4	28,200,000	28,900,000	30,412,708	31,112,708	30,120,505	30,820,505
BP5	29,600,000	30,750,000	31,812,708	32,962,708	31,520,505	32,670,507
Between BP4 -BP5	28,900,000	29,600,000	31,110,708	32,964,708	30,818,505	32,672,507

Table 5.2: Liftover locations obtained from UCSC when supplying an approximate value for the hg 17 breakpoint locations and then recording the hg19, hg38 locations. Locations are based on the position in hg17 as illustrated in the paper by Sharp et al[227]. Breakpoint 3 contains a gap in the assembly the locations for this breakpoint are split into "A" and "B" either side of this gap.

5.2 Aims

Sedaghatian-type spondylometaphyseal dysplasia (SSMD) was diagnosed in a child. To investigate the genetic contribution to the disease both parents and the affected child will be WES and WGS sequenced. Initial aCGH microarray analysis performed at the Wessex regional lab (Salisbury) suggested there was a possible micro-deletion on chromosome 15q13.3 which will be investigated from the NGS datasets using CNV calling. The *GPX4* gene is currently the only implicated cause in SSMD and will be checked for variation which could be causing a disorder.

5.3 Materials & methods

5.3.1 Family & sequencing

Blood samples were taken from both parents and the affected child by the attending clinician at Southampton General Hospital Dr David Hunt, see Figure 5.3. From the blood DNA was extracted by the Wessex regional genetics laboratory (Salisbury). WES was then performed at the Wellcome Trust Centre for Human Genetics (Oxford) and subsequently to BGI (Hong Kong) for WGS. Fastq files were returned from which all analyses in Chapter 4 and 5 were performed. Details of bioinformatic analyses performed on the samples were described in Chapter 4.

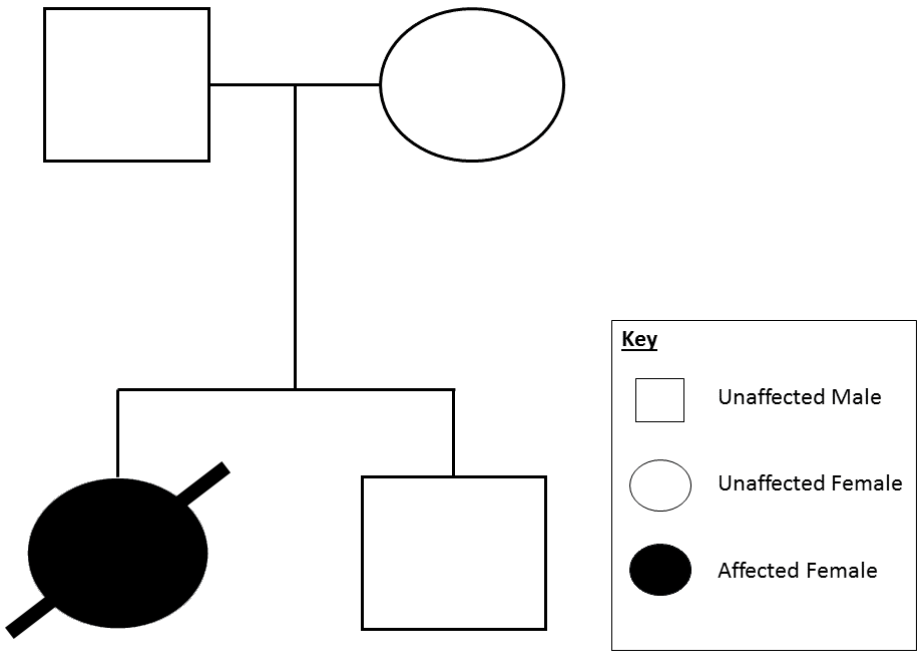

Figure 5.3: Pedigree diagram of skeletal dysplasia case with both parents, son who are unaffected and the affected daughter. Patients were of European ancestry, samples were available for analysis from both parents and the affected daughter for analysis. No sample was available for the unaffected son.

Exome samples

Exome sequencing was performed using the Agilent SureSelect Human All Exon (v5) exome selection kit spanning 50.4 Mb, targeting 357,999 exons over 21,522 genes performed at the Wellcome Trust Centre for Human Genetics (WTCHG). Read

length for exome samples was 100 base pairs.

Genome samples

Genome sequencing was performed at the Beijing Genomics Institute (BGI) using an Illumina Hiseq X10, sequencing was performed to try and achieve mean coverage of 30 across the genome. Reads from the X10 were 150 base pairs in length.

5.3.2 Variant calling

Variant calling was performed using GATK joint genotyping as described in more detail in Chapter 4. BAM files were combined per sample for WES and WGS using picard tools (v1.8.3) before variants calling to maximise depth and coverage and reduce false positives variants in due to low depths in either dataset. Variant calling of combined BAM files was performed using GATK HC (v3.7) to generate gVCFs per sample which were joint genotyped to create a single trio VCF. The VCF was annotated with VEP (v88) for GEMINI (v20.1). Results files from GEMINI were also annotated using ANNOVAR to provide additional annotations such as gnomAD AAFs and FATHMM-XF scores to prioritise non-coding variants of interest.

5.3.3 *GPX4* variant, transcription factors and binding sites

GPX4 and a one kilobase region either side of the gene was extracted using tabix commands, the region on chromosome 19 (hg38) extracted was 1,102,926 - 1,107,787 bp.

Using the region in and around *GPX4* visual inspection of variants called and binding sites were performed to check for potential alterations to transcription factor binding sites. Using IGV visual inspection was performed and searching for the binding motifs associated with NF-Y (CCAAT box) and GC rich regions likely to bind Sp1[218].

Transcription factor binding sites in the UTR of *GPX4* which affect post-transcriptional RNA were also inspected for abnormalities and variants likely to affect the regulation of *GPX4*. In particular the factor GRSF1 (AGGGGA motif) and DJ-1 (GC rich preference).

5.3.4 *GPX4* coverage

Coverage calculations were performed for *GPX4* using the SAMtools per base method as outlined fully in Chapter 4. Visualisation of coverage per base was then performed in R for the combined coverage and WES, WGS individually.

5.3.5 GEMINI variant analysis

Non-Mendelian inheritance

The first mode run in GEMINI was a check for the non-Mendelian transmission of alleles. This mode checks that one allele is inherited from each parent, in some cases both alleles come from one parent or new alleles may arise which are novel when comparing against parental genotypes. In this mode there are four possible violation variant types:

1. **Loss of Heterozygosity** - Child inherits one allele from one parent.
   $G/A + G/G \rightarrow A/A$

2. **Implausible *de-novo* mutations** - Child has two new alleles compared to parents.
   $G/G + G/G \rightarrow A/A$

3. **Plausible *de-novo* mutations** - Child has one new allele compared to parents.
   $T/T + T/T \rightarrow T/C$

4. **Uniparental Disomy** - Child's alleles are same as one parent.
   $A/A + G/G \rightarrow A/A$

For all of the variants identified as inherited in a non-Mendelian manner they were assigned a probability score. The probability scores are an estimate of the variant actually being a true positive. From the official documentation for GEMINI using a value above 0.99 is described as a reasonable filter from testing during development.

Compound heterozygotes

Compound heterozygotes occur when two heterozygous variants act together within the same gene, when possible this mode will phase by transmission. To consider a variant a compound heterozygote all affected individuals must be heterozygous at both sites and no unaffected individuals can be homozygous alternate at either site. Variants are assigned a priority describing if they are phaseable at both sites and heterozygous. Priority scores of two indicate that phasing data is missing for one or both parents. A score of three indicates that alleles are unphaseable as one or both parents are homozygous. Homozygous variants in either unaffected parents mean that the chance of the variant being causal are minimal (<1%)[229]. Each of the compound heterozygotes in the VCF file will have a "comp_het_id". The ID describes which variants are acting in tandem to cause the effect predicted.

Autosomal recessive

To fit a recessive model affected individuals must be homozygous alternate, no unaffected individuals can be homozygous alternate and parents must be unaffected, heterozygous for all affected children.

Tiered filtering

To prioritise variants for this case the annotations added to the GEMINI outputs from the ANNOVAR annotated trio VCFs were used. In total 5 tiers, encompassing a total of 430 genes as summarised in Appendix C - Table 8.7. Tiers 1 - 4 adopted a candidate gene approach while Tier-5 filtered variants using annotations.

Tier-1 focused on any variant annotated to be in a glutathione peroxidase gene, equivalent to the UCSC known gene field containing any entry equal to *GPX**. Currently there are seven known glutathione peroxidase genes and a suspected 8th gene, hence there are eight potential gene variants to be reported in this tier (*GPX1* - *GPX8*).

Tier-2 reported any variants in 182 genes identified from the HGMD genes matching the terms: 'skeletal dysplaysia', 'metaphyseal', 'spondylometaphyseal' and 'sedaghatian'. These terms were chosen as they matched with the phenotypic

description of the affected patient.

Tier-3 contained genes from the nosology paper published in 2015 listing the genes linked with skeletal dysplasia[215]. Additionally genes covered by two NHS gene panels for skeletal dysplasia covering 57[230] and 222 genes[231]. Removal to duplicate genes left a total of 327 unique genes. There was significant overlap between Tier-2 and Tier-3 with 89 of the genes shared between them.

Tier-4 contained all of the 18 SRPS related genes described in Table 5.1.

Tier-5 used filtering of annotations to identify variants that are both rare enough and predicted pathogenic. To each of the GEMINI results standardised filters were applied to subset results:

1. Func.knownGene - Select only : Exonic, splicing
2. ExonicFunc.knownGene: Exclude "Synonymous SNV"
3. Fuentes false postives : Exclude all

Autosomal recessive variants were filtered using a gnomAD Non-Finnish European (NFE) alternate allele frequency of 0.05 or less or missing along with a FATHMM-XF filter of above 0.5 .

Non-Mendelian inheritance was filtered first by gnomAD NFE AAFs below or equal to 0.05 or missing. Secondly a filter of 0.99 or above was implemented for the non-Mendelian violation probability in accordance with the recommended threshold. Violations types selected were for *de-novo* variants. Finally a FATHMM-XF score of above 0.5 or missing was implemented.

Compound heterozygotes were filtered using a more lenient gnomAD NFE AAF to remove all variants above 0.1 . A more lenient cut-off was used as variants without a compounding partner would theoretically be able to reach a higher population allele frequencies. Secondly as with non-Mendelian variants a FATHMM-XF score of 0.5 or above or missing was applied to select variants with some evidence of being damaging to function. Finally, compound heterozygotes have an ID assigned by GEMINI, both

variants for a pair need to be present after application of the previous filters to be considered as compounded.

Additionally samples were checked for ACMG (American College of Medical Genetics) gene variants, described in section 4.1.4, which are recommended to notify clinicians about as secondary findings due to the genes in the list being decided as highly actionable by panels of clinicians and geneticists.

5.3.6 Non-coding variants

Non-coding variants were also analysed for each of the GEMINI modes using the CADD-phred threshold of 20 and above or using the FATHMM-XF non-coding cut-off of 0.96 or above for high confidence pathogenic variants.

5.3.7 Splicing variants

SPIDEX[208] was used with ANNOVAR to annotate variants with scores predicting the effect of variants on splicing. SPIDEX is based on a machine learning model which extracts DNA sequence motifs and features along with cis-elements from the hg19 reference genome. RNA-seq expression data from the Illumina Human Body Map 2.0 project was also used to train the prediction model to estimate the **P**ercentage of transcripts with the central exon **S**pliced **I**n (PSI) . Training was performed using 10,689 exons with 1,393 sequence features extracted. For a SNP the model would estimate the predicted (PSI) for the 16 tissues before calculating the change in PSI relative to each of the reference and return the maximal change in a tissue termed ΔPSI. Z-scores for each variant ranks the ΔPSI score in terms of the difference to the mean divided by the standard deviation of the dataset.

SPIDEX variants of interest in both the original publication by *Xiong et al*[208] and *de Almeida et al.*[232] used a cut-off of $|\Delta PSI| \geq 5$. A value for ΔPSI lower than -5 is suggested to describe a variant which is deemed to sufficiently lower the percentage of transcripts with the central exon spliced in, so decreasing the efficiency of splicing. Conversely, a score above 5 indicates increase of splicing activity and possibly the gain of a splice site. Variants of interest were also cross-checked using the Human-Splice Finder (v3.1)[233].

5.3.8 Structural variant calling

LUMPY was used with WGS data only to generate a trio VCF as described previously in Chapter 4. Structural variants were converted into tab separated files for easier filtering and parsing. In the conversion process variants were also annotated using bedtools intersect to add gene names to variants should they overlap with a gene transcript. Using the gene names column a further annotation field was added identifying the genes which overlapped with any of the 430 genes identified from tiered filtering.

5.3.9 Copy number variation

CNVkit was run separately for WES and WGS. A pooled reference of the mother and father were used to more accurately call copy number variations in the affected child. Full details of settings used were described previously in Chapter 4.

5.3.10 Cross-dataset compound heterozygotes

GEMINI calculates possible compound heterozygotes variants from the calls generated by GATK HC. However this does not account for compound heterozygotes such as when there is a heterozygous deletion of a region and a deleterious SNP. Together the variants can cause the loss of both functional gene copies and hence not be called by any single tool as compounded. To compare variants the heterozygous, predicted pathogenic variants from each of the tools were extracted. Genes with two or more variants were analysed further by checking the phase of variants. A similar method was also used to look for di-genic recessive variants found in cases of SRPS which also needs to consider the phase of variants.

5.4 Results

5.4.1 *GPX4* coverage

Coverage over the *GPX4* gene was assessed for all samples targeted over the *GPX4* gene, returning per base coverage for WES and WGS samples separately and combined. Coverage was plotted as shown below in Figure 5.4.

Coverage for WES was higher over exons, reaching over 100x depth where exons were in close proximity to another exon. However none of the WES samples covered exon 1a sufficiently for variant calling. Mean coverage for WES was 38x, 35x and 53x for SD001, SD002 and SD003 respectively. Mean coverage for WGS samples over *GPX4* was 21x, 19x and 20x for SD001, SD002 and SD003 respectively. Combined WES and WGS mean depth of coverage was 51x, 55x and 69x for SD001, SD002 and SD003 respectively.

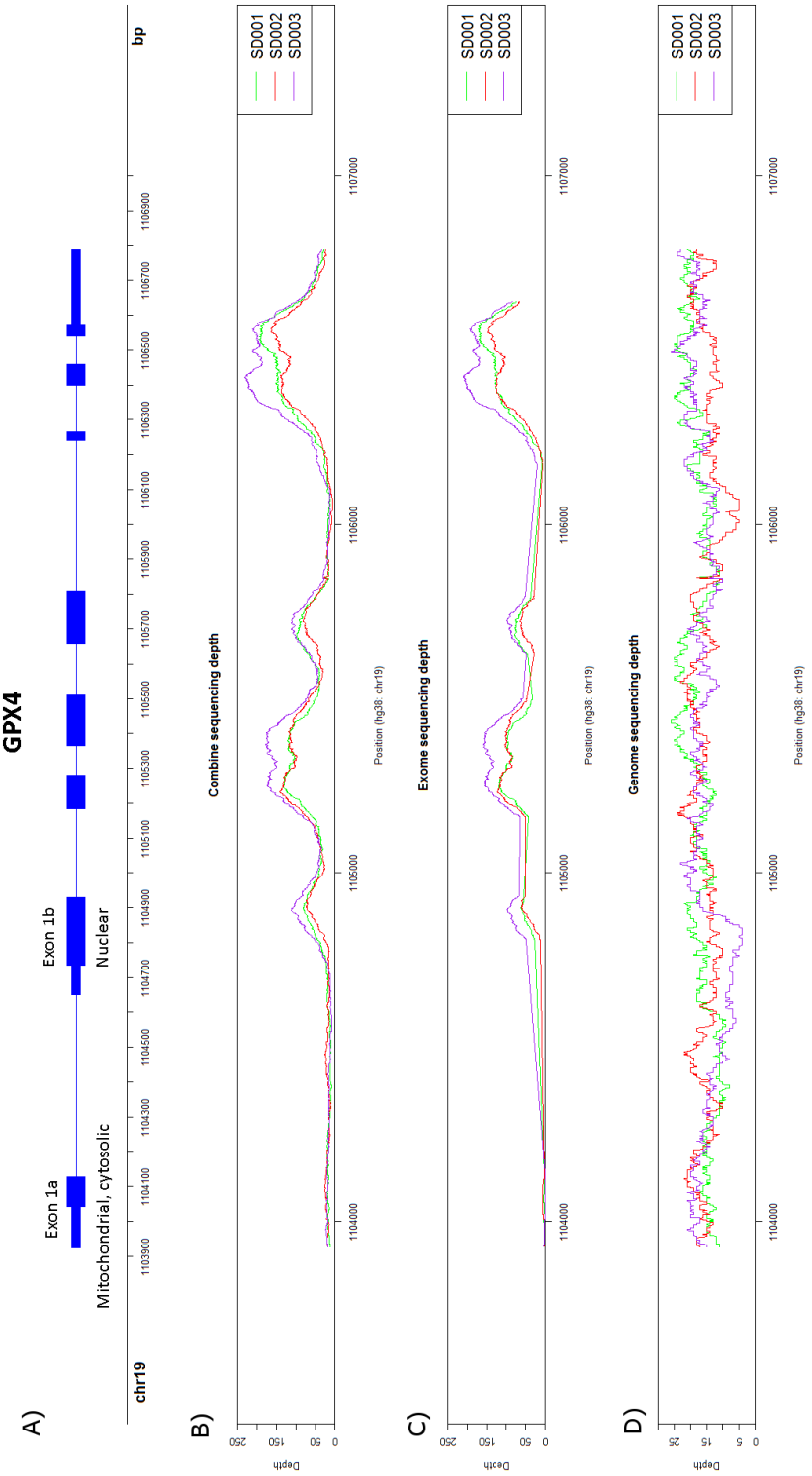

Figure 5.4: *GPX4* coverage for combined WES and WGS and individually. **A)** Collapsed transcript for *GPX4* is shown with alternate start exon 1b for the nuclear isoform of *GPX4*. **B)**Combined WES and WGS coverage over *GPX4* **C)** Exome sequencing coverage track showing the lack of coverage over exon 1a and finally genome sequencing is shown with the more consistent coverage over the entire of *GPX4* at around 20x depth. **D)** Genome sequencing coverage over the *GPX4* gene which does cover exon 1a and is more consistent in coverage across the gene but at lower depths than the exomes in panel C.

5.4.2 *GPX4* variants

GPX4 was the first gene analysed, all variants over *GPX4* and 1kb either side of the gene were extracted for SD003 and compared against parental genotypes. Eight variants found in the region of *GPX4* for sample SD003 are reported below in Table 5.3.

No.	Chr	Pos	Ref	Alt	SD001	SD002	SD003	Function	Gene	avsnp147	gnomAD ALL	gnomAD NFE
1	19	1103523	A	G	0/1	1/1	0/1	upstream	*GPX4*	rs8103283	0.7861	0.7052
2	19	1103769	A	T	0/0	0/1	0/1	upstream	*GPX4*	rs17526264	0.0598	0.0577
3	19	1103786	C	T	0/0	0/1	0/1	upstream	*GPX4*	rs17554931	0.0599	0.0578
4	19	1104439	G	C	0/0	0/1	0/1	UTR5	*GPX4*	rs2074450	0.0544	0.0568
5	19	1106478	G	C	0/0	0/1	0/1	intronic	*GPX4*	rs8178977	0.2601	0.2516
6	19	1106616	T	C	1/1	1/1	1/1	exonic	*GPX4*	rs713041	0.6097	0.5743
7	19	1106846	C	T	0/0	0/1	0/1	downstream	*GPX4,SBNO2*	rs2075710	0.3314	0.2525
8	19	1107336	A	G	0/1	1/1	0/1	downstream	*GPX4,SBNO2*	rs2075711	0.5885	0.5295

Table 5.3: Variants 1kb either side of *GPX4* from WGS data for the child (SD003) affected by SSMD. Only variant No. 6 is exonic, but is also synonymous, homozygous alternate in both parents and a common variant with alternate allele frequency in NFE populations of 0.57. No other variants around exons were rare enough or only in the affected child to be plausible as affecting splicing.

In total eight variants were called over *GPX4* or 1kb either side of the gene. The only exonic variant identified in *GPX4* for the affected child was rs713041. This variant could not be pathogenic alone as it is heterozygous and also shared with the child's mother, with a gnomAD all population allele frequency of 0.6097. In older versions of dbSNP rs713041 was called as a non-synonymous change causing the amino acid change p.L113S but has been revised in from dbSNP 147 onwards.

5.4.3 *GPX4* transcription factors and binding sites

From the analysis of *GPX4* and the surrounding regions there does not appear to be any alterations to transcription factor binding sites. To confirm the lack of changes the region around the 5' UTR was visualised. In the UTR both the *GRSF1* (AGGGGA motif) and *NF-Y* (CCAAT motif) binding site no variants or structural alterations. The lack of variants and the number of relatively high concentration of guanine and cytosine bases suggest both DJ-1 and Sp1 binding would not be inhibited.

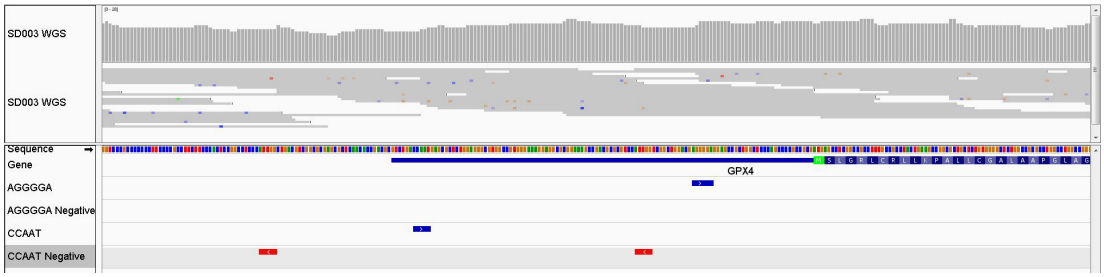

Figure 5.5: For the transcription factor binding site in the *GPX4* 5' UTR of exon 1a shows a lack of variants and the high G-C content in blocks make inhibition of GC-rich binding factors Sp1 and PARK7 unlikely. The preferential motifs for GRSF1 (AGGGGA motif) and NF-Y (CCAAT motif) can also be detected in the region as shown in the lower four tracks showing the motifs on forward strands in blue and reverse in red, suggesting no alterations in transcription factor binding.

Focusing on the specific RNA binding transcription factors GRSF1 and DJ-1 the genes were screened for variants. In total 179 variants were found but only one missense variant in *GRSF1* was found, though this variant had an AAF of 1 suggesting that the alternate allele is actually the reference allele for the NFE population. All other variants were intronic or in UTRs.

5.4.4 GEMINI variants

Summary totals of variants output from each GEMINI mode are shown in Table 5.4. Subsequent filtering for each mode was performed and detailed in the sections below for the combined trio dataset.

Gemini Mode/Tier	Exome	Genome	Combined
Autosomal Recessive	4,672	218,288	219,574
Non-Mendelian	28,582	201,354	195,909
Compound Hets	10,908	23,967	22,248

Table 5.4: Total number of variants detected to fall in each category by GEMINI for exome, genome sequencing and combined datasets. For Non-Mendelian and compound heterozygote variants the number of variants decreased compared to the genome.

Autosomal recessive analysis

A total of 219,574 autosomal recessive variants were identified. Tier-1 failed to identify any *GPX** variants. Tier-2 identified 1,401 variants matching to the HGMD terms, however only eight of these variants were coding as shown below in Table 5.5. All of the variants were too common to fit with the disease, with AAFs using gnomAD NFE between 0.28 and 0.77. Furthermore five of the eight variants are synonymous, of the three non-synonymous variants and none are predicted pathogenic by FATHMM-XF.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	rsID	KnownGene	gnomAD NFE	FATHMM-XF
AAGAB	chr15:67236036:T:G	T/G T/G G/G	74 66 0	86 55 187	rs7173826	nonsynonymous SNV	0.322	0.292
ACVR1	chr2:157780398:G:A	G/A G/A A/A	74 46 1	48 52 148	rs2227861	synonymous SNV	0.769	0.005
DNM2	chr19:10829116:T:C	T/C T/C C/C	30 41 3	32 38 88	rs2229920	synonymous SNV	0.279	0.006
JUP	chr17:41755893:T:A	T/A T/A A/A	28 30 0	28 39 82	rs1126821	nonsynonymous SNV	0.742	0.371
JUP	chr17:41769673:A:G	A/G A/G G/G	23 15 0	24 21 50	rs7405731	synonymous SNV	0.757	0.004
NFKBIA	chr14:35404564:G:A	G/A G/A A/A	58 66 1	59 47 126	rs1957106	synonymous SNV	0.332	0.008
SEC23A	chr14:39048721:T:C	T/C T/C C/C	20 16 0	18 22 51	rs11556216	synonymous SNV	0.381	0.005
STK4	chr20:45000494:G:A	G/A G/A A/A	52 33 0	52 41 119	rs17420378	nonsynonymous SNV	0.314	0.383

Table 5.5: Autosomal recessive Tier-2 identified eight variants in genes identified in genes which matched to HGMD terms. All variants were common with alternate allele frequency between 0.28 and 0.77 and none were predicted pathogenic by FATHMM-XF. Five of the eight variants were also synonymous SNPs.

Tier-3 identified 1,746 variants matching to the extended candidate list, only 14 of which were exonic. All 14 variants are too common to be causal in this case as gnomAD NFE AAFs ranged from 0.10 to 0.77 as shown in Table 5.6. 7 of the 14 variants were synonymous and 7 were non-synonymous though only the variant chr11:103287549:C:TS (rs10895391) in *DYNC2H1* had a FATHMM-XF score of 0.65, just above the 0.5 cut-off suggesting a low-confidence pathogenic variant.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	rsID	KnownGene	gnomAD NFE	FATHMM-XF
ACVR1	chr2:157780398:G:A	G/A G/A A/A	74 46 1	48 52 148	rs2227861	synonymous SNV	0.7692	0.005313
ALPL	chr1:21577638:T:C	T/C T/C C/C	29 24 1	31 25 67	rs34605986	nonsynonymous SNV	0.1077	0.128834
DYNC2H1	chr11:103287549:C:T	C/T C/T T/T	67 69 0	46 56 120	rs10895391	nonsynonymous SNV	0.3629	0.654937
GLI3	chr7:42048623:T:C	T/C T/C C/C	44 35 0	41 35 96	rs846266	nonsynonymous SNV	0.5841	0.038904
GUSB	chr7:65960907:A:G	A/G A/G G/G	29 27 0	22 23 69	rs9530	nonsynonymous SNV	0.5559	0.05435
NF1	chr17:31226467:G:A	G/A G/A A/A	59 78 2	45 54 159	rs2285892	synonymous SNV	0.3198	0.004488
PCNT	chr21:46391357:C:T	C/T C/T T/T	6 16 0	22 12 25	rs3737438	synonymous SNV	0.3703	0.004903
PCNT	chr21:46353189:C:A	C/A C/A A/A	63 60 2	61 51 118	rs2249057	synonymous SNV	0.3149	0.004994
POR	chr7:75980359:A:G	A/G A/G G/G	50 44 0	53 61 112	rs1135612	synonymous SNV	0.2511	0.013446
ROR2	chr9:91724039:C:T	C/T C/T T/T	34 30 0	32 34 84	rs10761129	nonsynonymous SNV	0.6813	0.047339
SALL4	chr20:51791427:C:T	C/T C/T T/T	39 53 0	57 34 108	rs13038893	synonymous SNV	0.3245	0.00757
SLC29A3	chr10:71322806:A:G	A/G A/G G/G	39 36 0	37 43 103	rs2277257	nonsynonymous SNV	0.4106	0.118187
SLC39A13	chr11:47413435:G:A	G/A G/A A/A	41 24 0	27 27 68	rs2293576	synonymous SNV	0.3358	0.007298
TCOF1	chr5:150392717:C:G	C/G C/G G/G	32 22 1	32 27 67	rs1136103	nonsynonymous SNV	0.2189	0.073854

Table 5.6: Autosomal recessive Tier-3 filter for extended candidate genes identified 14 exonic variants, split as seven non-synonymous and seven synonymous. All of the variants were too common to be of interest with AAFs which ranged from 0.10 to 0.77. Only the *DYNC2H1* variant was predicted by FATHMM-XF as pathogenic at 0.65.

Tier-4 identified genes listed in OMIM as associated with a form of SRPS, in total 256 variants were identified. One exonic variant was detected, the same *DYNC2H1* variant chr11:103287549:C:T (rs10895391) also identified in Tier-3 and shown in the Table 5.6.

Tier-5 applied filters to variants without using genes of interest, selection of coding variants (exonic or splicing) and below a gnomAD NFE AAF of 0.05 or missing identified 28 rare, coding variants. Removing synonymous variants reduces remaining variants to 16, a further five variants were removed due to being annotated as genes with false positive variants leaving 11 variants. Two further variants: rs7748778 and rs77127003 were then excluded as they has a gnomAD all populations frequency of 0.1458 and 0.0507 and are too common to be causal. 9 of the 11 variants had dbSNP IDs but only one had a ClinVar description for a phenotype of sprinting performance

deficiency and so two variants were left which were non-frameshift insertions and lacked annotations as shown in Table 5.7. Non-frameshift insertions were in the genes *MADCAM1* (Mucosal Vascular Addressin Cell Adhesion Molecule 1) associated with Inflammatory Bowel Disease and *SLC12A5* (Solute Carrier Family 12 Member 5) associated with epileptic encephalopathy, early infantile.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	rsID	KnownGene	gnomAD NFE	FATHMM-XF
MADCAM1	chr19:501743:-:N	T/. T/. N/N	-3	-3	.	nonframeshift insertion	.	.
SLC12A5	chr20:46022943:-:N	G/N G/. N/N	-3	-3	.	nonframeshift insertion	.	.

Table 5.7: Autosomal recessive Tier-5 filtered variants to identify rare (below 0.05 AAF), coding variants. 28 coding variants were identified, only the *SMAP1* variants was predicted pathogenic by FATHMM-XF. 12 variants were synonymous and excluded leaving 16 variants.

Finally, there were a total of 109 variants autosomal recessive variants overlapping with genes listed under the ACMG guidelines, however none of the variants were exonic.

Non-Mendelian inheritance - *de novo* variants

Tier-1 identified no variants annotated as *GPX** which violated Mendelian inheritance. Tier-2 identified 883 HGMD variants of which only one was exonic in the gene *HSPG2*, shown in Table 5.8. For this variant the father SD001 had 15 alternate allele reads compared with 69 reference reads, making the call close to a heterozygous call and so the variant may not actually be a Mendelian violation. Although rare by gnomAD this variant was also common by ExAC with a NFE AAF of 0.1498 and was below the FATHMM-XF pathogenicity predictions threshold of 0.5.

Tier-3 identified 1,230 variants, however only one exonic variant passed the violation probability of 0.99 or above, the same *HSPG2* variant as identified in Tier-2 was again identified shown in Table 5.8. Tier-4 identified 116 variants, though none were exonic and passed the violation probability of 0.99 or above.

Gene	Variant	Viol. Prob.	GT:SD00-1 2 3	gt ref depths	gt alt depths	rsID	KnownGene	gnomAD NFE	FATHMM-XF
HSPG2	chr1:21873389:C:G	0.99509	C/C C/C C/G	69 65 96	15 0 32	rs766722786	nonsynonymous SNV	0.0003	0.469073

Table 5.8: Non-Mendelian Tier-3 extended candidate gene variant. One exonic variant was found in the gene *HSPG2* which was rare using gnomAD genome alternate allele frequency for NFE but not in other populations as ExAC reported a NFE AAF of 0.1498.

Tier-5 involved filtering of variants to identify rare and coding variants, gnomAD NFE AAF below 0.05 and exonic or splicing, reduced variants to 404. Removing synonymous variants left 295 variants. When selecting only plausible *de-novo* and a violation probability of 0.99 or above 103 variants were left. Removal of Fuentes false positive gene variants[197] reduced the variants down to 61. Application of a FATHMM-XF filter of 0.5 or above or missing reduced variants to 26.

Gene	Variant	Viol. Prob.	GT:SD00-1 2 3	gt ref depths	gt alt depths	rsID	KnownGene	gnomAD NFE	FATHMM-XF
CCDC132	chr7:93294638:T:G	0.99985	T/T T/T T/G	74 72 90	18 8 32	rs75893203	.	0.0217	0.991548
COPB1	chr11:14479716:T:A	1	T/T T/T T/A	52 39 46	0 3 10	rs753555066	.	0.009	0.989688
RPRD1A	chr18:36031099:-:A	0.99277	T/T T/T T/TA	16 10 6	0 1 3	rs762367276	.	0.0136	.
ERVW-1	chr7:92469121:TTTC:-	1	TTTC/TTTC TTTC/TTTC TTTC/-	15 29 22	0 0 3	.	frameshift deletion	0.003	.
GOLGA6L2	chr15:23440320:T:-	0.99999	CT/CT CT/CT CT/C	67 64 33	0 0 4	rs745321745	frameshift deletion	0.0003	.
SPDYE5	chr7:75501469-75501543: * :-	1	*/* */* */-	18 16 4	0 0 5	.	frameshift deletion	.	.
ATP5A1	chr18:46099312:-:T	1	C/C C/C C/CT	18 26 11	0 0 4	.	frameshift insertion	0.0001	.
BOD1	chr5:173609434:-:TGAA	1	C/C C/C C/CTGAA	194 37 186	0 0 21	rs766153311	frameshift insertion	0.0013	.
MAP3K9	chr14:70809057:CCT:-	1	N/N N/N N/G	32 27 13	0 0 3	rs397840789	nonframeshift deletion	0.004	.
AC016753.2	chr2:85574861:-:TTA	0.99696	T/T T/T T/TTTA	9 8 5	0 0 3	.	nonframeshift insertion	.	.
KRTAP4-3	chr17:41168094:-:TTT	1	G/G G/G G/GTTT	37 74 93	0 0 13	rs770530211	nonframeshift insertion	0.0005	.
ATP6V1A	chr3:113805419:G:C	0.99998	G/G G/G G/C	98 88 123	13 14 34	rs768704096	nonsynonymous SNV	0.0011	0.97946
CNIH3	chr1:224684802:C:T	1	C/C C/C C/T	40 38 92	0 0 19	.	nonsynonymous SNV	.	0.500011
COG3	chr13:45493477:C:A	1	C/C C/C C/A	38 50 32	0 0 29	.	nonsynonymous SNV	.	0.91246
FP325317.1	chr9:43111392:T:C	1	T/T T/T T/C	38 30 32	0 0 8	.	nonsynonymous SNV	.	.
HSP90AB1	chr6:44249575:G:A	1	G/G G/G G/A	34 44 89	0 0 71	.	nonsynonymous SNV	.	0.957077
KIF1B	chr1:10365418:A:C	1	A/A A/A A/C	39 87 124	0 0 20	rs78662124	nonsynonymous SNV	0.0073	0.950458
LGALS3	chr14:55138328:A:C	1	A/A A/A A/C	75 33 90	20 0 36	rs200922957	nonsynonymous SNV	0.0033	0.696633
MCM8	chr20:5967525:G:A	1	G/G G/G G/A	94 72 104	15 11 26	rs753970338	nonsynonymous SNV	0.0011	0.838504
RASSF1	chr3:50338000:T:G	1	T/T T/T T/G	49 63 64	0 0 34	rs781352974	nonsynonymous SNV	0.0037	0.863434
RGPD4	chr2:107871378:G:A	1	G/G G/G G/A	8 16 7	0 0 4	rs201802537	nonsynonymous SNV	0.0115	0.617915
ZBTB7C	chr18:48029394:A:G	1	A/A A/A A/G	40 36 83	0 0 68	.	nonsynonymous SNV	.	0.626605
GOLGA6L22	chr15:22466823:T:C	0.99986	T/T T/T T/C	6 21 4	0 0 2	rs376943676	stoploss	.	.
TRBV5-4	chr7:142463532:C:T	1	C/C C/C C/T	114 162 231	11 0 43	rs55634899	synonymous SNV	.	.
KIR2DS4	chr19:54847281:-:A	1	C/C C/C C/CA	267 90 141	0 0 21	rs145114829	unknown	.	.
TCF25	chr16:89885977:-: **	1	C/C C/C C/**	38 35 21	0 0 25	.	unknown	0.0035	.

Table 5.9: Non-Mendelian - Tier-5 filtered variants. 26 variants were identified as passing filtering. All of the variants are in unique genes, with no gene containing more than one variant. * Reference allele = GGTTCCAGTTAGGCCGTTCCATGAACCCGAGGGCCAGGAAGAAGCGCTCTCGCATACCCTTGCTCCGTAAGCGTC ** Alternate allele = GCCCTTCTCTGTGGCTGCCCTTCTCTGCGGCT

For each of the 26 variants descriptions of genes and phenotype links were investigated using OMIM[223] via BioMart. No gene had more than one mutation. None of the 26 variants were in genes with attributable phenotypes or links to skeletal disease as shown in Table 5.10. A total of 284 variants were in genes listed in the ACMG gene guidelines, none of which were exonic.

Gene name	Gene description	Phenotype description - OMIM
AC016753.2	None	None
ATP5A1	ATP Synthase F1 Subunit Alpha	Combined Oxidative Phosphorylation Deficiency 22
ATP5A1	ATP Synthase F1 Subunit Alpha	Mitochondrial Complex V Deficiency, Nuclear Type 4
ATP6V1A	ATPase H+ transporting V1 subunit A	Autosomal recessive cutis laxa type 2 classic type
BOD1	biorientation of chromosomes in cell division 1	None
CCDC132	VPS50, EARP/GARPII Complex Subunit	None
CNIH3	cornichon family AMPA receptor auxiliary protein 3	None
COG3	component of oligomeric golgi complex 3	None
COPB1	coatomer protein complex subunit beta 1	None
ERVW-1	endogenous retrovirus group W member 1, envelope	None
FP325317.1	None	None
GOLGA6L2	golgin A6 family-like 2	None
GOLGA6L22	golgin A6 family-like 22	None
HSP90AB1	heat shock protein 90 alpha family class B member 1	None
KIF1B	kinesin family member 1B	Autosomal dominant Charcot-Marie-Tooth disease type 2A1
KIF1B	kinesin family member 1B	Pheochromocytoma
KIR2DS4	killer cell immunoglobulin like receptor, two Ig domains and short cytoplasmic tail 4	None
KRTAP4-3	keratin associated protein 4-3	None
LGALS3	galectin 3	None
MAP3K9	mitogen-activated protein kinase kinase kinase 9	None
MCM8	minichromosome maintenance 8 homologous recombination repair factor	Premature ovarian failure 10
RASSF1	Ras association domain family member 1	Lung Cancer Alveolar Cell Carcinoma Included
RGPD4	RANBP2-like and GRIP domain containing 4	None
RPRD1A	regulation of nuclear pre-mRNA domain containing 1A	None
SPDYE5	speedy/RINGO cell cycle regulator family member E5	None
TCF25	transcription factor 25	None
TRBV5-4	T cell receptor beta variable 5-4	None
ZBTB7C	zinc finger and BTB domain containing 7C	None

Table 5.10: Non-Mendelian - Tier-5 filtered gene descriptions and phenotypes associated with the gene as described by OMIM. None of the genes are attributed to skeletal or respiratory disease and two genes (*AC016753*.2 & *FP325317*.1) have yet to characterised.

Compound heterozygotes

For Tier-1 no *GPX** gene variants were found. Tier-2 identified 114 variants matching HGMD terms of which only 51 were exonic, exclusion of synonymous variants left only 25 variants. Variants with intact pair IDs with one variant inherited from each parent were then selected. Only three variant were remaining all in *HSPG2* forming two pairs as shown in Table 5.11. Of the three variants only rs372760688 is from the father SD001 and must be compounded with either of the other two variants, however whilst the variant is rare by gnomAD NFE AAF it was not predicted pathogenic by FATHMM-XF at a score of 0.107.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	ID	Comp. het. pair	rsID	KnownGene	gnomAD NFE	FATHMM-XF
HSPG2	chr1:21834781:C:A	C/A C/C C/A	33 37 70	28 0 62	52258	52258:52259	rs372760688	nonsynonymous SNV	6.67E-05	0.107858
HSPG2	chr1:21839494:G:A	G/G G/A A/G	38 35 48	0 35 57	52264	52258:52264	rs2291827	nonsynonymous SNV	0.1691	0.746044
HSPG2	chr1:21823627:C:T	C/C C/T T/C	47 56 81	0 49 68	52240	52240:52258	rs3736360	nonsynonymous SNV	0.1757	0.126472

Table 5.11: Compound heterozygotes - Tier-2 variants, Three heterozygous variants were identified all in *HSPG2* which form two pairs. Of the three variants only rs372760688 is from the only variant from the father SD001 and would require compounding with either of the other two variants from the mother.

For Tier-3 a total of 553 variants were identified of which 106 were exonic. Exclusion of synonymous variants reduced the total of variants down to 39 which identified four potential pairs over seven variants. Three of the variants were the same *HSPG2* variants as found in Tier-2. The other two pairs were in genes *DLL3* (Delta Like Canonical Notch Ligand 3) and *PCNT* (Pericentrin) as shown in Table 5.12. Both *PCNT* variants are too common at gnomAD NFE AAFs of 0.26 and 0.21. One of the *DLL3* variants rs1110627 is too common also at 0.5549 AAF in the NFE population but its pair has not frequency information and has a low FATHMM-XF pathogenicity score of 0.11.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	ID	Comp. het. pair	rsID	KnownGene	gnomAD NFE	FATHMM-XF
DLL3	chr19:39502943:C:G	C/G C/C C/G	8 19 4	8 0 6	5799622	5799622:5799623	.	nonsynonymous SNV	.	0.110297
DLL3	chr19:39504071:T:C	T/T T/C C/T	37 27 26	0 25 37	5799623	5799622:5799623	rs1110627	nonsynonymous SNV	0.5549	0.036184
HSPG2	chr1:21834781:C:A	C/C C/T T/C	47 56 81	0 49 68	52240	52258:52259	rs372760688	nonsynonymous SNV	6.67E-05	0.107858
HSPG2	chr1:21839494:G:A	C/A C/C C/A	33 37 70	28 0 62	52258	52258:52264	rs2291827	nonsynonymous SNV	0.1691	0.746044
HSPG2	chr1:21823627:C:T	G/G G/A A/G	38 35 48	0 35 57	52264	52240:52258	rs3736360	nonsynonymous SNV	0.1757	0.126472
PCNT	chr21:46443844:A:C	C/T C/C C/T	25 37 21	23 0 28	6132307	6132307:6132419	rs2070425	nonsynonymous SNV	0.2647	0.027976
PCNT	chr21:46416739:C:T	A/A A/C C/A	34 10 21	0 14 12	6132419	6132411:6132419	rs2073380	nonsynonymous SNV	0.21	0.042041

Table 5.12: Compound heterozygotes - Tier-3 variants, *HSPG2* were also reported in Tier-2. From other variants only the genes *DLL3* and *PCNT* have heterozygotes from both parents making them possible compound heterozygotes but have high allele frequencies and are not predicted pathogenic by FATHMM-XF.

Tier 4 identified only 32 variants of which 18 were exonic, removal of synonymous variants lowered this total to 5 variants. Of the 5 variants *DYNC2H1* was the only gene with 2 variants but were both shared with the unaffected mother, so no variant pairs were intact to be compounding.

For Tier-5 filtering using a gnomAD NFE AAF of 0.1 or below or missing reduced variants from 22,248 to 9,257. Exclusion of genes listed under the Fuentes false positives list reduced variants further to 4,927. Selection of exonic, splicing left 1,957 variants. Removal of synonymous variants reduced variants to 1,036. Using a FATHMM-XF scores of 0.5 or above or missing and selecting intact pairs left only 18 pairs totalling 30 variants covering 12 genes as summarised in Table 5.13.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	ID	Comp. het. ID	rsID	KnownGene	gnomAD NFE	FATHMM-XF
C9orf171	chr9:132499377:G:A	G/G G/A A/G	38 51 71	0 57 60	3596630	3596630:3596717	rs7047726	nonsynonymous SNV	0.0836	0.52233
C9orf171	chr9:132537613:N:-	N/C N/N N/C	20 35 11	19 0 18	3596717	3596630:3596717	rs542793318	nonframeshift deletion	0.0063	.
CES1	chr16:55811012:-:A	T/TA T/T T/TA	55 28 29	11 0 14	5263986	5263986:5264044	rs11307366	Splicing	0.0646	.
CES1	chr16:55821449:G:T	G/G G/T T/G	46 38 110	0 29 47	5264053	5263986:5264053	rs2307227	nonsynonymous SNV	0.0238	0.563911
CES1	chr16:55819586:G:C	G/G G/C C/G	41 37 102	0 28 35	5264043	5263986:5264043	rs114119971	nonsynonymous SNV	0.0062	0.527279
CNTNAP3B	chr9:42129091:C:A	C/C C/A A/C	15 10 10	0 3 7	3438586	3438585:3438586	rs62558062	nonsynonymous SNV	.	.
CNTNAP3B	chr9:42129032:G:T	G/T G/G G/T	5 6 9	9 0 6	3438585	3438585:3438586	rs372133350	nonsynonymous SNV	.	.
CNTNAP3B	chr9:41964577:A:C	A/C A/A A/C	13 35 25	14 0 8	3438137	3437818:3438137	rs141780724	nonsynonymous SNV	.	.
CRLF2	chrX:1193297:T:C	T/C T/T T/C	10 20 6	15 0 12	6248706	6248706:6248747	.	nonsynonymous SNV	.	.
CRLF2	chrX:1196817:C:T	C/C C/T T/C	35 23 28	0 25 35	6248747	6248706:6248747	rs151218732	nonsynonymous SNV	.	.
FTCD	chr21:46154209:C:G	C/G C/C C/G	41 37 31	35 0 23	6131318	6131311:6131318	.	nonsynonymous SNV	.	0.872734
FTCD	chr21:46145925:-:C	G/G G/GC GC/G	17 5 12	0 8 8	6131294	6131294:6131318	rs35208133	frameshift insertion	0.0056	.
GMPPA	chr2:219501888:G:T	G/G G/T T/G	34 55 50	0 49 61	951205	951205:951214	rs753112469	nonsynonymous SNV	6.68E-05	0.968875
GMPPA	chr2:219506305:C:T	C/T C/C C/T	39 38 48	31 0 45	951214	951205:951214	rs752643023	nonsynonymous SNV	.	0.871421
GMPPA	chr2:219501858:C:A	C/C C/A A/C	34 40 43	0 37 49	951204	951202:951204	rs767718283	nonsynonymous SNV	6.68E-05	0.961946
LILRB2	chr19:54279520:A:T	A/T A/A A/T	66 38 71	51 0 57	5837183	5837183:5837210	rs373032	nonsynonymous SNV	.	.
LILRB2	chr19:54279046:G:-	AG/AG AG/A A/AG	109 81 113	8 21 28	5837176	5837152:5837176	.	stopgain	.	.
MYOF	chr10:93333862:G:A	G/G G/A A/G	37 46 68	0 61 53	3835769	3835769:3835770	rs61861290	nonsynonymous SNV	0.0143	0.929518
MYOF	chr10:93333904:G:A	G/A G/G G/A	37 44 44	33 0 37	3835770	3835769:3835770	rs201634420	nonsynonymous SNV	0.0035	0.605349
MYOF	chr10:93347671:G:A	G/G G/A A/G	26 15 22	0 31 24	3835825	3835770:3835825	rs11187393	nonsynonymous SNV	0.0545	0.512846
NPFFR1	chr10:70255211:G:T	G/T G/G G/T	35 38 53	34 0 39	3792341	3792336:3792341	rs113487866	nonsynonymous SNV	0.039	0.622632
NPFFR1	chr10:70255817:T:G	T/T T/G G/T	37 15 22	0 21 25	3792343	3792341:3792343	rs3812694	nonsynonymous SNV	0.0613	0.898473
NVL	chr1:224300615:C:T	C/T C/C C/T	56 37 66	52 0 69	444873	444866:444873	.	nonsynonymous SNV	.	0.938902
NVL	chr1:224294382:C:T	C/C C/T T/C	40 39 41	0 21 36	444866	444866:444873	rs34631151	nonsynonymous SNV	0.037	0.652873
SDHA	chr5:236561:G:A	G/G G/A A/G	66 65 72	5 6 12	1872849	1872846:1872849	rs138277996	nonsynonymous SNV	6.68E-05	0.925202
SDHA	chr5:236513:C:T	C/T C/C C/T	62 67 72	5 4 7	1872846	1872846:1872849	rs201139275	nonsynonymous SNV	0	0.862645
SDHA	chr5:236504:T:C	T/C T/T T/C	54 60 73	4 4 8	1872845	1872845:1872849	rs201741295	nonsynonymous SNV	0	0.8268
SDHA	chr5:236534:C:T	C/T C/C C/T	67 78 74	9 5 9	1872847	1872847:1872849	rs76896145	nonsynonymous SNV	0	0.939419
TPTE	chr21:10552686:T:A	T/T T/A A/T	42 87 139	0 28 45	6041796	6041796:6041865	rs111657679	nonsynonymous SNV	.	.
TPTE	chr21:10569527:N:-	N/T N/N N/T	98 125 130	26 0 26	6041865	6041796:6041865	rs113444703	nonframeshift deletion	.	.

Table 5.13: Compound heterozygotes - Tier-5 genes identified as having multiple heterozygous variants after filtering with one variant inherited from each parent in the affected child. 18 variant pairs were identified totalling 30 variants distributed over 12 genes.

For each of the 12 genes in Table 5.13 the reported phenotype was queried to try and identify the more likely candidates as many variants lacked any annotations to prioritise further upon.

Gene name	Gene description	Phenotype description - OMIM
CES1	carboxylesterase 1	None
CNTNAP3B	contactin associated protein like 3B	None
CRLF2	cytokine receptor like factor 2	None
FTCD	formimidoyltransferase cyclodeaminase	Glutamate Formiminotransferase Deficiency
GMPPA	GDP-mannose pyrophosphorylase A	Glycosylation Disorder Characterized By Intellectual Disability And Autonomic Dysfunction
LILRB2	leukocyte immunoglobulin like receptor B2	None
NPFFR1	neuropeptide FF receptor 1	None
NVL	nuclear VCP-like	None
SDHA	succinate dehydrogenase complex flavoprotein subunit A	Leigh Syndrome
TPTE	transmembrane phosphatase with tensin homology	None

Table 5.14: Compound heterozygotes - Tier-5 gene descriptions and associated phenotypes. None of the the genes with heterozygotes from both parents were strong candidates to match with the observed phenotype of SD003.

36 variants were in genes listed in the ACMG guidelines, 28 of which were exonic, removing duplicated and synonymous left only five variants with three intact pairs all in the gene *APOB* shown below in Table 5.15. Two of the variants are listed as benign or likely benign in ClinVar both inherited from the father arguing against the variants being compounded for the gene *APOB*.

Gene	Variant	GT:SD00-1 2 3	gt ref depths	gt alt depths	ID	Comp. het. pair	rsID	KnownGene	gnomAD NFE	FATHMM-XF	ClinVar
APOB	chr2:21011127:T:C	T/C T/T T/C	94 36 131	95 0 107	550122	550117:550122	rs1801699	nonsynonymous SNV	0.0157	0.856306	Benign
APOB	chr2:21005955:C:T	C/T C/C C/T	33 35 46	42 0 47	550116	550116:550117	rs1801701	nonsynonymous SNV	0.0849	0.160134	Likely benign
APOB	chr2:21008406:G:A	G/G G/A A/G	66 65 86	0 55 63	550117	550117:550122	rs72653095	nonsynonymous SNV	0.0036	0.493991	Uncertain significance

Table 5.15: Compound heterozygotes Tier 5 - Three ACMG gene variants *APOB* were the only intact compound pairs when selecting exonic variants. Both variants from the father are annotated as benign or likely benign by ClinVar arguing against pathogenicity of either variant and compounding in the affected child.

5.4.5 Splicing variants

From the combined dataset a total of 147,155 variants had a reported score for SPIDEX, indicating the variants had effect of splicing variants. Filters were applied to identify any splicing variants in *GPX** genes, nine variants were identified shown below in Table 5.16. None of the variants were below a gnomAD NFE AAF of 0.05 or had a SPIDEX $|\Delta PSImax| \geq 5$. Only one variant, 6:28531790:T:C, in *GPX5* had a FATHMM-XF score above 0.5 but was only heterozygous in the unaffected sample SD001.

Gene	Variant	KnownGene	avsnp147	gnomAD NFE	SPIDEX-DPSI max	SPIDEX-Z-score	FATHMM-XF	GT:SD00-1, 2, 3
GPX4	19:1106478:G:C	intronic	rs8178977	0.2516	0.3872	1.006	0.050259	0/0 0/1 0/1
GPX7	1:52606782:T:C	exonic-synonymous	rs1970951	0.8346	0.2026	0.727	0.003807	1/1 1/1 1/1
GPX7	1:52607069:A:G	intronic	rs1970950	0.734	0.1292	0.565	0.042038	0/1 1/1 0/1
GPX7	1:52607124:G:C	intronic	rs1970949	0.7344	0.0495	0.296	0.020671	0/1 1/1 0/1
GPX5	6:28529213:T:A	intronic	rs113787419	0.0431	-0.6696	-1.245	0.028137	0/1 0/0 0/0
GPX5	6:28531790:T:C	exonic-nonsynonymous	rs58554303	0.0273	0.0442	0.271	0.583127	0/1 0/0 0/0
GPX3	5:151025096:G:A	intronic	rs3763012	0.2158	0.4702	1.102	0.019807	0/0 0/1 0/0
GPX3	5:151025142:G:A	intronic	rs3763011	0.2151	-0.2397	-0.758	0.014144	0/0 0/1 0/0
GPX3	5:151026811:G:A	intronic	rs869975	0.0745	0.1543	0.627	0.038721	0/0 0/1 0/0

Table 5.16: Splicing variants in *GPX** genes as annotated by SPIDEX, none of the variants were suggested as being significantly altering splicing with a $|\Delta PSImax| \geq 5$.

Using the extended candidate gene list of all 430 genes a total of 4,361 variants were identified with a recorded SPIDEX score, none of which had a SPIDEX $|\Delta PSImax| \geq 5$ making the effects on splicing not likely pathogenic.

Each of the GEMINI results files were then filtered to identify any variants with a $|\Delta PSImax| \geq 5$. For Autosomal recessive 49 variants were identified none of which were below a gnomAD NFE AAF of 0.05. No intact compound heterozygotes pairs were also found when using a 0.1 filter for gnomAD NFE AAF used in other compound heterozygote AAF filters.

For non-Mendelian *de-novo* a total of 31 variants were found with a $|\Delta PSImax| \geq 5$. 10 of the variants were below a gnomAD NFE AAF of 0.05 and also passed the violation probability of above 0.99. Four Variants in the gene *MUC19* were excluded as they are from mucosal genes with high false positive rates. The remaining six variants are shown in Table 5.17.

Gene	Variant	gts	gt ref depths	gt alt depths	Function	avsnp147	gnomAD NFE	spdiex dpsi max	spdiex Z score	fathmm xf coding	fathmm xf noncoding
AKR1C2	chr10:5003711:T:A	T/T T/T T/A	63 39 57	0 0 8	intronic	rs200856918	0.05	-7.8	-2.77	.	0.03
VPS50	chr7:93294638:T:G	T/T T/T T/G	74 72 90	18 8 32	splicing	rs75893203	0.02	-6.04	-2.62	.	0.99
CNIH3	chr1:224684802:C:T	C/C C/C C/T	40 38 92	0 0 19	exonic	.	.	-8.67	-2.84	0.5	.
CROCC	chr1:16944285:G:T	G/G G/G G/T	42 70 68	0 8 18	intronic	rs6684687	0	-18.84	-3.17	.	0.98
LILRA2	chr19:54575375:G:T	G/G G/G G/T	38 67 56	0 0 7	exonic	.	.	-73.37	-3.88	0.08	.
NBPF9	chr1:149058931:C:A	C/C C/C C/A	5 11 6	0 0 2	exonic	.	.	-5.48	-2.56	0.03	.

Table 5.17: Splicing variants as annotated by SPIDEX which are also *de novo* mutations genes with a $|\Delta PSImax| \geq 5$. A total of six variants were identified in six genes.

5.4.6 Non-coding variants

Each of the GEMINI files were filtered using the FATHMM-XF non-coding high confidence pathogenic threshold of 0.96 to identify potentially pathogenic non-coding variants. No autosomal recessive variants were found below a gnomAD NFE AAF of 0.05 or compound heterozygotes below 0.1 gnomAD NFE AAF. For non-Mendelian variants three variants were above the 0.96 FATHMM-XF non-coding threshold after applying the gnomAD NFE AAF below 0.05 and violation filter of 0.99 or above as shown in Table 5.18 The three non-Mendelian variants are in the genes *VPS50*, *CROCC* and *COPB1*.

Gene	Variant	gts	violation prob	Function	avsnp147	gnomAD NFE	spdiex dpsi-max	spidex Z-score	fathmmxf noncoding
VPS50	chr7:93294638:T:G	T/T T/T T/G	0.99985	splicing	rs75893203	0.022	-6.037	-2.62	0.992
CROCC	chr1:16944285:G:T	G/G G/G G/T	0.99793	intronic	rs6684687	0.001	-18.84	-3.166	0.976
COPB1	chr11:14479716:T:A	T/T T/T T/A	1	splicing	rs753555066	0.009	-1.132	-1.546	0.99

Table 5.18: FATHMM-XF non-coding variants passing filtering were identified from the GEMINI non-Mendelian output in three genes: *VPS50*, *CROCC* and *COPB1*.

5.4.7 Structural & copy number variants

LUMPY

Structural variants were called for all three samples using WGS data as discussed in Chapter 4. Variant calls were checked for any overlap with *GPX4* however none were detected in any sample. Filters were then applied to genotypes to identify possible variants of interest. In SD003 there were 57,681 variants called, from which 946 were identified as overlapping an exonic gene region. The 946 variants were then filtered to identify autosomal recessive variants. 27 variants were identified as shown below in Table 5.19.

Gene(s)	CHR	POS	END	SV-TYPE	SV-LENGTH	EVIDENCE	GT:SD00-1 2 3
LCE3C,LCE3B	1	152583065	152615321	DEL	-32,256	14	0/1 0/1 1/1
HLA-DRB5	6	32480913	32524704	DUP	43,791	8	0/1 0/0 1/1
HLA-A	6	29934324	29943681	DUP	9,357	12	0/1 0/1 1/1
HLA-DRB5,HLA-DRB1	6	32482947	32590565	DUP	107,618	27	0/1 0/1 1/1
HLA-DRB5	6	32505774	32544492	DEL	-38,718	36	0/1 0/1 1/1
HLA-DRB5	6	32518734	32554066	DUP	35,332	4	0/1 0/1 1/1
HLA-DRB5	6	32522915	32558938	DUP	36,023	35	0/1 0/1 1/1
HLA-DRB5	6	32523066	32559173	DEL	-36,107	36	0/1 0/1 1/1
HLA-DRB5	6	32523583	32559717	DUP	36,134	26	0/1 0/1 1/1
HLA-DRB5	6	32527106	32570789	DUP	43,683	8	0/1 0/1 1/1
HLA-DRB5	6	32528079	32571092	DEL	-43,013	5	0/1 0/1 1/1
HLA-DRB5	6	32528225	32571260	DUP	43,035	82	0/1 0/1 1/1
HLA-DRB5,HLA-DRB1	6	32528848	32588458	DUP	59,610	41	0/1 0/1 1/1
HLA-DRB1	6	32558480	32584924	DUP	26,444	9	0/1 0/1 1/1
TRGV5,TRGV4,TRGV3	7	38349652	38358764	DUP	9,112	6	0/0 0/0 1/1
Too many to display	10	46782619	87084562	DEL	-40,301,943	6	0/1 0/1 1/1
Too many to display	11	48536350	55339113	DUP	6,802,763	15	0/1 0/1 1/1
OR8G1	11	124250540	124251389	DEL	-849	80	0/1 0/1 1/1
PIK3C2G	12	18282465	18282467	DEL	-2	41	0/1 0/1 1/1
TPTE2,MPHOSPH8	13	19200585	19702108	DEL	-501,523	33	0/1 0/1 1/1
AHNAK2	14	104948922	104949352	DEL	-430	7	0/1 0/1 1/1
IGHV1-58,IGHV4-59,IGHV4-61,IGHV3-64,IGHV3-66	14	106594026	106676410	DUP	82,384	33	0/1 0/1 1/1
LCMT1	16	25055236	25126491	DUP	71,255	40	0/1 0/1 1/1
Too many to display	19	15586262	33263086	DUP	17,676,824	5	0/1 0/1 1/1
SMC1B	22	45349575	45350551	DUP	976	12	0/1 0/0 1/1
ADM2	22	50482734	50482737	DEL	-3	25	0/1 0/0 1/1
Too many to display	X	33481693	155835875	DUP	122,354,182	10	0/1 0/1 1/1

Table 5.19: Homozygous alternate variants called in SD003 by LUMPY, none of the genes which were overlapped by variants were skeletal associated. Multiple of the variants identified spanned over several megabases but no evidence of the variant spanning the entire of the described regions could be found. At the end of each variant the split or paired end reads could be identified but they were often too few to accurately call multi-Mb variants.

Four of the variants identified from LUMPY were improbably large: chr19:15586262 -

33263086 (17.6 MB Dup), chr10:46782619 - 87084562 (40. 3MB Del), chr11:48536350 - 55339113 (6.8 MB Dup) and chrX:33481693 - 155835875 (122.3 MB Dup) which between them covered 452 genes. The evidence from LUMPY for each of the variants is also low for the size of the supposed variants.

All genes overlapped by the other 24 variants were annotated with MIM morbid descriptions as shown below in Table 5.20. As shown in the Table below with MIM morbid descriptions of the genes with homozygous variants none have disease associations with similar phenotypes to this case. Most of the genes identified are associated with the immune system (HLA-*, IGHV*) or olfactory receptors (OR4*) where variants would be expected linked to the functions of those genes supporting the lack of good homozygous variant candidates.

Gene name	Gene description	MIM morbid description	MIM ID
HLA-A	major histocompatibility complex, class I, A	SUSCEPTIBILITY TOTOXIC EPIDERMAL NECROLYSIS	608579
SMC1B	structural maintenance of chromosomes 1B	.	.
ADM2	adrenomedullin 2	.	.
LCE3C	late cornified envelope 3C	.	.
LCE3B	late cornified envelope 3B	.	.
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	SARCOIDOSIS, SUSCEPTIBILITY TO,	181000
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	RHEUMATOID ARTHRITIS; SUSCEPTIBILITY TO	180300
HLA-DRB1	major histocompatibility complex, class II, DR beta 1	MULTIPLE SCLEROSIS, SUSCEPTIBILITY TO	126200
LCMT1	leucine carboxyl methyltransferase 1	.	.
IGHV1-58	immunoglobulin heavy variable 1-58	.	.
IGHV4-61	immunoglobulin heavy variable 4-61	.	.
IGHV4-59	immunoglobulin heavy variable 4-59	.	.
IGHV3-64	immunoglobulin heavy variable 3-64	.	.
IGHV3-66	immunoglobulin heavy variable 3-66	.	.
TRGV5	T-cell receptor gamma variable 5	.	.
TRGV3	T-cell receptor gamma variable 3	.	.
TRGV4	T-cell receptor gamma variable 4	.	.
TPTE2	transmembrane phosphoinositide 3-phosphatase and tensin homolog 2	.	.
TRIM49B	tripartite motif containing 49B	.	.
TRIM64C	tripartite motif containing 64C	.	.
MPHOSPH8	M-phase phosphoprotein 8	.	.
FOLH1	folate hydrolase 1	.	.
PIK3C2G	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 gamma	.	.
OR4C13	olfactory receptor family 4 subfamily C member 13	.	.
OR4C12	olfactory receptor family 4 subfamily C member 12	.	.
OR4C46	olfactory receptor family 4 subfamily C member 46	.	.
OR4A8	olfactory receptor family 4 subfamily A member 8 (gene/pseudogene)	.	.
OR4A5	olfactory receptor family 4 subfamily A member 5	.	.
TRIM48	tripartite motif containing 48	.	.
HLA-DRB5	major histocompatibility complex, class II, DR beta 5	.	.
AHNAK2	AHNAK nucleoprotein 2	.	.
OR8G1	olfactory receptor family 8 subfamily G member 1 (gene/pseudogene)	.	.

Table 5.20: Gene descriptions and associated phenotypes for genes overlapping with LUMPY called homozygous alternate variants. None of the genes identified were skeletal associated, most were either olfactory receptor or immune related.

Analysis of *de novo* structural variants in SD003 found 134 variants. 17 variants were larger than 10 Mb ranging from 19.3 Mb up to 91.1 Mb as shown in Appendix C, Table 8.9. No CNVs of this size were reported by CNVkit or array CGH arguing for these variants as false positives. The remaining 117 variants are split as: 65 deletions, 47 duplications and 5 inversions. To reduce the 117 variants all genes which variants overlapped were annotated to identify any overlapping candidate genes as shown in Table 5.21. Matches were identified for deletions in the genes *WDR60*, *SOST* and *XYLT1*. Gene duplications were called for *BMPER* and *SOST*. No genes of interest were identified from inversions. In total seven variants were identified over the four genes split as four deletions and three duplications.

Gene of Interest	Chr	Start	End	SV-Type	SV-length	Evidence	GT:SD00-1 2 3	Total no. of genes overlapped
XYLT1	16	15180157	18091724	DEL	-2911567	12	0/0 0/0 0/1	14
SOST	17	20740802	59950511	DUP	39209709	5	0/0 0/0 0/1	559
SOST	17	37872539	47455054	DUP	9582515	6	0/0 0/0 0/1	262
SOST	7	6842271	97951122	DEL	-91108851	5	0/0 0/0 0/1	396
BMPER	7	32635461	34969729	DUP	2334268	17	0/0 0/0 0/1	8
BMPER	7	32674049	35008540	DEL	-2334491	9	0/0 0/0 0/1	8
WDR60	7	158871384	158871386	DEL	-2	18	0/0 0/0 0/1	1

Table 5.21: LUMPY *de-novo* variants which overlapped four candidate genes. In total seven variants were identified split as four deletions and three duplications.

Only one of the heterozygote variants identified was clearly visible, this variant was called in *WDR60* as a two base deletion (chr7:158871384:N:-), this variant was investigated from the GATK calls as the deletion was small enough to be detected. The variant was also detected but was called as a non-frameshift deletion 7:158871384:158871386:GAA:- . A visualisation of the site shows that the deletion does actually appear to be 2bp, so the variant would be a frameshift deletion as shown in Figure 5.6. Pathogenicity is supported by *de novo* calculation from CADD which scored the INDEL using hg19 (7:158664075:GAA:G) as a raw score of 1.93 and phred-scaled score of 18.55.

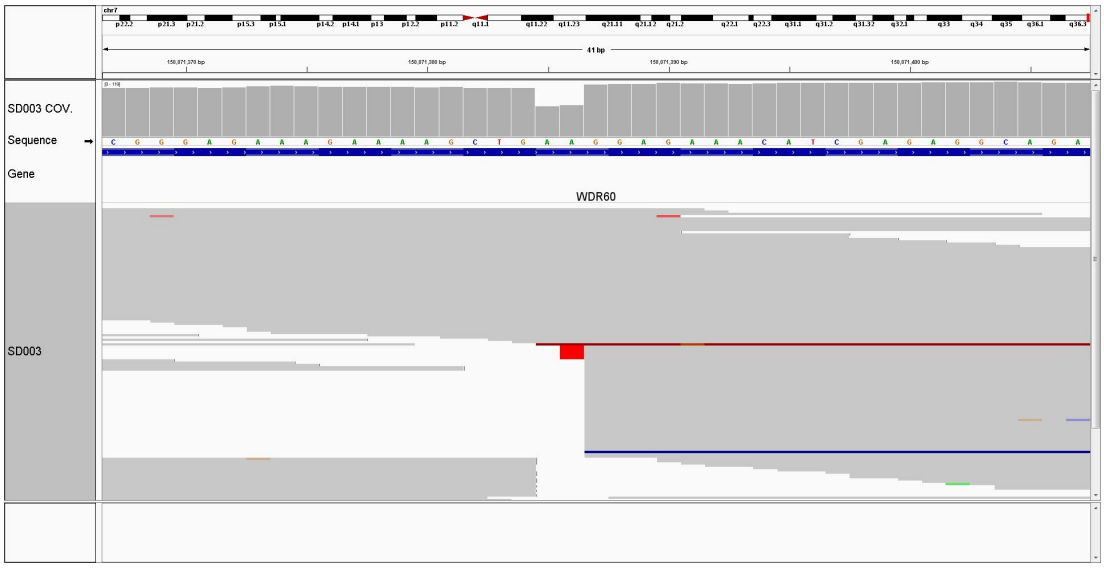

Figure 5.6: *WDR60* frameshift deletion called by LUMPY but as a non-frameshift deletion of GAA by GATK. Visually the deletion appears as a heterozygous two base deletion which is predicted pathogenic by CADD phred at 18.55 when using a pathogenicity cut-off of 15.

CNVkit

CNVkit results were analysed to identify non-neutral segment calls, first in WES for the sample SD003. Segments called as losses on chromosome Y or the length of X and alternate assemblies were ignored for the female samples leaving a total of nine non-neutral segments, five losses and four gains of copy as shown in Table 5.22. None of the variants called were found in parental samples or to overlap any of the 430 candidate genes.

Chr	start	end	log2	copy-number	depth	probes	weight
7	38353720	38354113	-25.7363	0	0	2	1.06481
7	144186908	144377327	0.543875	3	134.45	65	29.1501
8	7428949	8339682	0.206885	3	127.237	83	41.9722
12	52469015	52473779	-0.802813	1	58.8247	12	5.18401
13	88749578	90585242	0.575081	3	0.341835	11	7.68026
14	16053768	16053893	-0.417897	1	25.808	1	0.560284
15	30211343	30517664	-0.446799	1	57.311	15	8.11059
15	30518164	32490316	-0.970785	1	38.541	172	100.692
16	70120485	70177461	0.471669	3	189.084	32	16.4722

Table 5.22: CNVkit SD003 exome sequencing variants overlapping genes. Nine variants were identified including two deletions overlapping a known micro-deletion on chromosome 15 band q13.3 which was also identified previously by microarray for the sample SD003.

Many more non-neutral segments were identified in WGS for SD003, using the filtering criteria as used with WES identified 124 non-neutral segments. A filter was applied to remove variants with no overlap with of genes which reduced the total to 72 shown in Appendix C, Table 8.8. All genes overlapping CNVs were then screened against the extended candidate gene list which yielded no matches to any of the 430 genes except as with WES to the 15q13.3 region associated with disease SRTD1 (See Table 5.1).

Chr	start	end	log2	copy-number	depth	probes	weight
15	30105042	30379320	-0.330349	1	18.5934	143	113.769
15	30379320	30611401	-0.566578	1	15.0899	121	95.3238
15	30611401	32153493	-1.0079	1	11.1615	804	655.421
15	32153493	32494901	-0.556299	1	15.3231	178	140.214
15	32494901	32583130	-0.259396	1	19.7698	46	36.2495

Table 5.23: CNVkit WGS variants overlapping the 15q13.3 micro-deletion. More segments were called over the micro-deletion than by WES.

CNVkit was able to detect the deletion at the 15q13.3 loci in both the WES and WGS data using either a flat or pooled reference sequence as shown in Figure 5.7. In the figure segment calls with a log2 call below -0.4 called as a heterozygous loss and below -1.1 a homozygous loss. Both parents in Panel A for WGS data and Panel B for WES data show no obvious deletion in the parental samples SD001 and SD002 but a deletion in the BP4-5 region is evident from 30.8 to 32.6 Mb for SD003. Greater resolution is obtained in WGS due to coverage over the entire region instead of targeting exons, consequently there were more segment calls at the extremities of the deletion which better help resolve the start and end of events.

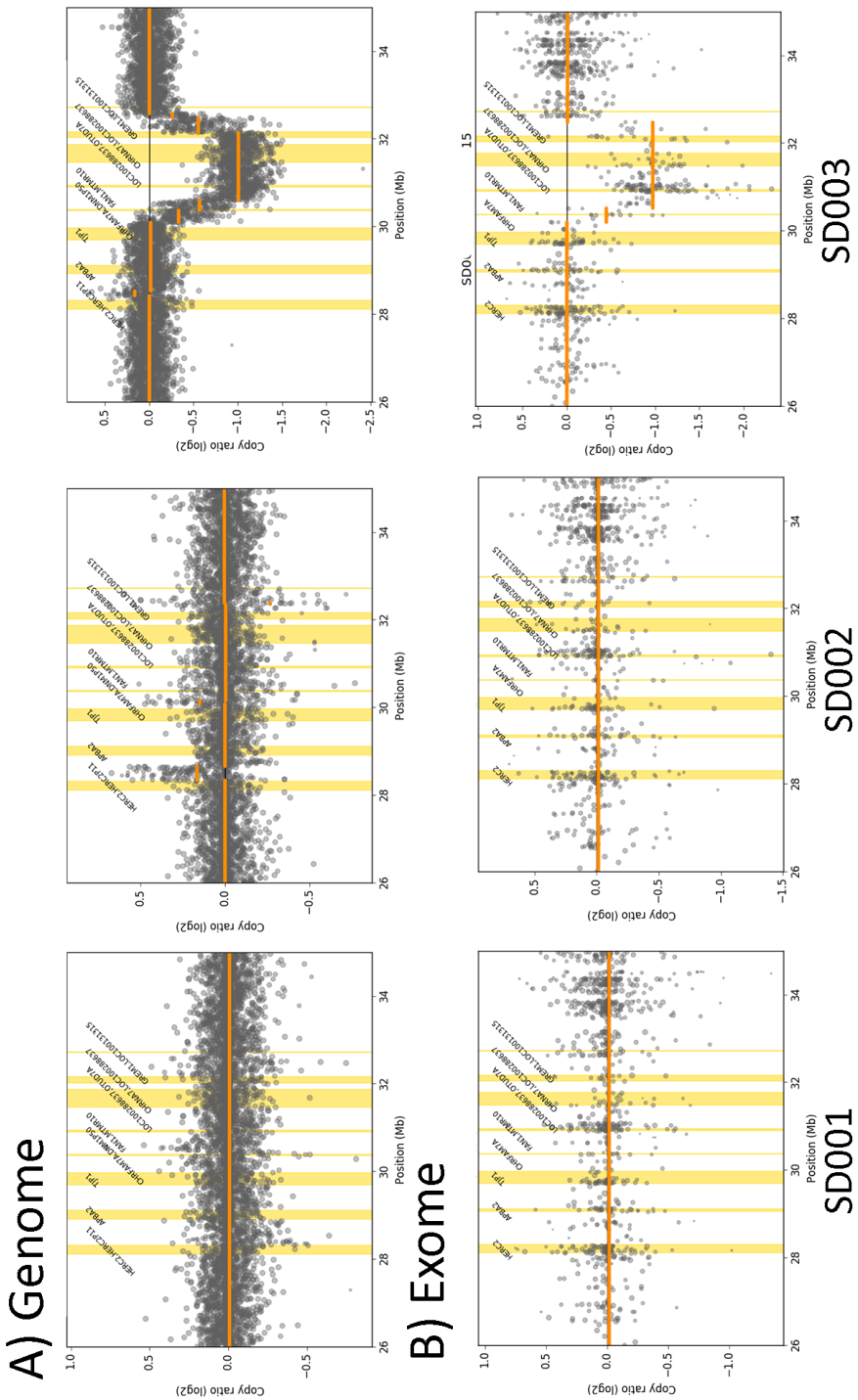

Figure 5.7: CNVkit calling of the 15q13.3 micro-deletion in all samples. In both exome and genome samples no micro-deletion was visible in SD001 or SD002 whilst a deletion was called in SD003. **A)** Whole genome sequencing shows the deletion in SD003 with 2 step segments at either side of the the deletion. **B)** Whole exome sequencing shows the deletion in SD003 though has fewer points and only calls one step segment on each side of the deletion.

Visualisation in IGV for SD003 over the affected region on chromosome 15 highlights three regions of abnormalities when a mapping quality of 0 is applied in WES and WGS as shown in Figure 5.8. These deviations are only visible around each of the breakpoints when the mapping quality filter is lowered to zero indicating the high homology between the these regions which would impair variant calling as discussed in more detail in Chapter 6. A deviation in the depth of coverage is visible between breakpoints 4 and 5 suggesting a loss of copy or deletion event.

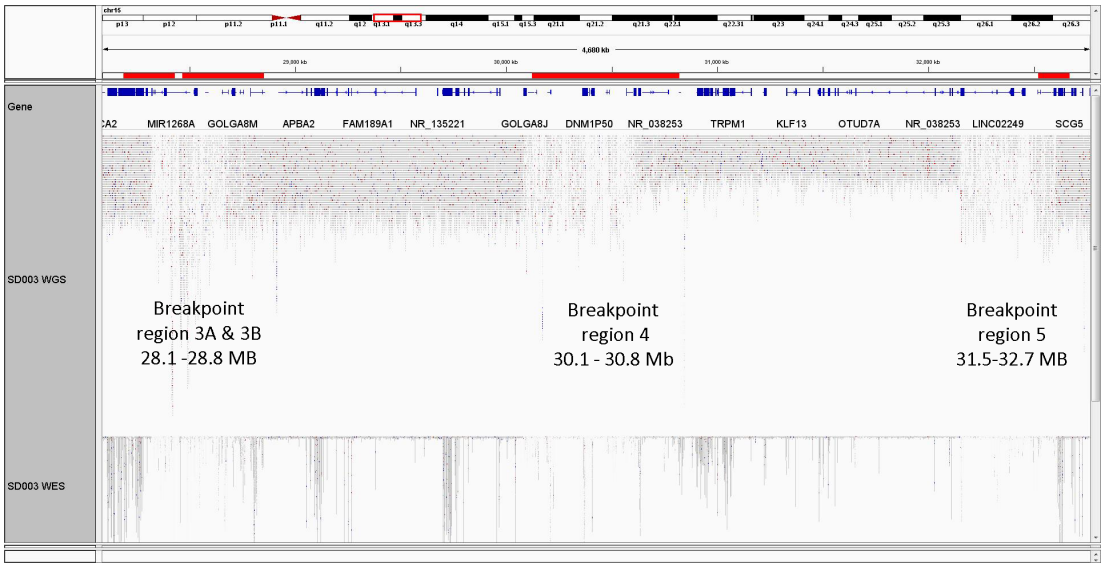

Figure 5.8: 15q13.3 micro-deletion in sample SD003 visualised in IGV with a mapping quality filter set to 0. Around each of the breakpoints patchy coverage is observed where due to the high sequence homology between the breakpoints it is not possible to align reads uniquely. The region in-between breakpoints 4 and 5 can be seen to be at lower depth than the surrounding regions which is prone to deletion with an inversion between breakpoints 4 and 5.

LUMPY fails to call an inversion in the 15q13.3 region, which is suggested to cause the loss of th region due to the loss of stability. All of the LUMPY calls over the 15q13.3 region are summarised in Table 5.24. Removing variants where parents were homozygous alternate reduced the number of variants to 26. 10 of the 26 were breakends with eight describing translocations between the breakpoints in the region. Three duplications were called, the largest of which is 1.99MB which encompasses the BP4-5 region which does not fit with the visualisation shown in Figure 5.8. The duplication called for 30,846,749 - 30,847,025 is debatable as with the duplication 31,391,499 - 31,392,728 which also cannot be easily identified from Figure 5.8.

POS	END	ID	MATEID	ALT	SVTYPE	SVLEN	Evidence	GT:SD00-1 2 3
28136307	28136307	98745-1	98745-2	N]chr15:30843417]	BND	.	20	0/1 0/1 0/1
28146403	28146403	98746-1	98746-2	[chr15:30832644[N	BND	.	21	0/1 0/1 0/1
28147445	28147445	98747-1	98747-2	N]chr15:30831705]	BND	.	34	0/1 0/1 0/1
28175205	28175207	32857	.	DEL	DEL	-2	17	0/1 0/0 0/1
28512050	28512050	98743-2	98743-1	[chr15:23178289[N	BND	.	7	0/1 0/1 1/1
28987103	28987105	32860	.	DEL	DEL	-2	17	0/0 0/1 0/1
29245500	29245502	32863	.	DEL	DEL	-2	16	0/0 0/0 0/1
29581067	29581069	32877	.	DEL	DEL	-2	13	0/0 0/1 0/1
29584192	29584194	32878	.	DEL	DEL	-2	18	0/0 0/1 0/1
29657979	29657981	32883	.	DEL	DEL	-2	7	0/1 0/0 0/1
29922965	29922967	32890	.	DEL	DEL	-2	7	0/0 0/1 0/1
30013404	30013406	32893	.	DEL	DEL	-2	22	0/1 0/1 1/1
30173089	30173089	136-2	136-1	N[chr1:202624589[	BND	.	4	0/1 0/1 0/1
30173126	30173126	137-2	137-1	]chr1:202625106]N	BND	.	28	0/1 0/1 0/1
30376137	30376137	98748-1	98748-2	N]chr15:32154879]	BND	.	29	0/1 0/1 0/1
30638153	32628195	96661	.	DUP	DUP	1990042	4	0/0 0/0 0/1
30831705	30831705	98747-2	98747-1	N]chr15:28147445]	BND	.	34	0/1 0/1 0/1
30832644	30832644	98746-2	98746-1	[chr15:28146403[N	BND	.	21	0/1 0/1 0/1
30843417	30843417	98745-2	98745-1	N]chr15:28136307]	BND	.	20	0/1 0/1 0/1
30846749	30847025	96662	.	DUP	DUP	276	5	0/1 0/0 1/1
31351792	31351794	32908	.	DEL	DEL	-2	13	0/0 0/0 1/1
31391499	31392728	96663	.	DUP	DUP	1229	4	0/1 0/0 0/1
31842622	31842624	32923	.	DEL	DEL	-2	18	0/0 0/1 1/1
32131968	32131970	32933	.	DEL	DEL	-2	14	0/0 0/0 1/1
32659213	32659215	32934	.	DEL	DEL	-2	11	0/0 0/1 0/1
32768774	32768776	32938	.	DEL	DEL	-2	24	0/0 0/1 0/1

Table 5.24: LUMPY variants called over the 15q13.3 region, all deletions called were too small to match CNVkit or microarray results. Duplications were also called in the intervening region where deletions should be called. Several breakends were detected across the region which suggest some cross-chromosomal events but most interestingly some of the variants are proximal to the breakpoint positions. Due to use of short read mapping the ability to resolve breakend events and their importance other than visualisation is currently limited.

Cross-dataset heterozygotes

As multiple analyses have been performed each have identified several heterozygous variants which suggested pathogenicity but did not fit with a recessive pattern or did not have a second variant in the gene to compound the first variant. It is also believed that SRPS, see Table 5.1, can be caused in a di-genic inheritance pattern which would not have been reported by a single tool, and alleles could be at a higher frequency since a single gene mutation may be tolerated.

No variants were found to overlap candidate genes for CNVkit WES or WGS or SPIDEX using the combined dataset, ruling out the variants from the tools as being important for di-genic SRPS CNVs or splice variants. The only tool (other than GEMINI) which identified variants overlapping candidate genes was LUMPY as was shown in Table 5.25. The variant most of interest was the potential frameshift deletion in *WDR60* which is exonic and only seen in the child. Other variants from LUMPY which overlapped genes could not be seen visually to support the calls over sizes ranging from 2.3 to 9.5 MB.

All heterozygous variants were extracted from the GATK calls. 9,557 variants were called, 87 were exonic or splicing, removing homozygous alternate and SD003 reference variants reduced the total to 33. Only one of the variants were predicted pathogenic by FATHMM-XF, a splicing variant also in *WDR60* (chr7:15886949:G:A). None of the exonic splice variants listed by Known Gene were predicted to be damaging to splicing with SPIDEX. The *WDR60* variant was not predicted to alter splicing in *WDR60* also by MaxEntScan annotations and was found to have an AAF of 0.18 in the NFE population by gnomAD.

Variant	Function	Gene	avsnp147	gnomAD NFE	fathmmxf nc	GTs	MaxEnt alt	MaxEnt diff	MaxEnt ref	MaxEnt diff (%)
chr7:158869491:GA	splicing	*WDR60*	rs842695	0.1866	0.847302	0/0 0/1 0/1	8.752	0.267	9.019	2.965

Table 5.25: *WDR60* Splicing variant of interest identified from filtering of heterozygous variants.

5.5 Discussion

Currently the only known cause of SSMD is via the loss of exons four and or five in *GPX4* leading to a non-functional transcript and loss of the ability to reduce membrane peroxides leading to cell death[216]. Due to the rarity and severity of SSMD there have been few documented cases and even fewer where sequencing was performed. In 2014 the identification of two SSMD cases allowed NGS to be performed which identified *GPX4* as the causal gene. In this study a child was diagnosed with SSMD prior to the identification of *GPX4* as a causal gene and so was initially exome sequenced along with both parents. In WES data the coverage across all three individuals was assessed for *GPX4*, from which exon 1a was identified as being poorly across all samples. Due to the lack of variants in the other exons of *GPX4* and the lack of coverage in exon-1a all individuals underwent WGS which was shown to have covered exon-1a. As shown previously in Chapter 4 WGS avoids the biases of WES and provides more consistent coverage with better depth of coverage in non-coding regions but due to lower mean depth of coverage may have also have false positives in exons covered at low depths. To correct the low depth calls for each sample the WES and WGS sample were combined to leverage the depth of each method to call variants. Eight variants were identified in or 1kb either side of *GPX4*, only one of which was exonic (rs713041). However this variant was synonymous and homozygous alternate in all samples in the trio, ruling the variant out as a disease cause. An intronic variant between exons six and seven (rs8178977) was heterozygous in both SD002 and SD003 but was also common with alternate allele frequency of 0.2516 in the Non-Finnish European population. A 5' UTR variant (rs2074450) of exon-1B, the nuclear targeted isoform, was identified but was heterozygous in both sample SD002 and SD003. Finally the intronic variant (rs8178977) called between exons six and seven was heterozygous in both SD002, SD003 and was too common to be causal. Comparisons of all *GPX4* and up/downstream variants show that all variants which were called in SD003 were also called in SD002, therefore *GPX4* could not have caused caused the phenotype of SD003 alone as the unaffected mother does not share the phenotype. From CNV and structural variant calling analysis no structural or copy variants were identified using WES or WGS datasets separately. All of the evidence therefore argues against loss of function of *GPX4*, as this was the

only gene known to cause SSMD this raised the question of the SSMD being caused by an alternative gene or a misdiagnosis of a similar condition(s) such as Short-Rib Polydactyly Syndromes.

Having ruled out *GPX4* as a potential cause in this case there are two possible conclusions, that SSMD is genetically heterogeneous and there are other genes which give the same or very similar phenotypes via a different mechanism. Alternatively the diagnosis of SSMD was incorrect and the patient has a related condition. Examples of misdiagnoses have previously been documented such as with the gene *XIAP* (X-linked inhibitor of apoptosis) where a child was diagnosed with an immune defect. NGS was able to identify a mutation in *XIAP* and its association in Crohn's disease, changing the patients diagnosis[234]. Skeletal dysplasias have many forms as shown by the 2015 nosology publication[215] which reported 456 forms of varying forms split into 40 categories. Diagnosis can also be complicated for SSMD as it shares large phenotypic overlap with Short-Rib Polydactyly Syndromes[221,222]. Therefore in this study candidate genes were then extended to include other forms of skeletal dysplasia and Short-Rib Polydactyly Syndromes in addition non-candidate based annotation filtering approaches were performed to try and identify pathogenic variants. Focus was initially on exonic variants due to the increased evidence and tools available to filter and prioritise variants.

GEMINI analysis of the trio files was performed in July 2018 to identify variants possible pathogenic that fit the inheritance patterns using genotypes. In total 219,547 autosomal recessive variants were identified, none of which were in any of the other *GPX** genes. None of the Tier-2 HGMD gene variants were of low enough AAFs in the NFE population by gnomAD as they ranged between 0.28 and 0.77. All Tier-2 variants were also predicted non-pathogenic by FATHMM-XF. Using the extended candidate gene list for all 430 genes in Tier-3 all variants were again too common with AAFs between 0.1 to 0.77. Only a single variant in *DYNC2H1* (chr11:103287549:C:T, rs10895391) which was predicted as a low confidence pathogenic variant by FATHMM-XF was identified to be of most interest from the tier, but the variant is listed benign on dbSNP and is too common at an AAF in the NFE population of 0.3629 by gnomAD. The same *DYNC2H1* variant was also the only SRPS variant

identified in Tier-4. Tier-5 filtering outside of candidate genes identified only two variants, both of which were non-frameshift insertions in the genes *MADCAM1* (Mucosal Vascular Addressin Cell Adhesion Molecule 1) associated with Inflammatory Bowel Disease[235] and *SLC12A5* (Solute Carrier Family 12 Member 5) associated with Epileptic Encephalopathy[236] neither of which have any description of lethality or relation to skeletal or respiratory disease.

Analysis of the *de novo* variants also identified no *GPX** variants for Tier-1. Tier-2 identified a sole variant in the gene *HSPG2* (heparan sulfate proteoglycan of basement membrane/ Perlecan) (chr1: 21873389: C: G, rs766722786) which passed the violation probability of 0.99 and was rare with a non-Finnish European AAF of 0.0003. However in ExAC the variant was found at an AAF of 0.1498. The variant also is not predicted pathogenic with FATHMM-XF at a score of 0.469 and no pathogenicity in the listed dbSNP entry. The variant also reported in sample SD001 with 69 reference and 15 variant alleles compared to 65 reference and 0 variant in SD002. For SD001 the variant allele had a frequency of 0.18 making it close to a heterozygous call so the variant may not actually be a Mendelian violation. Tier-3 also identified the same *HSPG2* variant as in Tier-2. No exonic Tier-4 variants were found. Tier-5 filtering of annotations identified 26 variants which passed filtering but as shown in Table 5.10 there were no genes with a phenotype which matched skeletal or respiratory phenotypes. Therefore, none of the variants were considered good candidates for the sample SD003.

Compound heterozygous variants were also investigated using GEMINI but failed to identify any *GPX** gene variants for Tier-1. Tier-2 identified only two intact pairs of compound heterozygote variants both in the gene *HSPG2*[237]. *HSPG2* binds to basement membrane protein and receptors and can cause two diseases: a heterozygous or dominant allele can lead to Schwartz-Jampel syndrome while a homozygous alternate variant or compound heterozygote variants cause Silverman-Handmaker type of Dyssegmental dysplasia[237–240]. Schwartz-Jampel syndrome is the milder of the diseases characterised by short stature and *pectus carinatum* and progressive disease onset in infancy but is non-lethal in neonates[240]. Silverman-Handmaker is neonatal lethal and characterised by short-limbed dwarfism,

small chest and pulmonary hypolasia giving this disease good overlap with the SSMD disease phenotype[237]. There are three variants in *HSPG2* with the only variant from the father (chr1:21834781:C:A, rs372760688) sufficiently rare with AAF of 6.67E-05 but of low predicted pathogenicity 0.1 by FATHMM-XF. Both other *HSPG2* variants shared maternally with SD002 are of higher gnomAD NFE AAFs of 0.169 (chr1:21839494:G:A, rs2291827) and 0.17 (chr1:21823627:C:T, rs3736360) but with pathogenicity scores of 0.74 (possibly pathogenic) and 0.126 (likely benign) respectively. Therefore these variants are not likely causal as both maternal variants are too common in the population and the variant in the father is not predicted pathogenic.

Tier-3 of compound heterozygotes contained seven variants, including the previously discounted three *HSPG2* variants and one pair of heterozygotes each in *DLL3* (Delta Like Canonical Notch Ligand 3)[241] and *PCNT* (Pericentrin)[242]. Neither of the genes are likely to be pathogenic as the *DLL3* variant chr19:39504071:T:C (rs1110627) has a AAF of 0.5549 in the NFE population by gnomAD and very low pathogenicity score of 0.03. Both *PCNT* variants are common with AAFs of 0.26 and 0.21 combined with low pathogenicity scores from FATHMM-XF of 0.02 and 0.04. No Tier-4 variants identified with intact pairs to be plausible compound heterozygotes.

Tier-5 filtering of compound heterozygotes found 30 variants but only 12 genes as shown in Table 5.9. 12 of the variants in this tier were present due to the lack of annotations either in gnomAD or FATHMM-XF. Genes were annotated to identify phenotype associations. Only the genes *FTCD* (formimidoyltransferase cyclodeaminase)[243], *SDHA* (succinate dehydrogenase complex flavoprotein subunit A)[244] and *GMPPA* (GDP-mannose pyrophosphorylase A)[245] had any phenotype link recorded but none were in keeping with skeletal or respiratory disease.

Across all three GEMINI modes the only exonic ACMG gene variants identified were in the gene *APOB* (Apolipoprotein B) as two potential compound heterozygotes caused by three variants. Two of the variants are listed as benign or likely benign in ClinVar and both were inherited from the father arguing against the variants being compounded and therefore there are no secondary findings from the ACMG genes.

Splicing variant analysis was performed using SPIDEX[208] first looking at any *GPX** gene variants but found none with a change in predicted splicing large enough to be pathogenic when using the recommended cut-off of $|\Delta PSImax| \geq 5$. No variants were also identified using the same threshold applied to the extended 430 gene candidate list. Finally SPIDEX was used to filter the GEMINI results which identified six variants in genes: *AKR1C2* (Aldo-keto reductase family 1, member C2) associated with sex reversal[246], *VPS50* (EARP/GARPII Complex Subunit) associated with golgi protein transportation[247], *CNIH3* (Cornichon Family AMPA Receptor Auxiliary Protein 3) associated also with golgi transportation, *CROCC* (Ciliary Rootlet Coiled-Coil, Rootletin) associated with cytoskeletal structure[248], *LILRA2* (Leukocyte Immunoglobulin Like Receptor A2) associated with monocyte receptors[249] and *NBPF9* (Neuroblastoma Breakpoint Family, Member 9) associated with a variety of diseases including neuroblastoma[250]. All six of these genes therefore have no direct relation of skeletal or respiratory diseases.

Non-coding variants were analysed to also identify any potential variants which could possibly be affecting gene expression using FATHMM-XF scores for non-coding variants. No high confidence pathogenic variants above the 0.96 threshold were found also with an AAF below 0.05 with the gnomAD NFE population for autosomal recessive GEMINI variants or compound heterozygotes with an AAF filter of below 0.1. Three *de novo* variants were identified but two were the *VPS50* and *CROCC* variants identified from SPIDEX which were not skeletal or respiratory associated. The only new variant was in the gene *COPB1* (Coatomer Protein Complex Subunit Beta 1) involved in golgi protein transport[251] also making this gene not directly linkable to skeletal or respiratory disease.

Analysis of samples to identify larger structural and copy number variants was performed with varying success. Many structural variants were called by LUMPY for SD003 totalling 57,681, some of which were implausibly large at upwards of 10MB deletions. Visualising the ends of these large variants showed split or paired reads could be seen only at either end of the events at the described locations not over the entire region. Recent uses of longer reads with long reads produced from nanopore or

other third generation methods have suggested that structural variant calls may only be partially resolved when using short reads and the exact rearrangements can only be seen with long reads[252]. 946 of the variants called by LUMPY overlapped a gene in SD003. Filtering then identified 27 autosomal recessively inherited variants, none of which were candidate genes or with a function or phenotype link to skeletal or respiratory disease. Identification of *de novo* variants found seven variants in candidate genes however six ranged from 2.3 to 9.5MB and as described previously the split reads could only be visualised at the ends of the variants not over the duration. A *WDR60* variant was called of size 2 bp as shown in Figure 5.6. The same variant is called as a non-frameshift deletion by GATK making this variant likely a miscall and actually a heterozygous frameshift deletion. GATK performs *de-novo* assembly of regions when calling and it is unclear why the re-assembly might have caused the variant to be miscalled[47]. As the *WDR60* variant is heterozygous and not seen in either parent it could be compounded to cause short-rib thoracic dysplasia 8, one of the SRPS shown in Table 5.1[224,225,253]. There is some evidence suggesting that SRPS can be caused by di-genic inheritance across multiple of the SRPS genes[221,226,254]. Attempts to look cross-dataset for mutations in candidate genes which would have been missed only identified a potential heterozygote splice variant also in *WDR60* from candidate genes, shown in Table 5.25 however the pathogenicity is questionable as it had a FATHMM-XF non-coding score of 0.84 but an NFE AAF of 0.18 and a small predicted change in splicing by MaxEntScan[233].

CNVkit also failed to detect any variants over any potential candidate or genes which had a partial phenotype match. However CNVkit was able to detect the suspected deletion in the 15q13.3 region. The deletion in the 15q13.3 region was found in both the WES and WGS data for the affected sample SD003 regardless of the reference type used (flat or pooled). As CNVkit only operates by comparing on and off-target read depth it can only detect deletions and duplications. Therefore complex events like breakend and translocations are not likely to be obvious and may not be detected as split and paired reads would be needed, hence LUMPY may be a better option for certain variant classes. LUMPY called multiple variants over the 15q13.3 region but all were too small to fit a deletion detected by previously by array CGH and CNVkit. LUMPY also called overlapping deletions and duplications making it not possible to

logically interpret the calls over the 15q13.3 region. The 15q13.3 region is associated with delayed neurological development via an inversion of the breakpoints four and five due to high homology between breakpoints which pre-disposes the BP4-5 interval region to deletion as was seen in SD003[227]. Some evidence suggested that the locus could also be involved in asphyxiating thoracic dystrophy (Short-rib thoracic dysplasia 1 - MIM:208500) but no mechanism or mutations within the locus have been able to be identified from the locus as causing Short-rib thoracic dysplasia one[255] and how to distinguish from the microdeletion in the same region causing Prader-Willi or Angelman syndrome[227]. Therefore whilst the microdeletion can be confirmed by CNVkit and can be clearly visualised it remains on unknown significance in this case.

5.6 Conclusions

No single variant was determined to be a cause of SSMD, the lack of *GPX4* variants suggests that SSMD may be caused by genetic heterogeneity via a novel gene or mechanism not yet known. Using an extended candidate gene list for other skeletal diseases identified a possible compound heterozygote involving a splicing mutation and a heterozygous frameshift deletion which could cause short-rib thoracic dysplasia eight[253].

All of the work performed in this study highlights the current problems in identifying causal variants amongst large genomic datasets such as WGS. Many of the variants identified as pathogenic in non-coding regions currently lack further annotations to prioritise upon. The number of coding variants predicted as pathogenic that are found in un-affected individuals also highlights the challenges of identifying causal variants. In future once support is available for hg38 and new methods are available annotations for both coding and non-coding variant predictions and our ability to filter variants will improve. As algorithms evolve and our understanding of the non-coding genome improves annotations will likely improve, enabling the annotation of more of the genome. INDELs were found to be amongst the most commonly un-annotated variants due to most being novel. In future *de novo* calculations of annotations could possibly calculate scores for variants missing annotations. Projects such as the 100,000 genomes also have a group for skeletal disease and should identify new targets in future to expand candidate genes further for skeletal diseases. The future re-analysis with newly identified targets could lead to a causal mutation for the observed disease and in future lead to faster diagnosis and possible treatments.

Chapter 6

Approaches to call SNVs and CNVs amongst high similarity FC-gamma receptors

6.1 Introduction

6.1.1 Fcγ receptor functions

Fcγ receptors (FcγRs) are a group of proteins expressed on the surface of leukocytes to facilitate binding to the Fc part of immunoglobulin G (IgG). There are six FcγRs expressed by humans divided into three classes with varying affinities for IgG, each class triggering different signalling pathways of the innate immune response depending upon the pathogen[256]. Each of the six receptor types are encoded by an individual gene as summarised below in Table 6.1 which also describes which cluster of differentiation (CD) antigen each of the receptors are expressed on.

FcγRs act to modulate the immune response by indirect regulation of cytokine production. Regulation is dependent upon the signalling from other receptors such as toll-like receptors (TLRs) which signals together with FcγRIIa to strongly amplify pro-inflammatory cytokine production or interleukin receptors IL6, IL12 or IL23, all of which antagonise cytokine production. FcγRIIB is the only inhibitory FcγR and balances against all other activating FcγRs to control the immune response. Exposure to pathogens alters the expression of the FCGR genes to produce FcγRs in favour of

Gene	Protein	IgG Affinity	CD	Cellular expression	FcγR Function
FCGR1A	FcγRI	High	CD64	Monocytes, macrophages	Activating
FCGR2A	FcγRIIa	Low	CD32A	Monocytes, granulocytes, B-cells, eosinophils	Activating
FCGR2B	FcγRIIb	Low	CD32B	Monocytes, granulocytes, B-cells, eosinophils	Inhibitory
FCGR2C	FcγRIIc	Low	CD32C	Monocytes, granulocytes, B-cells, eosinophils	Activating
FCGR3A	FcγRIIIa	Low	CD16A	Neutrophils, NK cells, Macrophages	Activating
FCGR3B	FcγRIIIb	Low	CD16B	Neutrophils, NK cells, Macrophages	Activating

Table 6.1: Six genes producing FcγReceptor proteins each have varying affinity for IgG and are expressed on various CD antigens on different immune cells as summarised. Upon binding to IgG the receptors will activate and up-regulate an immune response with the exception of FcγRIIb which acts to inhibit and modulate the immune response.

the activating, pro-inflammatory receptors. Under normal conditions IgG is unbound and in soluble form, which promotes inhibition of cytokine production. When IgG opsonises pathogens the complex stimulates pro-inflammatory cytokine production via receptors, including the pro-inflammatory FcγRs[257]. Pro-inflammatory FcγRs contain an immunoreceptor tyrosine motif (ITAM), in contrast to the inhibitory FcγRIIb which contains an immunoreceptor tyrosine inhibitory motif (ITIM). Pro-inflammatory FcγRs may also be inhibited by the binding of IgG monomers which prevent amplifications of inflammatory chemokine and cytokine production in combination with other receptors such as TLRs. Inhibition by soluble, circulating IgG is known as inhibitory ITAM signalling (ITAMi)[257].

Expression of FCGR genes must be tightly controlled to regulate the level of activating and inhibitory FcγRs, failure of control leads to loss of homoeostasis and autoimmune diseases such as rheumatoid arthritis and systemic lupus erythematosus (SLE)[257].

Cancer patients have variable responses to monoclonal antibody (mAbs) therapies with FcγRs crucial to mediating the efficacy[256]. mAbs are commonly IgG molecules which include an antigen-binding fragment that binds to tumour cell antigens. Also mAbs have a crystalline fragment (Fc) domain that binds FcγRs to trigger the immune response. Therefore binding to FcγRs will trigger an appropriate immune response targeting tumour cells[258]. Mutations in FCGR genes have been shown to alter the response in an organisms. An example of which is in the gene *FCGR3A*

where an amino acid of valine at position 158 is associated with strong response rates to rituximab in rheumatoid arthritis. However a SNP changing the valine at amino acid 158 to a phenylalanine has been shown in studies to reduce the efficacy of rituximab[256,259]. Other amino acid changes have been similarly identified to cause similar alterations in efficacy of other mAbs such as in *FCGR2A* the change of amino acid 131 to a histidine reduces the response to trastuzumab in HER2-positive breast cancer cases[256].

6.1.2 FCGR Genes - Function & SNVs

There are three groups of FCGR genes: FCGR1, FCGR2 and FCGR3 within which ancestral duplication and crossover events have created multiple genes as shown in Table 6.2.

Gene	Chr	Band	Transcript	Start	Stop	Strand
FCGR1A	1	q21.2	ENST00000369168.4	149,782,690	149,792,518	Plus
FCGR1B	1	q21.2	ENST00000616817.4	121,087,345	121,095,642	Plus
FCGR1C	1	q21.1	ENST00000579737.3	143,874,793	143,883,575	Plus
FCGR2A	1	q23.3	ENST00000367972.8	161,505,430	161,518,558	Plus
FCGR2B	1	q23.3	ENST00000358671.9	161,663,161	161,678,654	Plus
FCGR2C	1	q23.3	ENST00000543859.5	161,581,339	161,600,242	Plus
FCGR3A	1	q23.3	ENST00000436743.5	161,541,765	161,550,623	Minus
FCGR3B	1	q23.3	ENST00000367964.6	161,623,196	161,631,963	Minus

Table 6.2: Locations of the eight FCGR genes in the human genome reference hg38. Six of the FCGR genes are currently known to produce functional FcγReceptor proteins, *FCGR1B* and *FCGR1C* are not known to functional transcripts. However whilst *FCGR2C* is expressed, the majority of the European population contain a truncating mutation Q57X making the exact receptor function and importance unknown.

Ancestral FCGR gene duplications & predispositions to copy number variation

Multiple FCGR genes are believed to have originated due to duplications of an ancestral FCGR gene[260]. The creation of additional gene copies provided genomic and functional redundancy, allowing some copies to undergo structural changes and evolve new or modified functions[261]. This is evident for *FCGR2B* as it is the only FCGR receptor with an inhibitory ITIM motif while all other FCGRs have the activating ITAM motif. Also for *FCGR2C* it has been shown that the gene likely arose due to an unequal crossover of genes *FCGR2A* and *FCGR2B*[260,262]. Due to the ancestral duplications FCGR genes can be up to 98% similar[263]. Gene duplications are believed to have occurred at the time of division for great apes from new world monkeys[264–266]. The exact mechanism by which the original duplications of the ancestral FCGR gene occurred are unclear.

The high homology between the FCGR genes located on chromosome 1 also makes the locus more likely to facilitate copy number variants[262]. Changes in copy number involve the change of chromosomal structure so non-adjacent regions become juxtaposed[261]. Changes in the chromosomal structure can arise by two main methods either homologous recombination (HR) or non-allelic homologous recombination (NAHR)[261,262]. It has been shown that for genes in the FCGR locus that NAHR is the main cause copy number variations[266]. In NAHR the exchange of genomic material is always between homologous regions but located at different loci, either intra or cross-chromosomal, which can occur in both mitosis and meiosis[267]. Chromosomal imbalance or rearrangements caused by NAHR depend on factors such as the orientation of low copy repeats (LCRs). LCRs are region-specific DNA blocks between 10 and 300 kb in size and with similarities between 95 and 97%[267]. The interaction between low copy repeats with parallel orientation may cause deletions or duplications, whereas the pairing of low copy repeats with opposite orientations may generate inversions[261]. In the case of FCGR genes when grouped by family all FCGR1 and FCGR2 genes are on the forward strand, while FCGR3 are on the reverse strand. As all genes in a family are all in the same orientation it is expected duplications or deletions would be detected rather than inversions.

FCGR1

FCGR1 has three genes: *FCGR1A*, *FCGR1B* and *FCGR1C* though currently no receptors have been detected from transcripts for genes *FCGR1B* and *FCGR1C*[256]. Both *FCGR1B* and *FCGR1C* contain a six base pair insertion relative to *FCGR1A* proposed to have arisen during a gene duplication events from the ancestral FCGR1 gene. *FCGR1A* encodes the high affinity receptor FcγRI. The importance of the receptor *FCGR1A* remains unclear, under normal conditions it is proposed to be saturated by monomeric IgG and less responsive to changes in IgG-Immune complexes created upon pathogen opsonisation. Several healthy individuals have also been reported with stop gains in *FCGR1A*[256]. This highlights the potential redundancy that has arisen by the duplications of the ancestral FCGR genes. In other experiments cytokines have been shown to increase the binding of immune complexes by FcγRI and elevated FcγRI levels on monocytes in individuals with systemic lupus erythematosus and lupus nephritis[256].

FCGR2

FCGR2 has three coding genes: *FCGR2A*, *FCGR2B* and *FCGR2C* all located on chromosome band 1q23.3 each with eight exons. Each gene codes for an active FcγR which are particularly expressed on immune cells such as: monocytes, granulocytes, B-cells and eosinophils. Comparing human FCGR2 sequences with other mammals suggests that the two genes *FCGR2A* and *FCGR2B* also arose from an ancestral duplication event and the inhibitory ITIM motif was then obtained in further structural re-arrangements. An unequal crossover during homologous recombination of *FCGR2A* and *FCGR2B* is believed to have created *FCGR2C* at about the same time as humans diverged away from chimpanzee. The first five exons in *FCGR2C* originate from *FCGR2B* with exons 6-8 originating from *FCGR2A*. As a result of the duplication events there is sequence homology of around 97.9-99.8% between the FCGR2 genes when comparing exons as mentioned previously for two of the genes[256,260].

FCGR2A encodes FCγRIIa which upon bacterial infections leads to inflammatory cytokine production. FCγRIIa is the only receptor which strongly interacts with the IgG subtype IgG2[256]. Upon bacterial opsonization by IgG FcγRIIa signalling leads to

amplifications in cytokine production by increasing gene transcription of cytokine genes via caspase-1 which cleaves pro-IL-1β into its bio-active form[257]. This receptor is believed to be involved in cross-receptor signalling with TLRs which amplifies the production of cytokines such as IL-1β, IL-6, IL-23, and TNFα[268]. The SNP c.519:G>A (rs1801274) in *FCGR2A* causes the amino acid change H131R, in the membrane proximal Ig-like domain and has been shown to reduce the binding affinity to IgG2[256]. This polymorphism increases the susceptibility of an individual to bacterial infections as IgG2-mediated phagocytosis is impaired[257]. However in gnomAD the variant reports an AAF of 0.51 making the variant common amongst several populations.

FCGR2B encodes the only inhibitory receptor FcγRIIb, which has multiple isoforms. The isoform FCγRbII1 consists of all eight exons and is expressed on B-cells. The second isoform FcγRbII2 excludes exon six as with FcγR2a and FcγR2c and are expressed on myleloid cells. The expression of the isoforms are affected by two variants in the promoter regions of *FCGR2B* and *FCGR2C*, the sequence of these genes are identical due to the ancestral crossover and duplication events. The positions of interest are -386G>C (rs3219018) and -120A>T (rs34701572) for both genes giving rise to four *FCGR2B* haplotypes as shown in Table 6.3, one key difference though is the -120A allele is seen in only *FCGR2B* and not the *FCGR2C* promoter region[256].

Gene	Allele at position -386	Allele at position -120	Haplotype
FCGR2B	G	T	FCGR2B.1
FCGR2B	C	T	FCGR2B.2
FCGR2B	G	A	FCGR2B.3
FCGR2B	C	A	FCGR2B.4

Table 6.3: Haplotypes of *FCGR2B* created by two SNPs at positions -386 and -120. Both SNPs are in upstream promoter regions and have been associated with changes in expression between populations. FCGR2B.1 is the more commonly observed haplotype, while FCGR2B.4 is more commonly seen in European SLE cases.[256].

FCGR2C is suggested to have derived from an unequal crossover of *FCGR2A* and *FCGR2B*, the breakpoint is likely to have occurred downstream of exon six. *FCGR2C* is mostly expressed on the natural killer cells and is able to induce cytotoxic response upon binding IgG. Most Europeans are reported to contain a stop-gain Q57X which truncates FcγR2c, inhibiting the cell surface expression, making the function and importance of the receptor unclear due its absence in the majority of the European population[256].

FCGR3

FCGR3 contains two genes: *FCGR3A* and *FCGR3B*, expressed on chromosome 1q23.2 on the reverse strand, the reverse orientation compared to the FCGR1 and FCGR2 genes. Evidence suggests that an ancestral duplication around the point of human–chimpanzee divergence generated *FGCR3A* and *FCGR3B*[256].

FCGR3A encodes the receptor FcγRIIIa which is expressed on natural killer cells, monocytes and macrophages. Receptors mediate antibody dependent cell cytotoxicity upon stimulation by IgG in an immune complex[256]. SNPs have been documented which alter the FcγRIIIa affinity for IgG. An increase in affinity such as with the SNP c.559:T>G (rs396991), causing the amino acid change F158V in the second Ig-like domain. This SNP is associated with an increased risk of developing systemic lupus erythematosus[269] but also a stronger response to mAbs[256].

FCGR3B encodes the receptor FcγRIIIb expressed on neutrophils and basophils. Polymorphisms in exon three encode the human neutrophil antigen (HNA)-1 system in the Ig-like extracellular domain furthest from the membrane. HNA-1A and HNA-1B are the two most common HNA-1 haplotypes, which differ by six SNPs, inducing five amino acid changes. Each of the changes alter N-linked glycosylation sites increasing affinity for IgG1 and IgG3 with HNA-1A compared to HNA-1B which reduces the risk of developing SLE[256,270]. Each of the six SNPs are shown below in Table 6.4 along with the associated rsIDs and amino acid changes. HNA-1C is of unknown function though has been shown to be co-expressed with HNA-1A in samples with *FCGR3B* duplications[256].

rsID	**rs200688856**	**rs527909462**	**rs448740**	**rs5030738**	**rs147574249**	**rs2290834**
Codon	36	38	65	78	82	106
Amino Acid Change	S>R	L>L	S>N	A>D	N>D	I>V
HNA-1A (NA1)	AGG	CTC	AAC	GCT	GAC	GTC
HNA-1B (NA2)	AGC	CTT	AGC	GCT	AAC	ATC
HNA-1C (SH)	AGC	CTT	AGC	GAT	AAC	ATC

Table 6.4: FcγRIIIb is expressed on the surface of neutrophils, in exon three the human neutrophil antigen system is encoded using five SNPs, each highlighted in red for the HNA type the variant is found. Exon three encodes the Ig-like extracellular domain furthest from the membrane. HNA-1A and HNA-1B differ by five SNPs, four of which encode amino acid changes and are the two most common forms. HNA-1C is of unknown function and differs from HNA-1B by the SNP rs5030738. Codons described in the table are relative to the transcript used.

6.1.3 Copy number variants

Most of the FCGR genes currently in the human genome have arisen due to duplications, primarily through non-allelic homologous recombination (NAHR)[256]. NAHR most often occurs in unique sequence regions between low-copy repeats (LCRs) of sizes between 10 and 300kb[267,271]. Most regions with large LCRs greater than 10kb tend to be associated with genomic rearrangements and disease due to reductions in genomic stability, increasing the likelihood of NAHR[271]. Due to repetitive sequences incorrectly aligning during mitosis or meiosis genomic rearrangements can occur[267]. Genomic rearrangements are described as recurrent if breakpoints across individuals share the size and content, if the positions of the breakpoints vary by individual it is a non-recurrent rearrangement[271].

Four copy number regions, termed CNR 1-4, have been identified in the FCGR locus on chromosome 1 as shown below in Figure 6.1.

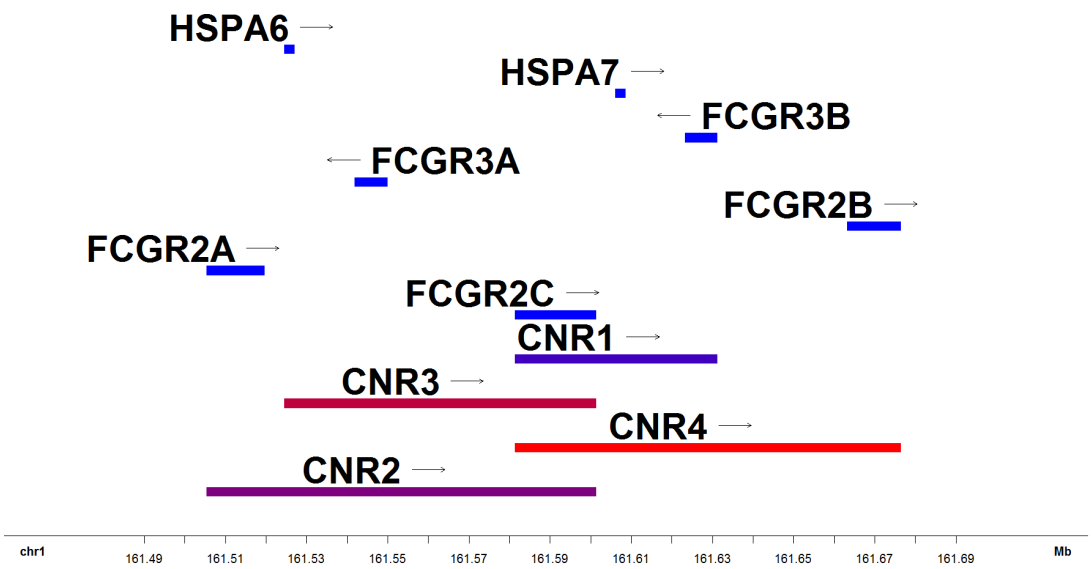

Figure 6.1: Visual illustrations of common regions of variable copy number and segmental duplications over the FCGR locus. Currently there are four known regions (CNR 1-4) in the FC gamma receptors locus where changes are most commonly reported. Exact base pair co-ordinates describing the positions of breakpoints most commonly observed in these genes have not been reported and regions shown here are based upon the genes they are known to overlap.

All four copy number regions overlap with *FCGR2C*, as shown in Figure 6.1, making this gene the most copy number variable, largely due to variations in *FCGR3B* and *FCGR3A*[256,262]. Breakpoints within the genes were also detected using multiple ligation-dependent probe amplification (MLPA), though breakpoints were rarer and non-recurrent. Some evidence shows that the copy number of FCGR genes could vary by population though there is not currently enough evidence for a correlation between copy number and infection susceptibility or autoimmune diseases[256].

6.1.4 Mapping and variant calling FCGR genes

Most next generation sequencing pipelines require reads to map with a minimum mapping quality of 20 to be counted in analyses such as coverage or variant calling. Mapping quality is assigned during alignment by the aligner based upon how likely a read is to be correctly mapped to that position. The value is expressed on the same scale as the PHRED quality system, therefore the minimum mapping quality of 20 indicates a 99% the read had been correctly mapped and a score of 30 would be 99.9% probability. Using a minimum mapping quality of 20 therefore helps to ensure that only reads with a high confidence of being correctly mapped and not misplaced are counted. As described previously for each of FCGR1, FCGR2 and FCGR3 there are multiple genes which have arisen due to proposed ancestral duplication events or structural rearrangements[256,259]. With duplication events between each of the genes in each FCGR group (1,2,3) sequences are largely shared resulting in high homology between the genes of around 98%[262].

When mapping short-reads an alignment program such as BWA[43] may be able to map reads equally well to multiple FCGR genes. As reads cannot be mapped uniquely the aligner has no confidence that the read can be placed in the correct location and by default then sets mapping quality to zero and randomly allocates reads between the matching sites. If an alignment was repeated the reads with a mapping quality of zero would be randomly distributed potentially leading to variable results. Due to reads with mapping quality below 20 being ignored by variant calling the loss of reads over the FCGR genes impairs variant calling and could result in potentially important variants being missed out in conventional NGS pipelines.

6.1.5 HaloPlex capture kits

The capture kit in this project used Agilents' HaloPlex technology. Kits operate by fragmenting DNA using restriction enzymes and denaturing to allow probe hybridisation and selection of targeted regions[272]. Each probe hybridises to both ends of the DNA fragment forming a circular molecule. All probes used are biotinylated, allowing the target to be retrieved with magnetic streptavidin beads. Circularised fragments are then closed by ligation before PCR enrichment of the circularised DNA fragments prior to sequencing using standard Illumina paired-end primers[273].

6.2 Aims

Traditionally FCGR genes are difficult to call variants over due to the high homology to other FCGR family genes. This study will attempt to call variants over the FCGR genes using short read targeted NGS with a custom HaloPlex capture kit designed to capture FCGR genes. From mapping and variant calls the study will assess the viability of short read mapping to call variants in the complicated FCGR genes.

Germline blood samples will be sequenced to call SNVs and CNVs over FCGR genes. Variant calls from NGS will be able to be compared against previously generated MLPA data for the samples and compare predictions for the HNA-1 sample haplotypes.

CNVs will also be called to determine if they fall into the existing described copy number regions described in the limited available literature. Should sufficient copy number events be recorded it will allow further refinement of the locations and boundaries of the copy number regions.

6.3 Materials & methods

6.3.1 Samples & sequencing

72 anonymised blood samples were obtained for this project. Samples were processed in two batches of 31 and 41 samples respectively, both batches were captured using the HaloPlex custom capture kit. Genomic DNA (gDNA) was extracted from the peripheral blood mononuclear cells. All samples were paired-end sequenced using an Illumina MiSeq sequencer. Reads from the MiSeq were 250bp for batch one samples and 150bp for the first run of batch two but 250bp for run two of batch two. Design of the capture kits and laboratory processing of samples, including MLPA was performed by Dr. Chantal Hargreaves and sequencing by other members of Professor Jon Strefford's Cancer Molecular Genetics group in the Faculty of Medicine, University of Southampton. Sequencing was performed at the Wessex regional genetics lab (Salisbury) and WISH laboratory (Southampton). Fastq files were provided for the analyses performed in this chapter along with the summarised results from MLPA for comparisons of SNVs and CNVs called from NGS.

The customised Agilent HaloPlex capture kit was comprised of 36,023 amplicons covering a total of 1,538,065 bp including 130 genes and 97 promoter loci, all genes and promoters targeted are listed in Appendix D, Table 8.10. Included amongst the genes were all eight FCGR receptors. The FCGR locus ranged from 161.49 - 161.69Mb on chromosome 1, covering in particular all FCGR2 and FCGR3 genes with 197,832 out of 200,000bp (98.8%) covered by the kit design. MLPA data was available for 66 samples which provided data for copy number predictions over the FCGR2 and FCGR3 locus and included six SNPs listed in Table 6.4 for *FCGR3B* comprising a samples HNA-1 haplotype and rs759550223 in *FCGR2C* encoding the X57Q polymorphism.

6.3.2 Quality control & mapping pipeline

A concise summary of the pipeline used to analyse samples is shown below in Figure 6.2. Raw fastq files were assessed for quality using FastQC (v 0.11.3)[32] as explained previously in Section 4.3.2 with aggregation performed using the MultiQC tool. The proprietary Agilent software SurecallTrimmer (v3.5.1.46) was used to perform read

trimming to remove adapters and filter low quality. Due to the short read lengths created by the small amplicons quality trimming was run with the low default threshold of five to try and preserve lengths for mapping FCGR genes. Reads were also only retained if they had a minimum fraction of the original read size between 30 and 100%.

BWA-MEM (v0.7.12)[43] was used for alignment against the hg38 reference genome[45]. BAM files were produced from SAM files using picard tools (v1.97)[49]. The number of reads successfully mapped to the genome was measured using SAMtools flagstats (v0.1.19)[46] and the insert sizes were calculated using picard tools. VerifyBamID[88] was also used to provide an estimate of the contamination.

Pear (v0.9.10)[274] was used to merge paired-end reads based on overlaps to avoid double counting, also potentially creating longer contiguous sequences which could aid mapping of repetitive sequences and duplicated genes such as the FCGR 1, 2 & 3. Pear was run with settings to require a minimum overlap of 10 bases and reads over 50 bp in length. Pear was run on fastq data generated from BAM files as shown in Figure 6.2. Three fastq files were generated: two paired-end unassembled files when reads did not overlap and a single-end file containing assembled reads. Overlaps were scored using the default assembly scoring algorithm of +1 for a match multiplied by base quality and -1 multiplied by base quality. For each read the highest assembly score is used for merging. Before merging the highest score is applied to a model to calculate if the score is statistically significant to proceed with merging. Merging will also not be performed if the overlap is smaller than the defined threshold of 10 bases[274]. Using all three fastq files (assembled and unassembled) from PEAR the reads were again aligned against the hg38 reference genome using BWA, generating two BAM files which were merged and sorted into a single BAM file per sample using picard (v1.97) tools from which coverage could be calculated.

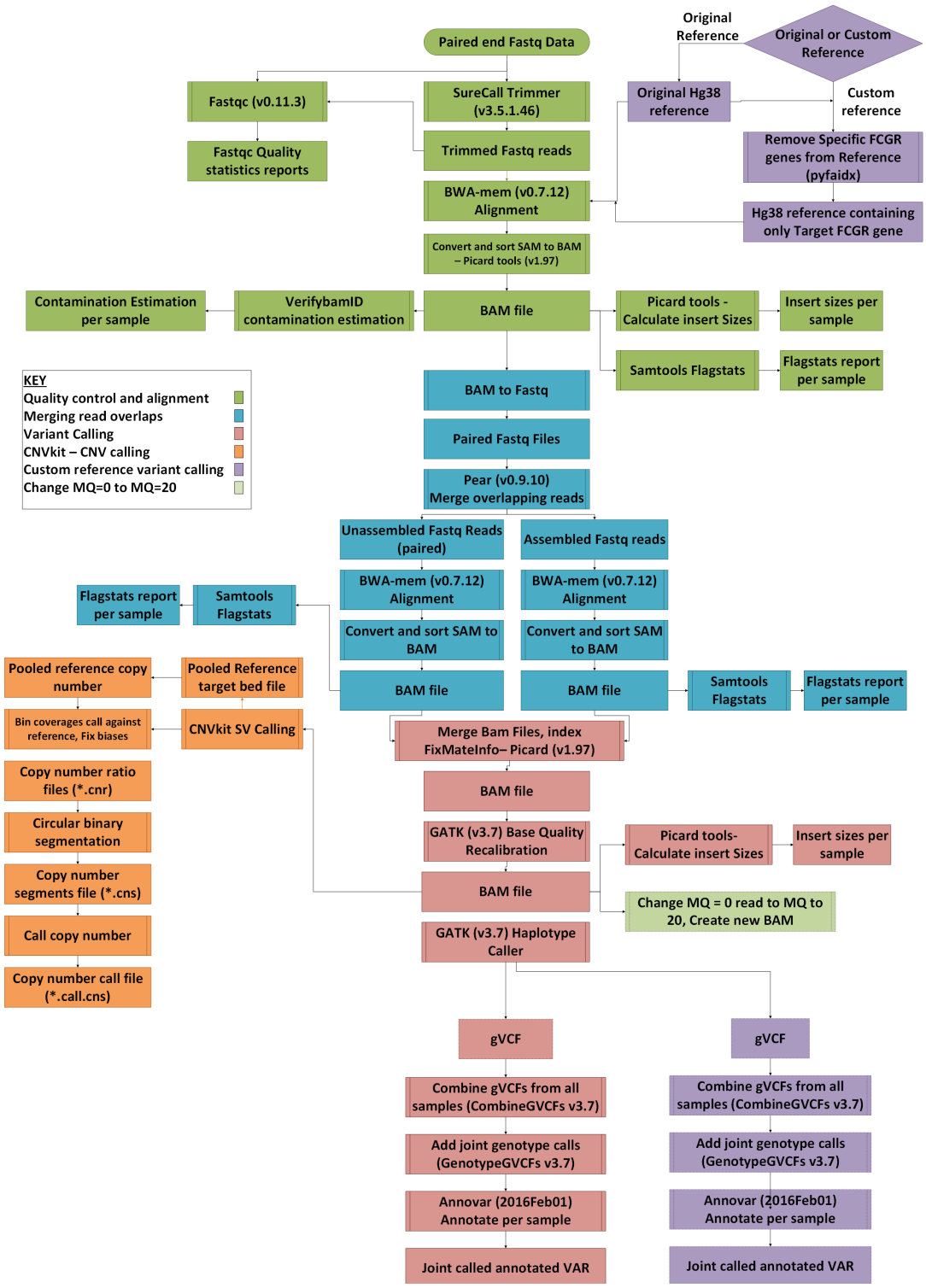

Figure 6.2: HaloPlex bioinformatic analysis pipeline flowchart going from raw fastq files to variant calls with each section of the pipeline colour coded. Variant calling with was performed twice using: the default hg38 human reference genome for alignment and with a custom reference for each of the FCGR 1, 2, 3 genes with only one of the FCGR genes for alignment and variant calling with the other non-FCGR targets.

6.3.3 Sample coverage

Coverage was calculated using SAMtools (v1.3.1)[42], bedtools (v2.21.0)[194] per sample and also averaged over all samples for each base covered by the capture kit before plotting using R (v3.30).

Coverage was calculated per sample based upon the locations covered by the HaloPlex kit using the GATK depth of coverage tool to calculate coverage over the targeted regions summarising coverage per gene. In both GATK and SAMtools coverage calculations required reads to be in the primary alignment and with a minimum mapping quality of 20.

A comparison was also performed to see how the median coverage calculated by SAMtools changes over the FCGR locus if the mapping quality filter was lowered to zero, which would help establish regions of problematic mapping due to high sequence homology with other FCGR genes.

6.3.4 Customised reference genomes for variant calling

Customised reference genomes were created to improve coverage and quality of calling in FCGR genes. For each of the FCGR genes a reference sequence was created using the original hg38 fasta sequence but with the sequence of the other FCGR genes in the family replaced by 'n' bases 'masking' out the sequence. 500 bases either side of the selected FCGR gene were also retained in addition to preserving non-FCGR gene sequences. For example the *FCGR2A* custom reference file would replace *FCGR2B* and *FCGR2C* genes and 500bp either side with 'n' characters for the nucleotides. Creation of the customised references was performed using python module pyfaidx[275].

The HNA-1 haplotype sites, as summarised in Table 6.5, were investigated to determine if the use of a custom reference improved variant calling and allowed the determination of each samples HNA-1 haplotype. For each of the sites listed in *FCGR3B* there are homologous positions in *FCGR3A* to which reads can map equally well. Due to the equal mapping score the aligner has no confidence of the read being able to be uniquely mapped and effectively results in the loss of read data due to the mapping quality of zero reads being ignored by variant callers. For example by

removing the homologous sites in *FCGR3A* from the reference all of the reads from *FCGR3A* and *FCGR3B* will be placed uniquely over *FCGR3B* with mapping qualities above zero and therefore usable to call variants.

As using a custom reference will pile-up reads between homologous sites the ratio of alleles must be used to determine the possible combinations of alleles between the sites. For example assuming both gene sites are copy neutral, if all alleles are reference or all variant this indicates both sites are either reference or variant respectively, an example is illustrated in Panel B of Figure 6.3. An equal number of reference and alternate alleles indicates that both sites are heterozygous or one site is reference whilst the other is homozygous for the alternate allele, shown in Panel C of Figure 6.3. A ratio of 0.25 to 0.75 of reference to alternate allele reads or vice versa indicates the presence of a heterozygous site and a homozygous reference or alternate site, illustrated in Panel D of Figure 6.3.

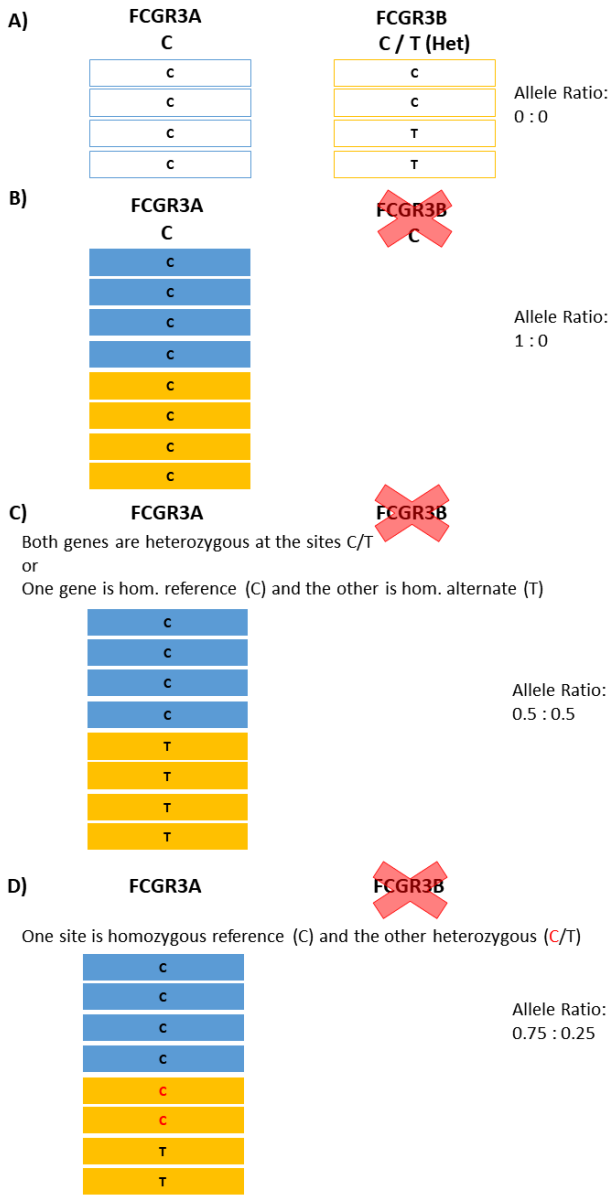

Figure 6.3: Examples of allele ratio calculations, all of the ratios shown here are based upon approximately similar sequencing depths and copy number being equal between genes **A)** Using the standard reference genome reads do not map uniquely due to high homology. Reads have mapping qualities of 0, indicated by transparent reads, and so are not counted in variant calling regardless of a variant call at either of the sites. Reads are randomly allocated between the two sites. **B)** Using a customised reference genome file to exclude the homologous region all reads are now counted. Read totals are representative for both sites and so ratios must be analysed. A ratio of 1:0 indicates both sites have only one allele present, so are both sites are either homozygous reference or alternate. **C)** Also using a custom reference genome file an equal ratio indicates that both sites are heterozygous or one site each is homozygous for the reference and alternate allele. Both cases assume similar depth of coverage at both sites and copy number. **D)** Also using a custom reference genome file should one site be heterozygous and the other homozygous a ratio of 0.75:0.25 will be observed again assuming equal coverage. Red C allele reads have originated from the heterozygous *FCGR3B* site.

By comparing the allele ratios at all six of the sites shown Table 6.5 the HNA-1 haplotype should be able to be determined given the known haplotypes present in the population.

		rs2290834	rs147574249	rs5030738	rs448740	rs527909462	rs200688856	
HNA-1		**I89V**	**N65D**	**A61D**	**N48S**	**L21L (C or T Nucleotide)**	**S19R**	**Amino Acids**
NA1	**A**	V	D	A	N	C	R	**VDANCR**
NA2	**B**	I	N	A	S	T	S	**INASTS**
SH	**C**	I	N	D	S	T	S	**INDSTS**
NA1+NA2	**A+B**	V	N	A	N	T	S	**VNANTS**
NA1+SH	**A+C**	V	N	D	N	T	S	**VNDNTS**
NA1*02	**NA1*02**	I	D	A	N	C	R	**IDANCR**
NA2*02	**NA2*02**	I	N	A	N	T	S	**INANTS**

Table 6.5: HNA-1 Haplotype alleles and amino acids for all six SNPs which make up the possible haplotypes including the less common haplotypes. For the variant rs527909462 the nucleotides present are C or T, both of which are synonymous for the amino acid L.

The amino acid change X57Q in FcγR2c caused by the SNP rs759550223 (NM_201563.5 - T268C) of *FCGR2C* was also investigated using the custom reference genome approach. However rs759550223 is annotated on dbSNP as falling in a non-coding transcript with the amino acid change described as "N/A" and non-polymorphic stop codon[256]. This variant has an equivalent site in *FCGR2B*, rs10917661 which has an ExAC global population frequency of 0.00003. For the FcγR2b variant the reference amino acid is Q (glutamine) at position 57, with the reference nucleotide being a C in *FCGR2B*. Hence the reference nucleotide and amino acid is the inverse of the site in *FCGR2C*. By default all T alleles will map to *FCGR2C* while all C alleles map to *FCGR2B*.

Using the ratios of each allele when using a custom reference genome file it would be possible to identify the *FCGR2C* Q57X variant status as illustrated in Figure 6.4. In a case of a homozygous C nucleotide variant in *FCGR2C* causing the X57Q amino acid change all reads will map to *FCGR2B* as shown in Panel A of Figure 6.4. By examining the ratios of alleles, if all alleles are C then the presence of the SNP rs759550223 can be confirmed as homozygous alternate and the *FCGR2B* as homozygous reference. It should also be possible to identify when there is a heterozygous change at either site assuming the copy number is equal and depth of coverage approximately equal at the sites. In this scenario as shown in Panel C of

Figure 6.4 the ratio will be 0.75:0.25. The limitation of this method is that homozygous reference alleles at both sites will undistinguishable to both sites being heterozygotes as both will yield a 0.5 to 0.5 ratio as shown in Panel B of Figure 6.2. Though in the European population it has been shown that most individuals are homozygous reference for this site in *FCGR2C* and *FCGR2B* is also reportedly non-polymorphic[256] .

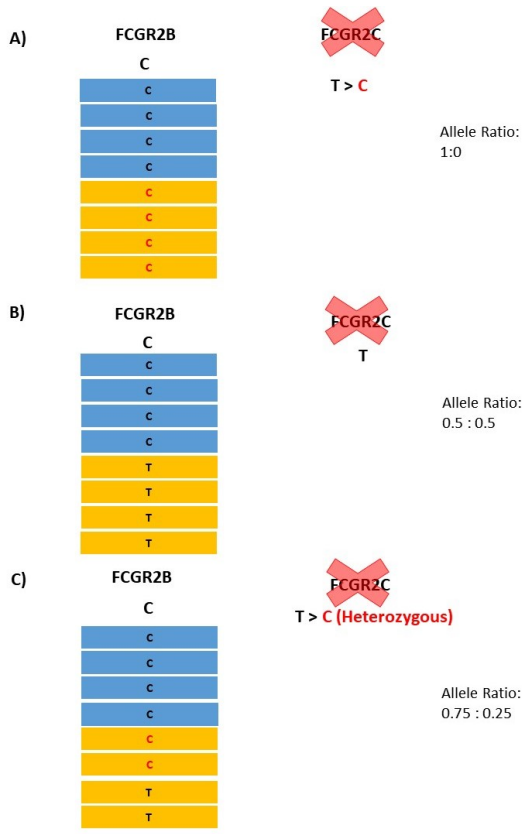

Figure 6.4: Examples illustrating how the allele ratios can be used to estimate the genotypes for the variant causing the Q57X variation between *FCGR2B* and *FCGR2C*. All of the ratios shown here are based upon approximately similar sequencing depths and copy number being equal between genes. **A)** A homozygous non-reference mutation in *FCGR2C*, hence the allele ratio is 1:0 as all alleles are nucleotide C. *FCGR2B* variants are reported in ExAC to be rare. All C alleles highlighted in red have originated from *FCGR2C*. **B)** The sites in both genes are homozygous reference at both sites. *FCGR2B* is reportedly non-polymorphic, making heterozygotes at both sites unlikely, but in such a case it would also give an equal number of alleles and makes identifying the exact genotypes impossible to determine without additional experiments. **C)** A single heterozygous mutation causes the ratio of alleles to change to 0.75 : 0.25. As the *FCGR2B* site is reported-non variant the T to C heterozygous polymorphism of *FCGR2C* is much more likely to be observed giving the 0.75 (C) : 0.25 (T) ratio. All C alleles highlighted in red have originated from *FCGR2C*

6.3.5 Variant analysis pipeline

Variant calling was performed using the GATK (v3.7) best practice guidelines by quality recalibrating the merged BAM file produced after PEAR. GATK tool BaseRecalibrator(v3.7) corrects for systemic biases in calling qualities from Illumina sequencers[50]. Variant calling using GATK HaplotyperCaller was performed producing a gVCF containing collapsed reference blocks rather than calling per base in the reference genome. A VCF file per sample was also produced containing only non-reference variants.

Each gVCF file was used in joint genotyping after merging all gVCF files produced into a multi-sample gVCF using the CombineGVCFs tool. Individual gVCF files were first combined to reduce the computational burden for the GenotypeGVCFs tool performed next. Joint genotyping aggregates information when applying the confidence model, firstly for a variant site all reference reads are compared against all alternate allele reads providing an estimation of confidence that a SNP does not exist. A similar model is then implemented for INDELs where all reads arguing against an INDEL are compared against supporting to give a confidence estimate an INDEL does not exist. From both of these estimates the genotype and probability likelihoods are calculated per variant for a SNP or INDEL with the most likely of the two variant types reported. A single VCF file is returned from joint genotyping with only the most likely variant reported.

Using the per sample VCFs pairwise comparisons were performed taking the chromosome, start, stop, reference and alternate alleles and finding matches between each of the samples using a bash script. The number of matches between a pair can then be used to calculate the percentage shared identity.

The output genotyped gVCF was then annotated using ANNOVAR (v2016FEB01)[56] with hg38 annotations for RefGene[195] and Known Gene[24]. Both databases provide descriptions of any gene overlaps or proximal variants with descriptions of the variant type (i.e. exonic or intronic and synonymous or non-synomymous). Amino acid changes are also provided, but often multiple are provided due to several possible transcripts for each gene.

Cross referencing of variants against dbSNP[19] was performed. For splicing predictions dbscSNV[276] was to flag variants. Variants with clinical association either positive or negative were annotated using ClinVar[65].

Allele frequency information was added from multiple sources: 1000 genomes[27], ExAC, gnomAD[63], Kaviar[277], nci60[278], haplotype reference consortium[26] and ESP[279] allele frequencies.

Pathogenicity predictions were provided from multiple tools: SIFT[60], Polyphen2[61], LRT[280], Mutation taster & assessor[281], FATHMM[62] & FATHMM-MKL[69], DANN[67], PROVEAN[282], VEST3[283], CADD[66], MetaSVM and MetaLR[284].

Measures of sequence conservation and structural protein domains were provided by: Fitcons[285], GERP++[286], phyloP[287], phastcons 20/7-way Mammalian/vertebrate[288], SiPhy 29-way with interpro domain prediction[289].

6.3.6 Copy number variant calling

Copy number variant calling was performed for all samples using CNVkit[80] by comparing the ratio of target and off-target reads to infer copy number against a reference. Initially sample depths were calculated over the target regions and then binned, the process is repeated for the off-target regions. Each of the binned files are then corrected against a reference, CNVkit recommends that samples be pooled to create a single reference consisting of averaged copy number for bins. The reference is used to correct each of the samples, generating copy number ratios normalised for GC-content of the reference genome, exon distribution and repeat sequences. Copy number ratios are scored on a log2 scale, which were then grouped to call copy number changes using a circular binary segmentation algorithm (DNAcopy). For each of the segments thresholds are then applied to call copy number based on the recommended thresholds as described previously in Chapter 4, Table 4.2.

CNVkit was optimised by changing parameters for bin size, segmentation algorithms and the reference sequence. CNVkit recommends to use the split command to create

target bin sizes of around 267 bases, the average size of a human exon. By reducing the bin size smaller CNVs can be called at the cost of increased noise in calls, consequently it is not recommended to reduce bin size to 100bp or below. Sequential reduction of bin size was investigated down to 150bp which achieved a balance of increased sensitivity and preventing excessive noise. Two alternative segmentation algorithms are provided with CNVkit: Fused-Lasso[290] and Haar. Fused-Lasso was tested in CNVkit development and was reported to perform better with some users WES and WGS samples and a smaller set of targeted panels[80]. Haar segmentation[291] was also tested in CNVkit development but was prone to over calling in all but small panels. Other algorithms were tested and compared with MLPA calls but were visibly worse and increased disagreements so the circular binary segmentation algorithm was also used in this study[80]. A pooled reference sequence created from all HaloPlex samples was compared against the use of CRA-25 reference, a known copy neutral sample from previous MLPA analysis.

6.4 Results

6.4.1 Quality control

Fastq files - Quality and trimming

FastQC analysis of all samples was performed before any trimming and filtering of reads. Three samples were identified with insufficient sequence reads to proceed with trimming and mapping as listed in Table 6.6. The majority of reads are also predicted to be duplicates by FastQC for all samples using the original fastq files as is expected for amplicon based sequencing.

SampleID	Batch	Total Sequences
CRA-39	1	12,746
CRA-117	1	98
CRA-139	1	1,755

Table 6.6: Three FCGR samples with an insufficient number of sequences to proceed with further in analyses are detailed. These samples were therefore excluded from future calculations and calling.

Trimming of read sequences led to the loss of reads from samples of variable amounts, on average total sequences per sample were reduced by 9.5%. CRA-134 was the only sample which reported greater than 20% of reads removed in trimming and filtering as shown below in Table 6.7. CRA-134 though still retained 1.4 million reads after trimming, filtering and so was retained for mapping and variant calling.

SampleID	Source	Batch	Read- pairs before	Read-pairs after	% discarded
CRA-134	Blood	1	1,872,943	1,413,025	24.56

Table 6.7: Statistics for CRA-134 which had the largest amount of data removed in quality filtering and trimming. In total 24.56% of reads were removed though the file still retained 1.4 million reads for analysis and was retained for mapping and variant calling.

The distribution of reads across all samples is shown below in Figure 6.5, with the average number of read pairs highlighted in blue at 1,342,189 for initial raw fastq files. Four samples were also identified as outliers with relatively high total sequences in the raw fastqs. These samples with higher reads pair totals (in descending order): CRA-90, CRA-35, CRA-147 and 699. Post trimming and filtering an average of 1,214,696 reads were remaining in samples, the trimmed boxplot visibly condenses as expected on the right of Figure 6.5. Due to the condensing of the plot four samples are high outliers but not in the same order (in descending order): CRA-90, CRA-147, 699 and CRA-35.

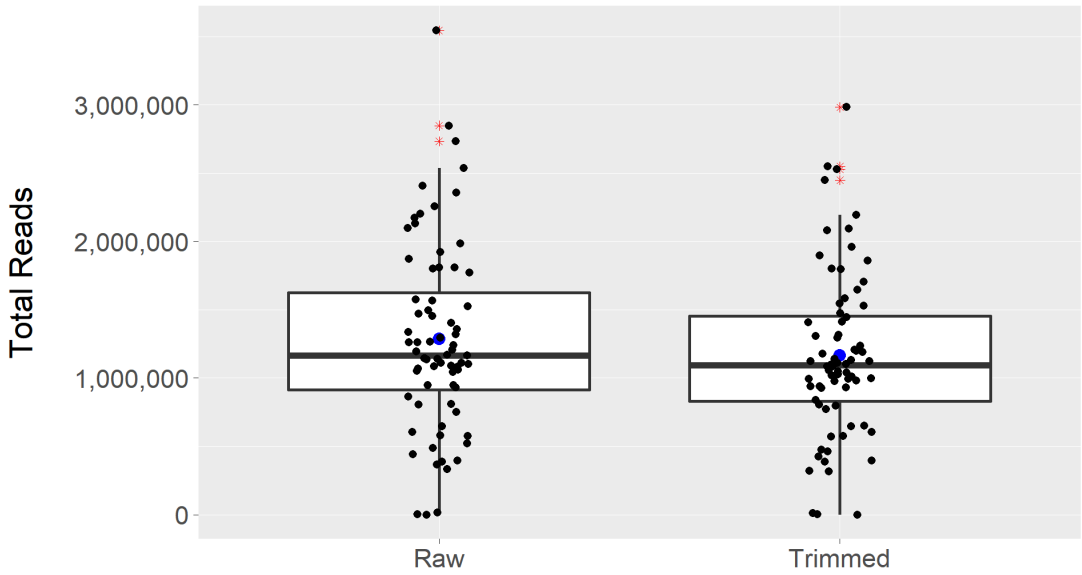

Figure 6.5: Boxplots of the total read sequences for the initial raw fastq files and totals post trimming and filtering. For raw fastq files the mean of samples, excluding four samples with low data, is shown by the blue point with a value of 1,342,189. The mean of trimmed and filtered samples reduces to 1,214,696. Filtering and trimming reduces the reads in samples leading to relative condensing and lower values of the trimmed boxplot. In both plots the same four samples are also classed as outliers with higher number of read pairs.

Insert sizes

Insert sizes of reads were calculated using picard-tools (v1.97) based on initial paired-end mapping of reads. The median insert size calculated over all samples is 142bp. The median insert sizes per sample are plotted in Figure 6.6. Insert sizes were smaller than the 250bp read length the MiSeq was producing for batch 1 and run 2 of batch 2, suggesting reads would be overlapping. An example of the insert size distribution from the median sample is shown below in panel B of Figure 6.7. For this sample the modal insert size was 102 bases in length.

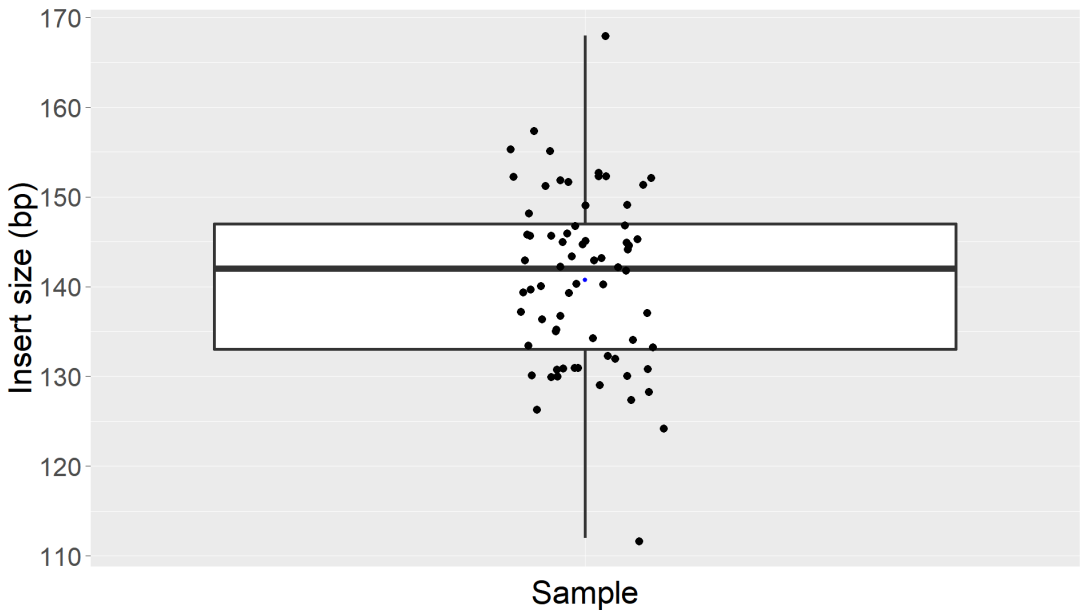

Figure 6.6: Median insert sizes of HaloPlex samples plotted as a boxplot. Over the entire cohort the median sample insert size was found to be 142bp.

Capture-kit design evaluation

The HaloPlex capture kit include a total of 130 genes and 97 promoter loci covered by a total of 36,023 amplicons. The size distributions for amplicons is shown below in Panel A of Figure 6.7. Whilst the sequencing length was 150 or 250 bp the modal amplicon size was 102 bp and a mean length of 199bp. In Panel B insert sizes for the median sample were plotted showing a similar trend to Panel A.

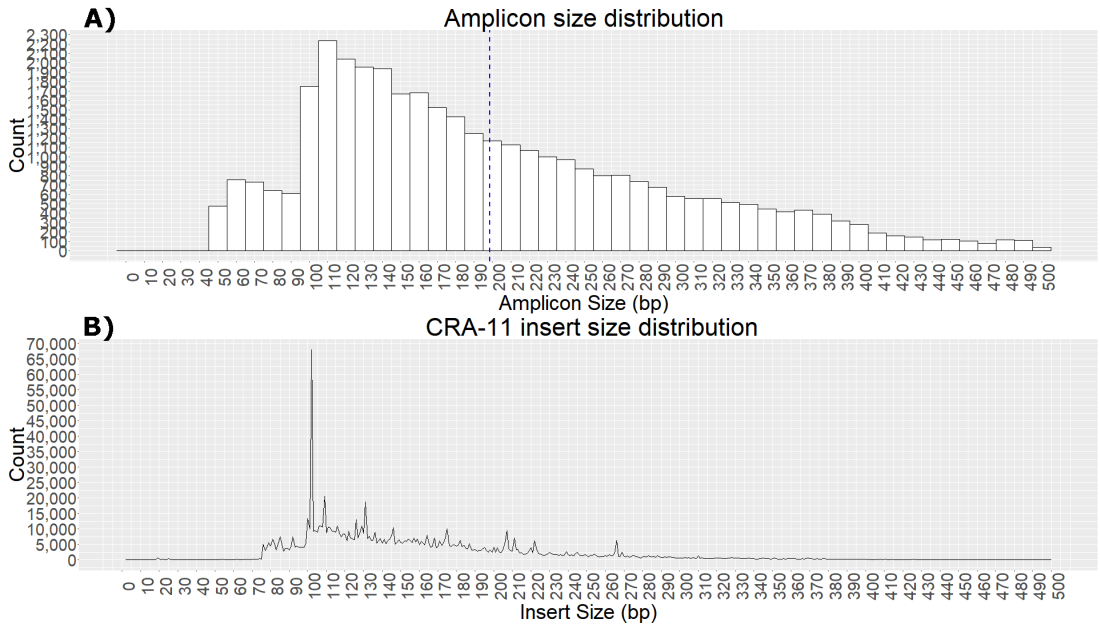

Figure 6.7: **A)** HaloPlex amplicon size histogram, in total there were 36,023 amplicons. Mean amplicon size is 199 bp and shown by the dotted blue line. **B)** For the median read sequences sample CRA-11 insert sizes were calculated from the initial mapping of paired end sequence data to the reference, showing that the distribution matches closely that above in Panel A for amplicon size with a peak at around 102bp for the most common insert size. The median sample insert size is 140bp for the sample.

Merging overlapping reads

Merging of overlapping was attempted due to insert sizes showing overlap of reads. Merging prevented counting the overlapping reads twice and introducing a bias into the results. When reads were merged by the program PEAR output read files were referred to as "assembled" or "unassembled" when reads remain as pairs but with insufficient overlap to merge. In Table 6.8 the percentage of merged reads averaged over all samples separated by batch is shown, all batches have a high assembled percentage at 98.5% and 95.3% for batch 1 and 2 respectively. The total number of assembled and unassembled reads for each of the samples are contained in Appendix D, Table 8.13. Both unassembled and assembled reads were retained for variant calling by generating two BAM files which could then be merged and sorted.

Assembled			
Batch	**Mean Reads (%)**	**Min No. Reads (%)**	**Max No. Reads (%)**
1	1,259,553 (98.5)	316,086 (96.1)	2,911,781 (99.5)
2	1,112,830 (95.3)	300,416 (92.7)	2,439,870 (99.6)
Unassembled			
Batch	**Mean Reads (%)**	**Min No. Reads (%)**	**Max No. Reads (%)**
1	19,973 (1.5)	4,073 (0.5)	71,214 (3.9)
2	48,978 (4.7)	4,036 (0.4)	170,706 (7.3)

Table 6.8: Summary of merging overlapping read pairs with PEAR. Shown are the maximum, minimum and average number of reads merged based on overlaps by PEAR. Most reads overlapped and hence the high percentages of assembled reads above 95% per batch and relatively few reads are unassembled for batches - 1 & 2.

Alignment with hg38 reference genome

Reads were mapped against the hg38 reference genome, both the number and percentage of reads mapped were averaged and recorded in Table 6.9. The number of mapped reads was calculated using flagstats on BAM files which contain both forward and reverse reads. Previous values for total sequences were obtained using FastQC were obtained from the forward read files, hence mapped reads was a larger number, closer to double the values obtained above from FastQC. The final median mapped total is approximately half the number of initial median reads for batches - 1 and 2 as it is comprised of the assembled, unassembled reads. Per sample statistics are shown in Appendix D, Table 8.11.

Batch	Samples included	Median reads	Initial median % mapped	Final median mapped
1	31/31	2,315,597	98.99	1,175,678
2	38/41	2,164,832	99.7	1,142,011

Table 6.9: Initial sample mapping summary statistics with median average by batch, mapped reads counts both reads in a pair separately allowing both paired and singleton mapped reads to be included. Batches 1 and 2 displayed similar counts of reads. Final median mapped totals are lower due to the file being comprised of assembled and unassembled reads.

Contamination estimation

VerifyBamID failed to identify any sample with a contamination requiring further investigations. None of the 69 samples were reported to have contamination above 0.6%. Individual freemix values and percentage contamination estimates are shown per sample in Appendix D, Table 8.12.

6.4.2 Sample depth of coverage

Coverage was calculated using the final BAM files from the merging of reads and GATK base quality recalibration. Mean depth of coverage per sample was calculated for all targeted bases as shown below in Figure 6.8. Mean depth was sufficient for most HaloPlex samples with all bar two above 20x depth. Two samples with below 20x mean depth of coverage are shown below in Table 6.10. Calculating the mean sample depth of coverage for HaloPlex using all the positions covered in the amplicon kit gives a mean depth of 86.8x.

Sample	Mean Depth	% bases>1x	% bases> 5x	% bases> 10x	% bases> 20x	% bases> 30x	% bases> 50x
CRA-129	17.72	93.4	78.1	61.7	34.9	18.2	4.9
PAC0133	19.91	91.8	74.8	58.1	35.1	21.2	8.5

Table 6.10: Averaged mean coverage over all samples was 86.8x across all targeted bases. Two samples returned mean depth of coverage below 20x which were CRA-129 and PAC0133. Mean depth of coverage for these samples was 17.7x and 19.9x respectively so they were still included in the study for variant calling.

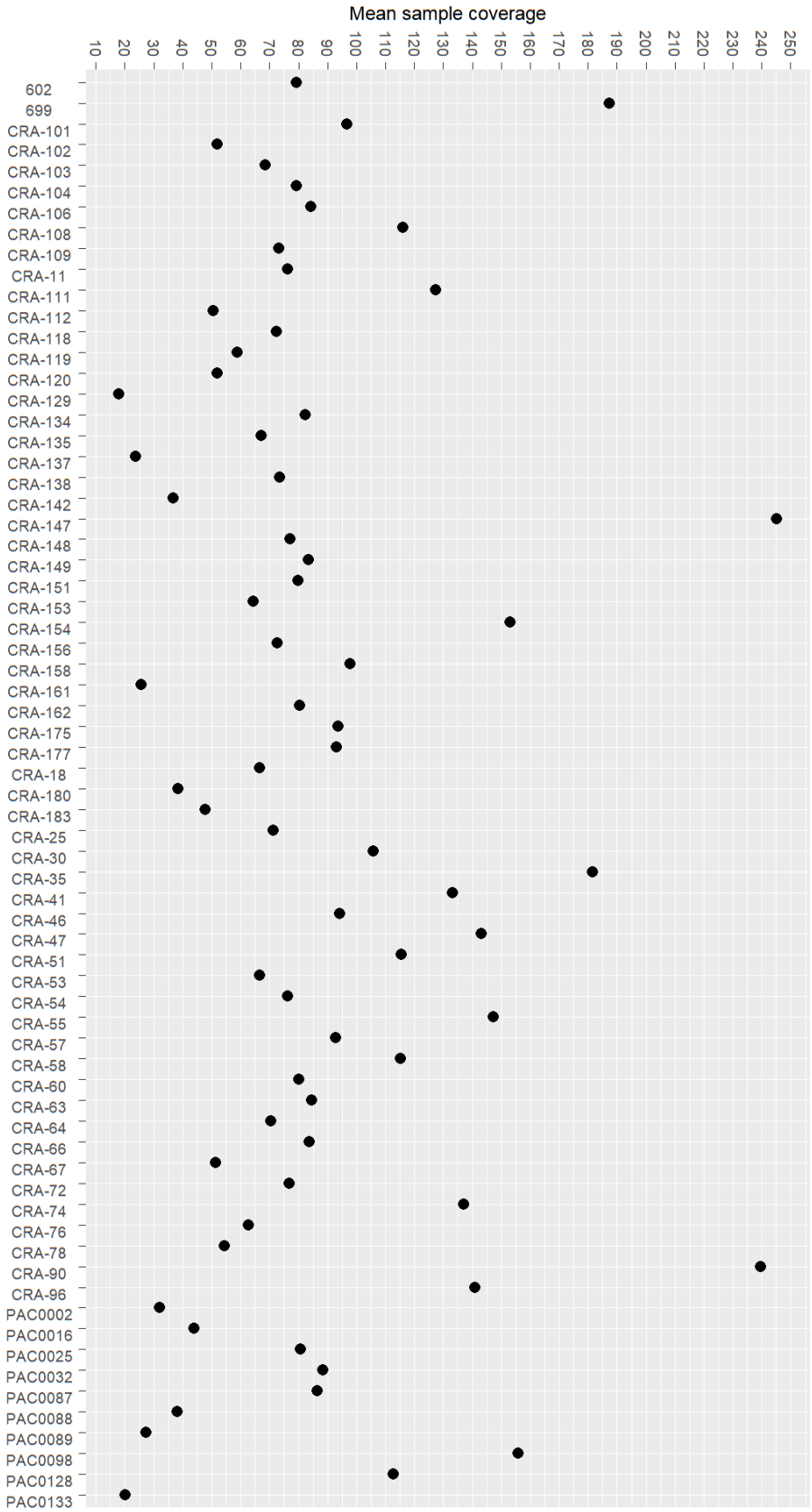

Figure 6.8: Average depth of coverage across samples. As indicated by the spread of the points samples displayed a range of depths of coverage ranging from on average 17.7x to 245x.

FCGR loci depth of coverage

Median depth of coverage per base over all samples was visualised over the FCGR low affinity locus containing all FCGR2 and FCGR3 genes in the region 161.49 to 161.69 Mb on chromosome 1 shown in Figure 6.9. In the figure Panel A shows the locations of genes over the region. Panel B shows the distribution of amplicon probes over genes was not even, with depth of amplicons describing how many amplicon targets overlapped at the base. Amplicon depth was not uniform and did not cover all bases in the region of the FCGR locus with Panel C describing the positions with no amplicon coverage. 2,375 amplicons overlap positions with no read coverage with a mapping quality of 20 as is used in variant calling. To investigate if this was caused by mapping quality of zero reads not being counted or if there was no coverage for the amplicons the mapping quality filter was disabled. In total 138 amplicons had positions of zero depth with a mapping quality of zero.

Panels D of Figure 6.9 describe the depth of coverage across the region with a mapping quality filter of 20, as is performed for variant calling, whilst Panel E shows depth of coverage with the filter lowered to include reads with mapping quality of zero. While depth of coverage appears generally good with the mapping quality filter set to include reads only above 20 as used in variant calling there are obvious gaps in coverage. Gaps are particularly large over *FCGR2C* and *FCGR2B* as shown in Panel D by the two red rectangles in Figure 6.9. Panel E shows that the gap in coverage for genes *FCGR2C* and *FCGR2B* is filled when mapping quality filters are set to zero.

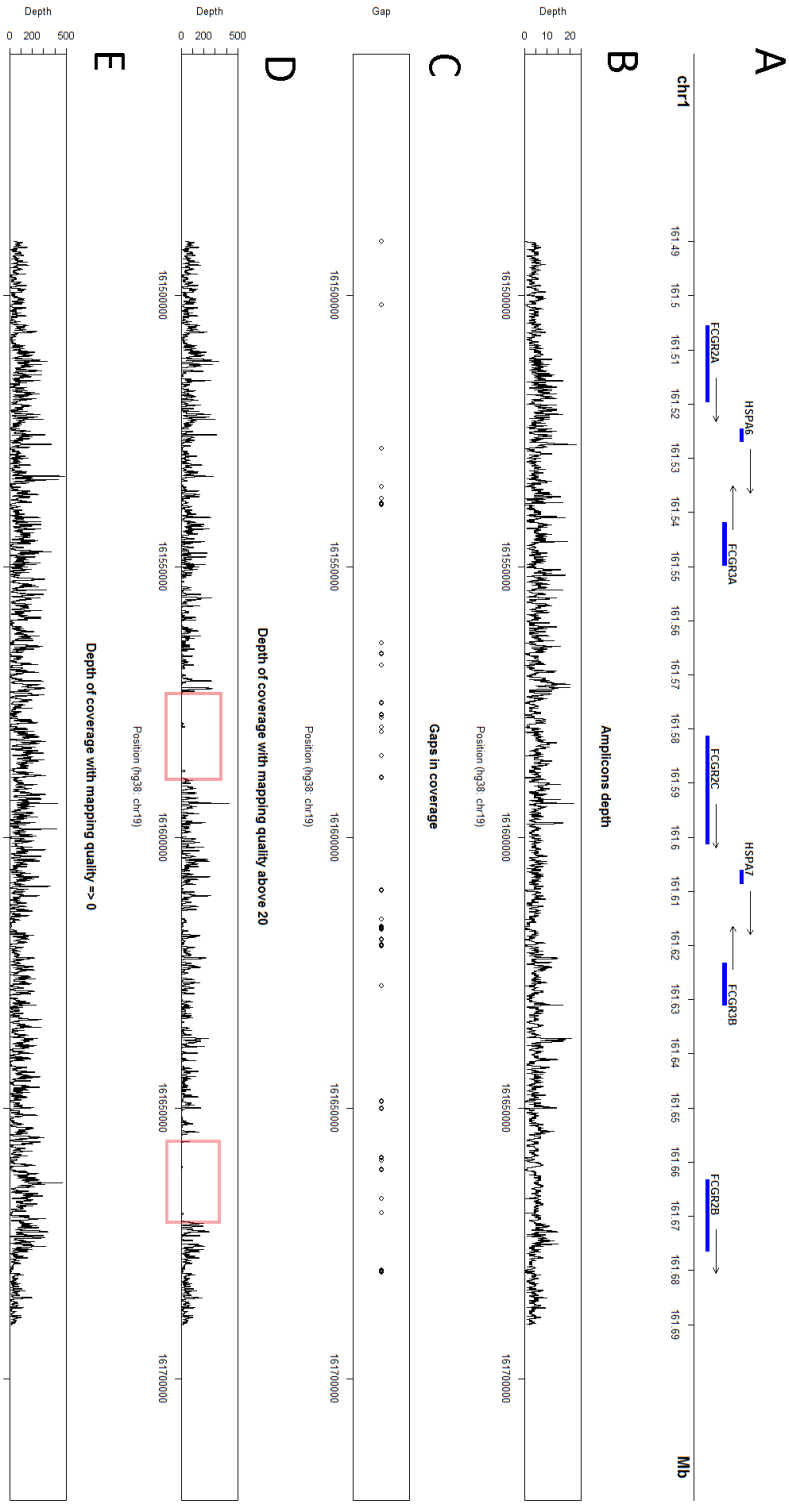

Figure 6.9: Median depth of coverage across the FCGR locus. A) Genes across the FCGR Locus on chr1:161.49-161.69 Mb. B) Depth of amplicon coverage across bases in the FCGR locus. C) Gaps in amplicon coverage across the FCGR locus. D) Median coverage across samples at each base across the FCGR locus with a mapping quality filter of above 20. Highlighted in red are large gaps spanning over part of the genes *FCGR2C* and also *FCGR2B*. E) Median coverage across samples at each base across the FCGR locus with mapping quality filter lowered to zero allowing reads mapped with qualities below the default threshold of 20.

6.4.3 Homology between FCGR genes

FCGR genes present a problem for short read mapping as the members of each family share large similarities due to the ancestral duplication and crossover events. Shown below in Figure 6.10 are global alignments for genes within FCGR 1, 2 & 3. Alignments show sequence similarity over the entire gene including introns and exons. Similarities and between the genes are also summarised in Table 6.11.

Gene	Length	Gene-2	Length-2	Coverage (%)	Similarity (%)
FCGR1A	9829	*FCGR1B*	8959	91	98
FCGR1A	9829	*FCGR1C*	8991	91	99
FCGR2A	13129	*FCGR2B*	15494	34	94
FCGR2A	13129	*FCGR2C*	18464	66	98
FCGR2B	15494	*FCGR2C*	18464	91	99
FCGR3A	8859	*FCGR3B*	8768	99	97

Table 6.11: Global Sequence alignment of FCGR genes showing overall similarity between genes over transcript regions including introns and exons. High homology can be seen within all FCGR1 and all FCGR3 genes but rearrangements of FCGR2 genes show divergence of *FCGR2A* and *FCGR2B* along with the unequal cross-over events which led to *FCGR2C*.

Within FCGR1 genes there is high similarity between *FCGR1A*, *FCGR1B* and *FCGR1C* at 98% and 99% respectively over 91% of either gene as shown in Panel A of Figure 6.10. Both plots support the hypothesis that the FCGR1 genes arose from ancestral gene duplications due to the high coverage and similarity between each other.

FCGR2 genes in Panel B of Figure 6.10 show the most divergence, particularly between *FCGR2A* and *FCGR2B* which only report 34% coverage but with 94% similarity. *FCGR2B* is 2365bp larger than *FCGR2A* despite being believed to have arisen from a duplication event with *FCGR2A*, however further large structural rearrangements have led to the development of the inhibitory ITIM motif unique to *FCGR2B* and low coverage relation to *FCGR2A*.

FCGR2C is believed to have derived from an equal cross-over event of *FCGR2A* and

FCGR2B. Similarity between *FCGR2B* and *FCGR2C* is shown with 91% of *FCGR2B* matching to *FCGR2C* with 99% similarity. This agrees with the claims that the first five exons of *FCGR2C* are identical to *FCGR2B*. The plot located in the lower right of Panel B shows similarity between *FCGR2A* and *FCGR2C* with 66% of *FCGR2A* covered at 98% similarity to *FCGR2C*. This agrees with evidence in literature that exons 6-8 of *FCGR2C* are from *FCGR2A* as part of the unequal cross over. Due to the ancestral duplications, cross-over events and subsequent rearrangements over each of the FCGR2 genes they have high similarity for some segments but others which have undergone further rearrangements match less strongly as indicated by the lower E-value segments and the lighter shaded regions of similarity in Figure 6.10.

FCGR3A and *FCGR3B* both report high similarities of 97% with high coverage of 99%. Differences have been recorded between the genes in the transmembrane regions with a reported 9 substitutions and 21 amino acids missing from cDNA. FcγRIIIa is attached to cells by a GPI anchor whereas FcγRIIIb proteins are transmembrane anchored[292].

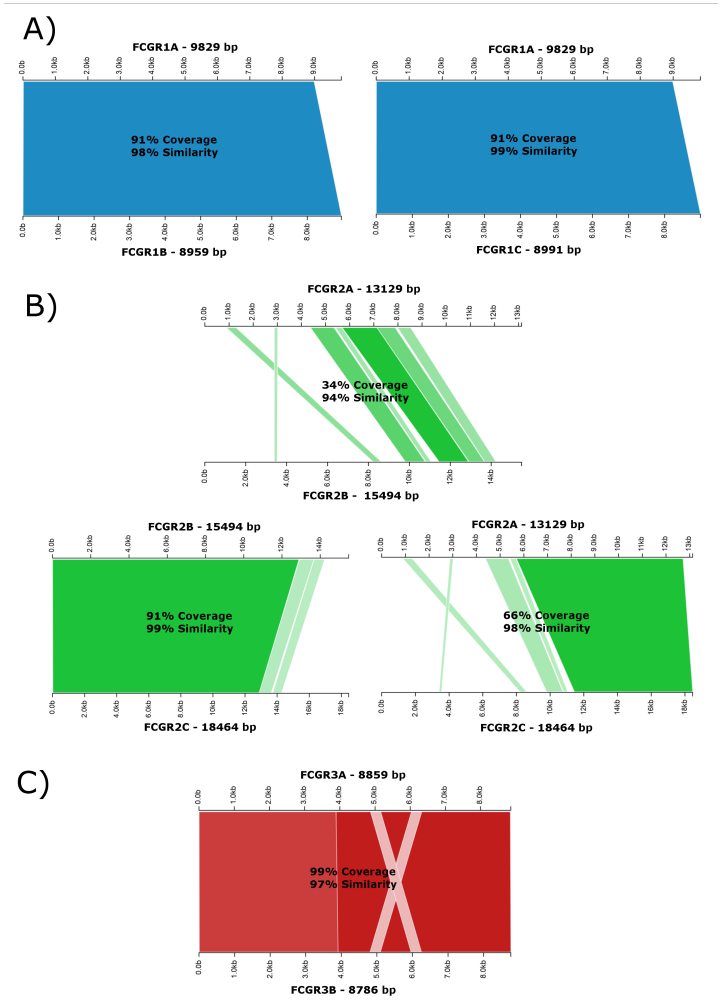

Figure 6.10: Visualisation of global alignments of FCGR genes. In each of the plots the gene nucleotide sequences were aligned using BLASTN. Regions of higher E-values score are indicated with darker shading and conversely lower scores as indicated by lighter shading. Where no sequence matches could be detected by BLAST between the sequences there is no shading. **A)** Similarity between FCGR1 genes shaded in blue is shown in two plots between *FCGR1A* - *FCGR1B* (left) and *FCGR1A* - *FCGR1C* (right). *FCGR1A* is longer than both *FCGR1B* and *FCGR1C* by approximately 1000bp (9%) and as plots are drawn to the same width a slant is visible in both plots due to the scale. Both *FCGR1B* and *FCGR1C* have 98% and 99% similarity to *FCGR1A* respectively both with 91%. **B)** Similarity between FCGR2 genes is shown in three plots coloured in green. In the upper plot similarity between *FCGR2A* and *FCGR2B* is shown. *FCGR2B* is 2365bp larger than *FCGR2A* hence the plot is slanted. There is high similarity at 94% for 34% of *FCGR2A* with *FCGR2B* and a fragmented match pattern with multiple segments of lower E-value alignments in lighter shading. In the lower-left plot similarity between *FCGR2B* and *FCGR2C* is shown with 91% of *FCGR2B* matching to *FCGR2C* with 99% similarity. The plot located in the lower right of Panel B shows similarity between *FCGR2A* and *FCGR2C* with 66% of *FCGR2A* covered at 98% similarity to *FCGR2C*. **C)** *FCGR3A* and *FCGR3B* also arose due to ancestral duplication events, hence the 99% coverage at 97% similarity, two rearrangement event appear visible as indicated by the lighter and crossed segments.. Alignments visualised using Kablammo (http://kablammo.wasmuthlab.org)

6.4.4 VCF pairwise identity matrices

From VCF files similarity was calculated using the number and percentage of shared variants per sample. 16 samples were identified with excessive similarity compared to most samples which ranged between 25 - 45% identical. The 16 samples displayed similarity to another sample of above 76% as shown in Figure 6.11.

	CRA-11	CRA-102	CRA-25	CRA-104	CRA-63	CRA-109	CRA-96	CRA-135	CRA-137	CRA-111	CRA-147	CRA-142	CRA-55	CRA-153	CRA-30	CRA-156
CRA-11	100	85.77	31.4	29.91	32.13	32.21	40.52	37.71	35.58	41.03	44.48	41.54	46.95	46.7	36.94	33.53
CRA-102	92.89	100	31.1	30.69	35.58	35.72	40.89	42.04	36.64	41.81	42.69	41.53	45.5	45.87	35.81	37.1
CRA-25	38.73	35.42	100	91.75	34.52	34.42	39.25	35.47	33.05	39.2	42.3	39.2	39.46	38.52	43.72	39.46
CRA-104	37.3	35.33	92.77	100	34.48	34.75	39.48	35.18	33	39.16	41.39	39.11	38.84	38.42	43.94	40.01
CRA-63	40.62	41.54	35.4	34.97	100	94.61	41.11	42.4	36.85	41.54	37.02	35.67	40.89	39.76	30.28	31.3
CRA-109	40.8	41.77	35.35	35.29	94.77	100	41.07	42.09	36.75	41.39	37.34	35.46	40.74	39.99	30.28	31.03
CRA-96	43.21	40.25	33.94	33.76	34.67	34.58	100	86.05	76.24	94.18	39.25	36.89	41.03	41.84	40.3	35.94
CRA-135	42.61	43.86	32.5	31.87	37.89	37.55	91.19	100	77.47	91.96	37.36	35.63	39.91	40.01	38.71	40.25
CRA-137	45.53	43.29	34.3	33.86	37.3	37.13	91.49	87.73	100	91.98	40.68	40.79	42.8	43.02	39.97	37.46
CRA-111	43.69	41.11	33.85	33.44	34.98	34.8	94.06	86.66	76.54	100	39.11	37.02	41.15	41.42	40.15	36.16
CRA-147	43.96	38.95	33.89	32.8	28.93	29.14	36.38	32.67	31.41	36.29	100	85.31	42.95	42.74	35.71	31.87
CRA-142	45.97	42.43	35.17	34.7	31.21	30.98	38.28	34.89	35.27	38.47	95.52	100	43.28	43.71	37.34	34.23
CRA-55	44.56	39.87	30.37	29.56	30.69	30.53	36.51	33.52	31.74	36.68	41.25	37.12	100	89.65	34.17	29.72
CRA-153	44.94	40.75	30.05	29.64	30.26	30.38	37.76	34.07	32.35	37.43	41.62	38.01	90.9	100	33.74	30.01
CRA-30	41.74	37.36	40.06	39.82	27.06	27.01	42.71	38.71	35.29	42.61	40.83	38.13	40.68	39.62	100	83.73
CRA-156	41.38	42.27	39.48	39.59	30.55	30.23	41.59	43.95	36.12	41.9	39.8	38.17	38.64	38.49	91.43	100

Figure 6.11: From pairwise comparisons of VCF files for each sample 16 samples were identified as excessively related with above 76% similarity to at least one other samples VCF file.

Samples with excessive similarity were able to be grouped as seven individuals as shown in 6.12. All individuals were sequenced twice bar individual-4 who was sequenced four times. Results therefore suggest unknown duplicate samples from the same individuals were present in the dataset.

Individual	Sample(Mapped reads)	Samples excluded...	..	..
1	CRA-11 (1,122,048)	CRA-102 (866,035)		
2	CRA-25 (1,159,596)	CRA-104 (1,030,241)		
3	CRA-63 (1,281,373)	CRA-109 (1,216,491)		
4	CRA-96 (1,817,345)	CRA-135 (1,122,382)	CRA-137 (327,744)	CRA-111 (1,710,477)
5	CRA-147 (2,591,376)	CRA-142 (472,006)		
6	CRA-55 (2,117,090)	CRA-153 (960,369)		
7	CRA-30 (1,550,694)	CRA-156 (1,096,739)		

Table 6.12: HaloPlex samples from same individuals. Grouping duplicate samples allows identification of several individuals who have been sequenced several times in the sample cohort.

Due to the presence of duplicate samples when performing pooled analyses such as joint variant calling or creation of pooled references for variant calling in CNVkit only the sample with the most reads from each individual was selected to be used. All samples excluded are summarised below in Table 6.13. In total 60 samples were remaining therefore to be used in pooled variant calling or in the creation of a pooled reference file for CNVkit.

SampleID	Batch/run	No. reads	Exclusion Reason
CRA-117	2	98	Insufficient data
CRA-139	2	1,755	Insufficient data
CRA-39	2	12,746	Insufficient data
CRA-102	2	866,035	Duplicate individual
CRA-104	1	1,030,241	Duplicate individual
CRA-109	2	1,216,491	Duplicate individual
CRA-135	2	1,122,382	Duplicate individual
CRA-137	1	327,744	Duplicate individual
CRA-111	1	1,710,477	Duplicate individual
CRA-142	1	472,006	Duplicate individual
CRA-153	1	960,369	Duplicate individual
CRA-156	2	1,096,739	Duplicate individual

Table 6.13: Samples excluded from the study due to insufficient reads or being a duplicate with fewer reads and hence not selected to be used in joint calling to avoid biases being introduced. A total of 60 samples were therefore remaining and used in variant calling.

6.4.5 Variant analysis

Joint genotyping of the 60 samples was performed to increase the power of variant calling and genotyping. The median number of variants called per sample was 1,477 over all targeted regions. For each of the samples the total number of variants showed large variance across the dataset. To better understand and explain the variability across the samples depth was plotted against the total variants per sample and the ratio of heterozygous to homozygous variants as shown in Figure 6.12. All 60 samples are organised by total number of variants per sample ranging from 2,377 to 12,559. From the plot it can be seen that samples with a fraction below 0.8 (80%) of bases covered at a minimum of 20x depth and below 6,000 total variants are those which have a more variable Het:Hom ratio.

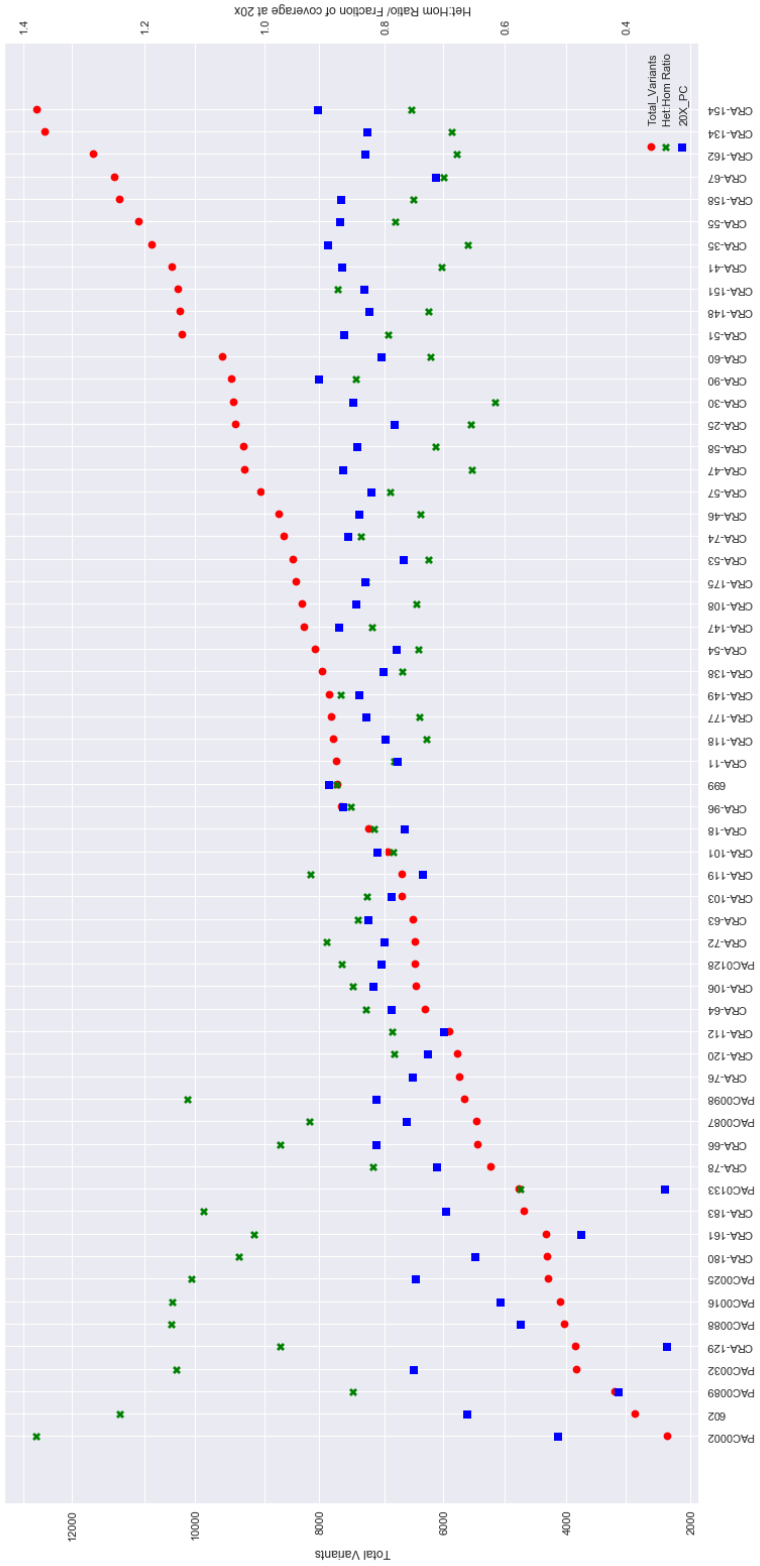

Figure 6.12: All 60 samples organised by total number of variants per sample (red) ranging from 2,377 to 12,559. The ratio of heterozygous to homozygous variants is shown on the secondary axis (green) along with the fraction of the sample covered at a minimum of 20x depth (blue). For lower totals the Het:Hom ratio is more variable, as total variants increase the ratio becomes more constant around 0.8 and percentage of the target covered with a depth of coverage of 20x also stabilises around 80% (0.8).

FCGR Variants

All FCGR low affinity locus (chr1:161.49 - 161.69MB) and any FCGR1 gene variants were filtered to retain any exonic variants. In total 1,982 FCGR locus variants were called across the 60 HaloPlex samples. Excluding variants which are not exonic or splicing returned only 47 variants, this was further reduced to 35 by excluding *HSPA6* variants as shown below in the Table 6.14.

There were no coding FCGR1 gene variants. Variants were divided between other genes as: 9 - *FCGR2A*, 6 - *FCGR2B*, 5 - *FCGR2C*, 11 - *FCGR3A*, 4 - *FCGR3B*. 16 of the 35 variants are synonymous, while 5 variants are either unknown or annotated ".".

In order to identify the HNA-1 haplotype of each sample there are six *FCGR3B* positions in which the allele must be known as described in Table 6.4. No variants were called at any of the six sites in any of the 60 samples. Therefore using the conventional reference genome for hg38 failed to call the variants necessary to determine the HNA-1 haplotypes using short read mapping.

Gene	Variant	Function	rsID	gnomAD	fathmm-MKL
FCGR2A	chr1:161506414:C:T	stopgain	rs9427397	0.1114	0.022
FCGR2A	chr1:161506415:A:G	nonsynonymous	rs9427398	0.1122	0.02
FCGR2A	chr1:161506590:C:T	synonymous	rs151051324	0.0009	.
FCGR2A	chr1:161509955:A:G	nonsynonymous	rs1801274	0.5018	0.009
FCGR2A	chr1:161510859:A:G	synonymous	rs11810143	0.1118	.
FCGR2A	chr1:161510888:C:T	nonsynonymous	rs199502630	0	0.018
FCGR2A	chr1:161518033:A:C	nonsynonymous	rs146883516	0.0019	0.007
FCGR2A	chr1:161518052:C:T	synonymous	rs145803059	0	.
FCGR2A	chr1:161518073:C:T	synonymous	rs12029217	0.1054	.
FCGR3A	chr1:161543028:T:A	nonsynonymous	rs759815530	.	0.052
FCGR3A	chr1:161543083:T:A	nonsynonymous	rs115866423	0.0099	0.036
FCGR3A	chr1:161544752:A:C	nonsynonymous	rs396991	0.3331	0.017
FCGR3A	chr1:161548424:T:C	nonsynonymous	rs148181339	0.2646	0.002
FCGR3A	chr1:161548434:T:C	synonymous	rs142368299	0.0051	.
FCGR3A	chr1:161548496:C:T	nonsynonymous	rs200727785	0.0166	0.002
FCGR3A	chr1:161548507:G:T	nonsynonymous	rs52820103	0	0.007
FCGR3A	chr1:161548524:C:T	synonymous	rs114535887	0.0433	.
FCGR3A	chr1:161548543:A:T	nonsynonymous	rs10127939	0.0687	0.018
FCGR3A	chr1:161548543:A:C	nonsynonymous	rs10127939	0.0404	0.019
FCGR3A	chr1:161549022:C:T	nonsynonymous	.	.	0.018
FCGR2C	chr1:161589781:C:T	unknown	rs138747765	0.303	0.063
FCGR2C	chr1:161591232:G:T	unknown	.	.	.
FCGR2C	chr1:161592145:G:A	unknown	rs777594828	0	.
FCGR2C	chr1:161595591:A:G	.	rs76277413	0.1154	0.001
FCGR2C	chr1:161599779:C:T	unknown	rs138731942	0.16	.
FCGR3B	chr1:161626224:A:G	synonymous	rs71632957	0.0003	.
FCGR3B	chr1:161626242:T:C	synonymous	rs114169903	0.0252	.
FCGR3B	chr1:161626250:G:A	nonsynonymous	rs71632958	0.0001	0.002
FCGR3B	chr1:161626282:T:C	nonsynonymous	rs71632959	0.0003	0.001
FCGR2B	chr1:161671594:G:A	synonymous	rs6665610	0.1721	.
FCGR2B	chr1:161673192:G:A	synonymous	rs146214041	0.0027	.
FCGR2B	chr1:161673195:G:A	synonymous	rs182968886	0.1206	.
FCGR2B	chr1:161674008:T:C	nonsynonymous	rs1050501	0.1645	0.043
FCGR2B	chr1:161675262:C:T	nonsynonymous	rs28651835	0.0047	0.025
FCGR2B	chr1:161675268:T:G	nonsynonymous	rs148534844	0.0001	0.053

Table 6.14: 35 exonic variants were called in FCGR genes across all 60 HaloPlex samples. gnomAD AAFs shown in the table are for the NFE group with functions from known gene annotations. There were no coding FCGR1 gene variants, variants were divided between other genes as: 9 - *FCGR2A*, 6 - *FCGR2B*, 5 - *FCGR2C*, 11 - *FCGR3A*, 4 - *FCGR3B*. 16 of the 35 variants are synonymous, while 5 variants are either unknown or ".".

6.4.6 Calling FCGR variants using customised reference genomes

Due to the high homology between FCGR genes of the same family many of the reads map with mapping quality of zero and so were excluded from being used in variant calling due to the lack of confidence. As shown in Figure 6.9 coverage is lacking over the FCGR low affinity locus, particularly over *FCGR2B* and *FCGR2C* but also at positions within *FCGR3B*. To enable use of the data contained within the samples mapping using customised reference genomes was performed, this enabled non-uniquely mapping reads to be recovered and assessed.

To adjust for the regions of homology between the FCGR genes which prevented mapping, single FCGR gene references were utilised to map all alleles between the homologous sites. Only mapping to a single gene will allow a pile-up of reads at a single site. From the pile-up it is then possible to interpret the possible combination of alleles for the two sites by considering the reference and alternate alleles at both sites (assuming roughly equal depth of coverage) and the copy number for the genes. By using all of the information it may be possible to determine the possible HNA-1 haplotype and also the prevalence of the *FCGR2C* rs759550223 variant status of each sample. Both of these cases are applicable as there are only two FCGR3 genes and the *FCGR2C* variant is in exon three, which is part of ancestral unequal cross-over inherited from *FCGR2B*.

HNA-1 Haplotype prediction

No variants over any sample were called for any of the six SNPs using the conventional hg38 reference genome initially. Using the per gene reference approach the six variants encoding the HNA-1 system in *FCGR3B* were investigated. For each sample using the customised *FCGR3B* gene reference with copy number data from MLPA and CNVkit the allele totals and combined amino acids per sample were used to infer the HNA-1 haplotype are shown in full under Appendix D - Table 8.15. To infer the haplotypes each of the six SNP sites were considered with the reads supporting the reference and alternate alleles counted.

An example of the HNA-1 haplotype determination is shown below in Table 6.15 in which the six sites are shown along with the total reference and alternate allele reads.

As this method uses the ratios of the alleles, changes in copy number will alter the expected ratios between sites. For this sample CRA-101 the copy number is two across both FCGR3 genes. For the sample the first variant returned 96 reference T allele reads and no alternate allele reads. In this case copy number is not required as both sites (*FCGR3A* and *FCGR3B*) can only be homozygous reference, producing the amino acid isoleucine (I). Considering the second variant site which returns an approximately equal number of reference (50) and alternate allele reads (47). In this scenario there are two possibilities; one site is homozygous reference and the other homozygous for the alternate allele, alternatively both sites are heterozygous. Without additional data the two sites cannot be distinguished further, so the site in *FCGR3B* could be an asparagine or aspartic acid residue (N/D). For the third site all reads (97) were homozygous for the reference allele and so both of the sites must be reference and the amino acid is alanine (A). Conversely for the fourth site all of the recorded reads (263) are homozygous for the alternate allele making the amino acid serine (S). For both the fifth and sixth variants both positions have approximately the same number of reference allele reads (125) and alternate allele reads (141). For both of these variants, both *FCGR3A* and *FCGR3B* sites could be heterozygous or one site homozygous reference and one homozygous alternate.

#	rs2290834		rs147574249		rs5030738		rs448740		rs527909462		rs200688856		#
Sample	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	#
Amino acid	**I**	**V**	**N**	**D**	**A**	**D**	**N**	**S**	**L**	**L**	**S**	**R**	#
Nucleotide	**T**	**C**	**T**	**C**	**G**	**T**	**T**	**C**	**A**	**G**	**G**	**C**	**Amino acids**
CRA-101	96	0	50	47	97	0	0	263	125	141	125	141	I, N/D, A, S, C/T, S/R

Table 6.15: Example of the method used to determine a samples HNA-1 haplotype from allele counts obtained when using a custom reference for *FCGR3B* with the removal of the *FCGR3A*. Copy number across the locus was neutral at 2 for the sample. From the ratios of alleles at the six variants which comprise the HNA-1 haplotype, three variant sites were able to identified as homozygous for either the reference or alternate allele can be determined, enabling the prediction of the HNA-1 haplotype as the HNA-1B for the sample.

For the above example sample all six variant sites were combined to give the haplotype of: I, N/D, A, S , C/T , S/R. Using the defined haplotypes as shown in Table 6.5 the possible haplotype can be determined. From the known variant sites: one (I), three (A) and four (S) the alleles can only fit with the HNA-1B haplotype,

this predicted haplotype is also supported by predictions made using MLPA data.

For all of the samples the calculation of the HNA-1 haplotypes were performed and were summarised as shown below in Table 6.16.

HNA-1 haplotypes	Number of Samples
A (NA1) only	6
B (NA2) only	19
A + B	27
A + C	3
C (SH) only	1
NA2*02 only	2
C + NA1*02	2
Total	60

Table 6.16: Summary of haplotypes predicted based upon allele ratios between six homologous sites. The most commonly identified haplotype was a combination of HNA-1A (NA1) + HNA-1B (NA2) forms which was found in 27 samples. 19 samples reported with the HNA-1B (NA2) haplotype. *Some samples have copy number gains or losses at this site so may have more or less than 2 of the indicated haplotypes.

The most commonly identified haplotype was a combination of HNA-1A + HNA-1B which was found in 27 samples followed by 19 HNA-1B samples. Samples containing both the HNA-1A + HNA-1B haplotype reported five of the six SNPs assessed as being heterozygous with an approximately equal allele ratio.

Some samples had copy number gains or losses over the FCGR3 genes resulting in more or less than two haplotypes. MLPA data for SNPs and copy number was only available for 57 of the 60 samples, agreement was found to be reasonable with 41 of the 57 (71.9%) samples in agreement.

Seven samples displayed copy number disagreements between CNVkit and MLPA which caused several minor disagreements in HNA-1 haplotype prediction as shown in Table 6.17. Haplotypes predicted were in agreement but differed in the numbers of copies predicted. These samples accounted for just over 12.2% of the mismatches. Hence if you counted haplotype matches without considering copy number the agreement would be around 84.1% between HaloPlex and MLPA.

Sample	**HaloPlex CN**	**HaloPlex HNA-1**	**MLPA CN**	**MPLA HNA-1**
CRA-46	2	BB	1	B
CRA-154	2	AA	1	A
PAC0032	3	AB + A/B	4	ABB
CRA-147	3	AB + A/B	2	AB
PAC0016	4	AB + A/B + A/B	3	ABB
PAC0133	5	AB + A/B + A/B + A/B	4	AABB
PAC0089	5	AB + A/B + A/B + A/B	4	AABB

Table 6.17: Seven samples differ between HaloPlex and MLPA for the HNA-1 haplotype predictions by the copy number called over the FCGR3B gene. The HaloPlex CN number indicates the copy number call made by CNVkit whereas the MLPA CN is the copy number from interpretation MLPA. All samples agree on the HNA-1 haplotypes but not the copy number.

Nine samples were found to disagree with the MLPA calls for the predicted haplotypes as shown in Table 6.18. These nine samples accounted for 15.7% of mismatches.

Sample	HaloPlex CN	HaloPlex HNA-1	MLPA CN	MPLA HNA-1
CRA-129	2	NA2*02 + NA2*02	2	AB
CRA-161	2	AB	2	BB
CRA-55	2	A+SH	3	AAB
CRA-60	2	NA2*02 + NA2*02	2	AB
PAC0098	2	BB	2	AA
CRA-175	3	SH + NA1*02 and/or NA2*02	2	AB
CRA-54	3	A + SH + A/SH	3	AAB
CRA-74	3	SH + NA1*02 and/or NA2*02	3	BB
CRA-119	4	A + SH + A/SH + A/SH	3	BBBB

Table 6.18: Nine samples displayed disagreements in the predicted HNA-1 haplotypes between MLPA and HaloPlex. Seven of the nine disagreements all involve the alternate, less common haplotypes of NA2*02, NA1*02 or SH (C). The other two disagreements involve the A and B haplotypes.

Seven of the nine samples are disagreements involving the HNA-1C (SH) and NA2*02 or NA1*02 forms which were identified in HaloPlex but not from MLPA. Over the entire dataset only the A and B haplotypes were reported by MLPA.

FCGR2C - Q57X variant calling

For the *FCGR2C* variant rs759550223, encoding the amino acid change Q57X, allele ratios per sample were calculated as shown in Appendix D, Table 8.16. 8 of 60 samples were homozygous for the C nucleotide, meaning that both *FCGR2B* and *FCGR2C* both produce transcripts with a glutamine (Q) at position 57. 7 out of the 8 samples agreed giving a concordance of 87.5% with the MLPA prediction of glutamines at both positions. 5 of the 7 samples also agreed with the copy number called between HaloPlex and MLPA data. The disagreements in copy number were over the calling of a single gain of copy, one sample CRA-55 did not call a gain in from HaloPlex data and CRA-74 was called with HaloPlex data as a multi-copy number gene with regions at copy of two and three.

For the other 52 samples alleles were split between C and T. In theory using allele ratios it could be used to interpret the possible combinations assuming *FCGR2B* was not polymorphic as reported in literature. However in practice it was impossible to distinguish an XX from a heterozygote XQ in *FCGR2C*. This was caused by the variable depths seen over the 52 samples and copy number changes between *FCGR2B* and *FCGR2C* within samples. Overall concordance between MLPA and custom reference predictions hence was low at only 30% when considering all samples.

6.4.7 Copy Number Variation - CNVkit

Optimisation of CNV calling

Initial results from CNVkit were gained using a pooled reference of the 60 HaloPlex samples with a target bin size of 267bp. Results were compared against MLPA data as summarised in Table 6.19 with the percentage of matches calculated per gene across the FCGR 2 & 3 locus.

CNVkit was tested using the copy neutral sample CRA-25 and pooled reference using 267bp and 150bp bins also shown in Table 6.19. A full comparison against MLPA by sample can be seen in Appendix D, Figure 8.1 for each of the references and bin size tested.

Reference Used	**Gene**						
Pooled- 267 bp	**FCGR2A**	**HSPA6**	**FCGR3A**	**FCGR2C**	**HSPA7**	**FCGR3B**	**FCGR2B**
Match	46	51	49	39	44	44	42
Multiple	5	0	1	6	0	0	9
Mismatch	9	9	10	15	16	16	9
Match (%)	76.67	85	81.67	65	73.33	73.33	70
Multiple (%)	8.33	0	1.67	10	0	0	15
Mismatch (%)	15	15	16.67	25	26.67	26.67	15
CRA-25- 267 bp	**FCGR2A**	**HSPA6**	**FCGR3A**	**FCGR2C**	**HSPA7**	**FCGR3B**	**FCGR2B**
Match	44	48	48	38	42	42	39
Multiple	7	0	1	3	0	0	14
Mismatch	9	12	11	19	18	18	7
Match (%)	73.33	80	80	63.33	70	70	65
Multiple (%)	11.67	0	1.67	5	0	0	23.33
Mismatch (%)	15	20	18.33	31.67	30	30	11.67
Pooled- 150bp	**FCGR2A**	**HSPA6**	**FCGR3A**	**FCGR2C**	**HSPA7**	**FCGR3B**	**FCGR2B**
Match	51	57	56	41	49	49	46
Multiple	6	0	0	8	0	0	11
Mismatch	3	3	4	11	12	12	3
Match (%)	85	95	93.33	68.33	81.67	81.67	76.67
Multiple (%)	10	0	0	13.33	0	0	18.33
Mismatch (%)	5	5	6.67	18.33	18.33	18.33	5
CRA-25 - 150 bp	**FCGR2A**	**HSPA6**	**FCGR3A**	**FCGR2C**	**HSPA7**	**FCGR3B**	**FCGR2B**
Match	47	53	51	38	47	47	43
Multiple	10	0	1	10	1	0	14
Mismatch	3	7	8	12	12	13	3
Match (%)	78.33	88.33	85	63.33	78.33	78.33	71.67
Multiple (%)	16.67	0	1.67	16.67	1.67	0	23.33
Mismatch (%)	5	11.67	13.33	20	20	21.67	5

Table 6.19: Summary of MLPA comparison for each of the references tested in CNVkit. For each of the references and size of bins tested the copy number over genes were compared with results from MLPA at the level of gains, losses or neutral. If MLPA and CNVkit results were in agreement they were counted as a match. Multiple matches were defined as when a gene had multiple copy number calls (e.g. gains and neutral segments) where the majority segment agreed with MLPA but due to at least one of the segment calls disagreeing between MLPA and CNVkit. Mismatches were when the copy number call from CNVkit did not match with MLPA (e.g. Gain by MPLA and neutral by CNVkit). Results show that the best performing reference combination was the pooled with a bin size of 150bp. This reference combination scored the highest match percentage for each of the seven genes in the FCGR locus. Across all of the references the worst performing genes were *FCGR2C*, *HSPA7* and *FCGR3B* with mismatch percentages from 18.3% up to 31.67%. Ambiguous matches were worst for *FCGR2C* as multiple segment calls were often made over the gene with multiple copy number states.

Using a pooled reference with 267bp bins variant calls achieved reasonable matches for *FCGR2A*, *HSPA6* and *FCGR3A* at 76.67%, 85% and 81.67 % respectively. However visualisations of CNV calls revealed obvious deviations in copy number ratios which should be called as segmentations with change in copy number but were not called suggesting a lack of sensitivity as illustrated in Panel A of Figure 6.13.

Decreasing bin sizes from 267bp to 150bp increased the agreement with MLPA in both the pooled and CRA-25 references, an example of which for the pooled reference is shown in Panel A of Figure 6.13. Agreement with MLPA is achieved for the sample shown in Panel A by decreasing bin size, however two smaller CNVs in between *FCGR3A* - *FCGR2C* and *FCGR3B* - *FCGR2B* are now called as highlighted by red ovals in the figure. Both of these small CNVs are beyond the resolution of MLPA and therefore difficult to validate. Also over *FCGR2B* there are now two segments called over the gene, in other samples when multiple copy number segments over a gene were identified but were not equal they were grouped under the multiple segment category.

While some samples improved in agreement others failed to improve as indicated in Panel B of Figure 6.13. MLPA data called a loss over the region from *FCGR2C* to *FCGR3B*, which can be seen clearly in both of the pooled reference calls yet is not called as a loss segment. In the 150bp reference of Panel B an increase in the number of copy number points ratio points can be seen but it is not sufficient to induce a call in the region of the visible CNV.

Improvements were obtained from decreases in bin size in all the references. The worst performing regions were from *FCGR2C* to *FCGR2B* as shown in Panel B of Figure 6.13 regardless of bin size and the reference being used (Pooled or CRA-25) as shown in Table 6.19.

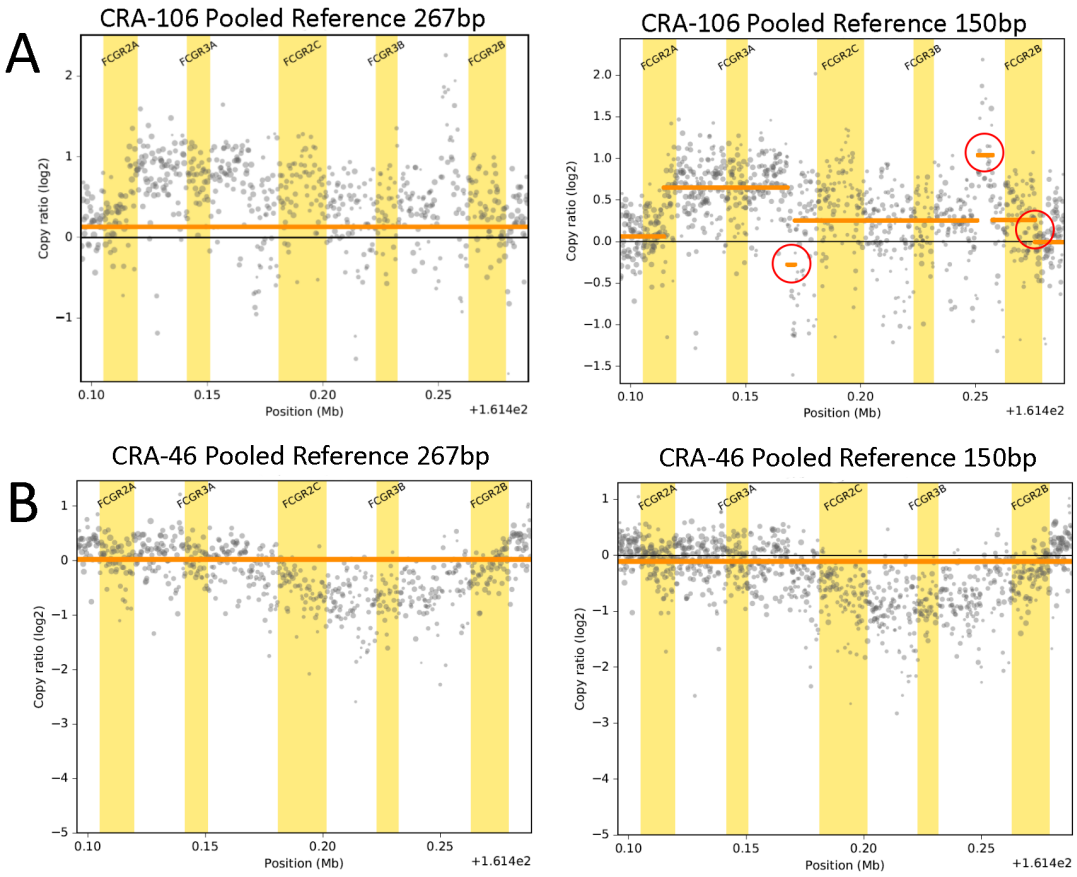

Figure 6.13: Examples of the effects of reducing bin size to increase the sensitivity of copy calling with a pooled reference. **A)** Sample CRA-106 with 267bp bins which fails to call a gain over the end of *FCGR2A* and *FCGR3A* but is called with 150bp bins in agreement with MLPA. Overall agreement increased as bin size was decreased however some samples also displayed over-calling of segments as highlighted in the three red circles on the right figure of Panel A. **B)** Not all obvious CNVs were called as a result of reducing bin size. Loss of copy can be observed over *FCGR2C* and *FCGR3B* which is not called with bin size of 267 or 150bp. Further reductions in bin size may call the CNV but would likely also increase false-positive calls due to insufficient depths of coverage leading to low data bins.

CNV calling of samples

Using the pooled reference with 150 bp bins and the CBS algorithm copy number variants were called. For HaloPlex samples copy number segment calls (gains, losses and neutral) are summarised per sample in Figure 6.14. Examination of total copy number calls per HaloPlex sample revealed that the majority of segments called were neutral for 46 of the 60 samples. 14 samples consisted of a majority of segments called as losses of which eight of the samples are named 'PAC*', all of which originated from the same sample collection batch. Gains of copy are the least called event with only sample PAC002 above a total of 5 gain segments called. Samples reporting the highest number of loss segments show little correlation with mean depth of coverage shown in Appendix D, Table 8.14. The lack of correlation with depth of coverage suggests that losses were not called due to lower read depths compared to the pooled reference.

Panel B of Figure 6.14 summarises CNV calls over the FCGR 2 and 3 genes in the FCGR locus. Results reveal all samples have at least one neutral segment in the region. 37 samples are entirely copy number neutral over the FCGR locus genes. 23 samples reported either gains or losses over the FCGR locus genes.

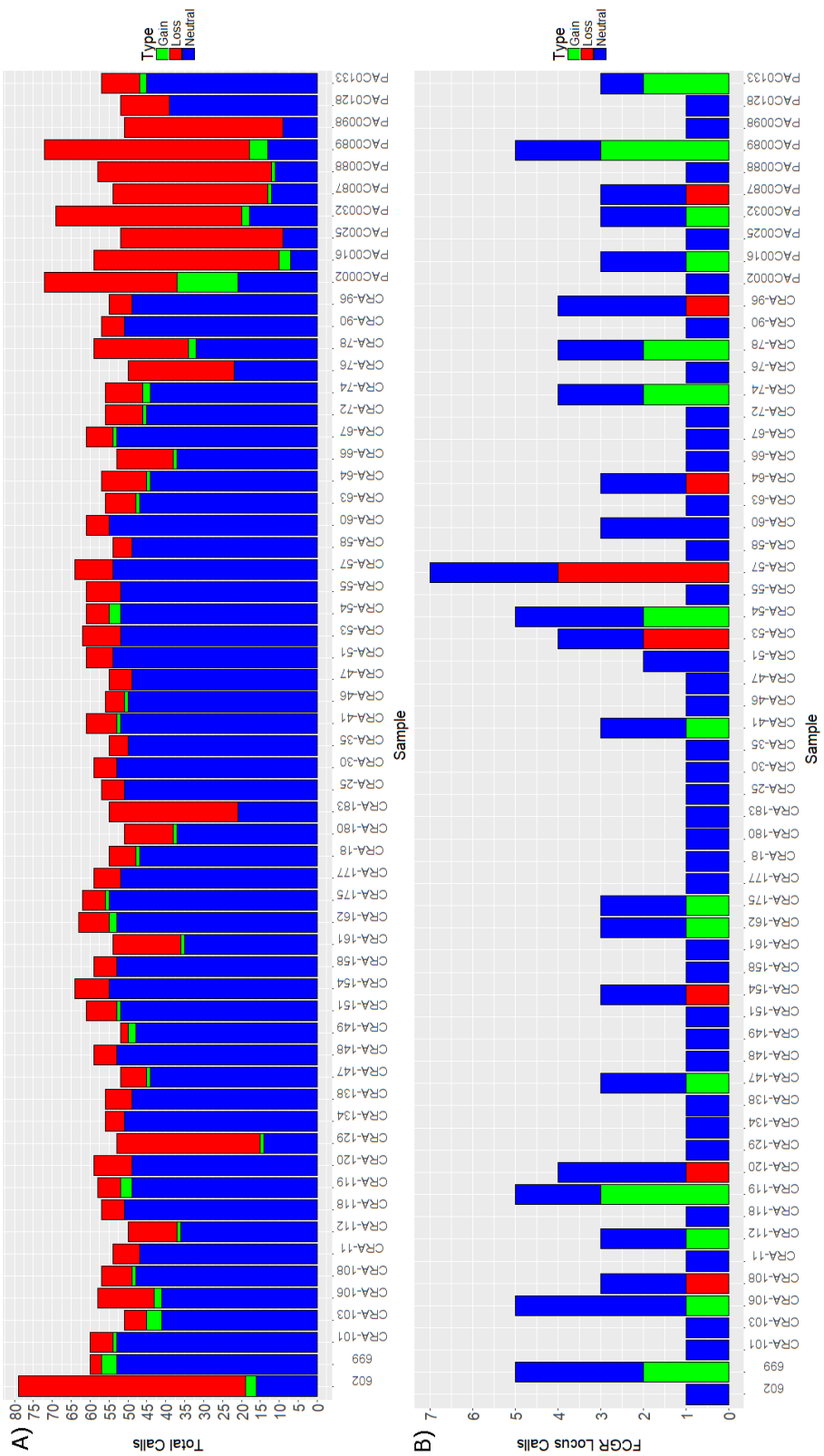

Figure 6.14: CNVkit Gains, losses and neutral segments per HaloPlex sample. **A)** Total CNV calls for each sample with 14 samples displaying the majority of segments as losses. Samples with majority losses all originate from the same batch. **B)** Restricting counts to just the FCGR genes showed all samples contained at least one neutral segment, which often spanned the entire FCGR locus.

Copy number variation over *FCGR3B*

To assess the prevalence of copy number changes which act in parallel with the HNA-1 haplotype all samples with CNVs called by CNVkit over the gene *FCGR3B* were reviewed. 18 samples were found with variants, of which 6 contain losses and 12 with gains of copy as summarised below in Table 6.20. Also shown in the table is the comparative predictions made using MLPA data.

Sample	CNVkit CN	HaloPlex HNA-1 haplotype	MLPA CN	MPLA HNA-1 haplotype
CRA-108	1	B	1	B
CRA-120	1	B	1	B
CRA-53	1	A	1	A
CRA-64	1	A	1	A
CRA-96	1	B	1	B
PAC0087	1	B	1	B
CRA-112	3	AB + A/B	3	AAB
CRA-147	3	AB + A/B	2	AB
CRA-175	3	SH + NA1*02 and/or NA2*02	2	AB
CRA-41	3	AB + A/B	3	ABB
CRA-54	3	A + SH + A/SH	3	AAB
CRA-74	3	SH + NA1*02 and/or NA2*02	3	BB
CRA-78	3	AB + A/B	3	AAB
PAC0032	3	AB + A/B	4	ABB
CRA-119	4	A + SH + A/SH + A/SH	3	BBBB
PAC0016	4	AB + A/B + A/B	3	ABB
PAC0089	5	AB + A/B + A/B + A/B	4	AABB
PAC0133	5	AB + A/B + A/B + A/B	4	AABB

Table 6.20: 18 Samples with copy number deviations from the expected diploid value of two were found when using CNVkit over *FCGR3B*. For each sample the predicted HNA-1 haplotypes are shown based on the calling method using customised reference genomes and also MLPA. All six samples with losses over the *FCGR3B* gene by MLPA and CNVkit showed agreement for the predicted haplotypes. Three samples showed agreement for the copy number call of three and possible haplotypes but MLPA predicted an exact combination. Five samples show agreement in the possible haplotypes but differ in the number of gains of copy predicted. Finally four samples with non HNA-1A or B haplotypes predicted by HaloPlex are not matched by MLPA which only had A and B forms called.

All six samples with a copy number of one show agreement of the predicted haplotypes between HaloPlex and MLPA . Three samples all with a copy number of three by both CNVkit and MLPA (CRA-112, CRA-41, CRA-78) show agreement. HaloPlex calls the three samples as AB + A/B but cannot resolve the extra copy further. MLPA for these samples predicts the haplotype as AAB, ABB and AAB for these samples respectively. Five samples (CRA-147, PAC0032, PAC0016, PAC0089, PAC0133) with copy gains show agreement in the possible haplotypes but disagree in

the copy number calls between methods. Four samples (CRA-175, CRA-54, CRA-74, CRA-119) had disagreements of haplotypes when alternative haplotypes to the A and B forms were identified from HaloPlex data.

Copy number regions in the FCGR locus

For the 26 HaloPlex samples which reported an alteration in copy number over the FCGR locus they were grouped based copy number regions (CNRs) breakpoints previously described in literature as CNRs 1 - 4. In addition two further breakpoints were defined, CNR-5 describes variants which span from *FCGR3A* to *FCGR2B* whilst CNR-6 describes copy number variants which span the almost the entire locus from *FCGR2A* to *FCGR2B*. Start and end positions of the CNVs and the CNRs are shown below in Figure 6.15 and grouped by copy number in Table 6.21. Exact start and end positions for the 26 samples are detailed in Appendix D, Table 8.17.

	Copy Number Region					
Copy number	**1**	**2**	**3**	**4**	**5**	**6**
0	0	0	0	0	0	0
1	1	2	0	5	0	1
3	0	2	0	8	0	1
4	0	0	0	2	0	1
5	0	0	0	1	1	1

Table 6.21: Copy number changes over the FCGR locus for the 26 Haloplex samples grouped by copy number regions. Results show only one loss for CNR-1. For CNR-2 an equal number of single gains and losses of copy were found with two samples each. No CNR-3 copy variations were found. For CNR-4 more gains were found, with 11 samples showing gains compared to six with a single copy loss. One sample was called with an extreme gain of three copies matching the CNR-5 description of covering the most of the FCGR locus. CNR-6 was for copy number variants which spanned the entire FCGR region. Four samples showed gains over the region and one with a single loss of copy.

1 of the 26 samples were grouped as a CNR-1 variant which started at *FCGR2C* and ended near the end of *FCGR3B* and was a single loss of copy. Four samples were identified as CNR-2 variants covering *FCGR2A* to *FCGR2C*, though one sample ended prior to the *FCGR2C* gene it was grouped into CNR2 as was the closest

matching CNR. CNR-2 variants were split as two single gains and two losses of copy. No samples were identified matching CNR-3 from *HSPA6* to *FCGR2C*. CNR-4 spanned from *FCGR2C* to *FCGR2B* which matched the locations for 16 samples split as 5 samples with losses and 11 with gains. Five samples did not quite fit the pattern of any of CNRs 1-4 hence additional CNRs were defined. A single sample overlapped *FCGR3A* to *FCGR2B*, which covers an additional gene compared to CNR-4 but not the entire region, this was termed CNR-5.

Using the described CNRs and the copy number calls from CNVkit the approximate breakpoints were estimated for each CNR as summarised in Table 6.22. Along with the estimated boundaries the expected genes the CNRs overlap consistent with current literature are also recorded.

CNR	**Start**	**End**	**Start-Gene**	**End-Gene**
1	161,564,172	161,637,469	*FCGR2C*	*FCGR3B*
2	161,514,434	161,599,992	*FCGR2A*	*FCGR2C*
3	N/A*	N/A*	*HSPA6*	*FCGR2C*
4	161,559,016	161,675,797	*FCGR2C*	*FCGR2B*
5	161,540,482	161,676,855	*FCGR3A*	*FCGR2B*
6	161,428,302	161,676,704	*FCGR2A*	*FCGR2B*

Table 6.22: CNR breakpoint locations estimated from CNV calls by CNVkit. For each of the CNRs the expected genes the CNR overlap based upon literature are also listed. * No CNR3 CNVs were called preventing estimation of the breakpoints.

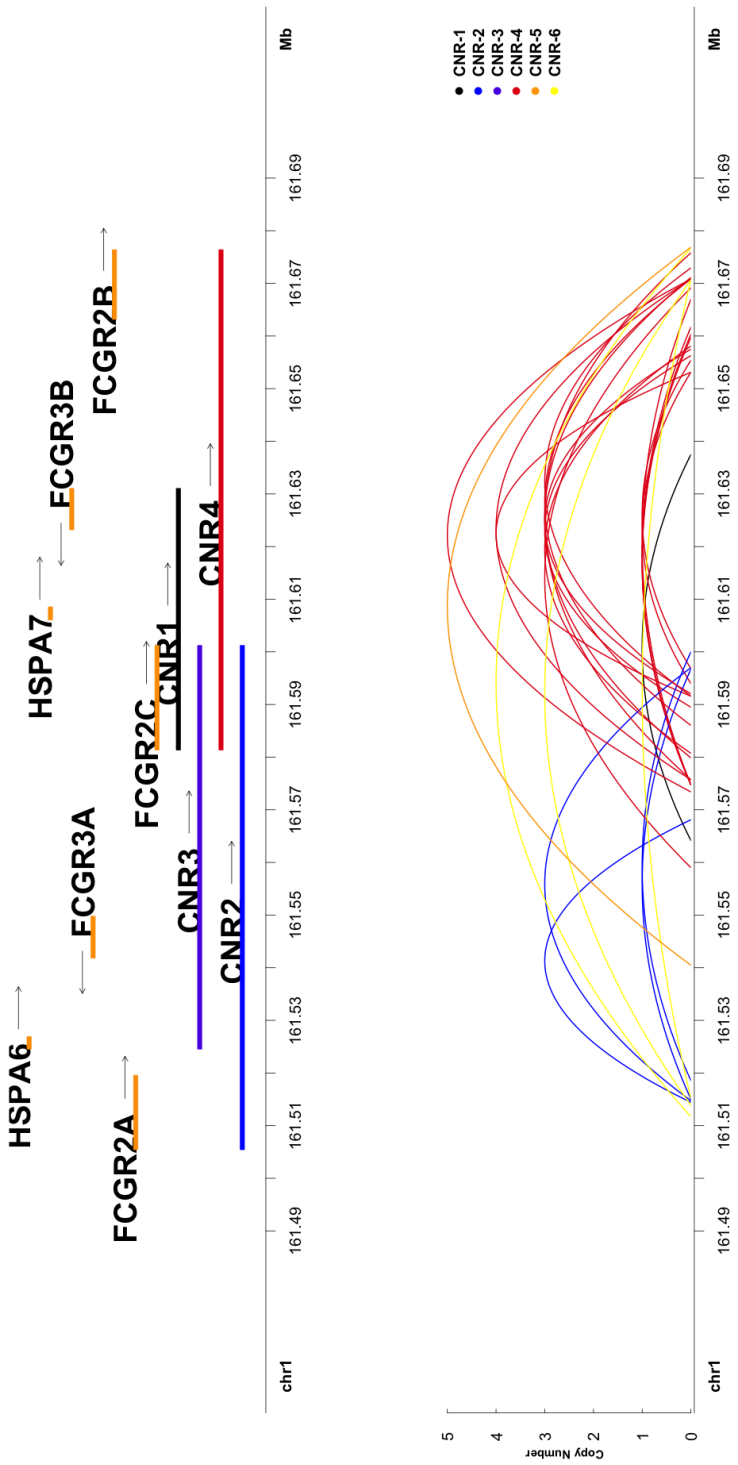

Figure 6.15: Each of the 26 samples grouped into copy number regions (CNRs) 1-4 as defined in the upper plot. In both plots colours describe a CNR as described in the key. Two additional CNRs were defined to encompass larger CNVs. CNR-5 describes a variant spanning from *FCGR3A* to *FCGR2B*. CNR-6 describes CNVs which spanned the entire region and is highlighted in yellow. In the lower plot the height of the arcs indicates the relative copy number with copy number above 2 indicating a gain and below a deletion.

Three samples spanned the almost the length FCGR locus and were termed CNR-6. It was possible that the region amplification could be called as a result of higher depths compared to the pooled reference across the entire of the samples. CNV segment plots for each of the samples were inspected to check that this was not the case as exemplified in Figure 6.16. The plot shows that each of the three CNVs are not across the entirety of the sample and appear to all have similar start and end coordinates.

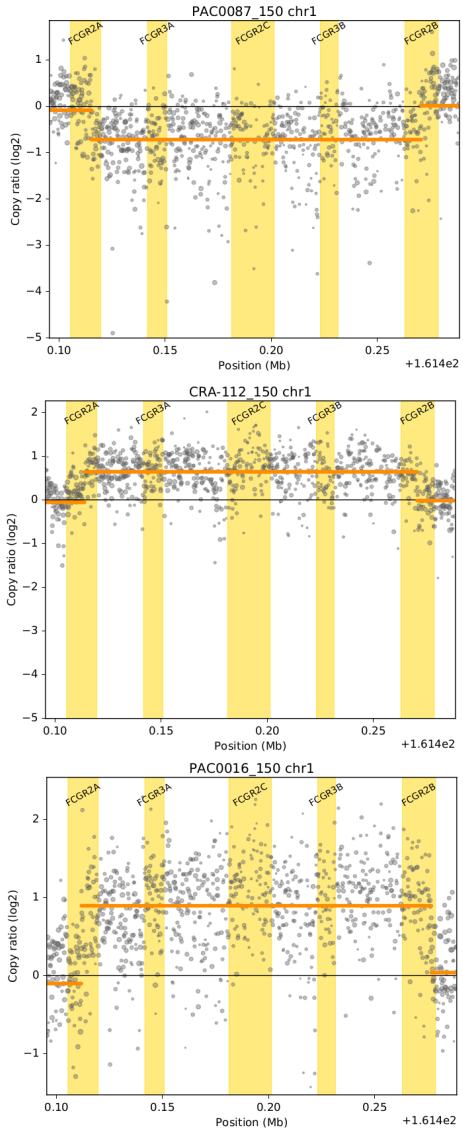

Figure 6.16: Visualisation of the CNR-6 samples. All three samples do not display sample-wide CNVs, arguing against technical reasons for the calls. Breakpoints for the CNVs fall within the same genes and are the most consistent of all the CNR groupings.

6.5 Discussion

FcγRs have a crucial role in the modulation of the immune response due to their expression on a range of leukocytes and the ability to facilitate binding to the Fc part of IgG[256]. Consequently when variants occur in FCGR genes they have the potential to affect the immune response generated in when challenged. It is the modulation of the immune response that makes the FCGR genes of interest for monoclonal antibody therapies, by understanding the variants present in normal and tumour samples treatments could then be tailored to specific patients[257]. Due to the high sequence similarity between FCGR genes conventional NGS pipelines struggle to align reads uniquely between the genes[259]. Consequently aligners such as BWA exclude many reads from variant calling by assigning them a mapping quality of zero and randomly allocating them between the matched sites[43]. In this study blood samples were sequenced after capture with a custom HaloPlex capture kit. The aim of sequencing was to assess if short read sequencing using a customised capture kit for FCGR genes enabled reads to be mapped uniquely and subsequently allow variant calling of the FCGR genes.

In this study 72 samples were captured using a customised HaloPlex capture kit. Initial analysis of samples using FastQC identified that three of the samples had insufficient sequence data to proceed and were subsequently excluded from all further analyses. Pairwise variant call comparisons between samples identified nine samples which were excessively similar with above 76% similarity to at least one other sample. Consequently only 60 samples were utilised for the final joint genotyped variant calls. For all samples upwards of 70% of reads were estimated to be duplicates as expected from amplicon based sequencing.

Initial mapping of samples produced seemingly good percentages of reads mapped to the genome being above on 97% for each batch. Small amplicon lengths however caused paired end reads to overlap as shown by the insert sizes in Figure 6.6. The effects of overlapping reads would be that each overlapped base would be counted twice. To avoid this bias overlapping reads were assembled to generate single end reads which also would increase read sizes where overlaps were not complete. By

merging reads based on overlaps it could increase the confidence for the overlapped regions. The mean percentages for assembled reads, where overlaps were merged, by batch were high at 98.5 and 95.3% for batches 1 and 2 respectively as expected due to the capture kit design.

Calculations of sample-wide depth of coverage over all targeted bases suggested that variants could be called for HaloPlex samples as the mean depth of coverage was 86.8x. Detailed analysis of coverage across the FCGR locus per base was performed as shown in Figure 6.9. Calculations of median sample depth across all samples for the FCGR low affinity locus with the default mapping quality of 20 applied showed several large regions of zero depth coverage. The regions of zero depth coverage were most obvious over the genes *FCGR2C* and *FCGR2B* which are suggested from literature to have arisen due to ancestral duplication and subsequent unequal crosses[256,260,262]. Figure 6.10 shows global alignments between the FCGR genes with around 98% similarity which also supports the ancestral duplications theory but also highlights the challenge of mapping reads uniquely between genes. In cases where reads map equally well to both sites genes BWA will align reads to one of the sites at random and set the mapping quality to zero[43]. Therefore reads are mapped between homologous sites but randomly allocated and rendered unusable in conventional variant calling pipelines due to the mapping quality. In total 59,334 bp across the FCGR low affinity locus had zero depth of coverage with a mapping quality filter of 20 or above applied.

Despite the limitations of short read mapping across the FCGR region attempts were made to call variants with GATK HaplotypeCaller. Results from variant calling were found to be limited over the FCGR locus due to the approximate 59.3kb of bases which were not callable, as expected from initial coverage analyses. HaplotypeCaller also uses a mapping quality threshold of 20, below which reads are not counted for variant calling causing the exclusion of large sections of the FCGR locus from variant calling[47] .

Analysis showed the number of variants called per sample varied greatly from around 2,200 to over 12,000. Most HaloPlex samples stabilised in Het:Hom ratio at around

0.8 once the fraction of coverage depth at 20x or more reached around 80% of bases per sample. It has been shown previously that to correctly call genotypes of variants the depth of coverage should be upwards of 30x for WGS[107–109]. All samples were still retained but additional depth would have been desirable for 26 of the samples below the 30x depth of coverage target.

The few variant calls which were able to be made over the FCGR locus were reviewed to ascertain if there was enough information to determine the HNA-1 haplotype of samples. Across HaloPlex samples a median of only five exonic FCGR gene variants were called, no variants were called at any of the six sites described in Table 6.5 for *FCGR3B*. The lack of variants called indicated that samples had insufficient data passing filters to allow determination the HNA-1 haplotype[256,257]. Only 35 positions with exonic or splicing variants were identified over the entire FCGR locus for the entire 60 sample dataset, only 21 of which were in FCGR 1, 2, or 3 genes.

From the results of the initial mapping and variant calling the FCGR genes were not well mapped and fewer than expected variants were identified. This highlights the problem of mapping FCGR genes using short read targeted technologies. Whilst mapping was problematic in this study, other short read based studies would also still likely have similar mapping issues.

Most of the sequence data mapped non-uniquely between FCGR genes, hence these reads were assigned a mapping quality of zero as the aligner could not place the reads uniquely. As was shown from the lowering of the mapping quality in coverage analyses the data to make variant calls for FCGR genes is contained within these files. To enable usage of the allele information between the homologous sites custom references were used to exclude other FCGR genes before repeating variant calling. For a variant identified the alleles would be for all homologous sites, therefore by using the copy number it would be possible to estimate the genotypes based on the ratios and alleles between them.

The method of custom reference variant calling was applied for the six HNA-1 variant positions in *FCGR3B* to identify the plausible haplotypes. As there are only two

FCGR3 genes the alleles piled-up will be for only two genes at a site. It was possible to predict the HNA-1 haplotypes for samples with reasonable accuracy. The most common haplotype was HNA-1A (NA1) + HNA-1B (NA2) which was found in 27 samples. In these 27 samples five of the six SNPs had two alleles present in an equal ratio. The HNA-1B haplotype was detected in 19 samples, this was the most commonly detected haplotypes in other studies. The third most prevalent haplotype was the NA1 form, also in agreement with the results from other studies[293–297]. All of the publications cited here used genotyping based experiments for the identification of HNA-1 haplotypes and not using amplicon based NGS approaches highlighting the novel nature of this approach to identify the haplotypes. The method was then compared to results obtained previously from MLPA data which found 71.9% agreement between the haplotypes predicted when considering exact matches. If samples with agreeing forms (i.e. A or B) but differ in the copy number between MLPA and HaloPlex then the agreement rises up to 84.1%. Most of the mismatches between the datasets involve the alternative C (SH), NA1*02 or NA2*02 forms which were not called at all over any of the 60 samples. MLPA as a method has several limitations such as its: inability to detect unknown point mutations, sensitivity to novel benign polymorphisms at or near a probe ligation site, sensitivity to chemical contaminants such as phenol or salt concentrations and MLPA probe signals are sensitive to small deletions, insertions, and mismatches. Any mismatch in the probe's target site can theoretically affect the probe's signal affecting the interpretation and accuracy of CNVs and or SNP[298–300].

HNA-1 haplotypes are of clinical importance for the regulation of immune responses. Each of the six variants in the HNA-1 haplotypes affect the N-linked glycosylation sites, with the HNA-1A form being of higher affinity in binding than the HNA-1B haplotype. Consequently the HNA-1A form has been more associated with the autoimmune disease systemic lupus erythematosus (SLE). Studies have also demonstrated that the risk of SLE from the HNA-1 haplotype is also conferred by the copy number state, with deletions of *FCGR3B* and the HNA-1A haplotype giving a much greater risk of SLE[256,294].

The customised reference method performed less well with the *FCGR2B/FCGR2C*

Q57X polymorphism (rs759550223). 8 of the 60 samples were identified with only C alleles, suggesting both the sites in in *FCGR2B* and *FCGR2C* were glutamine (Q). Seven of the eight samples agreed with the MLPA predictions of both sites being glutamines giving an agreement of 87.5%. Within the seven samples five reported the same copy number between HaloPlex and MLPA while two samples disagreed over a single gain of copy.

As illustrated in Figure 6.4 if both sites between genes are homozygous reference for C and T alleles respectively in *FCGR2B* and *FCGR2C* they will report an equal ratio of alleles. An equal ratio of alleles is not likely to indicate that the sites are both heterozygous as *FCGR2B* is believed to be non-polymorphic[256]. To obtain an unequal split of alleles both sites must be homozygous for the same allele as illustrated in Panel B of Figure 6.4 or the *FCGR2C* site could be heterozygous as shown in Panel C of Figure 6.4. 52 of the samples contained a mix of C and T alleles, however in practice it was impossible to distinguish an XX sample from a heterozygote XQ at *FCGR2C*. Allele ratios could in theory be used to interpret the possible combinations assuming *FCGR2B* was not polymorphic however interpretations were complicated by the variable depths over the sample cohort and copy number changes between genes in samples. Hence the overall concordance between MLPA and custom reference predictions hence was low at only 30% when considering all 60 samples. Therefore X57Q homozygous Q mutations in *FCGR2C* could be determined but not any other combination. The only method of determining the sites would be to make full use of longer reads which could span across the locus to assist mapping and variant calling such as SMRT or nanopore sequencing[13]. Both of these approaches were being performed at the time of writing but were available too late to analyse and include.

Optimisation of CNVkit was performed to increase sensitivity and led to the use of a pooled reference with a bin size to 150bp using the default circular binary segmentation. Comparing CNV calls to MLPA gave matches over the FCGR low affinity locus genes between 68.33% and 95% with HaloPlex samples. For the gene *FCGR3B* of most interest the agreement was 81.67% with MLPA. Using the CNVkit calls there were two samples (CRA-53, CRA-64) with the HNA-1A haplotype and also a loss of copy over *FCGR3B*. The combination of the HNA-1A haplotype and

copy number loss make this sample of higher risk for SLE. Four samples were also found with a copy number loss and the HNA-1B haplotype which are of lower risk than the HNA-1A haplotype[294]. It is suggested that the loss of copy over *FCGR3B* causes SLE by reduced receptor function due to expression or due to reduced ligand affinity[294]. For the 12 samples which had gains of copy, 3 of these samples all had a copy number of 3 which agreed between MLPA and HaloPlex. While both methods agreed that samples only had the HNA-1A and HNA-1B haplotypes the HaloPlex samples could not resolve the exact combination of A and B (i.e. ABB or AAB). 5 of the 12 samples show agreement in the possible A or B haplotypes but with disagreeing copy number with MLPA. 4 of the 12 samples disagreed with MLPA as the HaloPlex data suggested that the samples contained the alternative SH (C), NA2*02 or NA1*02 haplotypes. Over the entire dataset no haplotypes other than the HNA-1A or HNA-1B were identified by MLPA. To ascertain if this was a limitation of MLPA used or of the HaloPlex data further experimental data would be required such as from third generation sequencing.

Using the copy number calls across the FCGR locus all gains or losses identified were grouped based on the pre-existing breakpoints regions termed CNR 1-4[256] along with two additional breakpoints which were defined in this study and did not fit into the existing four described in literature[256,259]. The most commonly observed CNR has been reported to be CNR-1 which was not true in this dataset with only one sample identified[256]. Detection of the CNR-3 was not observed in this dataset and fits with its previous description of only being observed in South-east Asian populations[262]. Both of these findings help to validate CNV calling performed with CNVkit.

From the analysis of each CNR grouping the fit to the existing descriptions is mixed. Only one CNR-1 grouped variant was identified which was a single loss of copy. The start of the CNR-1 variant was around 161.58 Mb, there were several CNVs which started around this location but the endpoints exceeded the expected position at the end of *FCGR3B* for CNR-1. Instead these variants ended halfway to *FCGR2B* so were closer to the CNR-4 co-ordinates. Only four CNR-2 CNVs were identified, all of which fell within the CNR-2 locations defined in literature. No CNR-3 region variants were identified. 16 CNR-4 variants were found. 9 of the 16 CNR-4 variants were short

of the endpoint at the end of *FCGR2B* but were further outside of the endpoint of CNR-1 located at the end of *FCGR3B* and so were better fitted to the CNR-4 grouping. For the additionally defined CNR-5 the CNV identified was closest to CNR-4 but originated before *FCGR2C* close to *FCGR3A*, fitting neither the CNR-1 or CNR-4 locations. CNR-5 was only identified in one sample out of the 60. CNR-6 was identified in four samples which had CNVs which spanned across the majority of the FCGR locus. As shown in Figure 6.16 the CNVs are the most consistent of any of the CNRs and appear to have similar breakpoint coordinates. Additional samples would needed to to verify the prevalence and phenotype of both the CNR-5 and CNR-6 CNVs as currently the studies which have evaluated the breakpoints in the FCGR locus were for a limited number of samples and more resolution could be obtained from third generation sequencing reads[259].

6.6 Conclusions

In this study we sought to assess the viability of short read targeted approaches for identifying variants in the FCGR genes which are difficult to map due to high homology between the genes. From initial mapping and analyses of coverage it was shown that short read mapping was not able to resolve variant calls in the FCGR genes. It was however shown that it is possible to predict variants in the FCGR genes using custom reference genomes as shown for the determination of samples possible HNA-1 haplotype. Good concordance was then found for the HNA-1 haplotype predictions with the MLPA data. However for the rs759550223 variant between *FCGR2B* and *FCGR2C* the predictions did not display good agreement when the reference and alternate alleles are reversed between the genes. Therefore with the use of short reads it is possible to obtain variant calls for regions of high homology as demonstrated by the prediction of HNA-1 haplotypes in *FCGR3B* when looking over six SNPs. This information is lost when calling with only the conventional human reference. However the limitations of such an approach are shown when considering the results obtained from *FCGR2B* and *FCGR2C* for a single SNP. The differentiation of sites being homozygous reference or heterozygous is made impossible by the presence of both alleles C and T alleles due to the combination of unequal sequencing depth and copy number variants between the genes. When both of these factors are combined there are too many variables to make accurate predictions. Copy number analysis also was able to detect copy number variants, which were used also in the prediction of variant alleles between homologous sites. Copy number detection from HaloPlex data also showed good agreement with MLPA results and were able to identify samples with loss of copy and also the HNA-1A haplotype of highest risk of SLE.

The approaches documented to call SNVs in this project demonstrate the potential ability of short read targeted approaches to call complicated genomic loci such as the FCGR region. This chapter also highlights the failure of standard analysis pipelines to deal with these complex loci in which the non-uniquely mapped reads lead to the failure of variant calling. Whilst the results from custom references are an obvious improvement on standard pipelines the use of longer reads from third generation

sequencing offers potential further improvement with reads able to span across multiple genes or entire loci to aid mapping. At the time of writing third generation experiments involving SMRT and nanopore sequencing are currently being performed and analysed which can then be used to compare against the short read results.

Chapter 7

Conclusions

NGS technologies have improved in accuracy over the last decade whilst the associated costs have decreased. These two factors have led to the larger NGS studies utilising WES or WGS and the transition of these technologies from a research environment over to clinical settings. Projects such as the 100,000 genomes aimed to sequence the genomes of cancer and rare disease patients, generating a powerful resource for the investigation of the genetic contributions and mechanisms of disease. The knowledge gained from NGS sequencing has the potential to change medical genomics and lead into an era of personalised medicine with treatments designed and tailored for individual patients. With the increased number of NGS samples and the adoption of WGS leading to larger numbers of variants being identified there is the clear requirement for the improvement in algorithms, tools and annotations to extract the maximum information and crucially obtain accurate and reliable results. In this thesis five projects were undertaken each focusing on different aspects of NGS pipelines from quality control through to the prioritisation of variants using annotations from some of the latest tools.

Most NGS samples are outsourced to large sequencing centres which increases the chances of samples being contaminated. Contamination of samples will lead to a loss of accuracy of variant calls which could lead to the incorrect identification and association of variants disease, leading to the development and or application of incorrect treatments for a patient. Using only VCF files a regression model was developed to detect the level of contamination between samples of the same species. The regression model was trained on *in silico* contamination simulations. Best results

were obtained when training models with contamination simulations up to 10% contamination with an R^2 value of 0.909 obtained. Whilst simulations were performed up to 50% contamination this led to regression models overestimating the levels of contamination.

Application of the regression model with 245 WES sample VCFs and using the existing contamination estimation tool VerifyBamID[88] with BAM files enabled the prediction of contamination for samples. Results showed good agreement between the regression model and VerifyBamID with the same two samples only predicted above 5% contamination by the regression model and the VerifyBamID while all others were predicted below 5%. There is potential to improve the regression based prediction model further by training with a wider variety of samples and by capturing more variation such as with WGS. A wider variety of samples will also likely help correct any over-fitting of the model from the training dataset currently used. In future this method could also be extended to other diploid species and would not be limited to humans.

Contamination of samples may not be from the same species and so do not map to the human genome. These unmapped reads may provide information about the associated microbiome of the individual around the site of collection. Studies have previously shown that when using saliva captured exome samples it was possible to detect differences in the oral microbes between populations[89]. To assess the utility of unmapped reads 245 WES IBD samples were investigated and compared with 113 non-IBD WES samples to see if differences could be detected. For IBD samples microbiomes are of particular interest as the dysregulation of the immune system leads to the targeting of commensal bacteria and inflammation of the gut. Comparisons of IBD samples with controls failed to identify differences between the two groups. However the largest differences were detected between the collection methods with few bacterial reads detected in blood samples whilst saliva samples reported many bacteria associated with the upper respiratory tract. 57 samples were also found to contain traces of commonly sequenced plants such as *arabidopsis thaliana*, 55 of which were performed at the same sequencing centre. This indicates that unmapped reads can also be caused by environmental contamination from other

samples likely being sequenced also at the sequencing centres. Many high quality unmapped reads in each of the samples remained unclassifiable to a species, additional species databases may reduce the number. It is also possible that the reads are from regions of the human genome which are computationally challenging to map to or contain multiple variants which prevent mapping.

Whilst WES has been successful at uncovering genes associated with disease in the last few years as the cost of NGS has decreased the cost difference between WES and WGS has continued to narrow, leading to the increased use of WGS as demonstrated by the 100,000 genomes project. WGS offers broader coverage of the genome including non-coding regions leading to the large increase in the number variants called. Whilst our knowledge of non-coding regions and variants has improved due to projects such as ENCODE the interpretation of non-coding variants are currently a limiting factor.

Comparisons of WES and WGS were performed for a trio which showed that WGS returned additional variants from coding regions of all types from missense to stop gains. One of the largest relative increases was in the number of additional splicing and splice-region variants detected. The extra splice-region variants indicates that WGS may detect more of the variants on the exon-intron boundaries which may not be captured well or targeted by WES. These additional variants could be important in the identification of the genetic causes of a disease and so give WGS a greater diagnostic potential than WES. Using annotations to identify the most pathogenic variants of potential interest found that many variants currently identified as pathogenic are called in healthy individuals. The presence of predicted pathogenic mutations in healthy individuals makes the identification of causal variants for diseases challenging from WGS and is complicated by the large number of variants called. Without candidate genes, as is often the case in rare diseases, it is challenging to identify causal variants from WGS currently. However the advantage of having performed WGS is that all variants and regions should be covered enabling the future re-analysis of data to identify causal variants with improved annotations and methods.

A case of the rare, lethal disease SSMD in a child was diagnosed which was initially of unknown genetic causes at the outset of the study when sequencing was performed by WES. Sequencing was performed for the affected child and both unaffected parents with the aim of identifying the gene of interest for which potential for carrier testing and pre-implantation genetic diagnosis could be performed in future. No plausible candidates were identified from initial WES however the gene *GPX4* was subsequently identified as a cause of SSMD[216]. WGS sequencing was performed and combined with WES data to call variants. However the combined datasets failed to identify any potential variants in *GPX4* which could cause SSMD. Due to the overlap with other forms of skeletal dysplasia and skeletal disease an extended candidate gene list including 430 genes were assessed for variants in addition to a filter based approach using annotations to prioritise variants using the trio and phasing data. Results of the investigations failed to identify a definitive cause for this case. SSMD shares phenotypic overlap with short rib polydactyl syndromes which can be inherited in a di-genic pattern. Attempts were made to identify combined effects by looking at structural or copy variants with SNPs which identified a possible compound heterozygote in *WDR60*, the cause of a short-rib polydactyl syndrome. The combination was a deletion called by a structural variant caller and a heterozygote splicing variant with a pathogenic prediction, but only a small predicted change in splicing activity. It is possible that SSMD is genetically heterogeneous and the other genes have yet to be identified to cause the disease. This study highlights that WGS does not always provide the answer initially as to the genetic causes of rare disease due to either difficulty in interpretation or a lack of knowledge of the non-coding genome. Results for rare cases such as this may reveal answers therefore in time with further re-analysis and additional knowledge.

Whilst WGS is able to provide data for the majority of the genome not all of the regions are able to mapped and sequenced well. An example of this are the FCGR genes which encode for receptors expressed on the surface of a range of leukocytes. The receptors are crucial for the binding of monoclonal antibodies used in particular for the treatment of immune disease and cancers. Therefore mutations in these genes can lead to varying response to therapies. FCGR genes are problematic for mapping due to ancestral duplication events leading to around 98% similarity between regions

of the genes. This high similarity between the genes leads to reads mapping equally well between the genes and results in aligners not being able to place reads uniquely and consequently assigning a mapping quality of zero. The mapping quality of zero causes reads to be excluded from variant calling. The mapping of the FCGR locus and other regions of high homology in the genome therefore remains problematic with short read based approaches.

In the study a customised HaloPlex capture kit was tested for the ability to capture and call FCGR 2 and 3 gene variants. Of particular interest were six SNPs in *FCGR3B* which encode human neutrophil antigen-1 (HNA-1) haplotypes. By conventional mapping using the hg38 reference genome reads fail to map uniquely due to the homologous region in *FCGR3A*. By using a customised reference genome for just the *FCGR3B* gene it was possible to recover the read data and infer the haplotype based on knowledge of the homologous locus using the ratio of alleles. Predictions of the HNA-1 haplotypes showed 84.1% agreement with MLPA predictions. The FCGR low affinity locus containing FCGR 2 and 3 genes is copy number variable and so to predict HNA-1 haplotypes correct copy number calls were required. CNVs were called well across the locus with between 82 - 93% agreement with MLPA for *FCGR3A* and *FCGR3B*.

A solution to map FCGR and other high homology genes would be the use of third generation sequencing technologies such as SMRT or nanopore sequencing[13]. Both of these approaches were being performed at the time of writing but were available too late to analyse and include. Longer reads could potentially span across the entirety of genes or the locus, enabling them to be uniquely mapped and variants called.

In this thesis each project focused on a different aspect of NGS pipelines and showed that within-species and cross-species contamination can be detected from VCF and BAM files respectively. Analysis of WES and WGS samples were performed highlighting the additional benefits of WGS but also some of the challenges of identifying causal variants when looking genome-wide. Use of customised genome references can also improve variant calling in regions of the genome traditionally challenging to map to and provide additional variants which would be missed using

conventional mapping. All of these projects highlight that with the increased size and complexity of NGS samples there is additional information that can be extracted from datasets. As more samples and datasets from a variety of projects become available it will be possible to refine each of the approaches developed in this thesis. With additional knowledge, annotations and tools it could be possible to re-analyse the skeletal dysplasia trio and potentially identify the causal gene(s) and variant(s). As NGS methods continue to evolve they will likely become more integrated into clinical scenarios and lead to the development of personalised medicines transforming patient outcomes. However the research and development of better tools to analyse samples and prioritise variants are also needed to empower this revolution in healthcare.

Publications

Warwick, A., Gibson, J., Sood, R. & Lotery, A. A rare penetrant *TIMP3* mutation confers relatively late onset choroidal neovascularisation which can mimic age-related macular degeneration. Eye (Lond) (2015). doi:10.1038/eye.2015.204

Bibliography

1. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. en. *Nat Rev Genet* **12,** 745–755 (Nov. 2011).
2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. en. *PNAS* **74,** 5463–5467 (Dec. 1977).
3. Collins, F. S. The Human Genome Project: Lessons from Large-Scale Biology. *Science* **300,** 286–290 (Apr. 2003).
4. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* **30,** 418–426 (Sept. 2014).
5. Liu, L. *et al.* Comparison of Next-Generation Sequencing Systems. *J Biomed Biotechnol* **2012** (2012).
6. Metzker, M. L. Sequencing technologies — the next generation. en. *Nat Rev Genet* **11,** 31–46 (Jan. 2010).
7. Mardis, E. R. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry* **6,** 287–303 (2013).
8. Inc., I. *Illumina Sequencing Technology* Oct. 2013.
9. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24,** 133–141 (Mar. 2008).
10. Shendure, J. & Ji, H. Next-generation DNA sequencing. en. *Nature Biotechnology* **26,** 1135–1145 (Oct. 2008).
11. Miyamoto, M. *et al.* Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. en. *BMC Genomics* **15,** 699 (Aug. 2014).

12. Buermans, H. P. J. & den Dunnen, J. T. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. From genome to function* **1842,** 1932–1941 (Oct. 2014).

13. Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34,** 666–681 (Sept. 2018).

14. Beck, T. F., Mullikin, J. C., on behalf of the NISC Comparative Sequencing Program & Biesecker, L. G. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. en. *Clinical Chemistry* **62,** 647–654 (Apr. 2016).

15. Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. en. *Hum. Mol. Genet.* ddq416 (Sept. 2010).

16. Teng, J. L. L. *et al.* PacBio But Not Illumina Technology Can Achieve Fast, Accurate and Complete Closure of the High GC, Complex Burkholderia pseudomallei Two-Chromosome Genome. *Front Microbiol* **8** (Aug. 2017).

17. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. eng. *Hum. Mol. Genet.* **27,** R234–R241 (Aug. 2018).

18. England, G. Genomics England — 100,000 Genomes Project.

19. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. en. *Nucl. Acids Res.* **29,** 308–311 (Jan. 2001).

20. Gibbs, R. A. *et al.* The International HapMap Project. en. *Nature* **426,** 789–796 (Dec. 2003).

21. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* **118,** 1590–1605 (May 2008).

22. Consortium, T. E. P. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLOS Biol* **9,** e1001046 (Apr. 2011).

23. Park, P. J. ChIP-Seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10,** 669–680 (Oct. 2009).

24. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. en. *Genome Res.* **22,** 1760–1774 (Sept. 2012).

25. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. en. *Genome Res.* **22,** 1775–1789 (Sept. 2012).

26. Consortium, T. I. H. 3. Integrating common and rare genetic variation in diverse human populations. en. *Nature* **467,** 52–58 (Sept. 2010).

27. Consortium, T. 1. G. P. An integrated map of genetic variation from 1,092 human genomes. en. *Nature* **491,** 56–65 (Nov. 2012).

28. Delaneau, O. *et al.* Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5,** 3934 (June 2014).

29. Muddyman, D., Smee, C., Griffin, H., Kaye, J. & Project, t. U. Implementing a successful data-management framework: the UK10K managed access model. en. *Genome Med* **5,** 1–9 (Nov. 2013).

30. The International HapMap Consortium. A haplotype map of the human genome. en. *Nature* **437,** 1299–1320 (Oct. 2005).

31. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. en. *Nat Genet* **advance online publication** (Mar. 2015).

32. Andrews, S. *FastQC* Feb. 2015.

33. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. en. *Bioinformatics* **27,** 863–864 (Mar. 2011).

34. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. en. *EMBnet.journal* **17,** pp. 10–12 (May 2011).

35. FASTX-Toolkit. FASTX-Toolkit.

36. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. en. *Bioinformatics* **26,** 1783–1785 (July 2010).

37. GATK. *Read trimming* en.

38. GATK. *Trimming of Adaptor sequence?* en.

39. Novoalign. NovoAlign — Novocraft.

40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (Mar. 2012).

41. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18,** 821–829 (May 2008).

42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. eng. *Bioinformatics* **25,** 1754–1760 (July 2009).

43. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).

44. Hatem, A., Bozdağ, D., Toland, A. E. & Çatalyürek, Ü. V. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14,** 184 (June 2013).

45. Consortium, G. R. *Human genome assembly data - Genome Reference Consortium* Dec. 2013.

46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. en. *Bioinformatics* **25,** 2078–2079 (Aug. 2009).

47. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20,** 1297–1303 (Sept. 2010).

48. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. en (Nov. 2017).

49. Institute, B. Picard Tools - By Broad Institute.

50. DePristo, M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–498 (May 2011).

51. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22,** 568–576 (Mar. 2012).

52. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. en. *Science,* eaau3879 (Oct. 2018).

53. Danecek, P. *et al.* The variant call format and VCFtools. en. *Bioinformatics* **27,** 2156–2158 (Aug. 2011).

54. Hwang, S., Kim, E., Lee, I. & Marcotte, E. M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. en. *Scientific Reports* **5,** 17875 (Dec. 2015).

55. Firtina, C. & Alkan, C. On genomic repeats and reproducibility. en. *Bioinformatics,* btw139 (Mar. 2016).

56. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. en. *Nucleic Acids Research* **38,** e164–e164 (Sept. 2010).

57. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. eng. *Bioinformatics* **26,** 2069–2070 (Aug. 2010).

58. Flicek, P. *et al.* Ensembl 2014. en. *Nucl. Acids Res.* **42,** D749–D755 (Jan. 2014).

59. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6,** 80–92 (Apr. 2012).

60. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. en. *Nat. Protocols* **4,** 1073–1081 (June 2009).

61. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. eng. *Nat. Methods* **7,** 248–249 (Apr. 2010).

62. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. eng. *Hum. Mutat.* **34,** 57–65 (Jan. 2013).

63. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (Aug. 2016).

64. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. en. *Nucleic Acids Res* **43,** D805–D811 (Jan. 2015).

65. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. eng. *Nucleic Acids Res.* **46,** D1062–D1067 (Jan. 2018).

66. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (Mar. 2014).

67. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. eng. *Bioinformatics* **31,** 761–763 (Mar. 2015).

68. Van der Velde, K. J. *et al.* GAVIN: Gene-Aware Variant INterpretation for medical sequencing. *Genome Biol* **18** (Jan. 2017).

69. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31,** 1536–1543 (May 2015).

70. Rogers, M. F. *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. en. *Bioinformatics* **34,** 511–513 (Feb. 2018).

71. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nature Reviews Genetics* **17,** 93–108 (Jan. 2016).

72. Bell, R. J. A. *et al.* The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. en. *Science* **348,** 1036–1039 (May 2015).

73. Spielmann, M. & Mundlos, S. Looking beyond the genes: the role of non-coding variants in human disease. en. *Human Molecular Genetics* **25,** R157–R165 (Oct. 2016).

74. Diederichs, S. *et al.* The dark matter of the cancer genome: aberrations in regulatory elements, untranslated regions, splice sites, non-coding RNA and synonymous mutations. en. *EMBO Molecular Medicine,* e201506055 (Mar. 2016).

75. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. en. *Nature* **526,** 75–81 (Oct. 2015).

76. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. en. *Bioinformatics* **25,** 2865–2871 (Nov. 2009).

77. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. en. *Nature Methods* **6,** 677–681 (Sept. 2009).

78. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. en. *Bioinformatics* **28,** i333–i339 (Sept. 2012).

79. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. en. *Genome Biology* **15,** R84 (June 2014).

80. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. en. *PLOS Computational Biology* **12,** e1004873 (Apr. 2016).

81. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. en. *BMC Biology* **12,** 87 (Nov. 2014).

82. Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in genome sequencing projects. en. *PeerJ* **2,** e675 (Nov. 2014).

83. Langdon, W. B. Mycoplasma contamination in the 1000 Genomes Project. en. *BioData Mining* **7,** 3 (Apr. 2014).

84. Longo, M. S., O'Neill, M. J. & O'Neill, R. J. Abundant Human DNA Contamination Identified in Non-Primate Genome Databases. *PLoS ONE* **6,** e16410 (Feb. 2011).

85. Tae, H., Karunasena, E., Bavarva, J. H., McIver, L. J. & Garner, H. R. Large scale comparison of non-human sequences in human sequencing data. *Genomics* **104,** 453–458 (Dec. 2014).

86. Gouin, A. *et al.* Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads. en. *Heredity* **114,** 494–501 (May 2015).

87. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. en. *Standards in Genomic Sciences* **10,** 18 (2015).

88. Jun, G. *et al.* Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. *The American Journal of Human Genetics* **91,** 839–848 (Nov. 2012).

89. Kidd, J. M. *et al.* Exome capture from saliva produces high quality genomic and metagenomic data. en. *BMC Genomics* **15,** 262 (Apr. 2014).

90. Wang, Q., Jia, P. & Zhao, Z. VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. en. *Genome Medicine* **7,** 2 (Jan. 2015).

91. Cibulskis, K. *et al.* ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27,** 2601–2602 (Sept. 2011).

92. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting Somatic Mutations in Normal Cells. *Trends in Genetics* **34,** 545–557 (July 2018).

93. Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M. & Kang, H. M. Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *Am J Hum Genet* **97,** 284–290 (Aug. 2015).

94. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor–normal pairs. en. *Bioinformatics* **32,** 3196–3198 (Oct. 2016).

95. Consortium, T. G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. en. *Science* **348,** 648–660 (May 2015).

96. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. en. *Science* **348,** 880–886 (May 2015).

97. Guennec, K. L. *et al.* 17q21.31 duplication causes prominent tau-related dementia with increased *MAPT* expression. en. *Molecular Psychiatry* **22,** 1119–1125 (Aug. 2017).

98. Andreoletti, G. *et al.* AMMECR1: a single point mutation causes developmental delay, midface hypoplasia and elliptocytosis. en. *Journal of Medical Genetics,* jmedgenet–2016–104100 (Nov. 2016).

99. Pengelly, R. J. *et al.* A SNP profiling panel for sample tracking in whole-exome sequencing studies. en. *Genome Medicine* **5,** 89 (Sept. 2013).

100. The CNR-MAJ collaborators *et al.* SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. en. *Molecular Psychiatry* **21,** 831–836 (June 2016).

101. Mathur, P. *et al.* Whole exome sequencing reveals rare variants linked to congenital pouch colon. en. *Scientific Reports* **8** (Dec. 2018).

102. Helena Mangs, A. & Morris, B. J. The Human Pseudoautosomal Region (PAR): Origin, Function and Future. *Curr Genomics* **8,** 129–136 (Apr. 2007).

103. Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R. & Drăghici, S. Machine Learning and Its Applications to Biology. en. *PLOS Computational Biology* **3,** e116 (June 2007).

104. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. en. *MACHINE LEARNING IN PYTHON,* 6.

105. learn, S. *An introduction to machine learning with scikit-learn — scikit-learn 0.19.2 documentation* 2018.

106. Metsalu, T. & Vilo, J. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. en. *Nucleic Acids Res* **43,** W566–W570 (July 2015).

107. Ajay, S. S., Parker, S. C. J., Ozel Abaan, H., Fuentes Fajardo, K. V. & Margulies, E. H. Accurate and comprehensive sequencing of personal genomes. en. *Genome Research* **21,** 1498–1505 (Sept. 2011).

108. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. en. *Nature Reviews Genetics* **15,** 121–132 (Feb. 2014).

109. Wang, J. *et al.* The diploid genome sequence of an Asian individual. en. *Nature* **456,** 60–65 (Nov. 2008).

110. Inc., I. *Using a PhiX Control for HiSeq Sequencing Runs - hiseq-phix-control-v3-technical-note.pdf*

111. Camacho, C. *et al.* BLAST+: architecture and applications. eng. *BMC Bioinformatics* **10,** 421 (2009).

112. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. en. *Genome Biology* **15,** R46 (Mar. 2014).

113. Spisák, S. *et al.* Complete Genes May Pass from Food to Human Blood. *PLoS ONE* **8,** e69805 (July 2013).

114. Loohuis, L. M. O. *et al.* Transcriptome analysis in whole blood reveals increased microbial diversity in schizophrenia. en. *Translational Psychiatry* **8,** 96 (May 2018).

115. Kaser, A., Zeissig, S. & Blumberg, R. S. Inflammatory Bowel Disease. en. *Annual Review of Immunology* **28,** 573–621 (Mar. 2010).

116. Proctor, L. M. The National Institutes of Health Human Microbiome Project. en. *Seminars in Fetal and Neonatal Medicine* (June 2016).

117. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology* **13,** R79 (2012).

118. Docktor, M. J. *et al.* Alterations in Diversity of the Oral Microbiome in Pediatric Inflammatory Bowel Disease. *Inflamm Bowel Dis* **18,** 935–942 (May 2012).

119. Gevers, D. *et al.* The Treatment-Naive Microbiome in New-Onset Crohn's Disease. en. *Cell Host & Microbe* **15,** 382–392 (Mar. 2014).

120. Said, H. S. *et al.* Dysbiosis of Salivary Microbiota in Inflammatory Bowel Disease and Its Association With Oral Immunological Biomarkers. en. *DNA Research* **21,** 15–25 (Feb. 2014).

121. Miquel, S. *et al.* Faecalibacterium prausnitzii and human intestinal health. en. *Current Opinion in Microbiology* **16,** 255–261 (June 2013).

122. Sokol, H. *et al.* Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. en. *Proceedings of the National Academy of Sciences* **105,** 16731–16736 (Oct. 2008).

123. Gupta, V., Chaudhari, N. M., Iskepalli, S. & Dutta, C. Divergences in gene repertoire among the reference Prevotella genomes derived from distinct body sites of human. en. *BMC Genomics* **16,** 153 (2015).

124. Filippo, C. D. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. en. *PNAS* **107,** 14691–14696 (Aug. 2010).

125. Lucke, K. Prevalence of Bacteroides and Prevotella spp. in ulcerative colitis. en. *Journal of Medical Microbiology* **55,** 617–624 (May 2006).

126. Xu, J. A Genomic View of the Human-Bacteroides thetaiotaomicron Symbiosis. *Science* **299,** 2074–2076 (Mar. 2003).

127. Walker, A. W. *et al.* High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. en. *BMC Microbiology* **11,** 7 (2011).

128. Antharam, V. C. *et al.* Intestinal Dysbiosis and Depletion of Butyrogenic Bacteria in Clostridium difficile Infection and Nosocomial Diarrhea. en. *Journal of Clinical Microbiology* **51,** 2884–2892 (Sept. 2013).

129. Vital, M., Howe, A. C. & Tiedje, J. M. Revealing the Bacterial Butyrate Synthesis Pathways by Analyzing (Meta)genomic Data. en. *mBio* **5,** e00889–14–e00889–14 (Apr. 2014).

130. Park, S.-K., Kim, M.-S., Roh, S. W. & Bae, J.-W. Blautia stercoris sp. nov., isolated from human faeces. en. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* **62,** 776–779 (Apr. 2012).

131. Liu, T.-C. & Stappenbeck, T. S. Genetics and Pathogenesis of Inflammatory Bowel Disease. *Annual Review of Pathology: Mechanisms of Disease* **11,** 127–148 (2016).

132. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. eng. *J. Mol. Biol.* **215,** 403–410 (Oct. 1990).

133. Karlin, S. & Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. en. *PNAS* **87,** 2264–2268 (Mar. 1990).

134. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. eng. *Nucleic Acids Res.* **25,** 3389–3402 (Sept. 1997).

135. NCBI. *The Statistics of Sequence Similarity Scores*

136. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5,** e77 (Mar. 2007).

137. Huson, D. H., Buchfink, B. & Ruscheweyh, H.-J. An Efficient Pipeline for Microbiome Analysis (2015).

138. Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. en. *Genome Res.* **21,** 1552–1560 (Sept. 2011).

139. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* **2015,** 951–969 (Apr. 2015).

140. Régnier, P. & Marujo, P. E. *Polyadenylation and Degradation of RNA in Prokaryotes* en (Landes Bioscience, 2013).

141. Janda, J. M. & Abbott, S. L. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. en. *Journal of Clinical Microbiology* **45,** 2761–2764 (Sept. 2007).

142. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. eng. *Nucleic Acids Res.* **27,** 573–580 (Jan. 1999).

143. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to Identify Similar Regions in Large Sequence Sets. en. *Current Protocols in Bioinformatics* **00,** 10.3.1–10.3.18 (Jan. 2003).

144. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. en. *Bioinformatics* **28,** 464–469 (Feb. 2012).

145. Carver, T., Thomson, N., Bleasby, A., Berriman, M. & Parkhill, J. DNAPlotter: circular and linear interactive genome visualization. en. *Bioinformatics* **25,** 119–120 (Jan. 2009).

146. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. en. *Nature Methods* **7,** 335–336 (May 2010).

147. Kim, S., Kim, Y.-T., Yoon, H., Lee, J.-H. & Ryu, S. The complete genome sequence of Cronobacter sakazakii ATCC 29544T, a food-borne pathogen, isolated from a child's throat. en. *Gut Pathogens* **9** (Dec. 2017).

148. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. en. *Genome Biology,* 9 (2004).

149. Samuels, D. C. *et al.* Finding the lost treasures in exome sequencing data. en. *Trends in Genetics* **29,** 593–599 (Oct. 2013).

150. Whitacre, L. K. *et al.* What's in your next-generation sequence data? An exploration of unmapped DNA and RNA sequence reads from the bovine reference individual. en. *BMC Genomics* **16** (Dec. 2015).

151. GibbonsJun. 13, A., 2012 & Pm, 1. *Bonobos Join Chimps as Closest Human Relatives* en. June 2012.

152. Wong, K. *Tiny Genetic Differences between Humans and Other Primates Pervade the Genome* en.

153. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database (Oxford)* **2016** (Feb. 2016).

154. Naito, M. *et al.* Determination of the Genome Sequence of Porphyromonas gingivalis Strain ATCC 33277 and Genomic Comparison with Strain W83 Revealed Extensive Genome Rearrangements in P. gingivalis. *DNA Res* **15,** 215–225 (Aug. 2008).

155. Whatmore, A. M. *et al.* Genetic Relationships between Clinical Isolates of Streptococcus pneumoniae, Streptococcus oralis, and Streptococcus mitis: Characterization of "Atypical" Pneumococci and Organisms Allied to S. mitis HarboringS. pneumoniae Virulence Factor-Encoding Genes. en. *Infect. Immun.* **68,** 1374–1382 (Mar. 2000).

156. Peng, Z. *et al.* Identification of critical residues in Gap3 of Streptococcus parasanguinis involved in Fap1 glycosylation, fimbrial formation and in vitroadhesion. *BMC Microbiology* **8,** 52 (2008).

157. Kawamura, Y., Hou, X.-G., Sultana, F., Miura, H. & Ezaki, T. Determination of 16S rRNA Sequences of Streptococcus mitis and Streptococcus gordonii and Phylogenetic Relationships among Members of the Genus Streptococcus. en. *International Journal of Systematic Bacteriology* **45,** 406–408 (Apr. 1995).

158. Morita, E. *et al.* Predominant presence of Streptococcus anginosus in the saliva of alcoholics. en. *Oral Microbiology and Immunology* **20,** 362–365 (Dec. 2005).

159. Mitchell, T. J. The pathogenesis of streptococcal infections: from Tooth decay to meningitis. en. *Nat Rev Micro* **1,** 219–230 (Dec. 2003).

160. Todar, K. *Streptococcus pyogenes and streptococcal disease* 2012.

161. Nørskov-Lauritsen, N. Classification, Identification, and Clinical Significance of Haemophilus and Aggregatibacter Species with Host Specificity for Humans. *Clin Microbiol Rev* **27,** 214–240 (Apr. 2014).

162. Ratnayake, L., Olver, W. J. & Fardon, T. Aggregatibacter aphrophilus in a patient with recurrent empyema: a case report. *J Med Case Reports* **5,** 448 (Sept. 2011).

163. Gessner, B. D. & Berthe-Marie, N.-L. *Haemophilus influenzae- Infectious Disease and Antimicrobial Agents* 2014.

164. *Goldman's Cecil Medicine* en (Elsevier, 2012).

165. Chapalain, A. *et al.* Identification of quorum sensing-controlled genes in Burkholderia ambifaria. *Microbiologyopen* **2,** 226–242 (Apr. 2013).

166. Genomes, J. B. *Home - Burkholderia ambifaria MC40-6*

167. Rapsinski, G. J., Makadia, J., Bhanot, N. & Min, Z. Pseudomonas mendocina native valve infective endocarditis: a case report. *J Med Case Rep* **10** (Oct. 2016).

168. Lee, J., Lee, C. S., Hugunin, K. M., Maute, C. J. & Dysko, R. C. Bacteria from drinking water supply and their fate in gastrointestinal tracts of germ-free mice: A phylogenetic comparison study. *Water Research. Microbial ecology of drinking water and waste water treatment processes* **44,** 5050–5058 (Sept. 2010).

169. Brooke, J. S. Stenotrophomonas maltophilia: an Emerging Global Opportunistic Pathogen. *Clin Microbiol Rev* **25,** 2–41 (Jan. 2012).

170. Kaakoush, N. O. & Mitchell, H. M. Campylobacter concisus - A new player in intestinal disease. eng. *Front Cell Infect Microbiol* **2,** 4 (2012).

171. Ryan, M. P., Pembroke, J. T. & Adley, C. C. Ralstonia pickettii: a persistent Gram-negative nosocomial infectious organism. *Journal of Hospital Infection* **62,** 278–284 (Mar. 2006).

172. Coburn, B., Grassl, G. A. & Finlay, B. B. Salmonella, the host and disease: a brief review. en. *Immunol Cell Biol* **85,** 112–118 (Dec. 2006).

173. Woodward, D. L. Identification and characterization of Shigella boydii 20 serovar nov., a new and emerging Shigella serotype. en. *Journal of Medical Microbiology* **54,** 741–748 (Aug. 2005).

174. Fakruddin, M., Rahaman, M., Ahmed, M. M. & Hoque, M. M. Stress tolerant virulent strains of Cronobacter sakazakii from food. *Biol Res* **47** (Nov. 2014).

175. Li, F. *et al.* Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. eng. *Nat. Biotechnol.* **33,** 524–530 (May 2015).

176. Saski, C. A. *et al.* Sub genome anchored physical frameworks of the allotetraploid Upland cotton (Gossypium hirsutum L.) genome, and an approach toward reference-grade assemblies of polyploids. en. *Scientific Reports* **7,** 15274 (Nov. 2017).

177. Wu, G. A. *et al.* Genomics of the origin and evolution of *Citrus*. en. *Nature* **554,** 311–316 (Feb. 2018).

178. Alonso-Blanco, C. *et al.* 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166,** 481–491 (July 2016).

179. NHGRI. *Importance of Mouse Genome* en-US. 2000.

180. Mao, B. *et al.* In vitro fermentation of fructooligosaccharides with human gut bacteria. en. *Food Funct.* **6,** 947–954 (Mar. 2015).

181. Meynert, A. M., Ansari, M., FitzPatrick, D. R. & Taylor, M. S. Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics* **15,** 247 (July 2014).

182. Technologies, A. *SureSelect Human All Exon V7 exome — Agilent*

183. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22,** 1790–1797 (Sept. 2012).

184. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. en. *Nat Meth* **11,** 294–296 (Mar. 2014).

185. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. en. *Nat Genet* **48,** 214–220 (Feb. 2016).

186. Van der Velde, K. J. *et al.* Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. *Hum Mutat* **36,** 712–719 (July 2015).

187. Lionel, A. C. *et al.* Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. en. *Genetics in Medicine* **20,** 435–443 (Apr. 2018).

188. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. en. *PNAS* **112,** 5473–5478 (Apr. 2015).

189. Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation* **36,** 815–822.

190. Wenger, A. M., Guturu, H., Bernstein, J. A. & Bejerano, G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. en. *Genetics in Medicine* **19,** 209–214 (Feb. 2017).

191. Cornish, A. & Guda, C. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *Biomed Res Int* **2015** (2015).

192. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. en. *Nature Biotechnology* **32,** 246–251 (Mar. 2014).

193. ; on behalf of the ACMG Laboratory Quality Assurance Committee *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. en. *Genetics in Medicine* **17,** 405–423 (May 2015).

194. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. en. *Bioinformatics* **26,** 841–842 (Mar. 2010).

195. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. eng. *Nucleic Acids Res.* **44,** D733–745 (Jan. 2016).

196. Karolchik, D. The UCSC Genome Browser Database. *Nucleic Acids Research* **31,** 51–54 (Jan. 2003).

197. Fuentes Fajardo, K. V. *et al.* Detecting false-positive signals in exome sequencing. en. *Hum. Mutat.* **33,** 609–613 (Apr. 2012).

198. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. eng. *Bioinformatics* **30,** 1006–1007 (Apr. 2014).

199. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. *Hum Mutat* **32,** 894–899 (Aug. 2011).

200. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput Biol* **9** (July 2013).

201. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. eng. *Bioinformatics* **31,** 2202–2204 (July 2015).

202. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. en. *Nature Methods* **12,** 966–968 (Oct. 2015).

203. Seshan, V. E. & Olshen, A. B. DNAcopy: a package for analyzing DNA copy data (2010).

204. Huang, H. W., Mullikin, J. C. & Hansen, N. F. Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics* **16** (July 2015).

205. David, W. Mount. *Bioinformatics: sequence and genome analysis. Gold Spring Harbor Laboratory Press* (2001).

206. Shen, H. *et al.* Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. en. *PLoS ONE* **8** (ed Awadalla, P.) e59494 (Apr. 2013).

207. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. *CNVkit: Copy number detection and visualization for targeted sequencing using off-target reads* en. Tech. rep. biorxiv;010876v2 (Oct. 2014).

208. Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. en. *Science* **347,** 1254806–1254806 (Jan. 2015).

209. Gao, X. *et al.* Insertion/Deletion Polymorphisms in the Promoter Region of BRM Contribute to Risk of Hepatocellular Carcinoma in Chinese Populations. en. *PLoS ONE* **8** (ed Gorlova, O. Y.) e55169 (Jan. 2013).

210. Rančelis, T. *et al.* Analysis of pathogenic variants from the ClinVar database in healthy people using next-generation sequencing. en. *Genetics Research* **99** (2017).

211. Hehir-Kwa, J. Y. *et al.* A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. en. *Nature Communications* **7,** 12989 (Oct. 2016).

212. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. en. *Nature Communications* **8,** 14061 (Jan. 2017).

213. Warman, M. L. *et al.* Nosology and Classification of Genetic Skeletal Disorders: 2010 Revision. *Am J Med Genet A* **155,** 943–968 (May 2011).

214. Krakow, D. & Rimoin, D. L. The skeletal dysplasias. en. *Genet Med* **12,** 327–341 (June 2010).

215. Bonafe, L. *et al.* Nosology and classification of genetic skeletal disorders: 2015 revision. en. *Am. J. Med. Genet.* **167,** 2869–2892 (Dec. 2015).

216. Smith, A. C. *et al.* Mutations in the enzyme glutathione peroxidase 4 cause Sedaghatian-type spondylometaphyseal dysplasia. en. *J Med Genet* **51,** 470–474 (July 2014).

217. Ufer, C. & Wang, C. C. The Roles of Glutathione Peroxidases during Embryo Development. *Front Mol Neurosci* **4** (July 2011).

218. Ufer, C., Borchert, A. & Kuhn, H. Functional characterization of cis- and trans-regulatory elements involved in expression of phospholipid hydroperoxide glutathione peroxidase. *Nucleic Acids Res* **31,** 4293–4303 (Aug. 2003).

219. Ufer, C. *et al.* Translational regulation of glutathione peroxidase 4 expression through guanine-rich sequence-binding factor 1 is essential for embryonic brain development. *Genes Dev* **22,** 1838–1850 (July 2008).

220. Taylor, S. P. *et al.* Mutations in DYNC2LI1 disrupt cilia function and cause short rib polydactyly syndrome. en. *Nature Communications* **6,** 7092 (June 2015).

221. Cossu, C. *et al.* New mutations in DYNC2H1 and WDR60 genes revealed by whole-exome sequencing in two unrelated Sardinian families with Jeune asphyxiating thoracic dystrophy. *Clinica Chimica Acta* **455,** 172–180 (Apr. 2016).

222. Chen, H. en. in *Atlas of Genetic Diagnosis and Counseling* 199–212 (Springer New York, New York, NY, 2017).

223. Johns Hopkins University (Baltimore MD, M. N. I. o. G. M. *OMIM - Online Mendelian Inheritance in Man* May 2018.

224. Schmidts, M. *et al.* Exome sequencing identifies DYNC2H1 mutations as a common cause of asphyxiating thoracic dystrophy (Jeune syndrome) without major polydactyly, renal or retinal involvement. en. *Journal of Medical Genetics* **50,** 309–323 (May 2013).

225. McInerney-Leo, A. M. *et al.* Short-rib polydactyly and Jeune syndromes are caused by mutations in WDR60. eng. *Am J Hum Genet* **93,** 515–523 (Sept. 2013).

226. McInerney-Leo, A. *et al.* Whole exome sequencing is an efficient, sensitive and specific method for determining the genetic cause of short-rib thoracic dystrophies. en. *Clin Genet* **88,** 550–557 (Dec. 2015).

227. Sharp, A. J. *et al.* A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet* **40,** 322–328 (Mar. 2008).

228. Kalsner, L. & Chamberlain, S. J. Prader-Willi, Angelman, and 15q11-q13 duplication syndromes. *Pediatr Clin North Am* **62,** 587–606 (June 2015).

229. Kamphans, T. *et al.* Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees. *PLoS One* **8** (Aug. 2013).

230. NHS. *Skeletal Dysplasias, Perinatal, 57 Gene Panel* en.

231. NHS. *Skeletal Dysplasias 222 Gene Exome Panel* en.

232. De Almeida, R. M. *et al.* Whole gene sequencing identifies deep-intronic variants with potential functional impact in patients with hypertrophic cardiomyopathy. *PloS one* **12,** e0182946 (2017).

233. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. en. *Nucleic Acids Res* **37,** e67–e67 (May 2009).

234. Worthey, E. A. *et al.* Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. en. *Genetics in Medicine* **13,** 255–262 (Mar. 2011).

235. Gleave, S. B. *et al.* Splenic MAdCAM-1+ marginal reticular cells deliver antibody-inducing signals and confer gut-homing properties to human marginal zone B cells (IRC5P.622). en. *The Journal of Immunology* **194,** 58.5–58.5 (May 2015).

236. Stödberg, T. *et al.* Mutations in *SLC12A5* in epilepsy of infancy with migrating focal seizures. en. *Nature Communications* **6,** 8038 (Sept. 2015).

237. Arikawa-Hirasawa, E. *et al.* Dyssegmental dysplasia, Silverman-Handmaker type, is caused by functional null mutations of the perlecan gene. en. *Nature Genetics* **27,** 431–434 (Apr. 2001).

238. Arikawa-Hirasawa, E. *et al.* Structural and Functional Mutations of the Perlecan Gene Cause Schwartz-Jampel Syndrome, with Myotonic Myopathy and Chondrodysplasia. *The American Journal of Human Genetics* **70,** 1368–1375 (May 2002).

239. Nicole, S. *et al.* Perlecan, the major proteoglycan of basement membranes, is altered in patients with Schwartz-Jampel syndrome (chondrodystrophic myotonia). en. *Nature Genetics* **26,** 480–483 (Dec. 2000).

240. Rodgers, K. D., Sasaki, T., Aszodi, A. & Jacenko, O. Reduced perlecan in mice results in chondrodysplasia resembling Schwartz–Jampel syndrome. en. *Hum Mol Genet* **16,** 515–528 (Mar. 2007).

241. Bulman, M. P. *et al.* Mutations in the human *Delta* homologue, *DLL3*, cause axial skeletal defects in spondylocostal dysostosis. en. *Nature Genetics* **24,** 438–441 (Apr. 2000).

242. Willems, M. *et al.* Molecular analysis of pericentrin gene (PCNT) in a series of 24 Seckel/microcephalic osteodysplastic primordial dwarfism type II (MOPD II) families. en. *Journal of Medical Genetics* **47,** 797–802 (Dec. 2010).

243. Hilton, J. F. *et al.* The molecular basis of glutamate formiminotransferase deficiency. en. *Human Mutation* **22,** 67–73 (July 2003).

244. Alston, C. L. *et al.* Recessive germline SDHA and SDHB mutations causing leukodystrophy and isolated mitochondrial complex II deficiency. en. *Journal of Medical Genetics* **49,** 569–577 (Sept. 2012).

245. Koehler, K. *et al.* Mutations in GMPPA Cause a Glycosylation Disorder Characterized by Intellectual Disability and Autonomic Dysfunction. *The American Journal of Human Genetics* **93,** 727–734 (Oct. 2013).

246. Flück, C. E. *et al.* Why Boys Will Be Boys: Two Pathways of Fetal Testicular Androgen Biosynthesis Are Needed for Male Sexual Differentiation. *The American Journal of Human Genetics* **89,** 201–218 (Aug. 2011).

247. Schindler, C., Chen, Y., Pu, J., Guo, X. & Bonifacino, J. S. EARP is a multisubunit tethering complex involved in endocytic recycling. en. *Nature Cell Biology* **17,** 639–650 (May 2015).

248. Bahe, S., Stierhof, Y.-D., Wilkinson, C. J., Leiss, F. & Nigg, E. A. Rootletin forms centriole-associated filaments and functions in centrosome cohesion. en. *J Cell Biol* **171,** 27–33 (Oct. 2005).

249. Bleharski, J. R. *et al.* Use of Genetic Profiling in Leprosy to Discriminate Clinical Forms of the Disease. en. *Science* **301,** 1527–1530 (Sept. 2003).

250. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. & van Roy, F. A Novel Gene Family NBPF: Intricate Structure Generated by Gene Duplications During Primate Evolution. en. *Mol Biol Evol* **22,** 2265–2274 (Nov. 2005).

251. Duden, R., Griffiths, G., Frank, R., Argos, P. & Kreis, T. E. Beta-COP, a 110 kd protein associated with non-clathrin-coated vesicles and the golgi complex, shows homology to beta-adaptin. *Cell* **64,** 649–665 (Feb. 1991).

252. Sedlazeck, F. *Sniffles: Structural variation caller using third generation sequencing* original-date: 2015-10-25T18:32:47Z. Oct. 2017.

253. Schmidts, M. *et al.* Mutations in the Gene Encoding IFT Dynein Complex Component WDR34 Cause Jeune Asphyxiating Thoracic Dystrophy. *The American Journal of Human Genetics* **93,** 932–944 (Nov. 2013).

254. Thiel, C. *et al.* NEK1 Mutations Cause Short-Rib Polydactyly Syndrome Type Majewski. en. *The American Journal of Human Genetics* **88,** 106–114 (Jan. 2011).

255. Morgan, N. V. *et al.* A locus for asphyxiating thoracic dystrophy, ATD, maps to chromosome 15q13. en. *Journal of Medical Genetics* **40,** 431–435 (June 2003).

256. Hargreaves, C. E. *et al.* Fcg receptors: genetic variation, function, and disease. en. *Immunol Rev* **268,** 6–24 (Nov. 2015).

257. Vogelpoel, L. T. C., Baeten, D. L. P., de Jong, E. C. & den Dunnen, J. Control of Cytokine Production by Human Fc Gamma Receptors: Implications for Pathogen Defense and Autoimmunity. *Front Immunol* **6** (Feb. 2015).

258. Mellor, J. D., Brown, M. P., Irving, H. R., Zalcberg, J. R. & Dobrovic, A. A critical review of the role of Fc gamma receptor polymorphisms in the response to monoclonal antibodies in cancer. *Journal of Hematology & Oncology* **6,** 1 (Jan. 2013).

259. Hargreaves, C. E. *et al.* Evaluation of High-Throughput Genomic Assays for the Fc Gamma Receptor Locus. *PLOS ONE* **10,** e0142379 (Nov. 2015).

260. Warmerdam, P. A., Nabben, N. M., Graaf, S. A. v. d., Winkel, J. G. v. d. & Capel, P. J. The human low affinity immunoglobulin G Fc receptor IIC gene is a result of an unequal crossover event. en. *J. Biol. Chem.* **268,** 7346–7349 (Apr. 1993).

261. Hastings, P., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat Rev Genet* **10,** 551–564 (Aug. 2009).

262. Niederer, H. A. *et al.* Copy number, linkage disequilibrium and disease association in the FCGR locus. *Hum Mol Genet* **19,** 3282–3294 (Aug. 2010).

263. Bailey, J. A. Recent Segmental Duplications in the Human Genome. en. *Science* **297,** 1003–1007 (Aug. 2002).

264. Rogers, K. A., Scinicariello, F. & Attanasio, R. IgG Fc Receptor III Homologues in Nonhuman Primate Species: Genetic Characterization and Ligand Interactions. en. *The Journal of Immunology* **177,** 3848–3856 (Sept. 2006).

265. Nimmerjahn, F. & Ravetch, J. V. en. in *Advances in Immunology* 179–204 (Elsevier, 2007).

266. Machado, L. R. *et al.* Evolutionary History of Copy-Number-Variable Locus for the Low-Affinity Fcg Receptor: Mutation Rate, Autoimmune Disease, and the Legacy of Helminth Infection. en. *The American Journal of Human Genetics* **90,** 973–985 (June 2012).

267. Gu, W., Zhang, F. & Lupski, J. R. Mechanisms for human genomic rearrangements. *Pathogenetics* **1,** 4 (Nov. 2008).

268. Dunnen, J. d. *et al.* IgG opsonization of bacteria promotes Th17 responses via synergy between TLRs and FcgRIIa in human dendritic cells. en. *Blood* **120,** 112–121 (July 2012).

269. Wu, J. *et al.* A novel polymorphism of FcgammaRIIIa (CD16) alters receptor function and predisposes to autoimmune disease. *J Clin Invest* **100,** 1059–1070 (Sept. 1997).

270. Bux, J. Human neutrophil alloantigens. en. *Vox Sanguinis* **94,** 277–285 (May 2008).

271. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. en. *Nat Rev Genet* **17,** 224–238 (Apr. 2016).

272. Technologies, A. *HaloPlex Target Enrichment Manual* Jan. 2015.

273. Technologies, A. *HaloPlex - How it Works*

274. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (Mar. 2014).

275. Shirley, M. D., Ma, Z., Pedersen, B. S. & Wheelan, S. J. *Efficient "pythonic" access to FASTA files using pyfaidx* en. Tech. rep. e1196 (PeerJ Inc., Apr. 2015).

276. Jian, X., Boerwinkle, E. & Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. en. *Nucleic Acids Res* **42,** 13534–13544 (Dec. 2014).

277. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L. & Roach, J. C. Kaviar: an accessible system for testing SNV novelty. en. *Bioinformatics* **27,** 3216–3217 (Nov. 2011).

278. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. en. *Nature Reviews Cancer* **6,** 813–823 (Oct. 2006).

279. Tennessen, J. A. *et al.* Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. en. *Science* **337,** 64–69 (July 2012).

280. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. en. *Genome Res.* **19,** 1553–1561 (Sept. 2009).

281. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. en. *Nature Methods* **7,** 575–576 (Aug. 2010).

282. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. en. *PLOS ONE* **7,** e46688 (Oct. 2012).

283. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. eng. *BMC Genomics* **14 Suppl 3,** S3 (2013).

284. Kim, S., Jhong, J.-H., Lee, J. & Koo, J.-Y. Meta-analytic support vector machine for integrating multiple omics data. eng. *BioData Min* **10,** 2 (2017).

285. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. en. *Nature Genetics* **47,** 276–283 (Mar. 2015).

286. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. en. *PLOS Computational Biology* **6,** e1001025 (Dec. 2010).

287. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20,** 110–121 (Jan. 2010).

288. Siepel, A. & Haussler, D. Combining phylogenetic and hidden Markov models in biosequence analysis. eng. *J. Comput. Biol.* **11,** 413–428 (2004).

289. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. en. *Bioinformatics* **25,** i54–i62 (June 2009).

290. Tibshirani, R. & Wang, P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. en. *Biostatistics* **9,** 18–29 (Jan. 2008).

291. Ben-Yaacov, E. & Eldar, Y. C. A fast and flexible method for the segmentation of aCGH data. en. *Bioinformatics* **24,** i139–i145 (Aug. 2008).

292. Kurosaki, T. & Ravetch, J. V. A single amino acid in the glycosyl phosphatidylinositol attachment domain determines the membrane topology of FcgRIII. en. *Nature* **342,** 805–807 (Dec. 1989).

293. Adu, B. *et al.* Fc Gamma Receptor IIIB Polymorphisms Are Associated with Clinical Malaria in Ghanaian Children. en. *PLOS ONE* **7,** 10 (2012).

294. Morris, D. L. *et al.* Evidence for both copy number and allelic (NA1/NA2) risk at the FCGR3B locus in systemic lupus erythematosus. en. *European Journal of Human Genetics* **18,** 1027–1031 (Sept. 2010).

295. Siriboonrit, U. *et al.* Association of Fcg receptor IIb and IIIb polymorphisms with susceptibility to systemic lupus erythematosus in Thais. en. *Tissue Antigens* **61,** 374–383 (May 2003).

296. Nielsen, K. R., Koelbaek, M. D., Varming, K., Baech, J. & Steffensen, R. Frequencies of HNA-1, HNA-3, HNA-4, and HNA-5 in the Danish and Zambian populations determined using a novel TaqMan real time polymerase chain reaction method. en. *Tissue Antigens* **80,** 249–253 (Sept. 2012).

297. Changsri, K., Tobunluepop, P. & Al, E. Human neutrophil alloantigen genotype frequencies in Thai blood donors. en. *Blood Transfusion* (2014).

298. Saxena, D. *et al.* Utility and limitations of multiplex ligation-dependent probe amplification technique in the detection of cytogenetic abnormalities in products of conception. *J Postgrad Med* **62,** 239–241 (2016).

299. Technical support, M.-H. *What are the main advantages and limitations of MLPA? - Knowledgebase / MLPA Technique - MRC-Holland Technical Support*

300. Lips, E. H. *et al.* Quantitative copy number analysis by Multiplex Ligation-dependent Probe Amplification (MLPA) of BRCA1-associated breast cancer regions identifies BRCAness. *Breast Cancer Res* **13,** R107 (2011).

Chapter 8

Appendices

8.1 Appendix A

8.1.1 VerifyBamID analysis of control samples

Training Sample	Capture Kit	Source	Gender	VerifyBamID Freemix
1	agilent50	Germline	F	0.00948
2	agilent50	Germline	F	0.0053
3	agilent50	Germline	F	0.0045
4	agilent51	Germline	M	0.00376
5	agilent51	Germline	M	0.00329
6	agilent51	Germline	M	0.00464
7	agilent51	Germline	M	0.00468
8	agilent51	Germline	M	0.00228
9	agilent50	Germline	F	0.00061
10	agilent50	Germline	M	0.00123
11	agilent50	Germline	M	0.00215
12	agilent50	Germline	M	0.00142
13	agilent50	Germline	F	0.00147
14	aglient51 V5	Germline	M	0.00117
15	aglient51 V5	Germline	F	0.00305
16	aglient51 V5	Germline	M	0.0021
17	TrueSeq	Germline	M	0.0098
18	aglient51	Germline	F	0.00265
19	aglient51 V4	Germline	M	0.00532
20	aglient51 V4	Germline	M	0.00532
21	aglient51 V5	Germline	F	0.00705
22	aglient51 V5	Germline	F	0.00705
23	aglient51 V4	Germline	F	0.00195
24	aglient51 V4	Germline	F	0.00195
25	aglient51 V5	Germline	F	0.00162
26	aglient51 V5	Germline	F	0.00162
27	aglient51 V4	Germline	F	0.00523
28	aglient51 V4	Germline	F	0.00523
29	aglient51 V4	Germline	F	0.00452
30	aglient51 V4	Germline	M	0.0021
31	aglient51 V4	Germline	M	0.00562
32	aglient51 V4	Germline	F	0.0016
33	aglient51 V5	Germline	F	0.00093
34	aglient51 V5	Germline	F	0.00114
35	aglient51 V4	Germline	M	0.00586
36	aglient51 V4	Germline	M	0.0013
37	aglient51 V4	Germline	M	0.00147
38	aglient51 V4	Germline	M	0.00147
39	aglient51 V4	Germline	F	0.00324
40	aglient51 V5	Germline	M	0.00465
41	aglient51 V5	Germline	M	0.00662
42	aglient51 V5	Germline	M	0.00183
43	aglient51 V4	Germline	M	0.001
44	aglient51 V5	Germline	M	0.00155
45	aglient51 V5	Germline	M	0.00128

Table 8.1: 45 Samples used for generating contamination simulations. Samples were selected from a variety of exome capture kits: agilent50 (8), Agilent 51 (v4,5 or unspecified - 36) and TruSeq (1). All samples were from germline sources and were all predicted to be below one percent contaminated by VerifyBamID.

8.1.2 Analysis scripts

Bcftools Extractions of fields from VCF

Listing 8.1: Bcftools Extractions of fields from VCF

```
#!/bin/bash
#Add a header to the results file
echo "SAMPLE    CONTAMINATION   SKEW_ALL        KURTOSIS_ALL    SKEW_HOM
        KURTOSIS_HETS">skew_kurtosis.tsv

#Read a variables file listing the sample path and an output prefix
cat sample_list.txt|while read sample id; do

#Extract info from VCF
bcftools query -f '%CHROM\t%POS\t%END\t%REF\t%ALT\t%QUAL\t[%GT]\t
        [%INFO/DP4]\n' "$sample" > temp.vcf

#Calculate AAF
awk 'BEGIN {OFS=FS="\t"} $6 >=20 {print $0}' temp.vcf > temp2.vcf
awk 'BEGIN {OFS=FS="\t"} {print $8}' temp2.vcf|
        awk 'BEGIN {FS=","} {print(($3+$4)/($1+$2+$3+$4))*100}' >aaf.txt
awk 'BEGIN {OFS=FS="\t"} {print $8}' temp2.vcf|
        awk 'BEGIN{OFS=FS=","}{print $1+$2+$3+$4}' > total_dp.txt

#Add header to file for AAF calculations
echo "CHROM     POS     END     REF     ALT     QUAL
        GT      DP4     DP4_TOTAL       AAF" > "$id"_aaf.tsv

# Merge VCF with AAF calculations
paste temp2.vcf total_dp.txt aaf.txt >>"$id"_aaf.tsv
awk 'BEGIN {OFS=FS="\t"} ($9 >=10){print $0}' "$id"_aaf.tsv
        >"$id"_aaf_d10_p20.tsv

rm temp.vcf aaf.txt total_dp.txt temp2.vcf "$id"_aaf.tsv
done
```

Skew and Kurtosis measurements

Listing 8.2: Skew and kurtosis measurements

```
library(moments)
args<-commandArgs(TRUE)

data=read.csv(args[1],sep="\t")

hets=subset(data,AAF<88)
homs=subset(data,AAF>=88)

sk=skewness(homs$AAF)
kurt=kurtosis(hets$AAF)

k_all=skewness(data$AAF)
sk_all=kurtosis(data$AAF)
results=data.frame(args[2],sk_all,k_all,sk,kurt)

write.table(results, "skew_kurtosis.tsv"
, sep = "\t", col.names =F, append = T, row.names = F)
```

Het:Hom and deviation measurements

Listing 8.3: Het:Hom and deviation measurements

```
cat sample_list.txt | while read loc sample; do

no_var=$(wc -l "sample"|awk '{print $1-1}')
xvar=$(awk '{if ($1 == "X" || $1=="x")  count++} END {print count + 1}' "$sample")
xhets=$(awk '{OFS=FS="\t"}{if ($1 == "X" || $1=="x") print $5}' "$sample"|
        awk '{OFS=FS="\t"}{if ($1 == "0/1") count++} END {print count +1}')
pc_alt_hom=$(awk '{OFS=FS="\t"}{if ($5 == "1/1")print $8}' "$sample"|
        awk '{OFS=FS="\t"} { sum += $1; n++ } END { if (n > 0) print sum / n; }')
pc_alt_het=$(awk '{OFS=FS="\t"}{if ($5 == "0/1")print $8}' "$sample"|
        awk '{OFS=FS="\t"} { sum += $1; n++ } END { if (n > 0) print sum / n;}')
no_hets=$(awk '{if ($5 == "0/1") count++} END {print count}' "$sample")
cont="NULL"
echo "$sample $cont     $no_var $xvar   $xhets  $pc_alt_hom     $pc_alt_het
        $no_hets"> QC.sample.temp
sed -i 's/ /\t/g' QC.sample.temp

##. calculate % heterozyogtes on chrX
awk '{FS=OFS="\t"} {pc_Xhets = ($5/$4)*100; print $0,pc_Xhets}' QC.sample.temp\\
 > QC.sample.temp2
##. predict gender, male if <50 otherwise female
awk '{FS=OFS="\t"} {if ($9 < 50) print $0,"Male"; else print $0,"Female";}'
        QC.sample.temp2 > QC.sample.temp3
##. calulate % heterozygotes over all variants (usually ~61%)
awk '{FS=OFS="\t"} {pc_hets = ($8/$3)*100; print $0,pc_hets}'
        QC.sample.temp3 > QC.sample.temp4
##. highlight excess heterozygosity if >63% (arbitrary threshold)
awk '{FS=OFS="\t"} {if ($11 >= 63) print $0,"excess_hets"; else print $0,"ok";}'
 QC.sample.temp4        > QC.sample.temp5
##. highlight excess reference reads in alternative homozygotes if > 1% (arbitrary threshold)
awk '{FS=OFS="\t"} {if ((100-$6) >= 1) print $0,"excess_ref_reads_hom"; else print $0,"ok";}'
        QC.sample.temp5 > QC.sample.temp6
##. calculate the heterozygote to homozygote ratio (usually ~1.55)
awk '{FS=OFS="\t"} {hethomratio = ($8/($3-$8)); print $0,hethomratio}' QC.sample.temp6
         > QC.sample.temp7
##. calculate the % reference reads for homozygotes
```

```
awk '{FS=OFS="\t"} {pc_refallele_hom = (100-$6); print $0,pc_refallele_hom}'
        QC.sample.temp7 > QC.sample.temp8
## calculate the % reference reads for heterozygotes
awk '{FS=OFS="\t"} {pc_refallele_het = (100-$7); print $0,pc_refallele_het}'
        QC.sample.temp8         > QC.sample.temp9
##. calculate the deviation metric= sum of deviation from 0 for % reference reads in
#       homozygotes and the deviation from 50 for % reference reads in heterozygotes
awk '{FS=OFS="\t"} {deva = ($15-0); devb = ($16-50); deviation_metric =
        (deva >= 0 ? deva : 0 - deva) + (devb >= 0 ? devb : 0 - devb); print $0,
                deviation_metric,deva,devb}' QC.sample.temp9 > QC.sample.temp10
##. add header
echo "SAMPLE    CONTAM  NO_VAR  NO_X_VAR(NULL)  NO_X_HETS(NULL)
AVG_HOM_AAF     AVG_HET_AAF     NO_HETS PC_X_HETS(NULL) GENDER(NULL)
PC_HETS CONTAMINATION1  CONTAMINATION2  HET:HOM_RATIO
PC_REF_HOM      PC_REF_HET      DEV_METRIC      DEV_A   DEV_B"
> QC.sample.header

var=$((var+1))

if [ "$var" -eq "1" ]
then
cat QC.sample.header QC.sample.temp10 > QC_common_coding_autosome_vars.tsv
else
cat QC.sample.temp10 >> QC_common_coding_autosome_vars.tsv
fi

rm QC.sample.temp* QC.sample.header
done
```

Regressions application in SCIKIT-learn

Listing 8.4: Regressions application

```
import pandas as pd
import numpy as np
import seaborn as sns
import os
import matplotlib.pyplot as plt
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from scipy import stats

qc=pd.read_csv('QC_common_coding_autosome_vars.tsv',sep="\t")
qc=qc.drop_duplicates()
qc=qc.loc[qc['NO_VAR']>= 5000]
kurt=pd.read_csv('skew_kurtosis.tsv',sep="\t",names=['SAMPLE',
'SKEW_ALL','KURTOSIS_ALL','SKEW_HOMS','KURTOSIS_HETS'])
kurt=kurt.drop_duplicates()
t1=pd.merge(qc,kurt,on='SAMPLE')

t1.to_csv('germline_samples.txt',sep="\t",index=False)

#Load Training set
training=pd.read_csv('training_data.tsv',sep="\t")
training=training[training['SOURCE'] == ('Germline')]
training=training[training['CONTAMINATION'] <=10]
```

```
training_features=training[['SKEW_ALL','KURTOSIS_ALL','SKEW_HOMS',
'KURTOSIS_HETS','HET:HOM_RATIO','DEV_A','DEV_B','PC_X_HETS']]
contamination=training[['CONTAMINATION']]

#Load samples
measures=pd.read_csv('germline_samples.txt ',sep="\t")
features=measures[['SKEW_ALL','KURTOSIS_ALL','SKEW_HOMS','
KURTOSIS_HETS','HET:HOM_RATIO','DEV_A','DEV_B','PC_X_HETS']]
tmp=features.describe()
tmp.to_csv('all_vars_stats.tsv',sep="\t")


########################################
## Regressions
########################################
lm = linear_model.LinearRegression()
lm.fit(training_features,contamination)
lm_predict = lm.predict(features).reshape(-1,1)
measures['OLS-10']=lm_predict[:,0]


#OLS results to file
results.sort_values(by=['OLS-10'],ascending=False)
#temp=results.sort_values(by=['VBAMID'],ascending=False)
#temp.to_csv('germline_samples_prediction_ols_10.tsv',sep="\t",index=None)
```

8.1.3 Contamination estimation program results for 200 contamination simulations

SAMPLE CROSS	% X HETS	HET:HOM RATIO	DEV MET.	DEV A	DEV B	SKEW ALL	KURT. ALL	SKEW HOMS	KURT. HETS	CONTAM.
1	61.3468	1.58636	2.4462	1.0272	1.419	1.371456	0.13159	-2.579738	3.4681	1
1	61.1336	1.61958	2.5596	1.2207	1.3389	1.383238	0.136024	-2.288515	3.498466	2.5
1	61.1486	1.67207	2.6829	1.4569	1.226	1.403717	0.138828	-2.062231	3.414879	5
1	62.3816	1.78805	2.7675	1.7052	1.0623	1.445033	0.14253	-1.921941	3.12026	10
1	64.3202	2.00437	2.6336	1.7231	0.9105	1.507371	0.151163	-1.983909	2.709139	20
1	65.7746	2.15696	2.664	1.6224	1.0416	1.541561	0.161693	-2.114388	2.527901	30
1	65.769	2.24906	2.7464	1.5234	1.223	1.561071	0.169299	-2.253257	2.450948	40
2	64.4411	2.27517	2.7293	1.445	1.2843	1.563929	0.172782	-2.338012	2.446154	50
2	60.5946	1.67804	2.8881	0.8567	2.0314	1.390187	0.215185	-3.056416	3.876159	1
2	60.7918	1.7081	3.1268	1.1324	1.9944	1.403351	0.21511	-2.560905	3.91848	2.5
2	61.3147	1.77306	3.3767	1.4479	1.9288	1.430912	0.216982	-2.205592	3.771814	5
2	62.2422	1.89128	3.5043	1.7013	1.803	1.475073	0.216326	-2.044008	3.40253	10
2	64.0369	2.08947	3.5948	1.7685	1.8263	1.533516	0.214571	-2.083653	2.901778	20
2	65.1092	2.21236	3.7116	1.7746	1.937	1.563272	0.218966	-2.186235	2.686715	30
2	65.0456	2.29706	3.6996	1.7205	1.9791	1.579587	0.221825	-2.300824	2.587321	40
2	65.8734	2.32349	3.7943	1.7003	2.094	1.582953	0.226275	-2.310123	2.57587	50
3	62.7451	1.61361	2.533	0.5983	1.9347	1.365522	0.18711	-3.943752	3.750761	1
3	62.654	1.63508	2.8562	0.9281	1.9281	1.376125	0.189856	-2.939919	3.854087	2.5
3	62.9068	1.68306	3.1345	1.286	1.8485	1.397156	0.192428	-2.365911	3.865635	5
3	63.339	1.79569	3.3629	1.6808	1.6821	1.439719	0.196305	-1.980753	3.553993	10
3	64.1628	2.0427	3.4974	1.9027	1.5947	1.516526	0.200766	-1.951924	2.898265	20
3	65.135	2.26806	3.3327	1.7464	1.5863	1.567486	0.201137	-2.159012	2.546489	30
3	66.054	2.42194	3.2811	1.5834	1.6977	1.594391	0.208753	-2.399575	2.399486	40
3	66.1833	2.49472	3.3015	1.4878	1.8137	1.606409	0.215548	-2.506451	2.35882	50
4	7.80312	1.57269	3.0722	0.6217	2.4505	1.339497	0.196093	-3.620779	4.012419	1
4	10.782	1.5894	3.4905	1.0042	2.4863	1.347761	0.197929	-2.531198	4.160083	2.5
4	18.267	1.63782	3.8494	1.5037	2.3457	1.366911	0.200475	-1.835846	4.37398	5
4	38.4113	1.78011	4.1387	2.1277	2.011	1.421732	0.201161	-1.44218	4.066045	10
4	59.3173	2.23967	3.8893	2.2454	1.6439	1.545208	0.209951	-1.63938	2.786927	20
4	67.2859	2.73889	4.1806	1.8097	2.3709	1.629986	0.247577	-2.37538	2.280519	30
4	70.4512	3.14147	4.4944	1.286	3.2084	1.677724	0.287738	-3.179876	2.076035	40
4	70.8731	3.36427	4.6628	0.9392	3.7236	1.697869	0.306069	-3.580748	1.998029	50
5	8.34403	1.56123	2.8101	0.5458	2.2643	1.337023	0.186903	-4.003194	4.123482	1
5	9.06769	1.57553	3.1145	0.8305	2.284	1.344681	0.18799	-3.023267	4.221744	2.5

Table 8.2 continued from previous page

SAMPLE CROSS	% X HETS	HET:HOM RATIO	DEV MET.	DEV A	DEV B	SKEW ALL	KURT. ALL	SKEW HOMS	KURT. HETS	CONTAM.
5	11.2388	1.61775	3.3954	1.2039	2.1915	1.362116	0.191824	-2.3436	4.339803	5
5	20.7241	1.74028	3.6913	1.7318	1.9595	1.414084	0.193056	-1.830168	3.97203	10
5	37.4308	2.01979	3.8803	2.118	1.7623	1.506322	0.198837	-1.698076	3.07127	20
5	49.0385	2.33932	4.1544	2.0714	2.083	1.580839	0.213866	-1.939126	2.548556	30
5	54.3536	2.59475	4.2092	1.7921	2.4171	1.621387	0.232919	-2.24339	2.335255	40
5	55.4339	2.74827	4.2725	1.5652	2.7073	1.638638	0.244814	-2.44684	2.256128	50
6	8.80579	1.59026	3.134	0.589	2.545	1.347354	0.206642	-3.709251	4.029033	1
6	10.1083	1.60682	3.5173	0.9534	2.5639	1.353472	0.209408	-2.613083	4.209168	2.5
6	13.1894	1.64975	3.9635	1.4667	2.4968	1.372128	0.210256	-1.855675	4.433908	5
6	23.0409	1.79613	4.2165	2.1112	2.1053	1.422401	0.213067	-1.401225	4.174263	10
6	44.4328	2.24422	4.1619	2.3335	1.8284	1.542591	0.225688	-1.558102	2.896107	20
6	52.8069	2.75542	4.4418	1.8939	2.5479	1.631749	0.262623	-2.289029	2.339655	30
6	55.8348	3.18866	4.9223	1.3613	3.561	1.688455	0.311663	-3.16333	2.121332	40
6	56.6318	3.44992	5.0599	0.8998	4.1601	1.718333	0.337589	-3.599965	2.038882	50
7	7.66917	1.56051	2.9273	0.6385	2.2888	1.343034	0.177813	-3.44773	3.776516	1
7	8.53293	1.57959	3.2919	0.9671	2.3248	1.350811	0.180529	-2.566142	3.941415	2.5
7	12.426	1.63085	3.6651	1.4464	2.2187	1.368555	0.1849	-1.882032	4.164919	5
7	24.9639	1.78192	3.918	2.0539	1.8641	1.4207	0.188615	-1.467968	3.909592	10
7	45.3441	2.25543	3.6392	2.1197	1.5195	1.540107	0.20203	-1.66044	2.754062	20
7	54.4048	2.76842	3.9507	1.6225	2.3282	1.626186	0.247021	-2.411817	2.280456	30
7	57.3046	3.18498	4.3993	1.092	3.3073	1.682757	0.300276	-3.303688	2.108574	40
7	58.9857	3.40869	4.7237	0.7328	3.9909	1.709235	0.327357	-3.754867	2.04217	50
8	8.87949	1.59199	2.8664	0.6337	2.2327	1.350401	0.201259	-3.7208	4.272453	1
8	11.3169	1.61524	3.1804	0.968	2.2124	1.36185	0.202481	-2.762694	4.387745	2.5
8	16.6667	1.67213	3.4629	1.3666	2.0963	1.384911	0.205897	-2.157136	4.381768	5
8	29.2797	1.80269	3.8269	1.8635	1.9634	1.439807	0.207863	-1.729111	3.882858	10
8	51.4717	2.15067	4.3352	2.2824	2.0528	1.548548	0.218623	-1.728533	2.865971	20
8	61.1765	2.46056	4.4777	2.1355	2.3422	1.607142	0.235087	-2.005195	2.464348	30
8	65.4545	2.70703	4.491	1.8592	2.6318	1.639144	0.250982	-2.400192	2.263706	40
8	66.7962	2.80855	4.5149	1.6832	2.8317	1.651668	0.259826	-2.525313	2.220629	50
9	64.7812	1.69143	2.8471	0.6954	2.1517	1.408688	0.197791	-3.472104	3.731551	1
9	65.0346	1.75688	3.0052	0.9548	2.0504	1.428782	0.204202	-2.764304	3.754381	2.5
9	64.955	1.85886	3.2655	1.2254	2.0401	1.465469	0.210402	-2.314309	3.596099	5
9	65.2726	2.04914	3.6269	1.4333	2.1936	1.521933	0.219769	-2.098245	3.201478	10
9	65.6904	2.27203	4.2426	1.6264	2.6162	1.574181	0.233541	-2.059	2.800072	20
9	65.3846	2.4233	4.3615	1.5913	2.7702	1.596902	0.245162	-2.165091	2.62237	30

Table 8.2 continued from previous page

SAMPLE CROSS	% X HETS	HET:HOM RATIO	DEV MET.	DEV A	DEV B	SKEW ALL	KURT. ALL	SKEW HOMS	KURT. HETS	CONTAM.
9	65.1757	2.4876	4.4708	1.5912	2.8796	1.60823	0.252121	-2.221074	2.555976	40
9	63.773	2.48367	4.4984	1.6047	2.8937	1.609024	0.252986	-2.237702	2.561425	50
10	20.405	1.72173	3.4044	0.6276	2.7768	1.422354	0.202596	-3.58484	3.698717	1
10	21.2286	1.79443	3.6089	0.9051	2.7038	1.445064	0.207784	-2.748255	3.719536	2.5
10	25.9878	1.91977	3.8859	1.1556	2.7303	1.483216	0.217227	-2.29424	3.55011	5
10	32.9734	2.12169	4.4021	1.3926	3.0095	1.539845	0.230264	-2.044775	3.184605	10
10	41.522	2.42516	4.8573	1.4378	3.4195	1.599578	0.257414	-2.108122	2.78639	20
10	45.6975	2.60565	5.1272	1.414	3.7132	1.629237	0.279314	-2.223167	2.612207	30
10	48.3388	2.70349	5.2428	1.3735	3.8693	1.647209	0.296497	-2.297533	2.553084	40
10	49.7379	2.73098	5.246	1.3479	3.8981	1.652146	0.303848	-2.320465	2.554924	50
11	19.5601	1.71595	3.3257	0.6119	2.7138	1.421272	0.202373	-3.680326	3.710979	1
11	24.0991	1.78423	3.5082	0.8551	2.6531	1.444898	0.207046	-2.918097	3.718453	2.5
11	30.5516	1.89212	3.7827	1.0578	2.7249	1.479937	0.211915	-2.531951	3.543732	5
11	38.961	2.06142	4.1651	1.2372	2.9279	1.529532	0.217784	-2.26383	3.184	10
11	48.6535	2.27306	4.514	1.3295	3.1845	1.575495	0.231178	-2.227265	2.855661	20
11	53.9243	2.40597	4.741	1.3345	3.4065	1.598532	0.245747	-2.251214	2.725154	30
11	56.6245	2.4699	4.866	1.3478	3.5182	1.608673	0.249823	-2.271841	2.650512	40
11	58.7561	2.51922	4.8467	1.3155	3.5312	1.616789	0.256151	-2.282498	2.626981	50
12	21.5821	1.47182	3.2671	0.5949	2.6722	1.360397	0.062028	-3.618441	3.530619	1
12	26.4901	1.53146	3.3478	0.9091	2.4387	1.375005	0.071841	-2.638694	3.597041	2.5
12	36.0402	1.65424	3.459	1.2409	2.2181	1.408138	0.09056	-2.084015	3.507673	5
12	50.4148	1.91768	3.6196	1.5428	2.0768	1.470275	0.127341	-1.725342	3.123516	10
12	62.7233	2.33578	4.025	1.5271	2.4979	1.55424	0.18659	-1.774862	2.678751	20
12	67.9377	2.60173	4.464	1.3874	3.0766	1.606457	0.237099	-1.998648	2.505738	30
12	70.3718	2.78396	4.7647	1.2204	3.5443	1.646433	0.27943	-2.29938	2.448492	40
12	70.3002	2.87553	4.9243	1.0842	3.8401	1.669287	0.306749	-2.504239	2.43636	50
13	60.4906	1.81481	3.3897	0.5341	2.8556	1.455821	0.247303	-3.92656	3.663368	1
13	60.9171	1.88633	3.5488	0.7495	2.7993	1.477946	0.250646	-3.107198	3.657159	2.5
13	61.7556	2.00029	3.7982	0.9338	2.8644	1.51306	0.255465	-2.662761	3.489673	5
13	62.6211	2.1618	4.184	1.1121	3.0719	1.560324	0.254752	-2.358163	3.193862	10
13	64.3216	2.38171	4.5458	1.2397	3.3061	1.604108	0.259072	-2.243254	2.865454	20
13	64.2893	2.49018	4.7878	1.2536	3.5342	1.617969	0.26048	-2.275996	2.702262	30
13	64.15	2.56579	4.8097	1.2467	3.563	1.625927	0.265013	-2.320009	2.625445	40
13	63.8161	2.58317	4.7935	1.2236	3.5699	1.626589	0.261363	-2.341416	2.590852	50
14	20.7053	1.64168	3.1802	0.4366	2.7436	1.395237	0.170652	-4.145619	3.440043	1
14	22.1209	1.66353	3.496	0.7932	2.7028	1.401312	0.171797	-2.767326	3.523062	2.5

Table 8.2 continued from previous page

SAMPLE CROSS	% X HETS	HET:HOM RATIO	DEV MET.	DEV A	DEV B	SKEW ALL	KURT. ALL	SKEW HOMS	KURT. HETS	CONTAM.
14	25.2571	1.72816	3.7771	1.2771	2.5	1.417941	0.17629	-1.929957	3.715658	5
14	35.6044	1.96066	3.4847	1.7251	1.7596	1.468485	0.181624	-1.483424	3.573478	10
14	55.7915	2.6517	3.2595	1.4844	1.7751	1.595307	0.210974	-1.743497	2.585841	20
14	61.4912	3.32604	4.3106	0.9198	3.3908	1.680739	0.289575	-2.638011	2.198745	30
14	64.0239	3.79776	5.6427	0.5349	5.1078	1.746713	0.370056	-3.663175	2.07732	40
14	64.4068	3.95011	6.2933	0.4442	5.8491	1.775267	0.408057	-4.029179	2.06031	50
15	59.0792	1.72451	3.2568	0.5223	2.7345	1.413631	0.249396	-4.002751	3.972011	1
15	59.0698	1.75292	3.505	0.7997	2.7053	1.423772	0.249133	-2.931982	4.057884	2.5
15	59.1805	1.82804	3.743	1.1335	2.6095	1.448897	0.250275	-2.274785	4.120932	5
15	60.2497	2	4.0572	1.4619	2.5953	1.504996	0.250871	-1.903117	3.752079	10
15	62.2839	2.3847	4.4814	1.5818	2.8996	1.595196	0.259191	-1.906232	2.969045	20
15	64.1161	2.73139	4.9033	1.4418	3.4615	1.644387	0.279423	-2.150772	2.556353	30
15	65.283	2.98417	5.2157	1.2855	3.9302	1.670731	0.301448	-2.396584	2.356968	40
16	17.968	1.61486	3.3612	0.4963	2.8649	1.380233	0.176883	-3.838536	3.632134	1
16	21.5084	1.63583	3.7154	0.8771	2.8383	1.387808	0.177636	-2.590881	3.751234	2.5
16	31.1416	1.70388	3.9709	1.3767	2.5942	1.405051	0.182317	-1.838285	3.957116	5
16	50.5217	1.92984	3.7465	1.8957	1.8508	1.458168	0.189502	-1.397641	3.724604	10
16	65.6796	2.61587	3.3741	1.7491	1.625	1.592785	0.217976	-1.565025	2.653772	20
16	71.5633	3.32929	4.3113	1.1898	3.1215	1.690743	0.293835	-2.222061	2.241645	30
16	74.6222	3.88488	5.6602	0.7327	4.9275	1.768525	0.381634	-3.064938	2.104989	40
16	75.3049	4.17591	6.5914	0.5529	6.0385	1.818722	0.440492	-3.661359	2.081673	50
17	8.66242	1.53388	2.574	0.7196	1.8544	1.327669	0.168926	-3.31291	3.961431	1
17	11.3924	1.54968	3.0049	1.1036	1.9013	1.336061	0.171856	-2.399138	4.095238	2.5
17	17.3585	1.58926	3.5171	1.6568	1.8603	1.353224	0.174969	-1.746242	4.308017	5
17	34.189	1.72601	3.8296	2.3383	1.4913	1.408276	0.175382	-1.399649	3.877527	10
17	52.275	2.17056	3.3559	2.415	0.9409	1.532454	0.182622	-1.654645	2.656821	20
17	60.8963	2.64497	3.5633	1.9427	1.6206	1.618126	0.217222	-2.438591	2.195456	30
17	64.7718	3.04375	3.6748	1.2889	2.3859	1.666444	0.254333	-3.189751	2.020884	40
17	65.7471	3.19275	3.7043	0.9622	2.7421	1.68178	0.266209	-3.555385	1.966963	50
18	61.4061	1.6636	3.3102	0.5941	2.7161	1.399504	0.170462	-3.393852	3.502499	1
18	62.1905	1.70818	3.4371	0.8659	2.5712	1.410968	0.174577	-2.614803	3.61147	2.5
18	63.4328	1.8066	3.508	1.1757	2.3323	1.438135	0.178988	-2.083021	3.628012	5
18	66.0557	2.03262	3.6051	1.4225	2.1826	1.495654	0.19146	-1.792118	3.277616	10
18	70.0776	2.51437	4.0774	1.3775	2.6999	1.58735	0.227953	-1.896797	2.670148	20
18	73.3601	2.8961	4.6189	1.2097	3.4092	1.640113	0.267968	-2.159174	2.37997	30
18	74.5289	3.14109	4.9931	1.0539	3.9392	1.671938	0.300695	-2.407897	2.265283	40

Table 8.2 continued from previous page

SAMPLE CROSS	% X HETS	HET:HOM RATIO	DEV MET.	DEV A	DEV B	SKEW ALL	KURT. ALL	SKEW HOMS	KURT. HETS	CONTAM.
18	75.1282	3.2125	5.1321	0.9958	4.1363	1.684228	0.30816	-2.53823	2.22329	50
19	57.972	1.62635	2.7455	0.5503	2.1952	1.367472	0.205884	-3.884432	3.868174	1
19	57.8667	1.64206	3.0644	0.8615	2.2029	1.375059	0.207901	-2.902581	3.987079	2.5
19	58.7649	1.68903	3.4566	1.2673	2.1893	1.396817	0.211285	-2.262899	4.109785	5
19	60.9293	1.81335	3.7867	1.7717	2.015	1.444556	0.216476	-1.842053	3.791549	10
19	66.0714	2.13504	4.0759	2.0699	2.006	1.543207	0.224095	-1.841346	2.912612	20
19	68.6477	2.46761	4.2164	1.8551	2.3613	1.605276	0.237651	-2.105823	2.474801	30
19	70.8801	2.72761	4.3175	1.6012	2.7163	1.642429	0.254117	-2.404135	2.287981	40
19	72.4178	2.84384	4.3797	1.4633	2.9164	1.649726	0.26337	-2.484014	2.228611	50
20	60.4955	1.73329	3.3905	0.5207	2.8698	1.424575	0.211853	-3.780055	3.491618	1
20	60.7708	1.7635	3.7523	0.9163	2.836	1.431924	0.215283	-2.540713	3.597522	2.5
20	61.1111	1.84598	4.0502	1.4321	2.6181	1.45374	0.218539	-1.793358	3.754765	5
20	62.1697	2.12557	3.8997	1.8644	2.0353	1.516831	0.226768	-1.441138	3.453858	10
20	66.9204	2.88695	4.0624	1.5607	2.5017	1.642243	0.264843	-1.820347	2.507482	20
20	71.4286	3.61255	5.0605	0.9446	4.1159	1.7334	0.352982	-2.730935	2.195649	30
20	74.4548	4.04869	6.1797	0.6041	5.5756	1.800189	0.423082	-3.543794	2.107142	40
20	74.875	4.12753	6.3854	0.5591	5.8263	1.813292	0.432191	-3.630148	2.077409	50
21	55.1849	1.57468	2.4645	0.7061	1.7584	1.33976	0.190362	-3.335255	4.001317	1
21	55.2098	1.58566	2.9197	1.0951	1.8246	1.345987	0.191356	-2.385512	4.086374	2.5
21	55.7734	1.62223	3.5017	1.634	1.8677	1.364636	0.193858	-1.744398	4.28576	5
21	56.25	1.75096	3.914	2.3444	1.5696	1.417912	0.192615	-1.404678	3.932061	10
21	58.19	2.19539	3.5788	2.428	1.1508	1.539934	0.191518	-1.738698	2.655431	20
21	62.9491	2.69216	3.6966	1.8833	1.8133	1.62015	0.226388	-2.622246	2.186576	30
21	65.6727	3.06436	3.6648	1.2296	2.4352	1.66644	0.255428	-3.376903	2.009025	40
21	66.2222	3.2007	3.6936	0.9579	2.7357	1.683257	0.267259	-3.598898	1.962437	50
22	12.043	1.64951	3.2469	0.4361	2.8108	1.400598	0.150454	-4.075765	3.308304	1
22	12.9032	1.69275	3.3809	0.7027	2.6782	1.412137	0.153057	-2.991799	3.40107	2.5
22	16.4179	1.78023	3.4739	1.0124	2.4615	1.434757	0.160102	-2.3039	3.437751	5
22	29.7787	2.00248	3.474	1.2827	2.1913	1.486528	0.173161	-1.896391	3.179779	10
22	44.9123	2.47343	3.7354	1.3144	2.421	1.573514	0.208158	-1.891543	2.635235	20
22	54.3779	2.83697	4.3803	1.1701	3.2102	1.628175	0.247403	-2.12205	2.372849	30
22	57.91	3.08876	4.7277	1.0144	3.7133	1.661331	0.279193	-2.31154	2.257972	40
22	58.0645	3.17057	4.8107	0.9073	3.9034	1.671427	0.290266	-2.454857	2.220207	50
23	18.241	1.67935	3.4664	0.5782	2.8882	1.400909	0.192056	-3.545679	3.644537	1
23	20.7792	1.71946	3.6875	0.8749	2.8126	1.413867	0.194555	-2.673931	3.73772	2.5
23	25.3968	1.81064	3.811	1.1774	2.6336	1.442247	0.197998	-2.156379	3.762769	5

Table 8.2 continued from previous page

SAMPLE CROSS	% X HETS	HET:HOM RATIO	DEV MET.	DEV A	DEV B	SKEW ALL	KURT. ALL	SKEW HOMS	KURT. HETS	CONTAM.
23	37	2.03224	3.9132	1.4275	2.4857	1.504731	0.204809	-1.888833	3.369504	10
23	48.5608	2.48331	4.3817	1.4243	2.9574	1.596908	0.236616	-1.940958	2.717378	20
23	54.6776	2.84867	4.8963	1.2325	3.6638	1.649321	0.272962	-2.214417	2.430669	30
23	56.7111	3.08775	5.184	1.087	4.097	1.675342	0.298427	-2.415708	2.303358	40
23	57.6923	3.17245	5.4138	1.0613	4.3525	1.684097	0.311696	-2.526027	2.258858	50
24	23.0851	1.6528	3.1795	0.47	2.7095	1.390879	0.19627	-4.047868	3.621157	1
24	26.2055	1.67481	3.538	0.8268	2.7112	1.39727	0.19893	-2.724177	3.733364	2.5
24	33.9545	1.73764	3.8537	1.3011	2.5526	1.415974	0.201812	-1.943299	3.917167	5
24	50.4836	1.95303	3.7405	1.7835	1.957	1.471496	0.205357	-1.497128	3.734563	10
24	64.5186	2.56749	3.6542	1.6516	2.0026	1.593457	0.225629	-1.691032	2.684352	20
24	70.8305	3.23129	4.5802	1.1116	3.4686	1.680155	0.292238	-2.440038	2.253643	30
24	73.0238	3.74407	5.7338	0.6502	5.0836	1.740739	0.36745	-3.319682	2.094762	40
24	73.199	3.94443	6.36	0.5215	5.8385	1.769588	0.404418	-3.814142	2.049508	50
25	60.1533	1.68079	3.2421	0.47	2.7721	1.40892	0.191853	-4.065608	3.553762	1
25	60.1662	1.70376	3.6257	0.8405	2.7852	1.416257	0.195263	-2.715889	3.651507	2.5
25	60.4487	1.76726	3.9767	1.317	2.6597	1.432554	0.198863	-1.92285	3.790852	5
25	61.9819	1.9971	3.7609	1.7877	1.9732	1.483475	0.202627	-1.492441	3.568171	10
25	66.8835	2.66981	3.71	1.5999	2.1101	1.60524	0.229351	-1.733872	2.593024	20
25	70.7332	3.3419	4.5898	0.9854	3.6044	1.685017	0.301883	-2.615535	2.203951	30
25	72.8125	3.79238	5.8076	0.6052	5.2024	1.745357	0.375167	-3.501974	2.082015	40
25	73.3501	3.91103	6.2634	0.5272	5.7362	1.764653	0.398259	-3.817469	2.046733	50

Table 8.2: All 200 contamination simulation used to train regressions. All 9 features supplied into the regression for each sample are listed.

8.1.4 Contamination estimation program results for 245 germline samples

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
1	68469	61.9143	Female	1.61	0.75	1.77	2.52	1.39	0.13	-3.06	3.35	0.95	1.47
2	83252	60.1156	Female	1.64	0.54	2.4	2.94	1.38	0.19	-4.09	3.78	0.67	0.79
3	88772	60.4992	Female	1.68	0.53	2.16	2.69	1.4	0.21	-4.16	3.79	0.53	0.92
4	53541	70.6332	Female	2.95	1.87	0.96	2.83	1.67	0.21	-1.93	2.05	32.04	25.27
5	109246	23.3607	Male	1.67	0.53	1.89	2.42	1.41	0.17	-4.08	3.33	0.54	1.46
6	112864	21.8487	Male	1.74	0.38	2.01	2.39	1.43	0.22	-5.22	3.35	0.18	1.87
7	107763	23.9101	Male	1.76	0.39	1.99	2.38	1.44	0.22	-5.06	3.28	0.24	2.02
8	75693	7.68156	Male	1.6	0.26	2.75	3.01	1.37	0.18	-6.02	3.73	0.35	1.11
9	72213	5.58824	Male	1.68	0.33	2.51	2.84	1.39	0.21	-5	3.64	0.59	1.19
10	80526	7.18563	Male	1.54	0.24	2.66	2.9	1.34	0.17	-6.37	3.87	0.28	0.84
11	74907	7.74749	Male	1.65	0.24	2.77	3.01	1.38	0.22	-6.21	3.7	0.31	1.43
12	81692	7.15152	Male	1.59	0.28	2.69	2.97	1.36	0.19	-5.81	3.9	0.38	0.88
13	82679	7.26329	Male	1.57	0.26	2.58	2.84	1.36	0.18	-6.14	3.95	0.33	0.88
14	80294	58.8612	Female	1.67	0.23	2.47	2.7	1.39	0.21	-6.55	3.79	0.2	1.03
15	83273	58.6433	Female	1.66	0.25	2.71	2.96	1.39	0.21	-6.26	3.82	0.24	0.79
16	83266	8.21256	Male	1.6	0.3	2.67	2.97	1.36	0.2	-5.61	3.96	0.46	0.74
17	74577	7.0122	Male	1.58	0.28	2.56	2.84	1.36	0.17	-5.57	3.66	0.47	0.76
18	85838	7.66423	Male	1.63	0.24	2.58	2.82	1.38	0.2	-6.46	3.86	0.23	1.36
19	91586	64.6838	Female	1.67	0.38	2.18	2.56	1.4	0.2	-5.41	3.63	0.06	0.96
20	85288	18.7686	Male	1.73	0.28	2.67	2.95	1.43	0.19	-5.86	3.35	0.14	2.09
21	106355	19.2941	Male	1.69	0.32	2.81	3.13	1.41	0.2	-5.73	3.63	0.12	1.74
22	97451	16.4013	Male	1.68	0.33	2.78	3.11	1.41	0.2	-5.78	3.66	0.22	1.7
23	92975	19.7735	Male	1.45	0.29	2.76	3.05	1.35	0.06	-5.95	3.47	0.14	0.41
24	104377	61.1737	Female	1.73	0.33	2.78	3.11	1.43	0.23	-6.05	3.68	0.09	1.64
25	96505	60.5003	Female	1.78	0.27	2.9	3.17	1.45	0.24	-6.43	3.58	0.15	2
26	88037	62.6346	Female	1.69	0.15	2.77	2.92	1.41	0.21	-7.86	3.45	0.08	2.13
27	84816	19.59	Male	1.63	0.16	2.7	2.86	1.39	0.17	-7.45	3.41	0.12	2.13
28	72840	61.3798	Female	1.75	0.62	2.78	3.4	1.45	0.16	-3.37	3.08	0.91	2.41
29	80421	59.2154	Female	1.75	0.29	2.9	3.19	1.44	0.19	-5.79	3.22	0.05	2.15
30	107331	59.112	Female	1.71	0.27	2.71	2.98	1.41	0.25	-6.3	3.91	0.31	1.13
31	78444	19.0114	Male	1.64	0.17	2.84	3.01	1.39	0.17	-7.11	3.4	0.13	2.08
32	85891	18.1258	Male	1.66	0.21	2.78	2.99	1.4	0.19	-6.36	3.45	0.23	1.59
33	89249	17.7885	Male	1.37	0.24	2.19	2.43	1.29	0.06	-6.36	3.7	0.16	-0.02
34	85072	7.95073	Male	1.34	0.17	2.59	2.76	1.3	0.04	-7.52	3.8	0.15	0.34

Table 8.3 continued from previous page

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
35	86329	16.8009	Male	1.63	0.34	2.76	3.1	1.39	0.17	-5.07	3.5	0.49	1.11
36	81745	63.0319	Female	1.7	0.18	2.63	2.81	1.42	0.19	-7.08	3.39	0.1	1.86
37	78253	53.4419	Female	1.76	0.29	2.74	3.03	1.43	0.23	-5.4	3.4	0.45	1.5
38	82309	57.8947	Female	1.63	0.22	2.79	3.01	1.39	0.16	-6.36	3.41	0.19	1.13
39	75933	58.7222	Female	1.55	0.2	2.73	2.93	1.37	0.12	-6.57	3.29	0.16	0.84
40	72712	54.3379	Female	1.55	0.17	2.77	2.94	1.37	0.11	-6.97	3.28	0.13	1.1
41	79800	19.2757	Male	1.64	0.23	2.74	2.97	1.39	0.17	-6.16	3.41	0.26	1.52
42	83369	19.8606	Male	1.67	0.32	2.9	3.22	1.4	0.19	-5.05	3.47	0.49	1.28
43	81024	16.1176	Male	1.64	0.3	2.71	3.01	1.39	0.17	-5.28	3.44	0.41	1.24
44	113163	16.5144	Male	1.63	0.31	2.62	2.93	1.37	0.21	-5.75	4.06	0.28	0.9
45	105282	15.3343	Male	1.49	0.22	2.67	2.89	1.33	0.14	-7.05	4.05	0.19	0.88
46	110365	62.5638	Female	1.72	0.2	2.71	2.91	1.41	0.25	-7.19	3.88	0.07	1.44
47	108913	62.784	Female	1.7	0.26	2.8	3.06	1.4	0.25	-6.47	4.02	0.28	1.05
48	106780	63.3951	Female	1.7	0.22	2.75	2.97	1.4	0.25	-7.06	3.96	0.13	1.34
49	111031	16.6788	Male	1.6	0.27	2.77	3.04	1.36	0.2	-6.35	4.1	0.22	0.97
50	113453	63.255	Female	1.78	0.21	2.66	2.87	1.43	0.29	-7.03	4.15	0.11	1.43
51	105691	15.1899	Male	1.5	0.26	2.74	3	1.33	0.14	-6.33	4.01	0.21	0.56
52	91554	63.4863	Female	1.7	0.25	2.59	2.84	1.41	0.21	-7.31	3.58	0	2.13
53	94173	63.5478	Female	1.77	0.23	2.52	2.75	1.43	0.25	-7.34	3.67	0	2.31
54	80545	7.31383	Male	1.59	0.39	1.86	2.25	1.35	0.18	-5.02	3.77	0.42	0.88
55	90820	8	Male	1.57	0.42	2.04	2.46	1.35	0.19	-4.78	4.05	0.56	0.49
56	86189	58.9126	Female	1.6	0.38	2.07	2.45	1.36	0.19	-5.05	3.84	0.44	0.24
57	96096	21.3028	Male	1.78	0.44	2.91	3.35	1.44	0.21	-4.47	3.4	0.37	2.32
58	108127	22.5331	Male	1.74	0.42	2.86	3.28	1.43	0.2	-4.79	3.57	0.32	2
59	103301	22.1653	Male	1.71	0.53	3.48	4.01	1.41	0.21	-4.07	3.46	0.59	2.25
60	115932	60.6994	Female	1.77	0.37	1.95	2.32	1.44	0.23	-5.37	3.37	0.13	1.58
61	97885	27.907	Male	1.78	0.36	3.58	3.94	1.44	0.22	-5.09	3.25	0.17	2.68
62	137697	60.9934	Female	1.74	0.37	2.55	2.92	1.44	0.23	-5.44	3.66	0.24	1.24
63	137737	60.9251	Female	1.76	0.35	2.46	2.81	1.44	0.24	-5.63	3.62	0.19	1.35
64	84120	59.6219	Female	1.55	0.26	2.38	2.64	1.35	0.15	-5.86	3.78	0.27	0.02
65	83090	59.4203	Female	1.63	0.24	2.4	2.64	1.38	0.19	-6.37	3.73	0.14	0.82
66	87067	16.6484	Male	1.68	0.19	2.73	2.92	1.41	0.21	-7.12	3.47	0.2	2.14
67	86994	18.4735	Male	1.65	0.24	2.73	2.97	1.39	0.19	-6.28	3.51	0.29	1.54
68	139899	21.1144	Male	1.87	0.75	2.92	3.67	1.46	0.29	-3.69	4.09	0.72	3.62
69	142708	25.2106	Male	1.91	0.86	3.05	3.91	1.47	0.3	-3.35	4.02	1.1	4.42
70	69629	8.80399	Male	1.64	0.23	2.72	2.95	1.39	0.18	-6.37	3.53	0.09	1.71

Table 8.3 continued from previous page

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
71	73320	60.9756	Female	1.61	0.22	2.81	3.03	1.38	0.16	-6.54	3.48	0.11	1.17
72	69794	6.6129	Male	1.62	0.2	2.82	3.02	1.38	0.17	-6.67	3.51	0.12	1.87
73	78738	8.33333	Male	1.55	0.24	2.21	2.45	1.36	0.14	-5.89	3.48	0.29	0.75
74	74773	8.85341	Male	1.59	0.25	2.73	2.98	1.37	0.15	-6.03	3.59	0.14	1.22
75	76359	9.33148	Male	1.84	0.22	2.7	2.92	1.46	0.27	-6.4	3.56	0.07	2.69
76	71184	7.43034	Male	1.8	0.23	2.89	3.12	1.45	0.25	-6.37	3.46	0.16	2.6
77	78969	6.82148	Male	1.59	0.26	2.24	2.5	1.37	0.16	-5.9	3.51	0.31	1.05
78	69256	53.8188	Female	1.72	0.22	2.75	2.97	1.42	0.21	-6.49	3.46	0.15	1.83
79	98638	57.5668	Female	1.63	0.31	2.1	2.41	1.38	0.2	-5.5	3.97	0.3	0.22
80	96641	58.1311	Female	1.68	0.27	2.21	2.48	1.39	0.23	-6.23	3.96	0.16	0.8
81	91599	7.74487	Male	1.64	0.45	2.43	2.88	1.38	0.21	-4.25	3.88	0.88	0.72
82	80735	9.11565	Male	1.57	0.25	2.14	2.39	1.36	0.15	-5.99	3.61	0.28	0.88
83	82890	6.64011	Male	1.56	0.26	2.22	2.48	1.36	0.15	-5.85	3.63	0.27	0.81
84	84843	8.42246	Male	1.61	0.27	2.37	2.64	1.37	0.18	-5.9	3.68	0.24	1.08
85	82505	61.1588	Female	1.65	0.15	2.71	2.86	1.4	0.18	-7.63	3.35	0.08	1.9
86	87090	9.8401	Male	1.57	0.7	2.16	2.86	1.35	0.18	-3.32	4.03	1.43	1.12
87	80949	9.63542	Male	1.56	0.3	2.18	2.48	1.35	0.16	-5.7	3.78	0.27	0.78
88	81721	8.05627	Male	1.55	0.37	2.17	2.54	1.34	0.17	-5.04	3.83	0.53	0.6
89	78059	61.0945	Female	1.64	0.34	2.76	3.1	1.39	0.17	-4.83	3.44	0.71	0.63
90	69252	22.6629	Male	1.79	0.33	2.84	3.17	1.45	0.19	-5.43	3.06	0.08	3.1
91	85101	8.20707	Male	1.62	0.33	2.46	2.79	1.37	0.2	-5.24	3.87	0.47	0.75
92	80548	9.15301	Male	1.62	0.47	2.44	2.91	1.37	0.18	-4.19	3.7	0.97	0.88
93	102471	21.3012	Male	1.71	0.41	2.79	3.2	1.42	0.19	-5.09	3.55	0.04	2.13
94	74056	24.6914	Male	1.81	0.5	2.92	3.42	1.46	0.21	-3.98	3.2	0.8	2.79
95	87232	57.6021	Female	1.64	0.25	2.51	2.76	1.38	0.2	-6.21	3.77	0.2	0.72
96	69863	62.1803	Female	1.76	0.43	3.02	3.45	1.44	0.17	-4.44	3.08	0.46	2.18
97	81783	8.86571	Male	1.59	0.38	1.99	2.37	1.36	0.19	-5.06	3.84	0.43	0.79
98	78094	22.8145	Male	1.79	0.3	2.84	3.14	1.45	0.21	-5.81	3.23	0.05	2.94
99	81040	60.2476	Female	1.72	0.21	2.84	3.05	1.42	0.21	-6.76	3.45	0.16	1.88
100	87965	21.8779	Male	1.73	0.36	2.82	3.18	1.43	0.19	-5.43	3.41	0.06	2.28
101	89520	25.0689	Male	1.85	0.33	2.87	3.2	1.47	0.25	-5.6	3.36	0.05	2.95
102	80315	59.0731	Female	1.79	0.33	2.99	3.32	1.45	0.21	-5.36	3.2	0.27	2.32
103	79149	6.32411	Male	1.53	0.35	1.82	2.17	1.33	0.16	-5.27	3.78	0.38	0.52
104	81323	7.73006	Male	1.55	0.39	2.15	2.54	1.35	0.16	-4.84	3.82	0.52	0.52
105	77128	17.3684	Male	1.67	0.21	2.73	2.94	1.41	0.18	-6.63	3.32	0.22	2.03
106	79296	59.2566	Female	1.65	0.29	2.02	2.31	1.38	0.21	-5.85	3.7	0.27	0.71

Table 8.3 continued from previous page

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
107	67983	59.3277	Female	1.65	0.22	2.66	2.88	1.39	0.18	-6.38	3.43	0.26	1.19
108	72102	9.48148	Male	1.6	0.23	2.73	2.96	1.38	0.15	-6.27	3.54	0.17	1.49
109	82378	55.4103	Female	1.6	0.38	2.06	2.44	1.36	0.19	-5.05	3.84	0.52	0.27
110	82236	7.72251	Male	1.62	0.3	2.53	2.83	1.38	0.18	-5.78	3.7	0.23	1.31
111	82559	10.5834	Male	1.64	0.33	2.35	2.68	1.38	0.2	-5.26	3.69	0.44	0.98
112	90951	10.739	Male	1.6	0.29	2.42	2.71	1.37	0.18	-5.64	3.85	0.25	0.81
113	78925	7.45156	Male	1.64	0.33	2.02	2.35	1.37	0.21	-5.48	3.69	0.34	1.24
114	68581	55.5081	Female	1.66	0.29	2.63	2.92	1.39	0.18	-5.4	3.49	0.45	0.88
115	86185	7.81609	Male	1.59	0.4	1.97	2.37	1.35	0.2	-4.96	3.95	0.43	0.77
116	80344	6.68449	Male	1.6	0.28	1.99	2.27	1.35	0.19	-5.95	3.74	0.21	1.08
117	92425	18.8	Male	1.65	0.23	2.8	3.03	1.39	0.21	-6.55	3.68	0.2	1.57
118	78433	7.327	Male	1.55	0.3	2.65	2.95	1.35	0.14	-5.68	3.69	0.25	0.89
119	85995	7.83818	Male	1.6	0.46	2.03	2.49	1.36	0.18	-4.5	3.85	0.59	0.84
120	82974	8.21192	Male	1.61	0.39	2.28	2.67	1.36	0.2	-4.93	3.9	0.56	0.91
121	75418	57.2403	Female	1.65	0.26	2.92	3.18	1.4	0.17	-5.88	3.37	0.32	1.14
122	88079	60.144	Female	1.61	0.31	2.54	2.85	1.37	0.19	-5.56	3.88	0.21	0.48
123	89522	9.33489	Male	1.61	0.29	2.55	2.84	1.37	0.19	-5.75	3.85	0.3	0.98
124	93277	10.8369	Male	1.61	0.57	2.32	2.89	1.37	0.17	-3.78	3.87	1.07	1.01
125	87139	10.2564	Male	1.67	0.32	2.76	3.08	1.39	0.22	-5.45	3.94	0.21	1.3
126	73306	58.7338	Female	1.75	0.23	2.72	2.95	1.43	0.21	-6.04	3.26	0.28	1.69
127	86688	63.807	Female	1.71	0.34	2.6	2.94	1.4	0.24	-5.33	3.9	0.32	0.85
128	83870	55.6463	Female	1.69	0.31	2.52	2.83	1.4	0.22	-5.49	3.82	0.25	0.88
129	81110	56.3611	Female	1.7	0.25	2.84	3.09	1.41	0.2	-5.88	3.47	0.28	1.27
130	71472	22.6361	Male	1.67	0.21	2.76	2.97	1.4	0.17	-6.45	3.26	0.18	1.94
131	86223	7.52315	Male	1.61	0.47	2.63	3.1	1.36	0.19	-4.33	3.99	0.7	0.92
132	84568	9.39675	Male	1.57	0.31	2.32	2.63	1.36	0.15	-5.44	3.75	0.28	0.61
133	78308	9.57746	Male	1.66	0.33	2.68	3.01	1.38	0.21	-5.24	3.73	0.38	1.23
134	84396	7.72251	Male	1.55	0.37	1.84	2.21	1.34	0.16	-5.16	3.84	0.4	0.62
135	80661	23.7288	Male	1.68	0.21	2.78	2.99	1.4	0.19	-6.6	3.45	0.16	1.91
136	81242	19.8582	Male	1.62	0.24	2.91	3.15	1.39	0.16	-6.21	3.47	0.23	1.5
137	79204	60.4243	Female	1.58	0.36	1.82	2.18	1.34	0.18	-5.2	3.78	0.37	0.22
138	79316	61.3475	Female	1.58	0.25	2.11	2.36	1.35	0.18	-6.29	3.74	0.14	0.53
139	87116	61.2859	Female	1.67	0.52	2.69	3.21	1.39	0.21	-4.03	3.88	0.83	0.85
140	79702	7.28477	Male	1.57	0.38	2.15	2.53	1.35	0.17	-4.96	3.77	0.47	0.75
141	78559	7.29313	Male	1.55	0.45	2.18	2.63	1.34	0.17	-4.41	3.77	0.77	0.58
142	78098	54.4283	Female	1.65	0.19	2.75	2.94	1.39	0.19	-6.9	3.63	0.05	1.46

Table 8.3 continued from previous page

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
143	76341	59.2013	Female	1.67	0.21	2.85	3.06	1.4	0.19	-6.68	3.64	0.06	1.49
144	73585	55.3448	Female	1.66	0.18	2.78	2.96	1.39	0.19	-7.12	3.6	0.07	1.63
145	80685	57.4484	Female	1.61	0.31	2.46	2.77	1.38	0.16	-5.46	3.65	0.25	0.51
146	85397	59.2062	Female	1.71	0.19	2.75	2.94	1.42	0.21	-6.98	3.4	0.08	1.82
147	89225	10.7263	Male	1.57	0.29	2.44	2.73	1.36	0.16	-5.7	3.85	0.23	0.68
148	87025	59.7964	Female	1.67	0.19	2.73	2.92	1.41	0.19	-7.12	3.51	0.09	1.65
149	83537	7.97297	Male	1.57	0.29	2.31	2.6	1.36	0.15	-5.73	3.68	0.25	0.83
150	89441	61.8237	Female	1.67	0.2	2.76	2.96	1.4	0.21	-6.98	3.58	0.11	1.55
151	80757	7.83533	Male	1.56	0.37	2.07	2.44	1.34	0.17	-5	3.73	0.43	0.68
152	77917	54.4453	Female	1.53	0.41	1.81	2.22	1.34	0.15	-4.73	3.74	0.62	-0.01
153	82654	57.5691	Female	1.61	0.33	2.2	2.53	1.37	0.19	-5.52	3.84	0.31	0.36
154	82858	54.4267	Female	1.56	0.41	2.13	2.54	1.35	0.17	-4.75	3.8	0.62	0.07
155	75908	10.6703	Male	1.64	0.23	2.79	3.02	1.39	0.18	-6.38	3.63	0.04	1.79
156	78637	59.2357	Female	1.64	0.28	2.71	2.99	1.39	0.16	-5.49	3.43	0.39	0.75
157	86072	7.93238	Male	1.58	0.29	2.47	2.76	1.36	0.17	-5.6	3.9	0.33	0.63
158	84356	55.0147	Female	1.64	0.27	2.39	2.66	1.38	0.2	-5.85	3.77	0.31	0.58
159	92173	56.076	Female	1.7	0.27	2.48	2.75	1.4	0.23	-5.89	3.85	0.25	0.87
160	84599	12.9902	Male	1.59	0.44	2.59	3.03	1.37	0.15	-4.65	3.8	0.37	1.02
161	79310	53.2925	Female	1.62	0.32	1.94	2.26	1.36	0.2	-5.57	3.74	0.35	0.53
162	74343	8.30986	Male	1.62	0.37	2.3	2.67	1.37	0.19	-5.05	3.74	0.42	1.05
163	73902	12.0344	Male	1.61	0.29	2.68	2.97	1.38	0.16	-5.74	3.62	0.32	1.29
164	75836	22.9508	Male	1.68	0.31	2.83	3.14	1.41	0.18	-5.13	3.39	0.53	1.33
165	80855	52.1076	Female	1.65	0.35	1.94	2.29	1.37	0.21	-5.13	3.68	0.46	0.63
166	72668	59.4793	Female	1.63	0.2	2.81	3.01	1.39	0.17	-6.66	3.54	0.16	1.19
167	93763	57.8752	Female	1.71	0.19	2.86	3.05	1.42	0.23	-7.27	3.61	0.14	1.87
168	80992	58.7491	Female	1.69	0.3	2.86	3.16	1.41	0.2	-5.45	3.41	0.41	1.17
169	77854	12.5683	Male	1.62	0.24	2.86	3.1	1.38	0.18	-6.17	3.7	0.15	1.38
170	75426	11.157	Male	1.58	0.29	2.89	3.18	1.37	0.15	-5.58	3.58	0.34	1.09
171	83436	58.2724	Female	1.62	0.29	2.8	3.09	1.39	0.17	-5.28	3.43	0.51	0.5
172	75257	50.8563	Female	1.4	0.28	3.04	3.32	1.33	0.04	-5.8	3.62	0.24	-0.3
173	82191	59.3015	Female	1.66	0.26	2.94	3.2	1.4	0.19	-5.9	3.51	0.27	1.04
174	84619	55.8507	Female	1.66	0.26	2.75	3.01	1.39	0.2	-6.06	3.82	0.13	0.98
175	83702	9.97409	Male	1.59	0.27	2.81	3.08	1.37	0.16	-6.06	3.82	0.16	1.18
176	85075	12.4864	Male	1.65	0.34	2.72	3.06	1.38	0.19	-5.22	3.88	0.33	1.25
177	82686	59.4595	Female	1.7	0.27	2.84	3.11	1.4	0.21	-6.03	3.75	0.17	1.24
178	76249	11.0381	Male	1.66	0.25	3	3.25	1.39	0.2	-6.16	3.79	0.21	1.65

Table 8.3 continued from previous page

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
179	88644	57.7426	Female	1.64	0.29	2.78	3.07	1.38	0.19	-5.91	3.89	0.18	0.81
180	84906	59.3243	Female	1.68	0.44	2.75	3.19	1.39	0.21	-4.52	3.9	0.69	0.72
181	73913	58.7251	Female	1.7	0.36	2.85	3.21	1.42	0.18	-4.65	3.28	0.72	1.06
182	77312	19.2802	Male	1.67	0.23	2.8	3.03	1.41	0.18	-6.17	3.37	0.24	1.7
183	73403	54.9003	Female	1.63	0.25	2.72	2.97	1.38	0.18	-6.02	3.57	0.24	0.93
184	73194	56.5615	Female	1.64	0.28	2.71	2.99	1.39	0.18	-5.7	3.59	0.32	0.81
185	74110	8.62866	Male	1.7	0.25	2.76	3.01	1.4	0.21	-6.07	3.59	0.2	1.84
186	70742	54.8298	Female	1.7	0.27	2.75	3.02	1.41	0.2	-5.72	3.51	0	1.3
187	68517	60.8398	Female	1.66	0.21	2.94	3.15	1.4	0.16	-6.3	3.22	0.23	1.46
188	72807	57.9032	Female	1.62	0.31	2.64	2.95	1.38	0.17	-5.3	3.58	0.4	0.61
189	74199	9.0625	Male	1.65	0.27	2.77	3.04	1.39	0.19	-5.83	3.62	0.33	1.45
190	77238	60.0462	Female	1.73	0.22	2.94	3.16	1.42	0.22	-6.4	3.42	0.22	1.71
191	83999	16.8478	Male	1.66	0.32	2.92	3.24	1.4	0.19	-5.2	3.57	0.47	1.32
192	79210	17.8657	Male	1.69	0.27	2.94	3.21	1.41	0.19	-5.68	3.39	0.35	1.71
193	58058	66.5514	Female	1.69	0.16	2.97	3.13	1.41	0.2	-6.77	3.15	0.22	1.62
194	74851	9.67742	Male	1.64	0.28	2.88	3.16	1.39	0.19	-5.71	3.57	0.28	1.37
195	82620	59.1794	Female	1.68	0.26	2.76	3.02	1.41	0.19	-5.92	3.49	0.28	1.16
196	79358	9.16335	Male	1.66	0.34	2.79	3.13	1.39	0.2	-5.08	3.64	0.39	1.33
197	82300	15.1515	Male	1.66	0.15	2.61	2.76	1.4	0.18	-7.66	3.35	0.04	2.46
198	91060	21.322	Male	1.68	0.41	2.57	2.98	1.41	0.18	-4.44	3.49	0.66	1.23
199	92791	19.1083	Male	1.65	0.18	2.76	2.94	1.39	0.2	-7.17	3.54	0.15	1.9
200	86423	18.5622	Male	1.66	0.16	2.68	2.84	1.39	0.19	-7.73	3.49	0.06	2.43
201	87309	19.7572	Male	1.67	0.19	2.67	2.86	1.4	0.21	-7.07	3.52	0.19	2
202	89411	18.4615	Male	1.69	0.19	2.8	2.99	1.41	0.21	-7.13	3.54	0.18	2.16
203	91229	21.6043	Male	1.64	0.21	2.68	2.89	1.39	0.2	-6.66	3.58	0.18	1.6
204	84551	19.0739	Male	1.64	0.22	2.7	2.92	1.39	0.18	-6.24	3.46	0.3	1.4
205	89893	17.6471	Male	1.67	0.19	2.62	2.81	1.4	0.2	-7	3.46	0.17	2.01
206	82254	8.06878	Male	1.6	0.15	2.66	2.81	1.36	0.2	-8.15	3.77	0.08	2.25
207	80600	8.64714	Male	1.59	0.14	2.63	2.77	1.36	0.18	-8.18	3.67	0.08	2.26
208	91802	61.025	Female	1.74	0.28	2.82	3.1	1.42	0.24	-5.67	3.6	0.36	1.14
209	82706	20.5882	Male	1.62	0.26	2.67	2.93	1.38	0.16	-5.82	3.43	0.3	1.18
210	61702	5.98131	Male	1.59	0.17	2.69	2.86	1.38	0.13	-7.03	3.36	0.1	1.87
211	67358	61.165	Female	1.69	0.19	2.78	2.97	1.41	0.2	-6.64	3.44	0.1	1.53
212	65894	62.5988	Female	1.64	0.31	2.76	3.07	1.39	0.18	-5.24	3.45	0.53	0.71
213	70647	63.7125	Female	1.78	0.24	2.82	3.06	1.45	0.19	-5.99	3.08	0.13	2.38
214	80304	19.1964	Male	1.81	0.33	3.1	3.43	1.46	0.22	-5.47	3.24	0.11	2.9

Table 8.3 continued from previous page

SAMPLE	NO VAR	PC X HETS	GENDER	HET:HOM	DEV A	DEV B	DEV MET	SKEW ALL	KURT.ALL	SKEW HOMS	KURT. HETS	VBAMID	OLS
215	116939	22.7273	Male	1.75	0.4	2.81	3.21	1.43	0.23	-5.35	3.74	0.01	2.21
216	123054	22.9446	Male	1.76	0.39	2.89	3.28	1.43	0.24	-5.36	3.83	0.13	2.03
217	108880	27.6087	Male	1.72	1.38	2.69	4.07	1.42	0.18	-2.04	3.91	4.11	5.53
218	81420	22.4674	Male	1.72	0.34	2.83	3.17	1.43	0.17	-5.67	3.29	0	2.53
219	87849	61.9735	Female	1.82	0.34	2.93	3.27	1.46	0.23	-5.57	3.37	0.05	2.33
220	77653	22.8132	Male	1.76	0.33	3.14	3.47	1.44	0.19	-5.48	3.27	0.04	2.72
221	75375	25.0557	Male	1.77	0.35	3.02	3.37	1.45	0.19	-5.45	3.21	0.04	2.82
222	69904	62.0791	Female	1.74	0.44	2.9	3.34	1.43	0.17	-4.31	3.1	0.52	1.91
223	70726	58.9313	Female	1.8	0.35	3.02	3.37	1.46	0.19	-5.28	3.09	0.18	2.62
224	78840	61.9485	Female	1.74	0.3	3.05	3.35	1.44	0.18	-5.68	3.25	0.04	2.02
225	71468	20.6107	Male	1.74	0.52	2.91	3.43	1.44	0.18	-4	3.2	0.8	2.51
226	78884	19.8366	Male	1.7	0.39	2.89	3.28	1.42	0.16	-5.08	3.23	0.18	2.3
227	77112	23.5434	Male	1.78	0.33	3.1	3.43	1.45	0.2	-5.35	3.25	0.12	2.65
228	77417	62.7119	Female	1.79	0.29	3.03	3.32	1.45	0.21	-5.79	3.22	0.03	2.35
229	92040	21.7025	Male	1.74	0.39	2.85	3.24	1.43	0.19	-5.11	3.36	0.07	2.37
230	76782	25.2998	Male	2	0.37	2.8	3.17	1.53	0.29	-5.22	3.2	0.05	3.96
231	75308	22.6107	Male	1.8	0.34	2.84	3.18	1.46	0.2	-5.52	3.13	0.02	3.03
232	69049	22.7799	Male	1.73	0.32	3.04	3.36	1.43	0.16	-5.52	3.09	0.04	2.7
233	90634	20.9632	Male	1.78	0.39	2.79	3.18	1.44	0.21	-5.11	3.35	0.06	2.66
234	91076	61.2007	Female	1.78	0.31	2.9	3.21	1.44	0.22	-5.8	3.39	0.06	2.15
235	83460	60.6186	Female	1.74	0.16	2.7	2.86	1.43	0.23	-7.59	3.36	0.14	2.3
236	93989	19.3277	Male	1.67	0.19	2.85	3.04	1.39	0.22	-7.29	3.71	0.13	2.11
237	89231	60.0249	Female	1.68	0.17	2.81	2.98	1.4	0.21	-7.97	3.58	0.04	2.16
238	85182	16.8715	Male	1.63	0.16	2.66	2.82	1.38	0.18	-7.52	3.49	0.06	2.11
239	84443	17.7945	Male	1.62	0.2	2.78	2.98	1.38	0.17	-6.81	3.44	0.24	1.68
240	88445	17.3963	Male	1.65	0.16	2.8	2.96	1.39	0.19	-8.26	3.51	0.07	2.79
241	78941	18.3575	Male	1.63	0.22	2.75	2.97	1.39	0.16	-6.36	3.41	0.16	1.64
242	83991	57.1329	Female	1.66	0.21	2.81	3.02	1.4	0.19	-6.37	3.47	0.24	1.16
243	81775	60.1915	Female	1.7	0.17	2.76	2.93	1.41	0.21	-7.2	3.44	0.14	1.88
244	87402	16.8038	Male	1.7	0.18	2.72	2.9	1.41	0.21	-7.1	3.55	0.12	2.22
245	86502	19.5006	Male	1.68	0.18	2.7	2.88	1.41	0.2	-7.12	3.41	0.13	2.18

Table 8.3: Contamination predictions from OLS using 1-10% contamination training ranges and VerifyBamID for all 245 germline samples. All 9 features supplied into the regression for each sample are also listed.

8.2 Appendix B

8.2.1 Unmapped reads per sample

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED
PR0001	IBD	73068070	171	0.00023
PR0002	IBD	41633989	723	0.00174
PR0003	IBD	45394374	355	0.00078
PR0005	IBD	72174399	835	0.00116
PR0007	IBD	61447926	722	0.00117
PR0008	IBD	61677733	972	0.00158
PR0009	IBD	46347395	365	0.00079
PR0010	IBD	45145751	489	0.00108
PR0011	IBD	69945169	1327	0.0019
PR0012	IBD	61129022	715	0.00117
PR0014	IBD	44792748	555	0.00124
PR0015	IBD	44832026	600	0.00134
PR0018	IBD	68394731	1222	0.00179
PR0020	IBD	74648283	13563	0.01817
PR0021	IBD	41042809	609	0.00148
PR0022	IBD	74697660	128	0.00017
PR0023	IBD	62326825	224	0.00036
PR0025	IBD	49100329	659	0.00134
PR0026	IBD	51314457	541	0.00105
PR0027	IBD	47700087	645	0.00135
PR0028	IBD	44406560	1132	0.00255
PR0030	IBD	48245721	577	0.0012
PR0032	IBD	62363646	774	0.00124
PR0034	IBD	44799673	367	0.00082
PR0035	IBD	59033074	576	0.00098
PR0036	IBD	43753109	498	0.00114
PR0040	IBD	66529699	790	0.00119
PR0041	IBD	57549832	1127	0.00196
PR0043	IBD	47225631	1461	0.00309
PR0044	IBD	47689308	735	0.00154
PR0045	IBD	63114707	0	0
PR0046	IBD	46378065	1255	0.00271
PR0047	IBD	76851424	1146	0.00149
PR0048	IBD	57721128	170	0.00029
PR0049	IBD	47428682	121511	0.2562
PR0051	IBD	59148429	776	0.00131
PR0052	IBD	39877436	61	0.00015
PR0053	IBD	56591796	730	0.00129
PR0054	IBD	63318611	282	0.00045
PR0055	IBD	66965655	476	0.00071
PR0056	IBD	70209069	965	0.00137
PR0058	IBD	68407578	964	0.00141
PR0059	IBD	43127223	392	0.00091
PR0060	IBD	94006341	603	0.00064
PR0061	IBD	80862246	1223	0.00151
PR0062	IBD	48580969	1419	0.00292
PR0063	IBD	81540676	1674	0.00205
PR0064	IBD	86487618	37579	0.04345
PR0066	IBD	83719418	596	0.00071
PR0067	IBD	54103465	1551	0.00287
PR0068	IBD	79637773	558	0.0007
PR0069	IBD	51251335	603	0.00118
PR0070	IBD	38619292	373	0.00097
PR0071	IBD	87612013	428	0.00049
PR0074	IBD	73035730	130	0.00018
PR0075	IBD	43917090	738	0.00168

Table 8.4 continued from previous page

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED
PR0076	IBD	63007586	879	0.0014
PR0077	IBD	71202878	51149	0.07184
PR0079	IBD	49515885	634	0.00128
PR0080	IBD	50973560	414	0.00081
PR0081	IBD	59413081	557	0.00094
PR0082	IBD	57814718	674	0.00117
PR0083	IBD	88224099	601	0.00068
PR0084	IBD	59813794	1094	0.00183
PR0085	IBD	58663168	1199	0.00204
PR0086	IBD	62677905	886	0.00141
PR0087	IBD	57438805	769	0.00134
PR0089	IBD	54462333	821	0.00151
PR0091	IBD	49932407	392	0.00079
PR0092	IBD	77303999	1	0
PR0095	IBD	55187271	701	0.00127
PR0096	IBD	44783924	656	0.00146
PR0097	IBD	64527032	1007	0.00156
PR0098	IBD	54934620	882	0.00161
PR0099	IBD	60779501	182	0.0003
PR0100	IBD	49115069	627	0.00128
PR0102	IBD	64419715	694	0.00108
PR0103	IBD	68231002	927	0.00136
PR0104	IBD	73225207	2946	0.00402
PR0105	IBD	48270301	868	0.0018
PR0106	IBD	59064342	995	0.00168
PR0107	IBD	58214468	854	0.00147
PR0108	IBD	47495178	537	0.00113
PR0109	IBD	70224318	1319	0.00188
PR0110	IBD	65675670	856	0.0013
PR0111	IBD	78206442	812	0.00104
PR0112	IBD	77466010	2077	0.00268
PR0113	IBD	56113286	178	0.00032
PR0114	IBD	59221799	629	0.00106
PR0115	IBD	46592453	459	0.00099
PR0116	IBD	78011904	248	0.00032
PR0117	IBD	54024274	843	0.00156
PR0118	IBD	43408572	444	0.00102
PR0120	IBD	87449328	290	0.00033
PR0121	IBD	75839805	409	0.00054
PR0122	IBD	51407750	651	0.00127
PR0123	IBD	54777869	806	0.00147
PR0124	IBD	58923251	1818	0.00309
PR0125	IBD	61881083	548	0.00089
PR0126	IBD	50033355	614	0.00123
PR0128	IBD	51747786	777	0.0015
PR0129	IBD	51319152	403	0.00079
PR0130	IBD	70086125	371	0.00053
PR0131	IBD	49849811	556	0.00112
PR0132	IBD	95714113	1178	0.00123
PR0134	IBD	42110909	649	0.00154
PR0136	IBD	67722804	293	0.00043
PR0137	IBD	54614687	575	0.00105
PR0138	IBD	44682386	576	0.00129
PR0141	IBD	53350966	592	0.00111
PR0142	IBD	67917348	6521	0.0096
PR0144	IBD	46787024	31	0.00007
PR0145	IBD	78274941	781	0.001
PR0146	IBD	75299186	1394	0.00185
PR0149	IBD	72241705	943	0.00131
PR0150	IBD	73302362	935	0.00128
PR0151	IBD	87520116	1482	0.00169

Table 8.4 continued from previous page

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED
PR0153	IBD	77611820	742	0.00096
PR0155	IBD	41784553	268	0.00064
PR0156	IBD	46768958	316	0.00068
PR0158	IBD	51889105	702	0.00135
PR0159	IBD	53717391	299343	0.55726
PR0160	IBD	54748766	513	0.00094
PR0161	IBD	50770817	102345	0.20158
PR0165	IBD	53779930	696	0.00129
PR0166	IBD	41347486	5611	0.01357
PR0171	IBD	53417445	898	0.00168
PR0172	IBD	45949708	381	0.00083
PR0173	IBD	49160306	648	0.00132
PR0174	IBD	58917364	615	0.00104
PR0176	IBD	42625531	672	0.00158
PR0177	IBD	46195932	414	0.0009
PR0178	IBD	49676548	1836	0.0037
PR0179	IBD	40589345	533	0.00131
PR0180	IBD	49872804	847	0.0017
PR0182	IBD	44612713	284	0.00064
PR0186	IBD	54703478	511	0.00093
PR0187	IBD	51811987	458	0.00088
PR0188	IBD	62415886	494	0.00079
PR0191	IBD	46651750	93	0.0002
PR0192	IBD	57335120	813	0.00142
PR0193	IBD	56701247	585	0.00103
PR0194	IBD	52343337	1364	0.00261
PR0195	IBD	53310598	361	0.00068
PR0196	IBD	51067465	67	0.00013
PR0198	IBD	61660549	717	0.00116
PR0199	IBD	50728947	596	0.00117
PR0203	IBD	52316970	714	0.00136
PR0204	IBD	55984838	1593	0.00285
PR0206	IBD	42704082	5100	0.01194
PR0207	IBD	59143436	842	0.00142
PR0208	IBD	47683299	453	0.00095
PR0210	IBD	51981608	67	0.00013
PR0213	IBD	44305074	541	0.00122
PR0214	IBD	43360488	392	0.0009
PR0215	IBD	50528436	537	0.00106
PR0218	IBD	47774676	134	0.00028
PR0219	IBD	52101375	4492	0.00862
PR0220	IBD	41169805	131	0.00032
PR0221	IBD	62019551	50	0.00008
PR0241	IBD	70949103	669	0.00094
PR0260	IBD	46602995	1603	0.00344
SOPR0222	IBD	41354046	793	0.00192
SOPR0223	IBD	55340581	959	0.00173
SOPR0224	IBD	49429483	1015	0.00205
SOPR0226	IBD	56053371	1688	0.00301
SOPR0228	IBD	82664958	931	0.00113
SOPR0229	IBD	50977256	714	0.0014
SOPR0230	IBD	44583015	10448	0.02343
SOPR0232	IBD	53837673	1751	0.00325
SOPR0234	IBD	47122587	706	0.0015
SOPR0236	IBD	44722133	3500	0.00783
SOPR0237	IBD	41844549	744	0.00178
SOPR0238	IBD	74463590	235	0.00032
SOPR0243	IBD	54677500	399	0.00073
SOPR0244	IBD	49118705	520	0.00106
SOPR0245	IBD	49463348	609	0.00123
SOPR0246	IBD	50748749	602	0.00119

Table 8.4 continued from previous page

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED
SOPR0247	IBD	42052814	161	0.00038
SOPR0249	IBD	57172686	385	0.00067
SOPR0251	IBD	64255786	11169	0.01738
SOPR0253	IBD	41358944	610	0.00147
SOPR0254	IBD	48056682	606	0.00126
SOPR0255	IBD	86065407	151762	0.17633
SOPR0256	IBD	44243844	129	0.00029
SOPR0258	IBD	57243066	179	0.00031
SOPR0259	IBD	50082554	1761	0.00352
SOPR0261	IBD	48179574	680	0.00141
SOPR0267	IBD	44180432	774	0.00175
SOPR0268	IBD	49248967	103	0.00021
SOPR0270	IBD	56107702	1118	0.00199
SOPR0273	IBD	40561918	122	0.0003
SOPR0274	IBD	44786733	49	0.00011
SOPR0275	IBD	45640248	82	0.00018
SOPR0276	IBD	49154485	831	0.00169
SOPR0278	IBD	54947065	703	0.00128
SOPR0279	IBD	41292369	3698	0.00896
SOPR0283	IBD	45756342	626	0.00137
SOPR0284	IBD	44261711	435	0.00098
SOPR0285	IBD	42887417	330	0.00077
SOPR0286	IBD	53886983	82	0.00015
SOPR0287	IBD	45363102	694	0.00153
SOPR0289	IBD	53160290	930	0.00175
SOPR0290	IBD	43190093	757	0.00175
SOPR0294	IBD	42897151	42	0.0001
SOPR0301	IBD	51567096	980	0.0019
SOPR0303	IBD	57449548	197	0.00034
SOPR0306	IBD	45345634	1077	0.00238
SOPR0307	IBD	52920903	1374	0.0026
SOPR0309	IBD	65592208	1982	0.00302
SOPR0310	IBD	48769525	131	0.00027
SOPR0312	IBD	61748602	1006	0.00163
SOPR0313	IBD	59858582	183	0.00031
SOPR0314	IBD	56416935	190	0.00034
SOPR0316	IBD	47304610	1129	0.00239
SOPR0317	IBD	48637996	250	0.00051
SOPR0318	IBD	55894788	10471	0.01873
SOPR0320	IBD	52024278	131	0.00025
SOPR0323	IBD	60420992	832	0.00138
SOPR0325	IBD	40434565	719	0.00178
SOPR0326	IBD	51374396	688	0.00134
SOPR0327	IBD	50377953	69	0.00014
SOPR0333	IBD	88095599	23438	0.02661
SOPR0334	IBD	48978781	897	0.00183
SOPR0336	IBD	56252683	156	0.00028
SOPR0339	IBD	60607867	70	0.00012
SOPR0340	IBD	64959659	144	0.00022
SOPR0342	IBD	41342549	33	0.00008
SOPR0344	IBD	41431932	719	0.00174
SOPR0345	IBD	47847753	686	0.00143
SOPR0346	IBD	43148491	334	0.00077
SOPR0348	IBD	51948528	894	0.00172
SOPR0349	IBD	47356903	928	0.00196
SOPR0351	IBD	57708223	124	0.00021
SOPR0352	IBD	46787938	963	0.00206
SOPR0355	IBD	40929032	27	0.00007
SOPR0357	IBD	48057853	758	0.00158
SOPR0359	IBD	52204908	543	0.00104
SOPR0360	IBD	38989220	474	0.00122

Table 8.4 continued from previous page

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED
SOPR0361	IBD	51822178	793	0.00153
SOPR0362	IBD	51152807	1653	0.00323
SOPR0367	IBD	51493235	1158	0.00225
SOPR0368	IBD	52348002	991	0.00189
SOPR0370	IBD	50765993	741	0.00146
SOPR0372	IBD	48865902	1011	0.00207
0202378LB	CTRLS	46614281	545	0.00117
1114392LH	CTRLS	45992843	566	0.00123
13-E029-W1507179	CTRLS	84350605	1233	0.00146
13-E030-W1507180	CTRLS	86849919	1195	0.00138
1-5-18190	CTRLS	55457119	4123	0.00743
1-5-39217	CTRLS	61878965	72388	0.11698
3G	CTRLS	147709568	7501	0.00508
43 PD7584b	CTRLS	143037589	758	0.00053
43 PD7585b	CTRLS	136063838	734	0.00054
43 PD7587b	CTRLS	130399489	1131	0.00087
43 PD7597b	CTRLS	185905772	21437	0.01153
43 PD7598b	CTRLS	129960511	701	0.00054
5G	CTRLS	137733365	20932	0.0152
9G	CTRLS	130612598	7165	0.00549
ABW1514334	CTRLS	78820001	4143	0.00526
AK28A	CTRLS	82371091	1218	0.00148
AK28B	CTRLS	53271456	3195	0.006
AK30A	CTRLS	68494304	1109	0.00162
AK30B	CTRLS	43492458	1116	0.00257
AK51A	CTRLS	74027904	802	0.00108
AK51B	CTRLS	63081874	1488	0.00236
AK54A	CTRLS	86038768	1390	0.00162
AK54B	CTRLS	52945755	596	0.00113
AK68A	CTRLS	91387832	734	0.0008
AK68B	CTRLS	45278296	1220	0.00269
B530001	CTRLS	99815565	15524	0.01555
BAZEX0001	CTRLS	50042216	555	0.00111
CL003-3246m	CTRLS	43525898	491	0.00113
CL006-KP	CTRLS	46517223	592	0.00127
CL018-3868-Cous2	CTRLS	102767519	506	0.00049
DW001	CTRLS	59472882	546	0.00092
DW002	CTRLS	75782475	1244	0.00164
DWMB001	CTRLS	67690568	707	0.00104
DWSHFLD1	CTRLS	61691239	1120	0.00182
E7055	CTRLS	93428950	1418	0.00152
IKTCC01	CTRLS	51161029	705	0.00138
IKTCC02	CTRLS	47932827	484	0.00101
JB-W1511944	CTRLS	81805961	1073	0.00131
JL-W1511514	CTRLS	76210993	882	0.00116
JLW1511693	CTRLS	71506849	21452	0.03
KD003	CTRLS	60849114	727	0.00119
MC0001	CTRLS	46687835	435	0.00093
MCAD002	CTRLS	44279019	1852	0.00418
MCAD003	CTRLS	47848714	2187	0.00457
NEPH007	CTRLS	51455947	486	0.00094
NEPH008	CTRLS	56008309	425	0.00076
NF PCA 02	CTRLS	48010857	181	0.00038
NF PsDys003	CTRLS	45180269	265	0.00059
NG149-2	CTRLS	45598110	109998	0.24123
NG151-2	CTRLS	48758804	23713	0.04863
NG156	CTRLS	85806487	3086	0.0036
NG-PE	CTRLS	43354444	1248	0.00288
NRS	CTRLS	46970966	149399	0.31807
PD-W1204043	CTRLS	94626648	1238	0.00131
PID0016A	CTRLS	49562051	428	0.00086

Table 8.4 continued from previous page

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED
PID0020A	CTRLS	55912247	754	0.00135
PID0024A	CTRLS	46825845	545	0.00116
PID005PR	CTRLS	71510296	711	0.00099
PID008SM	CTRLS	63509463	648	0.00102
PID039A	CTRLS	41319558	542	0.00131
PID040A	CTRLS	42206229	503	0.00119
PID045A	CTRLS	46799259	611	0.00131
PMP220G	CTRLS	42832247	487	0.00114
PT001	CTRLS	52213465	610	0.00117
PT002	CTRLS	48089354	421	0.00088
PT003	CTRLS	53587621	572	0.00107
PT006	CTRLS	52758529	697	0.00132
PT007	CTRLS	50474169	616	0.00122
PT008	CTRLS	56046798	529	0.00094
PT009	CTRLS	59319432	14088	0.02375
PT010	CTRLS	53106843	20232	0.0381
RC091	CTRLS	65632815	186	0.00028
RC142	CTRLS	61165164	430	0.0007
RNAR1	CTRLS	46064009	78264	0.1699
SD001	CTRLS	50711354	582	0.00115
SD002	CTRLS	44129566	450	0.00102
Self-Colob001	CTRLS	52918059	577	0.00109
VFW1515559	CTRLS	76491281	2850	0.00373
VW-15-1	CTRLS	54819471	447	0.00082
VW-15-71	CTRLS	43790483	292	0.00067
VW-7-1	CTRLS	46113239	414	0.0009
VW-7-23	CTRLS	50807769	490	0.00096
VW-7-30	CTRLS	53407168	489	0.00092
W0202380NB	CTRLS	49242082	547	0.00111
W0203178EM	CTRLS	44831249	452	0.00101
W0701628JM	CTRLS	53644248	454	0.00085
W0802815 AS	CTRLS	46272909	500	0.00108
W0909021 RM	CTRLS	66150395	467	0.00071
W1003129JI	CTRLS	52283384	700	0.00134
W10054516JL	CTRLS	40721049	559	0.00137
W1006127 LB	CTRLS	56190277	592	0.00105
W1008996 WA	CTRLS	49420420	810	0.00164
W1102779ME	CTRLS	44245930	495	0.00112
W1103547 JD	CTRLS	45943136	478	0.00104
W1104401SF	CTRLS	43892519	527	0.0012
W1107214LH	CTRLS	44372264	615	0.00139
W1200013 MK	CTRLS	52099247	593	0.00114
W1200014PR	CTRLS	53344499	440	0.00082
W1202920	CTRLS	48495975	553	0.00114
W1204042CD	CTRLS	49591323	473	0.00095
W1209378CH	CTRLS	47556490	526	0.00111
W12132810W	CTRLS	45616071	657	0.00144
W1214012JW	CTRLS	46326350	462	0.001
W1215091 MH	CTRLS	48701796	773	0.00159
W1305468 NL	CTRLS	45156801	577	0.00128
W1308445SD	CTRLS	53781002	611	0.00114
W1312876MW	CTRLS	49774424	609	0.00122
W1400858MS	CTRLS	48996343	536	0.00109
W1403112KR	CTRLS	51597602	502	0.00097
W1416409LA	CTRLS	55925988	28469	0.0509
W1502963LR	CTRLS	41764764	415	0.00099
Wexcon001	CTRLS	68216768	1084	0.00159
ZB-W1505055	CTRLS	94099936	1394	0.00148

Table 8.4 continued from previous page

SAMPLE	GROUP	NO. READS	NO. UNMAPPED	% UNMAPPED

Table 8.4: Number of reads in BAM files per samples and the number of high quality reads passing filtering for all 358 exome samples. 2 groups are shown in the table: SOTON CTRLS and SOTON IDB.

8.2.2 Unmapped reads by kingdom or domain

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
0202378LB	CONTROL	14	1	0	0	1	500	0	12226	Unknown
1114392LH	CONTROL	15	3	0	0	0	508	3	10863	Unknown
13-E029-W1507179	CONTROL	19	1	0	0	0	1120	5	2340	Unknown
13-E030-W1507180	CONTROL	23	7	0	0	0	1052	4	3240	Unknown
1-5-18190	CONTROL	1533	1	0	0	1	52	0	8547	Unknown
1-5-39217	CONTROL	33122	4	0	0	17	407	7	78716	Unknown
3G	CONTROL	2309	0	0	0	6	1288	11	32009	Unknown
43 PD7584b	CONTROL	0	0	0	0	3	629	7	758	Unknown
43 PD7585b	CONTROL	0	0	0	0	1	573	1	734	Unknown
43 PD7587b	CONTROL	22	2	0	0	1	958	6	1131	Unknown
43 PD7597b	CONTROL	2323	29	8	0	9	1655	22	21437	Unknown
43 PD7598b	CONTROL	1	0	0	0	0	567	8	701	Unknown
5G	CONTROL	6612	7	0	0	6	1234	8	55335	Unknown
9G	CONTROL	2294	2	0	0	0	1195	12	43992	Unknown
ABW1514334	CONTROL	916	10	0	0	15	2034	90	8977	Unknown
AK28A	CONTROL	700	0	0	0	12	354	0	35165	Unknown
AK28B	CONTROL	114	0	0	0	9	2971	0	14775	Unknown
AK30A	CONTROL	555	0	0	0	11	423	0	30041	Unknown
AK30B	CONTROL	467	0	0	0	11	418	0	11029	Unknown
AK51A	CONTROL	367	0	0	0	10	301	0	40634	Unknown
AK51B	CONTROL	328	0	2	0	17	579	0	14202	Unknown
AK54A	CONTROL	515	0	0	0	13	696	0	31217	Unknown
AK54B	CONTROL	186	1	0	0	1	370	2	14193	Unknown
AK68A	CONTROL	193	0	0	0	6	409	0	33116	Unknown
AK68B	CONTROL	511	0	0	0	7	361	1	19777	Unknown
B530001	CONTROL	6530	3	3	0	2	135	1	15692	Unknown
BAZEX0001	CONTROL	53	29	0	0	0	416	7	20125	Unknown
CL003-3246m	CONTROL	32	1	0	0	0	438	1	17906	Unknown
CL006-KP	CONTROL	60	0	0	0	0	432	4	23540	Unknown
CL018-3868-Cous2	CONTROL	48	87	0	0	3	317	2	29428	Unknown
DW001	CONTROL	45	0	0	0	1	399	4	5295	Unknown
DW002	CONTROL	283	0	0	0	4	544	2	9316	Unknown
DWMB001	CONTROL	17	0	0	0	1	640	1	707	Unknown
DWSHFLD1	CONTROL	60	0	235	0	470	35	0	1120	Unknown
E7055	CONTROL	78	2	0	0	0	1177	11	9690	Unknown

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
IKTCC01	CONTROL	29	4	0	0	1	591	6	16043	Unknown
IKTCC02	CONTROL	29	2	0	0	1	427	0	18867	Unknown
JB-W1511944	CONTROL	6	33	0	0	2	947	3	3562	Unknown
JL-W1511514	CONTROL	21	2	0	0	1	722	4	1690	Unknown
JLW1511693	CONTROL	4434	5	1	0	6	928	4	22557	Unknown
KD003	CONTROL	244	3	0	0	71	379	7	11478	Unknown
MC0001	CONTROL	22	0	0	0	0	372	3	13603	Unknown
MCAD002	CONTROL	990	2	0	0	1	286	1	3520	Unknown
MCAD003	CONTROL	1185	0	0	0	3	319	3	4332	Unknown
NEPH007	CONTROL	1	0	0	0	0	453	0	2215	Unknown
NEPH008	CONTROL	3	1	0	0	1	362	2	2801	Unknown
NF PCA 02	CONTROL	35	2	0	0	0	131	2	15235	Unknown
NF PsDys003	CONTROL	39	3	0	0	0	212	0	19217	Unknown
NG149-2	CONTROL	29542	10	0	0	35	481	4	124115	Unknown
NG151-2	CONTROL	7304	31	0	1	12	300	2	32708	Unknown
NG156	CONTROL	751	1	0	0	6	871	0	26110	Unknown
NG-PE	CONTROL	331	3	0	0	2	352	4	12024	Unknown
NRS	CONTROL	67711	42	0	0	10	443	0	173269	Unknown
PD-W1204043	CONTROL	12	9	0	0	0	1124	7	2711	Unknown
PID0016A	CONTROL	40	0	0	0	0	330	5	5441	Unknown
PID0020A	CONTROL	36	2	0	0	2	658	4	6414	Unknown
PID0024A	CONTROL	48	1	0	0	0	423	5	7736	Unknown
PID005PR	CONTROL	98	4	0	0	0	518	2	9616	Unknown
PID008SM	CONTROL	100	1	0	0	1	453	2	8350	Unknown
PID039A	CONTROL	27	2	0	0	2	444	1	14618	Unknown
PID040A	CONTROL	26	2	0	0	1	455	2	13705	Unknown
PID045A	CONTROL	28	0	1	0	1	535	4	19590	Unknown
PMP220G	CONTROL	19	0	1	0	0	445	2	14767	Unknown
PT001	CONTROL	34	2	0	0	1	543	2	13827	Unknown
PT002	CONTROL	37	0	0	0	1	362	4	16101	Unknown
PT003	CONTROL	77	0	0	0	7	374	4	15828	Unknown
PT006	CONTROL	13	245	0	0	0	410	5	8696	Unknown
PT007	CONTROL	22	1	0	0	2	573	3	13715	Unknown
PT008	CONTROL	23	9	1	0	0	466	1	18838	Unknown
PR0011	IBD	62	0	0	0	0	1016	4	4746	Blood
PR0018	IBD	611	3	0	0	157	406	8	12396	Blood

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
RC091	CONTROL	52	0	0	0	1	125	0	10861	Unknown
RC142	CONTROL	23	0	0	0	0	387	5	6777	Unknown
PR0020	IBD	5	0	0	0	11241	1198	0	13563	Blood
SD001	CONTROL	43	1	1	0	0	481	5	15793	Unknown
SD002	CONTROL	38	0	0	0	0	376	5	13750	Unknown
Self-Colob001	CONTROL	30	14	0	0	0	501	4	20218	Unknown
PR0028	IBD	0	0	0	0	1037	56	0	1132	Blood
VW-15-1	CONTROL	9	0	0	0	3	323	2	16235	Unknown
VW-15-71	CONTROL	8	1	0	0	0	232	3	14669	Unknown
VW-7-1	CONTROL	8	7	0	0	1	328	2	14770	Unknown
VW-7-23	CONTROL	22	1	0	0	3	354	4	18662	Unknown
VW-7-30	CONTROL	6	5	0	0	1	386	8	14031	Unknown
W0202380NB	CONTROL	7	3	0	0	0	504	2	17001	Unknown
W0203178EM	CONTROL	3	0	0	0	0	424	0	1721	Unknown
W0701628JM	CONTROL	17	0	0	0	1	392	4	2208	Unknown
W0802815 AS	CONTROL	24	0	0	0	0	451	1	15370	Unknown
W0909021 RM	CONTROL	35	9	0	0	1	343	2	17026	Unknown
W1003129JI	CONTROL	14	1	0	0	1	570	9	10258	Unknown
W10054516JL	CONTROL	24	0	0	0	1	469	6	10304	Unknown
W1006127 LB	CONTROL	19	14	0	0	2	526	5	13004	Unknown
W1008996 WA	CONTROL	18	11	0	0	1	670	6	12244	Unknown
W1102779ME	CONTROL	8	9	0	0	0	445	6	1865	Unknown
W1103547 JD	CONTROL	30	1	0	0	0	407	0	22223	Unknown
W1104401SF	CONTROL	58	1	0	0	0	427	2	16203	Unknown
W1107214LH	CONTROL	41	0	0	0	4	513	3	12356	Unknown
W1200013 MK	CONTROL	24	15	0	0	0	501	2	20705	Unknown
W1200014PR	CONTROL	2	1	0	0	0	410	2	13696	Unknown
W1202920	CONTROL	33	0	0	0	1	474	2	13775	Unknown
W1204042CD	CONTROL	1	1	0	0	2	396	4	3806	Unknown
W1209378CH	CONTROL	3	2	0	0	0	484	4	24788	Unknown
W12132810W	CONTROL	128	0	0	0	3	420	1	15065	Unknown
W1214012JW	CONTROL	26	16	0	0	2	388	3	14426	Unknown
W1215091 MH	CONTROL	27	30	0	0	0	665	5	12524	Unknown
W1305468 NL	CONTROL	34	5	0	0	0	504	0	12846	Unknown
W1308445SD	CONTROL	17	8	0	0	0	545	2	2048	Unknown
W1312876MW	CONTROL	26	1	1	0	0	523	2	16140	Unknown

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
W1400858MS	CONTROL	26	0	0	0	2	453	4	14308	Unknown
W1403112KR	CONTROL	8	4	0	0	1	449	2	12743	Unknown
PR0041	IBD	0	0	0	0	607	42	0	1127	Blood
W1502963LR	CONTROL	16	8	0	0	0	371	1	11717	Unknown
PR0043	IBD	191	0	0	0	509	34	0	1461	Blood
PR0046	IBD	124	0	0	0	553	61	0	1255	Blood
PR0001	IBD	38	0	0	0	0	115	0	4351	Blood
PR0002	IBD	1	0	0	0	643	52	0	723	Blood
PR0003	IBD	0	0	0	0	269	35	0	355	Blood
PR0005	IBD	110	1	0	0	3	563	6	7613	Blood
PR0007	IBD	101	1	0	0	4	501	2	11318	Blood
PR0008	IBD	77	0	0	0	0	711	1	10573	Blood
PR0009	IBD	42	0	0	0	0	288	3	19907	Blood
PR0010	IBD	49	0	0	0	0	420	3	6049	Blood
PR0047	IBD	307	1	0	0	78	647	4	4703	Blood
PR0012	IBD	247	5	0	0	70	328	1	8263	Blood
PR0014	IBD	69	0	0	0	0	467	0	9875	Blood
PR0015	IBD	74	0	0	0	0	504	0	10512	Blood
PR0049	IBD	35372	9	4	6	88	287	0	121511	Saliva
PR0061	IBD	428	0	0	0	95	571	3	12264	Blood
PR0021	IBD	73	0	0	0	0	513	2	3999	Blood
PR0022	IBD	29	1	0	0	0	82	0	3082	Blood
PR0023	IBD	13	0	0	0	0	181	2	1443	Blood
PR0025	IBD	80	0	0	0	0	550	8	10186	Blood
PR0026	IBD	45	0	0	0	2	462	3	22344	Blood
PR0027	IBD	64	1	0	0	1	514	10	6285	Blood
PR0062	IBD	125	0	0	0	699	35	0	1419	Blood
PR0030	IBD	87	1	0	0	0	476	3	7260	Blood
PR0032	IBD	127	2	0	0	5	491	4	7911	Blood
PR0034	IBD	41	0	0	0	2	285	0	20382	Blood
PR0035	IBD	77	1	0	0	1	410	3	9842	Blood
PR0036	IBD	30	0	0	0	0	384	2	498	Blood
PR0040	IBD	86	0	0	0	1	577	2	9416	Blood
PR0063	IBD	392	2	0	0	74	900	7	12597	Blood
PR0064	IBD	13086	9	0	0	5	459	4	51240	Saliva
PR0044	IBD	0	2	0	0	619	23	1	735	Blood

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
PR0045	IBD	0	0	0	0	0	0	0	1	Blood
PR0067	IBD	2	0	1	0	767	82	0	1551	Blood
PR0077	IBD	17118	21	3	0	36	351	0	67237	Saliva
PR0048	IBD	11	0	0	0	0	125	1	1552	Blood
PR0084	IBD	68	2	0	0	2	783	1	7697	Blood
PR0051	IBD	106	0	0	0	1	528	2	7329	Blood
PR0052	IBD	5	0	0	0	0	51	0	61	Blood
PR0053	IBD	5	0	139	0	349	15	0	730	Blood
PR0054	IBD	13	0	0	0	1	250	1	6658	Blood
PR0055	IBD	44	0	0	0	0	366	3	13536	Blood
PR0056	IBD	102	1	0	0	3	632	3	10946	Blood
PR0058	IBD	89	0	1	0	1	689	4	9254	Blood
PR0059	IBD	32	0	0	0	2	341	0	18167	Blood
PR0060	IBD	44	1	0	0	1	506	6	12849	Blood
PR0085	IBD	63	1	0	0	2	889	4	11639	Blood
PR0097	IBD	265	2	0	0	53	548	0	5290	Blood
PR0104	IBD	2786	5	0	0	0	92	0	2946	Blood
PR0109	IBD	82	0	0	0	1	964	6	6133	Blood
PR0066	IBD	47	0	0	0	1	500	2	16443	Blood
PR0112	IBD	739	3	0	0	159	763	10	5349	Blood
PR0068	IBD	36	0	0	0	1	481	1	12913	Blood
PR0069	IBD	37	0	0	0	0	549	3	22556	Blood
PR0070	IBD	27	0	0	0	0	297	4	14495	Blood
PR0071	IBD	52	11	1	0	0	262	2	19525	Blood
PR0074	IBD	41	0	0	0	0	75	0	4383	Blood
PR0075	IBD	0	1	0	0	684	33	0	738	Blood
PR0076	IBD	249	1	0	0	70	470	3	4072	Blood
PR0124	IBD	198	0	0	0	642	74	0	1818	Blood
PR0079	IBD	32	0	0	0	0	538	9	15981	Blood
PR0080	IBD	43	1	0	0	0	332	3	19222	Blood
PR0081	IBD	80	0	0	0	1	406	6	7057	Blood
PR0082	IBD	80	1	0	0	2	515	2	6584	Blood
PR0083	IBD	49	1	0	0	2	514	6	14867	Blood
PR0132	IBD	112	0	0	0	12	653	11	10641	Blood
PR0142	IBD	1981	8	0	0	0	155	1	12123	Saliva
PR0086	IBD	7	1	0	0	2	814	9	890	Blood

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
PR0087	IBD	19	2	0	0	1	688	11	773	Blood
PR0089	IBD	9	0	0	0	0	728	3	824	Blood
PR0091	IBD	10	0	0	0	0	360	5	6505	Blood
PR0092	IBD	0	0	0	0	0	0	0	2	Blood
PR0095	IBD	43	0	0	0	1	633	3	15984	Blood
PR0096	IBD	0	0	0	0	587	45	0	656	Blood
PR0146	IBD	162	2	0	0	2	854	2	1395	Blood
PR0098	IBD	48	19	0	0	0	723	2	18195	Blood
PR0099	IBD	16	0	0	0	1	140	1	2045	Blood
PR0100	IBD	0	0	0	0	5	534	0	627	Blood
PR0102	IBD	80	2	0	0	1	499	5	9864	Blood
PR0103	IBD	62	0	0	0	1	702	2	8089	Blood
PR0151	IBD	124	0	0	0	1	1042	6	1482	Blood
PR0105	IBD	0	0	0	0	815	32	1	868	Blood
PR0106	IBD	2	0	0	0	795	60	2	995	Blood
PR0107	IBD	32	2	0	0	1	595	6	855	Blood
PR0108	IBD	44	1	2	0	0	466	3	17718	Blood
PR0159	IBD	120378	116	21	0	39	568	0	303033	Saliva
PR0110	IBD	316	1	0	0	58	453	2	6019	Blood
PR0111	IBD	316	1	1	0	64	394	5	11816	Blood
PR0161	IBD	41516	57	0	4	468	709	4	106947	Saliva
PR0113	IBD	15	0	1	0	0	140	1	1810	Blood
PR0114	IBD	116	0	0	0	3	438	4	7334	Blood
PR0115	IBD	1	0	0	0	381	61	0	459	Blood
PR0116	IBD	39	0	0	0	0	192	2	7957	Blood
PR0117	IBD	27	3	0	0	0	724	8	844	Blood
PR0118	IBD	47	0	0	0	0	350	7	16195	Blood
PR0120	IBD	76	1	0	0	0	201	1	7862	Blood
PR0121	IBD	65	0	0	0	0	330	0	8151	Blood
PR0122	IBD	11	0	0	0	0	589	5	656	Blood
PR0123	IBD	3	1	0	0	683	13	0	806	Blood
PR0166	IBD	1498	0	134	0	301	13	0	5611	Saliva
PR0125	IBD	41	0	0	0	0	488	0	23578	Blood
PR0126	IBD	4	0	0	0	0	608	0	614	Blood
PR0128	IBD	1	0	0	0	675	74	0	777	Blood
PR0129	IBD	1	0	0	0	340	48	0	403	Blood

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
PR0130	IBD	54	0	0	0	1	305	0	6856	Blood
PR0131	IBD	50	0	0	0	0	473	3	20220	Blood
PR0178	IBD	79	0	0	0	5	1479	34	31752	Blood
PR0134	IBD	0	2	0	0	617	18	0	649	Blood
PR0136	IBD	47	1	0	0	0	237	0	6495	Blood
PR0137	IBD	8	0	0	0	0	489	8	578	Blood
PR0138	IBD	9	0	0	0	0	550	2	10847	Blood
PR0141	IBD	41	0	1	0	0	528	3	19254	Blood
PR0194	IBD	34	0	0	0	1	1266	2	19138	Blood
PR0144	IBD	1	0	0	0	0	29	0	31	Blood
PR0145	IBD	49	0	0	0	0	550	10	781	Blood
PR0204	IBD	1501	5	0	0	0	58	0	1593	Blood
PR0149	IBD	51	1	0	0	0	747	3	943	Blood
PR0150	IBD	61	0	0	1	1	651	3	935	Blood
PR0206	IBD	2376	0	0	0	372	30	0	5100	Saliva
PR0153	IBD	70	3	0	0	0	486	6	743	Blood
PR0155	IBD	26	0	1	0	0	219	4	11396	Blood
PR0156	IBD	18	0	0	0	1	253	2	13677	Blood
PR0158	IBD	135	0	0	0	0	329	1	3471	Blood
PR0219	IBD	12	0	1	0	3020	1250	18	4492	Blood
PR0160	IBD	69	0	0	0	0	343	1	6145	Blood
PR0260	IBD	1508	5	0	0	0	60	0	1603	Blood
PR0165	IBD	65	3	0	0	0	543	4	3319	Blood
PT009	CONTROL	9463	18	0	0	248	900	3	26542	Unknown
PR0171	IBD	0	0	0	0	743	25	0	898	Blood
PR0172	IBD	27	0	0	0	0	337	4	17613	Blood
PR0173	IBD	2	1	0	0	519	41	0	648	Blood
PR0174	IBD	31	1	0	0	0	537	2	13775	Blood
PR0176	IBD	2	0	0	0	620	28	0	672	Blood
PR0177	IBD	33	1	0	0	0	334	3	16130	Blood
PT010	CONTROL	13711	12	0	0	390	1351	2	31980	Unknown
PR0179	IBD	1	0	0	0	493	23	0	533	Blood
PR0180	IBD	0	3	0	0	804	14	0	847	Blood
PR0182	IBD	11	0	0	0	0	262	0	284	Blood
PR0186	IBD	45	1	1	0	0	329	2	3838	Blood
PR0187	IBD	17	0	0	0	0	421	3	13597	Blood

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
PR0188	IBD	62	1	0	0	0	282	3	4403	Blood
PR0191	IBD	0	0	0	0	0	90	1	93	Blood
PR0192	IBD	36	44	0	0	0	649	8	21121	Blood
PR0193	IBD	54	0	0	0	0	504	4	22470	Blood
RNAR1	CONTROL	20728	25	0	4	7	403	0	101593	Unknown
PR0195	IBD	35	0	0	0	2	303	2	18322	Blood
PR0196	IBD	4	0	0	0	0	59	0	67	Blood
PR0198	IBD	2	0	0	0	0	636	5	3910	Blood
PR0199	IBD	24	0	0	0	0	511	4	14367	Blood
PR0203	IBD	32	0	0	0	1	649	3	17660	Blood
SOPR0224	IBD	2	0	0	0	809	61	0	1015	Blood
SOPR0226	IBD	194	1	0	0	629	26	0	1688	Blood
PR0207	IBD	38	10	0	0	2	739	5	24579	Blood
PR0208	IBD	27	11	0	0	1	379	1	15136	Blood
PR0210	IBD	1	0	0	0	0	63	0	67	Blood
PR0213	IBD	32	1	0	0	0	448	3	11043	Blood
PR0214	IBD	19	0	0	0	1	344	5	13902	Blood
PR0215	IBD	16	4	0	0	0	476	5	16497	Blood
PR0218	IBD	7	0	0	0	0	115	1	134	Blood
SOPR0230	IBD	4935	2	120	0	329	26	0	10448	Saliva
PR0220	IBD	2	0	0	0	0	125	0	131	Blood
PR0221	IBD	2	0	0	0	0	31	0	50	Blood
PR0241	IBD	2	0	0	0	2	551	10	669	Blood
SOPR0232	IBD	0	1	2	0	922	75	0	1751	Blood
SOPR0222	IBD	1	0	0	0	650	40	0	793	Blood
SOPR0223	IBD	1	0	8	0	0	92	0	959	Blood
SOPR0236	IBD	939	0	1	0	387	19	0	3500	Saliva
SOPR0251	IBD	3	0	0	0	9248	1005	0	11169	Blood
SOPR0228	IBD	15	4	0	0	1	790	2	2890	Blood
SOPR0229	IBD	3	0	2	0	0	70	0	714	Blood
SOPR0255	IBD	49278	28	0	0	5	850	0	152951	Saliva
SOPR0259	IBD	508	1	0	0	1	344	2	15105	Blood
SOPR0234	IBD	2	1	0	0	577	22	0	706	Blood
SOPR0270	IBD	0	0	0	0	918	43	0	1118	Blood
SOPR0237	IBD	0	0	0	0	645	65	0	744	Blood
SOPR0238	IBD	0	0	0	0	0	104	1	235	Blood

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
SOPR0243	IBD	10	0	0	0	0	350	2	9074	Blood
SOPR0244	IBD	22	0	0	0	1	469	4	10545	Blood
SOPR0245	IBD	17	0	0	0	0	522	5	12036	Blood
SOPR0246	IBD	17	1	0	0	0	543	1	11048	Blood
SOPR0247	IBD	45	0	0	0	3	103	1	15588	Blood
SOPR0249	IBD	6	3	0	0	0	310	2	22093	Blood
SOPR0279	IBD	1221	1	128	0	325	34	0	3698	Blood
SOPR0253	IBD	0	0	0	0	563	34	0	610	Blood
SOPR0254	IBD	28	1	0	0	2	525	3	18040	Blood
SOPR0306	IBD	2	0	0	0	2	67	0	1077	Blood
SOPR0256	IBD	1	0	0	0	0	46	0	129	Blood
SOPR0258	IBD	1	0	0	0	0	79	0	179	Blood
SOPR0307	IBD	2	0	0	0	1090	102	0	1374	Blood
SOPR0261	IBD	1	0	0	0	456	24	0	680	Blood
SOPR0267	IBD	0	0	0	0	717	31	0	774	Blood
SOPR0268	IBD	1	0	0	0	0	94	0	103	Blood
SOPR0309	IBD	244	1	0	0	683	64	0	1982	Blood
SOPR0273	IBD	0	0	0	0	0	63	0	122	Blood
SOPR0274	IBD	1	0	0	0	0	47	0	49	Blood
SOPR0275	IBD	3	0	0	0	0	74	1	82	Blood
SOPR0276	IBD	0	0	0	0	1	699	0	831	Blood
SOPR0278	IBD	14	0	142	0	336	17	0	703	Blood
SOPR0312	IBD	9	0	9	0	0	53	0	1006	Blood
SOPR0283	IBD	3	0	0	0	0	623	0	626	Blood
SOPR0284	IBD	0	0	0	0	342	40	0	435	Blood
SOPR0285	IBD	2	0	0	0	258	30	0	330	Blood
SOPR0286	IBD	1	0	0	0	1	77	0	82	Blood
SOPR0287	IBD	1	0	4	0	0	70	0	694	Blood
SOPR0289	IBD	6	0	4	0	0	91	0	930	Blood
SOPR0290	IBD	0	2	2	0	0	45	0	757	Blood
SOPR0294	IBD	0	0	0	0	0	40	0	42	Blood
SOPR0301	IBD	2	0	5	0	0	45	0	980	Blood
SOPR0303	IBD	25	0	0	0	0	132	0	197	Blood
SOPR0316	IBD	1	0	0	0	0	649	0	1129	Blood
SOPR0318	IBD	27	1	1	0	8559	871	0	10471	Blood
SOPR0333	IBD	23	0	0	0	18413	4039	0	23438	Saliva

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION
SOPR0310	IBD	5	0	0	0	0	112	1	131	Blood
SOPR0362	IBD	0	0	0	0	1404	54	0	1653	Blood
SOPR0313	IBD	13	0	0	0	0	131	1	183	Blood
SOPR0314	IBD	5	0	0	0	0	164	0	190	Blood
SOPR0367	IBD	0	0	0	0	3	65	0	1158	Blood
SOPR0317	IBD	13	0	0	0	0	139	1	250	Blood
SOPR0372	IBD	1	0	6	0	0	95	0	1011	Blood
SOPR0320	IBD	2	0	0	0	0	120	1	131	Blood
SOPR0323	IBD	1	0	11	0	3	81	0	832	Blood
SOPR0325	IBD	2	0	0	0	668	33	0	719	Blood
SOPR0326	IBD	0	0	0	0	0	687	0	688	Blood
SOPR0327	IBD	3	0	0	0	0	56	2	69	Blood
VFW1515559	CONTROL	1257	3	0	0	0	746	5	3782	Unknown
SOPR0334	IBD	4	0	0	0	548	23	0	897	Blood
SOPR0336	IBD	8	0	0	0	0	145	0	156	Blood
SOPR0339	IBD	5	0	0	0	0	57	1	70	Blood
SOPR0340	IBD	4	0	0	0	2	129	0	144	Blood
SOPR0342	IBD	1	0	0	0	0	28	0	33	Blood
SOPR0344	IBD	1	0	1	0	3	74	1	719	Blood
SOPR0345	IBD	3	0	0	0	0	680	2	686	Blood
SOPR0346	IBD	1	0	0	0	268	18	0	334	Blood
SOPR0348	IBD	3	0	8	0	0	82	0	894	Blood
SOPR0349	IBD	5	0	5	0	0	114	0	928	Blood
SOPR0351	IBD	4	0	0	0	0	111	0	124	Blood
SOPR0352	IBD	2	0	0	0	0	27	0	963	Blood
SOPR0355	IBD	0	0	0	0	0	26	0	27	Blood
SOPR0357	IBD	0	0	0	0	604	60	0	758	Blood
SOPR0359	IBD	5	0	0	0	396	79	0	543	Blood
SOPR0360	IBD	0	0	0	0	0	470	0	474	Blood
SOPR0361	IBD	0	1	0	0	0	788	0	793	Blood
W1416409LA	CONTROL	9062	2	0	0	101	448	2	40719	Unknown
Wexcon001	CONTROL	3	0	0	0	0	1024	3	1084	Unknown
SOPR0368	IBD	1	0	0	0	0	984	0	991	Blood
SOPR0370	IBD	0	0	0	0	0	740	0	741	Blood
ZB-W1505055	CONTROL	10	6	0	0	0	1260	3	2450	Unknown

Table 8.5 continued from previous page

SAMPLE	GROUP	BACTERIAL	VIRAL	FUNGAL	ARCHAEA	PLANTS	METAZOA	HUMAN	UNCLASSIFIED	COLLECTION

Table 8.5: Unmapped reads totals from all 358 exome samples mapped at kingdom or domain levels

8.2.3 *CABIN1* Average coverage per exon

CHR	START	END	EXON	AVERAGE COVERAGE
chr22	24011314	24011367	1	1.68889
chr22	24035443	24035520	2	82.1667
chr22	24036088	24036181	3	124.958
chr22	24038347	24038461	4	97.925
chr22	24041138	24041273	5	147.65
chr22	24042903	24043084	6	117.808
chr22	24049090	24049220	7	177.619
chr22	24050824	24050974	8	134.072
chr22	24054872	24055159	9	297.261
chr22	24056191	24056360	10	152.114
chr22	24059226	24059363	11	114.269
chr22	24059923	24060141	12	167.45
chr22	24061946	24062025	13	154.733
chr22	24062958	24063146	14	203.861
chr22	24064034	24064187	15	185.281
chr22	24066986	24067181	16	241.136
chr22	24070799	24071042	17	299.35
chr22	24072353	24072510	18	189.094
chr22	24076168	24076284	19	134.914
chr22	24083227	24083389	20	66.9472
chr22	24084578	24084785	21	254.153
chr22	24085005	24085151	22	188.672
chr22	24087451	24087713	23	179.267
chr22	24091582	24091843	24	298.869
chr22	24095930	24096082	25	135.928
chr22	24098013	24098192	26	164.15
chr22	24113565	24113748	27	159.569
chr22	24119366	24119698	28	170.244
chr22	24134301	24134415	29	286.289
chr22	24164399	24164563	30	215.419
chr22	24165529	24165626	31	132.247
chr22	24166638	24167313	32	428.239
chr22	24168446	24168521	33	84.4917
chr22	24171712	24171995	34	543.089
chr22	24176110	24176275	35	243.253
chr22	24177503	24177817	36	314.608
chr22	24178052	24178628	37	160.172

Table 8.6: *CABIN1* mean exon coverage calculated for all 358 WES samples

8.3 Appendix C

8.3.1 Extended 430 candidate gene list for skeletal trio

Gene	Gene Description
ACAN	aggrecan
ACP5	acid phosphatase 5, tartrate resistant
ACVR1	activin A receptor type 1
ADAMTS10	ADAM metallopeptidase with thrombospondin type 1 motif 10
ADAMTS17	ADAM metallopeptidase with thrombospondin type 1 motif 17
ADAMTSL2	ADAMTS like 2
AGA	aspartylglucosaminidase
AGPS	alkylglycerone phosphate synthase
AHI1	Abelson helper integration site 1
AKT1	AKT serine/threonine kinase 1
ALPL	alkaline phosphatase, liver/bone/kidney
ALX1	ALX homeobox 1
ALX3	ALX homeobox 3
ALX4	ALX homeobox 4
AMER1	APC membrane recruitment protein 1
ANKH	ANKH inorganic pyrophosphate transport regulator
ANO5	anoctamin 5
ANOS1	anosmin 1
ANTXR2	anthrax toxin receptor 2
ARHGAP31	Rho GTPase activating protein 31
ARSB	arylsulfatase B
ARSE	arylsulfatase E (chondrodysplasia punctata 1)
ATP6V0A2	ATPase H+ transporting V0 subunit a2
AXIN2	axin 2
B3GALT6	beta-1,3-galactosyltransferase 6
B3GAT3	beta-1,3-glucuronyltransferase 3
B4GALT7	beta-1,4-galactosyltransferase 7
BHLHA9	basic helix-loop-helix family member a9
BICC1	BicC family RNA binding protein 1
BMP1	bone morphogenetic protein 1
BMP2	bone morphogenetic protein 2
BMP4	bone morphogenetic protein 4
BMP7	bone morphogenetic protein 7
BMPER	BMP binding endothelial regulator
BMPR1B	bone morphogenetic protein receptor type 1B
CA2	carbonic anhydrase 2
CACNA1G	calcium voltage-gated channel subunit alpha1 G
CANT1	calcium activated nucleotidase 1
CASR	calcium sensing receptor
CC2D2A	coiled-coil and C2 domain containing 2A
CCBE1	collagen and calcium binding EGF domains 1
CCDC8	coiled-coil domain containing 8
CCNQ	cyclin Q
CDC6	cell division cycle 6
CDH3	cadherin 3
CDKN1C	cyclin dependent kinase inhibitor 1C
CDT1	chromatin licensing and DNA replication factor 1
CEP120	centrosomal protein 120
CEP290	centrosomal protein 290
CHST14	carbohydrate sulfotransferase 14
CHST3	carbohydrate sulfotransferase 3
CHSY1	chondroitin sulfate synthase 1
CKAP2L	cytoskeleton associated protein 2 like
CKAP4	cytoskeleton associated protein 4
CLCN5	chloride voltage-gated channel 5
CLCN7	chloride voltage-gated channel 7
CNTNAP2	contactin associated protein like 2

Table 8.7 continued from previous page

COG1	component of oligomeric golgi complex 1
COL10A1	collagen type X alpha 1 chain
COL11A1	collagen type XI alpha 1 chain
COL11A2	collagen type XI alpha 2 chain
COL1A1	collagen type I alpha 1 chain
COL1A2	collagen type I alpha 2 chain
COL2A1	collagen type II alpha 1 chain
COL9A1	collagen type IX alpha 1 chain
COL9A2	collagen type IX alpha 2 chain
COL9A3	collagen type IX alpha 3 chain
COMP	cartilage oligomeric matrix protein
CREB3L1	cAMP responsive element binding protein 3 like 1
CREBBP	CREB binding protein
CRTAP	cartilage associated protein
CTSA	cathepsin A
CTSK	cathepsin K
CUL7	cullin 7
CYP26B1	cytochrome P450 family 26 subfamily B member 1
CYP26C1	cytochrome P450 family 26 subfamily C member 1
DAG1	dystroglycan 1
DDR2	discoidin domain receptor tyrosine kinase 2
DDX41	DEAD-box helicase 41
DEPDC5	DEP domain containing 5
DHCR24	24-dehydrocholesterol reductase
DHODH	dihydroorotate dehydrogenase (quinone)
DLL3	delta like canonical Notch ligand 3
DLX3	distal-less homeobox 3
DLX5	distal-less homeobox 5
DLX6	distal-less homeobox 6
DMP1	dentin matrix acidic phosphoprotein 1
DNM2	dynamin 2
DOCK6	dedicator of cytokinesis 6
DSC2	desmocollin 2
DSG2	desmoglein 2
DSP	desmoplakin
DSPP	dentin sialophosphoprotein
DSTYK	dual serine/threonine and tyrosine protein kinase
DVL1	dishevelled segment polarity protein 1
DYM	dymeclin
DYNC2H1	dynein cytoplasmic 2 heavy chain 1
DYNC2LI1	dynein cytoplasmic 2 light intermediate chain 1
EBP	emopamil binding protein (sterol isomerase)
EDA	ectodysplasin A
EDA2R	ectodysplasin A2 receptor
EDAR	ectodysplasin A receptor
EDARADD	EDAR associated death domain
EFNB1	ephrin B1
EFTUD2	elongation factor Tu GTP binding domain containing 2
EIF2AK3	eukaryotic translation initiation factor 2 alpha kinase 3
ENPP1	ectonucleotide pyrophosphatase/phosphodiesterase 1
EOGT	EGF domain specific O-linked N-acetylglucosamine transferase
EP300	E1A binding protein p300
ERCC1	ERCC excision repair 1, endonuclease non-catalytic subunit
ERCC6	ERCC excision repair 6, chromatin remodeling factor
ERF	ETS2 repressor factor
ESCO2	establishment of sister chromatid cohesion N-acetyltransferase 2
EVC	EvC ciliary complex subunit 1
EVC2	EvC ciliary complex subunit 2
EXT1	exostosin glycosyltransferase 1
EXT2	exostosin glycosyltransferase 2
EYA1	EYA transcriptional coactivator and phosphatase 1
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit

Table 8.7 continued from previous page

FAM111A	family with sequence similarity 111 member A
FAM20C	FAM20C, golgi associated secretory pathway kinase
FBLN1	fibulin 1
FBN1	fibrillin 1
FBN2	fibrillin 2
FBXW4	F-box and WD repeat domain containing 4
FERMT3	fermitin family member 3
FGF10	fibroblast growth factor 10
FGF16	fibroblast growth factor 16
FGF23	fibroblast growth factor 23
FGF8	fibroblast growth factor 8
FGF9	fibroblast growth factor 9
FGFR1	fibroblast growth factor receptor 1
FGFR2	fibroblast growth factor receptor 2
FGFR3	fibroblast growth factor receptor 3
FIG4	FIG4 phosphoinositide 5-phosphatase
FKBP10	FK506 binding protein 10
FLNA	filamin A
FLNB	filamin B
FMN1	formin 1
FOXC1	forkhead box C1
FOXF1	forkhead box F1
FREM1	FRAS1 related extracellular matrix 1
FUCA1	alpha-L-fucosidase 1
FZD2	frizzled class receptor 2
FZD6	frizzled class receptor 6
GALNS	galactosamine (N-acetyl)-6-sulfatase
GALNT3	polypeptide N-acetylgalactosaminyltransferase 3
GATA2	GATA binding protein 2
GATA3	GATA binding protein 3
GDF3	growth differentiation factor 3
GDF5	growth differentiation factor 5
GDF6	growth differentiation factor 6
GJA1	gap junction protein alpha 1
GJB6	gap junction protein beta 6
GLB1	galactosidase beta 1
GLI3	GLI family zinc finger 3
GNAS	GNAS complex locus
GNPAT	glyceronephosphate O-acyltransferase
GNPTAB	N-acetylglucosamine-1-phosphate transferase subunits alpha and beta
GNPTG	N-acetylglucosamine-1-phosphate transferase subunit gamma
GNS	glucosamine (N-acetyl)-6-sulfatase
GORAB	golgin, RAB6 interacting
GPC6	glypican 6
GPX1	glutathione peroxidase 1
GPX2	glutathione peroxidase 2
GPX3	glutathione peroxidase 3
GPX4	glutathione peroxidase 4
GPX5	glutathione peroxidase 5
GPX6	glutathione peroxidase 6
GPX6	glutathione peroxidase 6
GPX7	glutathione peroxidase 7
GPX8	glutathione peroxidase 8 (putative)
GREM1	gremlin 1, DAN family BMP antagonist
GRHL2	grainyhead like transcription factor 2
GSC	goosecoid homeobox
GUSB	glucuronidase beta
HDAC4	histone deacetylase 4
HDAC8	histone deacetylase 8
HES7	hes family bHLH transcription factor 7
HESX1	HESX homeobox 1
HGSNAT	heparan-alpha-glucosaminide N-acetyltransferase

Table 8.7 continued from previous page

HNF1B	HNF1 homeobox B
HOXA11	homeobox A11
HOXA13	homeobox A13
HOXC13	homeobox C13
HOXD13	homeobox D13
HPGD	15-hydroxyprostaglandin dehydrogenase
HSPG2	heparan sulfate proteoglycan 2
ICK	intestinal cell kinase
IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial
IDS	iduronate 2-sulfatase
IDUA	iduronidase, alpha-L-
IFITM5	interferon induced transmembrane protein 5
IFT122	intraflagellar transport 122
IFT140	intraflagellar transport 140
IFT172	intraflagellar transport 172
IFT43	intraflagellar transport 43
IFT52	intraflagellar transport 52
IFT80	intraflagellar transport 80
IFT88	intraflagellar transport 88
IHH	indian hedgehog
IKBKG	inhibitor of nuclear factor kappa B kinase subunit gamma
IL1RN	interleukin 1 receptor antagonist
IMPAD1	inositol monophosphatase domain containing 1
INPPL1	inositol polyphosphate phosphatase like 1
ITGA8	integrin subunit alpha 8
JSRP1	junctional sarcoplasmic reticulum protein 1
JUP	junction plakoglobin
KAT6B	lysine acetyltransferase 6B
KIAA0586	KIAA0586
KIF14	kinesin family member 14
KIF22	kinesin family member 22
KIF7	kinesin family member 7
KRT74	keratin 74
LAMA1	laminin subunit alpha 1
LAMB1	laminin subunit beta 1
LBP	lipopolysaccharide binding protein
LBR	lamin B receptor
LEMD3	LEM domain containing 3
LFNG	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase
LIFR	LIF receptor alpha
LMBR1	limb development membrane protein 1
LMNA	lamin A/C
LMX1B	LIM homeobox transcription factor 1 beta
LONP1	lon peptidase 1, mitochondrial
LPIN2	lipin 2
LRP4	LDL receptor related protein 4
LRP5	LDL receptor related protein 5
LTBP2	latent transforming growth factor beta binding protein 2
MAB21L2	mab-21 like 2
MAFB	MAF bZIP transcription factor B
MAN2B1	mannosidase alpha class 2B member 1
MAN2C1	mannosidase alpha class 2C member 1
MATN3	matrilin 3
MEGF8	multiple EGF like domains 8
MEOX1	mesenchyme homeobox 1
MESP2	mesoderm posterior bHLH transcription factor 2
MGP	matrix Gla protein
MKS1	Meckel syndrome, type 1
MMP1	matrix metallopeptidase 1
MMP13	matrix metallopeptidase 13
MMP2	matrix metallopeptidase 2

Table 8.7 continued from previous page

MMP9	matrix metallopeptidase 9
MNX1	motor neuron and pancreas homeobox 1
MSH2	mutS homolog 2
MSX2	msh homeobox 2
MYCN	MYCN proto-oncogene, bHLH transcription factor
MYH10	myosin heavy chain 10
MYH7	myosin heavy chain 7
NAGLU	N-acetyl-alpha-glucosaminidase
NECTIN1	nectin cell adhesion molecule 1
NECTIN4	nectin cell adhesion molecule 4
NEK1	NIMA related kinase 1
NEK8	NIMA related kinase 8
NEU1	neuraminidase 1
NF1	neurofibromin 1
NFIX	nuclear factor I X
NFKBIA	NFKB inhibitor alpha
NIN	ninein
NIPBL	NIPBL, cohesin loading factor
NKX3-2	NK3 homeobox 2
NLRP3	NLR family pyrin domain containing 3
NOG	noggin
NOS2	nitric oxide synthase 2
NOTCH2	notch 2
NPHP3	nephrocystin 3
NPPC	natriuretic peptide C
NPR2	natriuretic peptide receptor 2
NSD1	nuclear receptor binding SET domain protein 1
NSDHL	NAD(P) dependent steroid dehydrogenase-like
OBSL1	obscurin like 1
OFD1	OFD1, centriole and centriolar satellite protein
ORC1	origin recognition complex subunit 1
ORC4	origin recognition complex subunit 4
ORC6	origin recognition complex subunit 6
OSTM1	osteopetrosis associated transmembrane protein 1
P3H1	prolyl 3-hydroxylase 1
P4HB	prolyl 4-hydroxylase subunit beta
PAM16	presequence translocase associated motor 16
PAPSS2	3'-phosphoadenosine 5'-phosphosulfate synthase 2
PCNT	pericentrin
PCYT1A	phosphate cytidylyltransferase 1, choline, alpha
PDE3A	phosphodiesterase 3A
PDE4D	phosphodiesterase 4D
PEX7	peroxisomal biogenesis factor 7
PGM3	phosphoglucomutase 3
PHEX	phosphate regulating endopeptidase homolog X-linked
PIGV	phosphatidylinositol glycan anchor biosynthesis class V
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PITX1	paired like homeodomain 1
PKP1	plakophilin 1
PKP2	plakophilin 2
PLEKHM1	pleckstrin homology and RUN domain containing M1
PLOD2	procollagen-lysine,2-oxoglutarate 5-dioxygenase 2
PLS3	plastin 3
POC1A	POC1 centriolar protein A
POLR1C	RNA polymerase I and III subunit C
POLR1D	RNA polymerase I and III subunit D
POP1	POP1 homolog, ribonuclease P/MRP subunit
POR	cytochrome p450 oxidoreductase
PPIB	peptidylprolyl isomerase B
PRKAR1A	protein kinase cAMP-dependent type I regulatory subunit alpha
PROKR2	prokineticin receptor 2
PTDSS1	phosphatidylserine synthase 1

Table 8.7 continued from previous page

PTH1R	parathyroid hormone 1 receptor
PTHLH	parathyroid hormone like hormone
PTPN11	protein tyrosine phosphatase, non-receptor type 11
PYCR1	pyrroline-5-carboxylate reductase 1
RAB23	RAB23, member RAS oncogene family
RAB33B	RAB33B, member RAS oncogene family
RAD21	RAD21 cohesin complex component
RASGRP2	RAS guanyl releasing protein 2
RBM8A	RNA binding motif protein 8A
RBPJ	recombination signal binding protein for immunoglobulin kappa J region
RECQL4	RecQ like helicase 4
RET	ret proto-oncogene
RMRP	RNA component of mitochondrial RNA processing endoribonuclease
RNU4ATAC	RNA, U4atac small nuclear (U12-dependent splicing)
ROR2	receptor tyrosine kinase like orphan receptor 2
RPGRIP1L	RPGRIP1 like
RUNX2	runt related transcription factor 2
RYR2	ryanodine receptor 2
SALL1	spalt like transcription factor 1
SALL4	spalt like transcription factor 4
SBDS	SBDS, ribosome maturation factor
SCN5A	sodium voltage-gated channel alpha subunit 5
SEC23A	Sec23 homolog A, coat complex II component
SEC24D	SEC24 homolog D, COPII coat complex component
SERPINF1	serpin family F member 1
SERPINH1	serpin family H member 1
SETD2	SET domain containing 2
SF3B4	splicing factor 3b subunit 4
SGSH	N-sulfoglucosamine sulfohydrolase
SH3BP2	SH3 domain binding protein 2
SH3PXD2B	SH3 and PX domains 2B
SHH	sonic hedgehog
SHOX	short stature homeobox
SIX1	SIX homeobox 1
SIX2	SIX homeobox 2
SKI	SKI proto-oncogene
SLC17A5	solute carrier family 17 member 5
SLC25A12	solute carrier family 25 member 12
SLC25A3	solute carrier family 25 member 3
SLC26A2	solute carrier family 26 member 2
SLC26A4	solute carrier family 26 member 4
SLC29A3	solute carrier family 29 member 3
SLC34A3	solute carrier family 34 member 3
SLC35D1	solute carrier family 35 member D1
SLC39A13	solute carrier family 39 member 13
SLCO5A1	solute carrier organic anion transporter family member 5A1
SMAD3	SMAD family member 3
SMAD4	SMAD family member 4
SMARCAL1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a like 1
SMC1A	structural maintenance of chromosomes 1A
SMC3	structural maintenance of chromosomes 3
SNIP1	Smad nuclear interacting protein 1
SNRPB	small nuclear ribonucleoprotein polypeptides B and B1
SNX10	sorting nexin 10
SOST	sclerostin
SOX9	SRY-box 9
SP7	Sp7 transcription factor
SRP72	signal recognition particle 72
STK4	serine/threonine kinase 4
SULF1	sulfatase 1
SUMF1	sulfatase modifying factor 1
TAX1BP3	Tax1 binding protein 3

Table 8.7 continued from previous page

TBCE	tubulin folding cofactor E
TBX15	T-box 15
TBX3	T-box 3
TBX4	T-box 4
TBX5	T-box 5
TBX6	T-box 6
TBXAS1	thromboxane A synthase 1
TCF12	transcription factor 12
TCIRG1	T cell immune regulator 1, ATPase H+ transporting V0 subunit a3
TCOF1	treacle ribosome biogenesis factor 1
TCTEX1D2	Tctex1 domain containing 2
TCTN3	tectonic family member 3
TET2	tet methylcytosine dioxygenase 2
TGDS	TDP-glucose 4,6-dehydratase
TGFB1	transforming growth factor beta 1
TGFB2	transforming growth factor beta 2
TGFB3	transforming growth factor beta 3
TGFBR1	transforming growth factor beta receptor 1
TGFBR2	transforming growth factor beta receptor 2
THPO	thrombopoietin
TLR1	toll like receptor 1
TMC6	transmembrane channel like 6
TMC8	transmembrane channel like 8
TMCO1	transmembrane and coiled-coil domains 1
TMEM216	transmembrane protein 216
TMEM38B	transmembrane protein 38B
TMEM43	transmembrane protein 43
TMEM67	transmembrane protein 67
TNFRSF11A	TNF receptor superfamily member 11a
TNFRSF11B	TNF receptor superfamily member 11b
TNFSF11	TNF superfamily member 11
TP63	tumor protein p63
TRAF6	TNF receptor associated factor 6
TRAPPC2	trafficking protein particle complex 2
TREM2	triggering receptor expressed on myeloid cells 2
TRIP11	thyroid hormone receptor interactor 11
TRPS1	transcriptional repressor GATA binding 1
TRPV4	transient receptor potential cation channel subfamily V member 4
TTC21B	tetratricopeptide repeat domain 21B
TWIST1	twist family bHLH transcription factor 1
TWIST2	twist family bHLH transcription factor 2
TYROBP	TYRO protein tyrosine kinase binding protein
UPF3B	UPF3B, regulator of nonsense mediated mRNA decay
UPK3A	uroplakin 3A
VLDLR	very low density lipoprotein receptor
WDR19	WD repeat domain 19
WDR34	WD repeat domain 34
WDR35	WD repeat domain 35
WDR60	WD repeat domain 60
WISP3	WNT1 inducible signaling pathway protein 3
WNT1	Wnt family member 1
WNT10A	Wnt family member 10A
WNT10B	Wnt family member 10B
WNT3	Wnt family member 3
WNT4	Wnt family member 4
WNT5A	Wnt family member 5A
WNT6	Wnt family member 6
WNT7A	Wnt family member 7A
XYLT1	xylosyltransferase 1
XYLT2	xylosyltransferase 2
ZBTB16	zinc finger and BTB domain containing 16
ZMPSTE24	zinc metallopeptidase STE24

Table 8.7 continued from previous page

ZSWIM6	zinc finger SWIM-type containing 6

Table 8.7: 430 candidate genes used in prioritising variants for the SSMD trio.

8.3.2 CNVkit WGS filtered variants

Sample	chromosome	start	end	log2	cn	depth	probes	weight
SD003_WGS	chr1	109687856	109697446	-17.1142	0	0.525991	3	2.06506
SD003_WGS	chr1	143541337	143779171	0.485238	3	33.822	124	97.4003
SD003_WGS	chr1	143779171	144074545	0.296543	3	29.2842	154	122.694
SD003_WGS	chr1	152584770	152613540	-19.4289	0	0.0131039	15	9.704
SD003_WGS	chr1	196752780	196796895	-1.00974	1	11.0007	23	17.3736
SD003_WGS	chr1	196796895	196835255	-11.7715	0	1.395	19	11.4073
SD003_WGS	chr1	196835255	196917730	-0.952325	1	11.5982	43	33.4823
SD003_WGS	chr1	207526261	207564621	-0.483977	1	13.4083	20	13.9295
SD003_WGS	chr1	248574337	248633794	-0.908045	1	11.8945	31	23.5286
SD003_WGS	chr5	105097282	105166330	-0.995037	1	11.5188	36	28.2695
SD003_WGS	chr6	255637	382226	0.515881	3	33.6512	66	50.0625
SD003_WGS	chr6	32510852	32570310	-0.568176	1	13.172	27	18.1474
SD003_WGS	chr8	6974579	7022530	0.567318	3	41.9399	25	17.2885
SD003_WGS	chr8	7348599	7548077	0.320107	3	35.827	103	73.6956
SD003_WGS	chr8	8091047	8121738	-0.726919	1	11.5085	16	10.74
SD003_WGS	chr8	8121738	8208057	0.245242	3	39.8277	45	28.3436
SD003_WGS	chr8	8208057	8229157	-0.837327	1	10.9483	11	8.2028
SD003_WGS	chr8	12155699	12224754	0.329269	3	36.2988	36	25.4259
SD003_WGS	chr8	12349558	12606575	0.225355	3	40.2656	132	90.596
SD003_WGS	chr9	10000	38769	-0.307959	1	18.5053	15	11.4173
SD003_WGS	chr9	15818019	15821855	-5.41315	0	0.366003	2	1.26287
SD003_WGS	chr9	40780574	42477989	0.20807	3	31.8856	875	627.361
SD003_WGS	chr9	60554823	61132394	-0.345584	1	17.628	209	128.379
SD003_WGS	chr9	62270853	62610621	0.314866	3	30.5259	177	132.999
SD003_WGS	chr9	63252862	63317979	0.273378	3	31.6983	34	25.9071
SD003_WGS	chr9	64395736	64556884	0.21236	3	27.8459	84	64.5044
SD003_WGS	chr9	66153525	66212990	0.524923	3	36.1158	31	22.7949
SD003_WGS	chr9	138154425	138223473	0.339022	3	33.6674	36	26.0097
SD003_WGS	chr11	5762164	5789016	0.75741	4	22.237	14	8.71758
SD003_WGS	chr11	55597178	55689237	0.796666	4	23.9675	48	30.7123
SD003_WGS	chr11	89767465	90074347	-0.429398	1	22.38	154	113.645
SD003_WGS	chr13	88780995	90549456	0.470138	3	34.0316	922	724.999
SD003_WGS	chr14	18653182	19394573	0.395386	3	54.5319	308	195.743
SD003_WGS	chr14	19661580	19955031	0.351735	3	58.5315	153	99.2956
SD003_WGS	chr14	73530015	73556867	0.834441	4	20.3489	14	8.72002
SD003_WGS	chr14	79640705	79648377	-19.6793	0	0.022201	4	2.07147
SD003_WGS	chr15	20470851	20520674	-0.406107	1	16.9216	26	20.7563
SD003_WGS	chr15	20823644	20973115	-0.687581	1	11.984	76	47.8632
SD003_WGS	chr15	20973115	21193490	-0.288489	1	20.9977	113	80.0442
SD003_WGS	chr15	21828502	21939801	-0.572706	1	13.2723	55	35.3997
SD003_WGS	chr15	21966667	21991613	-0.665233	1	13.3273	11	7.00363
SD003_WGS	chr15	21991613	22127859	0.242464	3	12.4139	58	28.3428
SD003_WGS	chr15	22219969	22302485	0.439867	3	31.1937	43	33.6995
SD003_WGS	chr15	23163596	23320988	0.40854	3	37.0665	56	38.2948
SD003_WGS	chr15	24337541	24535098	0.503349	3	34.8875	103	79.5895
SD003_WGS	chr15	30105042	30379320	-0.330349	1	18.5934	143	113.769
SD003_WGS	chr15	30379320	30611401	-0.566578	1	15.0899	121	95.3238
SD003_WGS	chr15	30611401	32153493	-1.0079	1	11.1615	804	655.421
SD003_WGS	chr15	32153493	32494901	-0.556299	1	15.3231	178	140.214
SD003_WGS	chr15	32494901	32583130	-0.259396	1	19.7698	46	36.2495
SD003_WGS	chr15	84185672	84515703	-0.314877	1	15.3045	146	107.63
SD003_WGS	chr16	10000	61786	-0.302632	1	15.4626	27	19.1822
SD003_WGS	chr16	55761676	55788528	-1.02305	1	11.3669	14	11.3406

Table 8.8 continued from previous page

Sample	chromosome	start	end	log2	cn	depth	probes	weight
SD003_WGS	chr16	55807708	55828807	0.519222	3	33.5698	11	8.49549
SD003_WGS	chr16	70121829	70165943	0.52814	3	41.5164	23	17.1765
SD003_WGS	chr16	74360635	74427765	0.455057	3	37.3155	35	25.9145
SD003_WGS	chr16	90165050	90228345	0.244003	3	32.5784	32	23.428
SD003_WGS	chr17	16753475	16845538	0.45447	3	31.9932	48	36.7649
SD003_WGS	chr17	18437463	18562132	-0.810036	1	13.1881	62	48.0463
SD003_WGS	chr17	45496368	45521302	-0.662615	1	15.2038	13	8.7251
SD003_WGS	chr17	46133141	46292334	0.456176	3	41.1059	83	56.9736
SD003_WGS	chr17	46489887	46560852	0.373518	3	31.45	37	27.2981
SD003_WGS	chr19	43196599	43259892	-0.874737	1	11.8348	33	25.2461
SD003_WGS	chr20	25783985	25845360	-0.57977	1	11.0726	32	20.6545
SD003_WGS	chr20	26010307	26073600	-0.556333	1	11.7706	33	22.3608
SD003_WGS	chr21	13591071	13681217	-0.761079	1	12.142	42	29.135
SD003_WGS	chr21	43396740	43450451	-0.540543	1	8.7515	25	13.7148
SD003_WGS	chr22	11462696	11977555	0.244223	3	34.6292	159	100.36
SD003_WGS	chr22	15664273	15694948	0.700791	4	64.0874	16	9.13133
SD003_WGS	chr22	18212289	18425808	-0.799514	1	11.503	56	39.7751
SD003_WGS	chr22	18642341	18933964	-0.378894	1	14.4693	116	76.4806
SD003_WGS	chr22	21168383	21202906	0.27216	3	27.4091	18	13.9087

Table 8.8: CNVkit WGS variant calls for SD003 which overlap at least one gene.

8.3.3 LUMPY - Large heterozygous variants

CHR	POS	END	SV-TYPE	SV-LENGTH	EVIDENCE	GT:SD00-1 2 3
chr16	29299273	88273118	DUP	58,973,845	4	0/0 0/0 0/1
chr16	29214730	88180630	DUP	58,965,900	4	0/0 0/0 0/1
chr17	20740802	59950511	DUP	39,209,709	5	0/0 0/0 0/1
chr17	16814147	41575068	DUP	24,760,921	13	0/0 0/0 0/1
chr2	35710442	58983817	DUP	23,273,375	6	0/0 0/0 0/1
chr17	45585459	64919622	DUP	19,334,163	8	0/0 0/0 0/1
chr9	42971794	64485627	DEL	-21,513,833	5	0/0 0/0 0/1
chr6	129223231	155789998	DEL	-26,566,767	9	0/0 0/0 0/1
chr6	22375581	52730953	DEL	-30,355,372	4	0/0 0/0 0/1
chr4	119333163	165218756	DEL	-45,885,593	6	0/0 0/0 0/1
chr6	110502314	165461184	DEL	-54,958,870	5	0/0 0/0 0/1
chr16	29153630	88115747	DEL	-58,962,117	4	0/0 0/0 0/1
chr16	29167269	88129447	DEL	-58,962,178	9	0/0 0/0 0/1
chr16	29224385	88190298	DEL	-58,965,913	15	0/0 0/0 0/1
chr16	29196652	88162813	DEL	-58,966,161	6	0/0 0/0 0/1
chr3	85398317	154534862	DEL	-69,136,545	4	0/0 0/0 0/1
chr7	6842271	97951122	DEL	-91,108,851	5	0/0 0/0 0/1

Table 8.9: LUMPY-SV large heterozygous variants called in sample SD003 WGS

8.4 Appendix D

8.4.1 Genes and promoters targeted by HaloPlex capture kit

Feature	Chromosome	Start	End	Feature	Chromosome	Start	End
F1BPROM	1	12105864	12107047	CCL19	22	34690813	34691562
TNFRSF1B	1	12166689	12209509	CCL21PROM	22	34707765	34710350
C1QAPROM	1	22308952	22310298	TDRD7SNP	22	97428015	97428860
C1QCPROM	1	22316040	22317216	C5PROM	22	118188915	118190356
C1QBPROM	1	22325483	22327118	C5	22	120951951	121075383
C1QA	1	22636330	22639863	PTGESPROM	22	126974908	126976259
C1QC	1	22643445	22648251	PTGES	22	129738042	129753372
C1QB	1	22652386	22661860	C8GPROM	22	134078938	134080147
CD52PROM	1	25990285	25991452	C8G	22	136945016	136947250
CD52	1	26317674	26320668	IL2RAPROM	2	5967645	5968920
C8APROM	1	56387828	56389266	IL2RA	2	6010485	6062648
C8BPROM	1	56462237	56463748	CXCL12PROM	2	43873491	43874847
C8A	1	56854535	56918371	CXCL12	2	44297257	44386619
C8B	1	56928962	56966293	HPSE2SNP	2	98458910	98460010
IL12RB2PROM	1	66840447	66841848	SOX6SNP	3	16111404	16112353
IL12RB2	1	67307076	67397072	CD59PROM	3	33680207	33681326
GSTM1PROM	1	109143986	109145321	CD59	3	33697881	33736587
GSTM1	1	109687580	109694021	MS4A1	3	60455590	60471026
FCGR1B	1	143874348	143884668	MS4A1PROM	3	60687088	60688471
FCGR1BPROM	1	121344983	121345456	GSTP1	3	67583433	67586853
FCGR1CPROM	1	143825391	143826778	GSTP1PROM	3	67815101	67816042
FCGR1C	1	143874348	143884009	IL10RA	3	117986122	118001727
FCGR1A	1	149782299	149792646	IL10RAPROM	3	118114396	118116017
FCGR1APROM	1	149810138	149811212	CXCR5	3	118883435	118897980
IL6R	1	154404909	154469580	CXCR5PROM	3	119011903	119013136
IL6RPROM	1	154431357	154432865	WNK1SNP	4	884310	885216
FCRL5	1	157507407	157552775	TNFRSF1APROM	4	6218514	6219733
FCRL4	1	157573555	157598341	TNFRSF1A	4	6328568	6342355
FCRL4PROM	1	157602665	157604041	C1SPROM	4	6950221	6951782
FCRL3	1	157676229	157701107	C1RPROM	4	6969887	6971158
FCRL3PROM	1	157707237	157708770	C1S	4	6987958	7071376
FCRL2	1	157745379	157777343	C1R	4	7079983	7092673
FCRL2PROM	1	157774741	157776042	C3AR1PROM	4	7904694	7905797
FCRL1	1	157794187	157820369	C3AR1	4	8058014	8066787
FCRL1PROM	1	157823430	157824817	KLRD1PROM	4	10150809	10152191
FCRL6	1	159800204	159816387	KLRK1PROM	4	10218611	10219982
FCRL6PROM	1	159829441	159830840	KLRD1	4	10225865	10317419
FCGRLOCUS	1	161354978	161905425	KLRK1	4	10372041	10390328
AITPRXSNP	1	161568309	161575047	LRP1PROM	4	56733472	56734937
AITDISINS	1	161648705	161650096	LRP1	4	57128150	57213503
AITDISSNP	1	161650139	161659148	IFNGPROM	4	67759810	67761248
AXDND1SNP	1	179550834	179551844	IFNG	4	68154665	68160032
CFH	1	196651700	196747758	FREM2SNP	5	38858925	38859900
CFHPROM	1	196681697	196682900	L2HGDHSNP	6	50302457	50303227
IL10PROM	1	206593157	206594299	SLC12A6SNP	7	34236317	34237253
IL10	1	206767233	206772645	IL4RPROM	8	27301331	27302721
CD55PROM	1	207146946	207148326	IL4R	8	27313595	27364962
CD55	1	207321242	207361098	AARSSNP	8	70269217	70270128
CD46PROM	1	207577497	207578867	CCL3	9	36088166	36090525
CD46	1	207751700	207795675	CCL4	9	36103424	36105817
TGFB2PROM	1	218170767	218172057	CCL3PROM	9	37727068	37728249
TGFB2	1	218345225	218444869	CCL4PROM	9	37850825	37852270
IL1R1PROM	12	101452646	101454171	CCR7	9	40553494	40565627
IL1R1	12	102064356	102179996	CCR7PROM	9	42400568	42401877
IL1BPROM	12	112070957	112072318	PECAM1	9	64319261	64330031
IL1B	12	112829534	112837092	PECAM1PROM	9	66322103	66323518
CXCR4PROM	12	135355664	135356963	COG1SNP	9	73201174	73202273

Table 8.10 continued from previous page

CXCR4	12	136114143	136118511	LAMA3SNP	10	23833315	23834347
ABCB11SNP	12	168932393	168932859	MYO5BSNP	10	49929138	49930014
CD28PROM	12	202840612	202842010	C3PROM	11	6676402	6678643
CD28	12	203706270	203739305	C3	11	6678843	6730844
CXCR2PROM	12	217259418	217260666	DNMT1SNP	11	10156053	10157017
CXCR2	12	218125146	218137448	CALRPROM	11	12826509	12828091
COL4A4SNP	12	227031774	227032878	CALR	11	12938448	12944757
SUMF1SNP	16	4361491	4362695	IL12RB1PROM	11	17946964	17948365
CD47	16	108042675	108091340	IL12RB1	11	18059250	18087202
CD47PROM	16	108323155	108324561	TGFB1PROM	11	40823471	40824717
CD200	16	112333008	112362971	TGFB1	11	41330670	41354053
CD200PROM	16	112612970	112614352	LILRB3PROM	11	53711993	53713197
IL12A	16	159988449	159996268	LILRA6PROM	11	53732100	53733550
IL12APROM	16	160269618	160271379	LILRB5PROM	11	53746103	53747393
EVCSNP	17	5747613	5748875	LILRB2PROM	11	53769317	53770741
IL8PROM	17	72873692	72874902	LILRA5PROM	11	53802732	53804101
IL8	17	73740336	73744004	LILRA4PROM	11	53828991	53830354
CXCL13PROM	17	76589490	76590757	LILRA2PROM	11	54068589	54070014
CXCL13	17	77511513	77612089	LILRA1PROM	11	54089478	54090472
IL2PROM	17	121529202	121530623	LILRB1PROM	11	54112303	54113824
IL2	17	122451188	122456980	LILRB4PROM	11	54157977	54159296
C9PROM	18	39282959	39285559	LILRB3	11	54216222	54217742
C9	18	39288336	39425040	KIR2DS3	11	54224151	54233365
C7PROM	18	40908301	40910038	KIR2DL2PROM	11	54233540	54234318
C7	18	40928192	40983083	KIR2DL2	11	54234480	54285346
C6PROM	18	41140888	41143251	KIR2DS1	11	54285355	54341791
C6	18	41149108	41261535	KIR2DL4	11	54300286	54303754
IL6ST	18	55934778	55995143	KIR2DS5	11	54341811	54355871
IL6STPROM	18	56638218	56639432	LILRA2	11	54572638	54587664
VCANSNP	18	83538258	83539398	LILRA1	11	54593396	54602680
IRF1	18	132481377	132490993	LILRB1	11	54616716	54637868
IL13	18	132656015	132661285	LILRB4	11	54643586	54670867
IL4	18	132673692	132682856	KIR3DL1	11	54724084	54867357
IRF1PROM	18	133144651	133146021	KIR3DL3	11	54724961	54736637
IL13PROM	18	133321313	133322629	KIR2DL3	11	54738322	54764223
IL4PROM	18	133337078	133338367	KIR2DL1	11	54769710	54784603
IL12B	18	159314507	159331217	KIR2DS4	11	54832519	54848751
IL12BPROM	18	159886643	159887989	KIR3DL2	11	54850252	54859290
TNF	19	31575438	31578455	FERMT1SNP	13	6118851	6119971
TNFPROM	19	31606722	31608027	CD40	13	46117971	46130143
C2	19	31897421	31952226	CD40PROM	13	47488402	47489692
C2PROM	19	31929706	31930241	IL10RBPROM	14	31892749	31894232
CFBPROM	19	31977009	31978252	IFNGR2PROM	14	32029466	32030824
C4A	19	31981925	32035555	IL10RB	14	33266149	33297411
C4APROM	19	32012981	32014522	IFNGR2	14	33384719	33479672
C4B	19	32014663	32017596	NDUFV3SNP	14	42903125	42904069
FOXO3PROM	19	108237316	108238869	PI4KA	15	19972722	20859648
FOXO3	19	108559642	108685202	PI4KAPROM	15	20352289	20353182
IFNGR1PROM	19	136875301	136876130	PI4KASNP	15	20786555	20787659
IFNGR1	19	137197248	137219572	MIFPROM	15	23551244	23552305
GRM1SNP	19	146433494	146434657	MIF	15	23893680	23895474
IL6PROM	20	22686344	22687735	PIGAPROM	23	15300177	15301626
IL6	20	22725565	22732068	PIGA	23	15319373	15335731
ABCA13SNP	20	48409978	48410932	IL13RA2	23	115003819	115020248
PDP1SNP	21	93923254	93924213	IL13RA2PROM	23	115886657	115887162
CCL19PROM	22	34688316	34690564	IL13RA1	23	118727236	118794799
CCL19	22	34690813	34691562	IL13RA1PROM	23	119592507	119593974

Table 8.10: HaloPlex capture kit targeted genes and promoters. 130 genes were targeted and 97 gene promoters. All promoters are annotated with the "prom" suffix and SNPs of interest with the"SNP" suffix.

8.4.2 Initial mapping statistics for HaloPlex samples

Sample	Batch	Total Reads	Mapped	Mapped (%)	Properly paired	Singletons
CRA-104	1	2,090,394	2,031,358	97.18	2,005,022	7,540
CRA-101	1	2,674,347	2,613,104	97.71	2,580,003	8,517
CRA-111	1	3,462,375	3,390,935	97.94	3,349,379	11,074
CRA-74	1	3,779,498	3,719,871	98.42	3,679,657	9,313
CRA-90	1	6,060,104	5,965,546	98.44	5,898,535	17,301
CRA-142	1	945,137	931,927	98.6	922,052	2,335
CRA-96	1	3,658,038	3,610,318	98.7	3,571,499	9,368
CRA-60	1	2,463,846	2,431,918	98.7	2,396,832	6,500
CRA-67	1	2,042,024	2,016,401	98.75	1,983,980	4,812
CRA-149	1	2,240,143	2,214,449	98.85	2,189,556	5,142
CRA-108	1	3,144,076	3,110,427	98.93	3,077,947	7,064
CRA-134	1	2,895,982	2,865,465	98.95	2,828,686	5,357
CRA-137	1	656,738	649,876	98.96	642,234	1,646
CRA-112	1	1,328,697	1,314,955	98.97	1,301,350	3,051
CRA-138	1	2,030,053	2,009,323	98.98	1,986,775	4,772
CRA-103	1	1,887,400	1,868,287	98.99	1,851,672	3,758
CRA-53	1	2,030,241	2,014,153	99.21	1,985,978	4,554
CRA-25	1	2,323,417	2,306,271	99.26	2,274,319	5,048
CRA-51	1	3,232,614	3,210,729	99.32	3,167,844	7,744
CRA-151	1	2,307,777	2,294,026	99.4	2,267,561	3,404
CRA-54	1	2,213,076	2,200,099	99.41	2,172,083	4,004
CRA-148	1	2,407,135	2,394,723	99.48	2,368,852	3,335
CRA-154	1	4,273,470	4,252,142	99.5	4,210,952	5,541
CRA-106	1	2,062,577	2,052,274	99.5	2,031,946	4,141
CRA-147	1	5,203,365	5,177,856	99.51	5,137,764	6,728
CRA-57	1	2,639,833	2,627,206	99.52	2,596,426	4,047
CRA-162	1	2,455,003	2,444,448	99.57	2,416,403	2,878
CRA-153	1	1,917,864	1,909,694	99.57	1,888,955	2,368
CRA-118	1	2,000,811	1,992,669	99.59	1,971,607	3,263
CRA-76	1	1,571,549	1,565,334	99.6	1,549,594	2,937
CRA-158	1	2,879,109	2,868,131	99.62	2,839,353	3,514
PAC0089	2	799,546	795,154	99.45	787,704	512
PAC0016	2	1,222,338	1,215,752	99.46	1,205,692	692
CRA-120	2	1,695,269	1,687,329	99.53	1,674,832	1,039
CRA-41	2	3,858,105	3,842,039	99.58	3,810,393	3,422
CRA-58	2	3,350,726	3,336,505	99.58	3,308,517	2,804
CRA-78	2	1,611,133	1,604,932	99.62	1,591,371	986
CRA-183	2	1,305,521	1,300,651	99.63	1,291,381	1,249
PAC0133	2	969,326	965,792	99.64	953,706	393
CRA-156	2	2,065,319	2,057,981	99.64	2,043,616	1,020
CRA-102	2	1,632,601	1,626,688	99.64	1,614,006	871
CRA-119	2	1,902,668	1,895,811	99.64	1,879,583	927
CRA-135	2	2,120,898	2,113,366	99.64	2,097,174	1,014
CRA-30	2	3,113,932	3,102,922	99.65	3,076,242	2,733
PAC0088	2	1,160,707	1,156,722	99.66	1,147,968	638
CRA-129	2	642,630	640,428	99.66	634,634	294
CRA-180	2	1,157,362	1,153,508	99.67	1,145,034	519
PAC0087	2	2,499,063	2,491,116	99.68	2,471,826	1,735
CRA-64	2	2,103,432	2,097,035	99.7	2,081,546	950
CRA-109	2	2,294,226	2,287,427	99.7	2,269,573	1,110
699	2	5,112,927	5,097,551	99.7	5,064,024	3,210
CRA-11	2	2,245,940	2,239,265	99.7	2,219,918	1,792
CRA-18	2	1,985,006	1,979,208	99.71	1,961,313	1,593
CRA-55	2	4,235,074	4,222,703	99.71	4,188,848	3,441
CRA-63	2	2,408,959	2,402,004	99.71	2,386,495	1,072
CRA-177	2	2,929,908	2,921,399	99.71	2,900,008	1,180
CRA-175	2	2,991,451	2,982,915	99.71	2,959,503	1,324
CRA-47	2	3,984,550	3,973,542	99.72	3,942,496	3,336
PAC0098	2	4,428,806	4,416,410	99.72	4,388,489	2,614

Table 8.11 continued from previous page

Sample	Batch	Total Reads	Mapped	Mapped percent	Properly paired	Singletons
CRA-35	2	4,982,481	4,968,813	99.73	4,926,039	3,982
CRA-66	2	2,247,993	2,241,987	99.73	2,228,363	976
CRA-161	2	862,808	860,507	99.73	852,745	364
PAC0025	2	2,194,167	2,188,202	99.73	2,174,020	951
CRA-72	2	2,276,361	2,270,523	99.74	2,253,657	1,007
PAC0128	2	3,623,920	3,614,665	99.74	3,587,275	1,710
PAC0002	2	779,439	777,426	99.74	772,787	432
CRA-46	2	2,661,014	2,654,401	99.75	2,634,161	2,008
602	2	1,876,321	1,871,803	99.76	1,863,846	720
PAC0032	2	2,135,496	2,130,750	99.78	2,120,023	853

Table 8.11: HaloPlex initial mapping statistic per sample sorted by percentage of mapped read in ascending order.

8.4.3 Estimating contamination of HaloPlex samples

Sample	Freemix	Contamination Estimate(pc)
602	0.00006	0.006
699	0.00008	0.008
CRA-101	0.00014	0.014
CRA-102	0.00014	0.014
CRA-103	0.00038	0.038
CRA-104	0.00001	0.001
CRA-106	0.00007	0.007
CRA-108	0.00078	0.078
CRA-109	0.00046	0.046
CRA-11	0.00032	0.032
CRA-111	0	0
CRA-112	0.00111	0.111
CRA-118	0	0
CRA-119	0	0
CRA-120	0.0001	0.01
CRA-129	0.00002	0.002
CRA-134	0	0
CRA-135	0.00034	0.034
CRA-137	0.00519	0.519
CRA-138	0.00015	0.015
CRA-142	0.00039	0.039
CRA-147	0.00003	0.003
CRA-148	0.00022	0.022
CRA-149	0.00121	0.121
CRA-151	0.00334	0.334
CRA-153	0.00401	0.401
CRA-154	0.00131	0.131
CRA-156	0	0
CRA-158	0.00098	0.098
CRA-161	0.00029	0.029
CRA-162	0.00111	0.111
CRA-175	0.00017	0.017
CRA-177	0.00007	0.007
CRA-18	0.00004	0.004
CRA-180	0.00248	0.248
CRA-183	0	0
CRA-25	0.00186	0.186
CRA-30	0.00016	0.016
CRA-35	0.0001	0.01
CRA-41	0.00011	0.011
CRA-46	0.00009	0.009

Table 8.12 continued from previous page

Sample	Freemix	Contamination Estimate(pc)
CRA-47	0.00149	0.149
CRA-51	0.00016	0.016
CRA-53	0.00014	0.014
CRA-54	0.00018	0.018
CRA-55	0.00251	0.251
CRA-57	0.00011	0.011
CRA-58	0.00011	0.011
CRA-60	0.00011	0.011
CRA-63	0.00013	0.013
CRA-64	0	0
CRA-66	0.00008	0.008
CRA-67	0.00009	0.009
CRA-72	0	0
CRA-74	0.00004	0.004
CRA-76	0.00002	0.002
CRA-78	0.00003	0.003
CRA-90	0.00057	0.057
CRA-96	0.00056	0.056
PAC0002	0.00015	0.015
PAC0016	0.00007	0.007
PAC0025	0.00015	0.015
PAC0032	0	0
PAC0087	0	0
PAC0088	0.00011	0.011
PAC0089	0.00017	0.017
PAC0098	0	0
PAC0128	0.00008	0.008
PAC0133	0.00025	0.025

Table 8.12: VerifyBamID results for HaloPlex and HaloPlex-HS samples

8.4.4 Merging of overlapping reads using Pear

Sample	Batch	Assembled reads	Assembled reads (pc)	Unassembled reads	Unassembled reads (pc)
CRA-101	1	1,274,682	96.78	42,448	3.22
CRA-103	1	910,473	98.09	17,707	1.91
CRA-104	1	987,762	96.09	40,191	3.91
CRA-106	1	1,004,363	99.3	7,076	0.7
CRA-108	1	1,519,452	98.2	27,810	1.8
CRA-111	1	1,653,683	97.04	50,459	2.96
CRA-112	1	638,707	97.95	13,392	2.05
CRA-118	1	975,415	99.37	6,225	0.63
CRA-134	1	1,391,474	98.48	21,551	1.53
CRA-137	1	316,086	98.08	6,194	1.92
CRA-138	1	975,552	97.99	20,059	2.02
CRA-142	1	451,211	97.42	11,940	2.58
CRA-147	1	2,530,391	99.26	18,768	0.74
CRA-148	1	1,166,675	99.11	10,540	0.9
CRA-149	1	1,080,425	98.37	17,853	1.63
CRA-151	1	1,112,316	98.99	11,409	1.02
CRA-153	1	929,960	99.19	7,589	0.81
CRA-154	1	2,074,470	99.15	17,809	0.85
CRA-158	1	1,398,318	99.3	9,896	0.7
CRA-162	1	1,187,270	99.17	9,994	0.84
CRA-25	1	1,127,863	98.97	11,742	1.03
CRA-51	1	1,568,012	99.03	15,358	0.97
CRA-53	1	983,177	98.89	10,999	1.11
CRA-54	1	1,076,184	99.15	9,209	0.85
CRA-57	1	1,286,105	99.32	8,828	0.68
CRA-60	1	1,186,220	98.38	19,482	1.62
CRA-67	1	981,426	98.33	16,630	1.67
CRA-74	1	1,816,297	97.61	44,506	2.39
CRA-76	1	767,848	99.47	4,073	0.53
CRA-90	1	2,911,781	97.61	71,214	2.39
CRA-96	1	1,762,558	97.88	38,200	2.12
602	2	867,899	93.19	63,454	6.81
699	2	2,356,849	93.25	170,706	6.75
CRA-102	2	753,792	93.54	52,085	6.46
CRA-109	2	1,059,831	93.58	72,760	6.42
CRA-11	2	1,099,418	99.59	4,485	0.41
CRA-119	2	880,816	93.83	57,932	6.17
CRA-120	2	784,130	93.51	54,426	6.49
CRA-129	2	300,416	94.64	17,013	5.36
CRA-135	2	978,355	93.38	69,361	6.62
CRA-156	2	949,545	93.06	70,826	6.94
CRA-161	2	403,906	94.82	22,086	5.19
CRA-175	2	1,378,953	93.44	96,770	6.56
CRA-177	2	1,353,833	93.55	93,321	6.45
CRA-18	2	971,182	99.59	4,036	0.41
CRA-180	2	533,042	93.26	38,513	6.74
CRA-183	2	599,467	92.98	45,258	7.02
CRA-30	2	1,522,300	99.54	7,055	0.46
CRA-35	2	2,439,870	99.61	9,569	0.39
CRA-41	2	1,886,522	99.46	10,194	0.54
CRA-46	2	1,302,747	99.64	4,777	0.37
CRA-47	2	1,953,342	99.62	7,498	0.38
CRA-55	2	2,073,151	99.57	8,914	0.43
CRA-58	2	1,638,330	99.48	8,648	0.53
CRA-63	2	1,110,609	93.27	80,165	6.73
CRA-64	2	973,192	93.65	66,014	6.35
CRA-66	2	1,033,730	93.16	75,944	6.84
CRA-72	2	1,055,026	93.79	69,883	6.21
CRA-78	2	749,282	94.11	46,895	5.89

Table 8.13 continued from previous page

Sample	Batch	Assembled reads	Assembled reads (pc)	Unassembled reads	Unassembled reads (pc)
PAC0002	2	362,515	93.83	23,850	6.17
PAC0016	2	570,985	94.57	32,759	5.43
PAC0025	2	1,027,307	94.49	59,926	5.51
PAC0032	2	980,504	92.69	77,357	7.31
PAC0087	2	1,174,143	95	61,798	5
PAC0088	2	544,932	94.83	29,683	5.17
PAC0089	2	379,607	96.07	15,541	3.93
PAC0098	2	2,068,916	94.22	126,978	5.78
PAC0128	2	1,711,538	95.26	85,224	4.74
PAC0133	2	457,539	95.92	19,471	4.08

Table 8.13: Percentages and totals of reads successfully merged based on overlaps between pairs per HaloPlex sample. Unassembled reads are those either mapped as singletons or with insufficient overlap between pairs.

8.4.5 HaloPlex mean sample coverage

Sample	Mean	Bases > 1x	Bases > 5	Bases > 10	Bases > 20	Bases > 30	Bases > 50
CRA-129	17.72	93.4	78.1	61.7	34.9	18.2	4.9
PAC0133	19.91	91.8	74.8	58.1	35.1	21.2	8.5
CRA-137	23.43	95	83.8	71	48.4	30.7	10.1
CRA-161	25.32	94.9	83.8	71.1	49.2	31.7	12
PAC0089	26.96	91.2	75.7	62.1	42.7	29.1	14.9
PAC0002	31.88	92.4	80.5	69.8	52.6	39	20.5
CRA-142	36.55	96.3	89.5	81	64.2	49.5	26.3
PAC0088	37.74	92	81.5	72.3	58.4	46.6	27.8
CRA-180	38.12	96.3	89	80.7	66	52.1	29.1
PAC0016	43.56	94.5	85.3	76.4	61.8	48.8	29.9
CRA-183	47.44	96.2	89.9	83	70.8	59.3	38.8
CRA-112	50.16	96.8	91.1	84.2	71.2	59	38.1
CRA-67	51.08	97.2	91.8	85.1	72.3	60.9	41.1
CRA-102	51.68	96.4	91.6	85.4	73.7	62.6	42
CRA-120	51.72	96.4	91.2	85.3	73.7	62.9	43
CRA-78	54.07	96	90.4	83.9	72.1	61.8	42.5
CRA-119	58.48	96.4	91.5	85.3	74.3	64.4	46
CRA-76	62.54	97	92	86.8	76.2	66.6	49.3
CRA-153	64.1	97.8	93.9	89.4	80	70.4	52.5
CRA-53	66.37	97.5	92.9	87.7	77.5	68.1	51.7
CRA-18	66.39	97.6	92.7	87	77.4	68.8	52.9
CRA-135	66.91	96.7	93.1	88.5	79.5	70.7	54
CRA-103	68.22	97.1	92.8	88	79.4	70.5	54.5
CRA-64	70.27	96.8	92.8	88.1	79.3	71.1	55.7
CRA-25	71.12	97.9	93.6	88.8	78.9	69.9	54.1
CRA-118	72.2	97.6	94	89.3	80.4	72	56
CRA-156	72.52	96.6	92.6	88.1	80	72.2	57.2
CRA-109	72.88	96.9	93.4	88.8	80.5	72.4	57.5
CRA-138	73.22	97.2	93.3	89.4	80.9	72.6	56.9
CRA-11	76.03	97.5	92.7	87.5	78.4	70.6	56.3
CRA-54	76.09	97.5	93.2	88.4	78.4	70.3	55.2
CRA-72	76.64	97.2	93.3	88.9	80.5	72.7	58.1
CRA-148	76.82	97.5	94.4	90.8	82.9	75.1	59.9
CRA-104	78.9	97.8	93.5	88.9	80.5	72.2	57.3
602	78.92	89.6	80.5	74.7	66.6	59.9	48.8
CRA-151	79.65	98.2	95.1	91.3	83.8	76.2	61.3
CRA-60	79.97	97.8	94.2	89.6	81.1	73	58.8
CRA-162	80.23	98	95	91.3	83.8	76.2	61.5
PAC0025	80.43	94.5	88.9	83.7	75.2	67.8	55
CRA-134	82.09	98.1	95.1	91.1	83.3	76	61.4
CRA-149	83.08	98.4	95.2	91.9	84.6	77.4	63.1
CRA-66	83.54	96.8	93.3	89	81.6	74.7	61.3
CRA-106	84.09	97.7	94.1	90.2	82.5	74.5	60.5
CRA-63	84.39	96.9	93.6	89.8	83.1	76.5	63.8
PAC0087	86.14	95.7	90.3	85.1	76.8	69.8	57.4
PAC0032	88.16	95.2	89.3	84.2	75.7	68.9	56.7
CRA-57	92.45	97.6	94.1	90.2	82.7	75.5	62.6
CRA-177	92.75	97	94.3	90.7	83.4	76.6	63.7
CRA-175	93.54	97.1	94.4	90.6	83.5	76.7	64
CRA-46	93.93	98.1	95.1	91.5	84.7	77.8	65.5
CRA-101	96.4	97.7	93.7	89.4	81.7	74.2	61.4
CRA-158	97.48	98.6	96.3	93.5	87.6	81.2	69.1
CRA-30	105.55	98	95.2	91.7	85.7	79.4	68.2
PAC0128	112.45	95	91.2	87.6	80.8	74.9	64.8
CRA-58	114.85	97.9	94.5	91	84.7	79.1	68.8
CRA-51	115.28	98.2	95.7	93	87.1	81.1	70.2
CRA-108	115.92	98	94.7	91.1	85.1	78.9	68.1
CRA-111	127.1	97.9	94.8	91.8	86.2	80.8	70.6
CRA-41	132.81	98.3	95.8	92.7	87.3	82.3	73.3

Table 8.14 continued from previous page

Sample	Mean	Bases > 1	Bases > 5	Bases > 10	Bases > 20	Bases > 30	Bases > 50
CRA-74	136.88	98.1	95.2	92.2	86.3	81.1	71.1
CRA-96	140.59	97.8	94.9	92.2	87.1	82.4	73
CRA-47	142.78	98.3	95.3	92.5	87.1	82.2	73.6
CRA-55	147.15	98.4	95.3	92.7	87.6	82.9	74.7
CRA-154	152.89	98.5	96.6	94.8	91.3	87.6	79.6
PAC0098	155.64	92.2	88.3	85.6	81.5	77.5	70.2
CRA-35	181.28	98.5	96.3	94	89.6	85.5	78.2
699	187	97.6	95.3	93.2	89.4	85.6	79.1
CRA-90	239.5	98.3	96.3	94.5	91	88.1	82.2
CRA-147	245.02	97.8	94.9	92.1	87.9	83.8	77.2

Table 8.14: HaloPlex mean coverage per sample and percentage of samples at: 1, 5, 10, 20, 30, 50x depth of coverage

8.4.6 Custom calling using per gene references to determine HNA-1 haplotypes

.	1	.	2	.	3	.	4	.	5	.	6	.	AAs	CNVkit	Prediction	MLPA CN	MLPA Hap.
Sample	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**Ref**	**Alt**	**#**	**#**	**#**	**#**	**#**
Amino Acid	**I**	**V**	**N**	**D**	**A**	**D**	**N**	**S**	**L**	**C/T**	**S**	**R**	**#**	**#**	**#**	**#**	**#**
Nucleotide	**T**	**C**	**T**	**C**	**G**	**T**	**T**	**C**	**A**	**G**	**G**	**C**	**Amino acids**	**CNVkit CN**	**Haplotypes predicted**	**MLPA CN**	**MPLA Haplotype**
602	93	0	39	54	70	23	0	440	331	312	331	312	INDSTS	2	SH+SH	N/A	N/A
699	494	136	44	208	256	0	90	378	150	421	150	421	VNANTS	2	AB	2	AB
CRA-101	96	0	50	47	97	0	0	263	125	141	125	141	INASTS	2	BB	N/A	BB
CRA-103	107	0	59	47	106	0	0	226	118	97	118	97	INASTS	2	BB	2	BB
CRA-106	345	0	50	115	168	0	0	321	126	170	126	171	INASTS	2	BB	2	BB
CRA-108	93	0	29	64	92	0	0	249	83	159	83	159	INASTS	1	B	1	B
CRA-11	201	63	12	28	42	0	30	112	53	135	53	135	VNANTS	2	AB	2	AB
CRA-112	199	43	17	90	112	0	88	173	35	216	35	216	VNANTS	3	AB+A/B	3	AAB
CRA-118	109	0	46	63	109	0	0	236	115	139	115	139	INASTS	2	BB	2	BB
CRA-119	218	32	52	41	80	13	30	194	133	112	133	112	VNDNTS	4	A+SH+A/SH+A/SH	3	BBBB
CRA-120	54	0	17	37	54	0	0	105	25	69	25	69	INASTS	1	B	1	B
CRA-129	42	0	12	9	21	0	16	31	24	23	24	24	INANTS	2	NA2*02+NA2*02	2	AB
CRA-134	121	0	61	59	121	0	1	284	154	138	154	138	INASTS	2	BB	2	BB
CRA-138	103	0	45	59	105	0	0	271	117	138	117	138	INASTS	2	BB	2	BB
CRA-147	760	186	107	338	454	0	106	527	143	199	143	199	VNANTS	3	AB+A/B	2	AB
CRA-148	139	0	60	79	139	0	0	237	97	98	97	98	INASTS	2	BB	2	BB
CRA-149	219	89	27	98	127	0	64	180	58	217	58	218	VNANTS	2	AB	2	AB
CRA-151	177	65	28	83	112	0	48	145	42	161	42	161	VNANTS	2	AB	2	AB
CRA-154	132	262	1	176	184	0	232	119	0	408	0	408	VDANCR	2	AA	1	A
CRA-158	241	93	36	89	130	0	60	186	64	198	64	198	VNANTS	2	AB	1	AB
CRA-161	60	26	7	17	24	0	13	42	16	42	16	42	VNANTS	2	AB	2	BB
CRA-162	245	60	54	68	125	0	40	176	98	139	98	139	VNANTS	2	AB	3	ABB
CRA-175	146	0	68	78	114	32	58	245	147	191	147	191	INDNTS	3	SH+ NA1*02 and/or NA2*02	2	AB
CRA-177	249	69	27	98	126	0	63	184	69	224	69	224	VNANTS	2	AB	3	AB
CRA-18	191	57	8	30	41	0	28	76	39	110	39	110	VNANTS	2	AB	2	AB
CRA-180	93	45	11	32	44	0	26	50	14	70	14	70	VNANTS	2	AB	2	AB
CRA-183	131	41	9	47	56	0	26	75	31	88	31	88	VNANTS	2	AB	3	AB
CRA-25	92	0	53	39	92	0	2	245	129	119	129	121	INASTS	2	BB	2	BB
CRA-30	101	0	57	44	101	0	0	238	148	127	148	127	INASTS	2	BB	2	BB
CRA-35	303	326	1	101	105	0	168	163	0	445	0	445	VDANCR	2	AA	2	AA
CRA-41	358	110	47	98	152	0	66	311	169	308	169	308	VNANTS	3	AB+A/B	3	ABB

Table 8.15 continued from previous page

.	1	.	2	.	3	.	4	.	5	.	6	.	AAs	CNVkit	Prediction	MLPA CN	MLPA Hap.
CRA-46	52	0	13	39	52	0	0	169	56	158	56	158	INASTS	2	BB	1	B
CRA-47	117	0	55	62	117	0	1	362	213	198	213	198	INASTS	2	BB	2	BB
CRA-51	277	94	37	114	153	0	91	284	112	295	113	295	VNANTS	2	AB	2	AB
CRA-53	101	60	0	71	72	0	50	88	0	149	0	149	VDANCR	1	A	1	A
CRA-54	282	62	23	114	114	23	117	161	66	215	66	216	VNDNTS	3	A+SH+A/SH	3	AAB
CRA-55	255	61	44	142	145	41	189	316	123	490	123	491	VNDNTS	2	A+SH	3	AAB
CRA-57	173	66	40	59	101	0	75	180	104	177	104	177	VNANTS	2	AB	2	AB
CRA-58	287	82	14	56	73	0	48	159	64	187	64	188	VNANTS	2	AB	2	AB
CRA-60	93	0	26	67	93	0	56	196	59	200	59	200	INANTS	2	NA2*02+NA2*02	2	AB
CRA-63	74	85	0	78	79	0	83	85	0	185	0	185	VDANCR	2	AA	2	AA
CRA-64	113	66	0	73	73	0	59	91	0	159	0	159	VDANCR	1	A	1	A
CRA-66	131	97	0	77	79	0	97	104	0	241	0	241	VDANCR	2	AA	2	AA
CRA-67	70	0	34	37	71	0	0	141	63	73	63	73	INASTS	2	BB	2	BB
CRA-72	167	72	13	82	100	0	50	150	50	165	50	165	VNANTS	2	AB	2	AB
CRA-74	198	0	81	115	158	39	118	442	200	369	203	369	INDNTS	3	SH+ NA1*02 and/or NA2*02	3	BB
CRA-76	97	0	50	48	98	0	0	194	95	95	95	95	INASTS	2	BB	2	BB
CRA-78	128	87	11	54	71	0	65	94	30	156	30	156	VNANTS	3	AB+A/B	3	AAB
CRA-90	533	193	57	212	296	0	144	418	129	447	129	447	VNANTS	2	AB	2	AB
CRA-96	155	0	54	100	155	0	0	299	95	212	93	214	INASTS	1	B	1	B
PAC0002	53	17	14	30	44	0	30	101	38	119	38	119	VNANTS	2	AB	2	AB
PAC0016	166	44	23	49	72	0	35	160	55	158	56	157	VNANTS	4	AB+A/B+A/B	3	ABB
PAC0025	141	38	25	78	104	0	65	190	69	234	69	234	VNANTS	2	AB	2	AB
PAC0032	233	52	52	96	149	0	85	354	238	311	238	311	VNANTS	3	AB+A/B	4	ABB
PAC0087	34	0	30	31	61	0	0	139	75	77	75	77	INASTS	1	B	1	BB
PAC0088	55	0	11	19	30	0	0	91	31	82	31	82	INASTS	2	BB	?	AB
PAC0089	118	55	6	11	18	0	17	41	21	58	21	58	VNANTS	5	AB+A/B+A/B+A/B	4	AABB
PAC0098	158	0	143	146	289	0	0	559	335	368	333	368	INASTS	2	BB	2	AA
PAC0128	345	102	36	79	116	0	68	237	83	291	83	291	VNANTS	2	AB	2	AB
PAC0133	167	24	8	3	11	0	5	23	20	19	20	19	VNANTS	5	AB+A/B+A/B+A/B	4	AABB

Table 8.15: HNA-1 haplotypes predicted from using single gene references

8.4.7 Custom calling using per gene references to determine *FCGR2C* X57Q variants

Sample	HALO FCGR2C	T	C	HALO CN 2C	HALO CN 2B	MLPA CN	MLPA	HALO	Comment
602	0/1	74	435	2	2	N/A	N/A	XX	Cannot tell XQ from XX
CRA-108	1/1	0	346	1	2	1	Q	Q	N/A
CRA-57	1/1	0	277	1	1,2	1	Q	Q	N/A
CRA-96	1/1	0	349	1	2	1	Q	Q	N/A
CRA-183	0/1	55	110	2	2	3?	QQ	XX	Cannot tell XQ from XX
CRA-47	1/1	0	607	2	2	2	QQ	QQ	N/A
PAC0128	1/1	0	459	2	2	2	QQ	QQ	N/A
CRA-54	0/1	38	239	2,3	3,2	3	QQQ	XX,XXX	Cannot tell XQ from XX
CRA-55	1/1	0	774	2	2	3	QQQ	QQ	N/A
CRA-74	1/1	0	457	2,3	2	3	QQQ	QQ,QQQ	N/A
CRA-120	0/1	42	134	2,1	2	1	X	XX,X	Cannot tell XQ from XX
CRA-154	0/1	90	365	1,2	2	1	X	X,XX	Cannot tell XQ from XX
CRA-158	0/1	113	222	2	2	1	X	XX	Cannot tell XQ from XX
CRA-46	0/1	66	245	2	2	1	X	XX	Cannot tell XQ from XX
CRA-53	0/1	57	107	1	2	1	X	X	Cannot tell XQ from XX
CRA-64	0/1	32	140	2,1	2	1	X	XX,X	Cannot tell XQ from XX
CRA-11	0/1	39	262	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-118	0/1	47	264	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-147	0/1	167	178	3	3,2	2	XQ	XXX	Cannot tell XQ from XX
CRA-148	0/1	37	169	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-161	0/1	29	51	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-177	0/1	50	330	2	2	3	XQ	XX	Cannot tell XQ from XX
CRA-18	0/1	28	234	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-30	0/1	146	269	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-35	0/1	114	615	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-51	0/1	68	320	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-58	0/1	66	344	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-72	0/1	52	205	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-76	0/1	33	153	2	2	2	XQ	XX	Cannot tell XQ from XX
CRA-101	0/1	102	185	2	2	N/A	XX	XX	Cannot tell XQ from XX
CRA-103	0/1	64	108	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-134	0/1	107	196	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-138	0/1	60	160	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-149	0/1	102	221	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-151	0/1	81	145	2	2	2	XX	XX	Cannot tell XQ from XX

Table 8.16 continued from previous page

Sample	HALO FCGR2C	T	C	HALO CN 2C	HALO CN 2B	MLPA CN	MLPA	HALO	Comment
CRA-175	0/1	49	405	3	3,2	2	XX	XXX	Cannot tell XQ from XX
CRA-180	0/1	46	83	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-25	0/1	89	157	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-60	0/1	51	215	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-63	0/1	79	223	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-66	0/1	97	196	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-67	0/1	32	150	2	2	2	XX	XX	Cannot tell XQ from XX
CRA-90	0/1	229	431	2	2	2	XX	XX	Cannot tell XQ from XX
PAC0025	0/1	110	204	2	2	2	XX	XX	Cannot tell XQ from XX
PAC0087	1/1	0	220	1	1,2	1	XX	Q	N/A
PAC0088	0/1	33	104	2	2	1	XX	XX	Cannot tell XQ from XX
CRA-129	0/1	14	31	2	2	2	XXQ	XX	Cannot tell XQ from XX
PAC0098	0/1	75	459	2	2	2	XXQ	XX	Cannot tell XQ from XX
699	0/1	310	442	3,2	3,2	3	XXX	XXX,XX	Cannot tell XQ from XX
CRA-106	0/1	113	166	2	2	3	XXX	XX	Cannot tell XQ from XX
CRA-112	0/1	68	91	3	3,2	4	XXX	XXX	Cannot tell XQ from XX
CRA-162	0/1	159	147	1,2	2	3	XXX	X,XX	Cannot tell XQ from XX
CRA-41	0/1	243	388	3	2	3	XXX	XXX	Cannot tell XQ from XX
CRA-78	0/1	63	114	3	3,2	3	XXX	XXX	Cannot tell XQ from XX
PAC0002	0/1	52	91	2	2	2	XXX	XX	Cannot tell XQ from XX
PAC0032	0/1	201	252	3	3,2	3	XXX	XXX	Cannot tell XQ from XX
CRA-119	0/1	119	137	3,4	3,2	3	XXXX	XXX,XXXX	Cannot tell XQ from XX
PAC0016	0/1	51	79	4	4,2	4	XXXX	XXXX	Cannot tell XQ from XX
PAC0089	0/1	41	35	5	5,4,2	4	XXXX	XXXXX	Cannot tell XQ from XX
PAC0133	0/1	24	11	5	5,2	4	XXXX	XXXXX	Cannot tell XQ from XX

Table 8.16: *FCGR2C* haplotype predictions compared with MLPA for the X57Q variant

8.4.8 Comparing HaloPlex - CNVkit references against MLPA CNV calls

	MLPA						
Sample	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B
602	3						
699		3	3	3	2	2	2
CRA-101							
CRA-103	2	2	2	2	2	2	2
CRA-106	3	3	3	3	2	2	2
CRA-108	2	2	2	1	1	1	2
CRA-11	2	2	2	2	2	2	2
CRA-112	3	3	3	4	3	3	2
CRA-118	2	2	2	2	2	2	2
CRA-119	2	2	2	3	3	3	2
CRA-120	2	2	2	1	1	1	2
CRA-129	2	2	2	2	2	2	2
CRA-134	2	2	2	2	2	2	2
CRA-138	2	2	2	2	2	2	2
CRA-147	2	2	2	2	2	2	2
CRA-148	2	2	2	2	2	2	2
CRA-149	2	2	2	2	2	2	2
CRA-151	2	2	2	2	2	2	2
CRA-154	2	2	2	1	1	1	2
CRA-158	2	2	2	1	1	1	2
CRA-161	2	2	2	2	2	2	2
CRA-162	2	2	2	3	3	3	2
CRA-175	2	2	2	2	2	2	2
CRA-177	2	2	2	3	3	3	2
CRA-18	2	2	2	2	2	2	2
CRA-180	2	2	2	2	2	2	2
CRA-183	2	2	2	3	3	3	2
CRA-25	2	2	2	2	2	2	2
CRA-30	2	2	2	2	2	2	2
CRA-35	2	2	2	2	2	2	2
CRA-41	2	2	2	3	3	3	2
CRA-46	2	2	2	1	1	1	2
CRA-47	2	2	2	2	2	2	2
CRA-51	2	2	2	2	2	2	2
CRA-53	2	2	2	1	1	1	2
CRA-54	2	2	2	3	3	3	2
CRA-55	2	2	2	3	3	3	2
CRA-57	1	1	1	1	2	2	2
CRA-58	2	2	2	2	2	2	2
CRA-60	2	2	2	2	2	2	2
CRA-63	2	2	2	2	2	2	2
CRA-64	2	2	2	1	1	1	2
CRA-66	2	2	2	2	2	2	2
CRA-67	2	2	2	2	2	2	2
CRA-72	2	2	2	2	2	2	2
CRA-74	2	2	2	3	3	3	2
CRA-76	2	2	2	2	2	2	2
CRA-78	2	2	2	3	3	3	2
CRA-90	2	2	2	2	2	2	2
CRA-96	2	2	2	1	1	1	2
PAC0002	2	2	2	2	2	2	2
PAC0016	3	3	3	4	3	3	2
PAC0025	2	2	2	2	2	2	2
PAC0032	2	2	2	3	3	4	3,2
PAC0087	1	1	1	1	1	1	1,2
PAC0088	2	2	2	1	1	?	3
PAC0089	2	2	2	4	4	4	2
PAC0098	2	2	2	2	2	2	2
PAC0128	2	2	2	2	2	2	2
PAC0133	2	2	2	4	4	4	2

	Haloplex CNVkit calls - Pooled Reference - 267 bp						
Sample	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B
602	2	2	2	2	2	2	2
699	3	3	3	3	3	3	3
CRA-101	2	2	2	2	2	2	2
CRA-103	2	2	2	2	2	2	2
CRA-106	2	2	2	2	2	2	2
CRA-108	2	2	2	1	1	1	1,2
CRA-11	2	2	2	2	2	2	2
CRA-112	2,4	4	4	4	4	4	4,2
CRA-118	2	2	2	2	2	2	2
CRA-119	3	3	3	4,5	5	5	4
CRA-120	2	2	2	1	1	1	2
CRA-129	2	2	2	2	2	2	2
CRA-134	2	2	2	2	2	2	2
CRA-138	2	2	2	2	2	2	2
CRA-147	2	2	2	4	4	4	4
CRA-148	2	2	2	2	2	2	2
CRA-149	2	2	2	2	2	2	2
CRA-151	2	2	2	2	2	2	2
CRA-154	2,1	1	1	1	2	2	2
CRA-158	2	2	2	2	2	2	2
CRA-161	2	2	2	2	2	2	2
CRA-162	3	3	3	3,4	4	4	2
CRA-175	3	3	3	4	4	4	4,2
CRA-177	2	2	2	2	2	2	2
CRA-18	2	2	2	2	2	2	2
CRA-180	2	2	2	2	2	2	2
CRA-183	2	2	2	2	2	2	2
CRA-25	2	2	2	2	2	2	2
CRA-30	2	2	2	2	2	2	2
CRA-35	2	2	2	2	2	2	2
CRA-41	2	2	2	2	2	2	2
CRA-46	2	2	2	2	2	2	2
CRA-47	2	2	2	2	2	2	2
CRA-51	2	2	2	2	2	2	2
CRA-53	2	2	2	1	1	1	2
CRA-54	2	2	2	2,4	4	4	2
CRA-55	2	2	2	2	2	2	2
CRA-57	2,1	1	1	1,2	2	2	2
CRA-58	2	2	2	2	2	2	2
CRA-60	2	2	2	2	2	2	2
CRA-63	2	2	2	2	2	2	2
CRA-64	2	2	2	2	2	2	2
CRA-66	2	2	2	2	2	2	2
CRA-67	2	2	2	2	2	2	2
CRA-72	2	2	2	2	2	2	2
CRA-74	2	2	2	2,4	4	4	3
CRA-76	2	2	2	2	2	2	2
CRA-78	2	2	2,3	3,4	4	4	4,2
CRA-90	2	2	2	3	3	3	3
CRA-96	2	2	2	1	1	1	2,3
PAC0002	3	3	3	3	3	3	3
PAC0016	2,5	5	5	5	5	5	5,2
PAC0025	2	2	2	2	2	2	2
PAC0032	2	2	2	4	4	4	4,2
PAC0087	2,1	1	1	1	1	1	1,2
PAC0088	2	2	2	2	2	2	2
PAC0089	2	2	4	6	6	6	6,3
PAC0098	2	2	2	2	2	2	2
PAC0128	2	2	2	2	2	2	2
PAC0133	3	3	3	5	5	5	5,2

	Haloplex CNVkit calls - CRA-25 Ref - 267 bp						
Sample	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B
602	2	2	2	2	2	2	2
699	2,4	4	4	4	2	2	2
CRA-101	2	2	2	2	2	2	2
CRA-103	2	2	2	2	2	2	2
CRA-106	2	2	2	2	2	2	2
CRA-108	2	2	2	2	2	2	2
CRA-11	2	2	2	2	2	2	2
CRA-112	2,4	4	4	4	4	4	4,2
CRA-118	2	2	2	2	2	2	2
CRA-119	2	2	2	5	5	5	5,2
CRA-120	2	2	2	2	2	2	2
CRA-129	3	3	3	3	3	3	3
CRA-134	2	2	2	2	2	2	2
CRA-138	2	2	2	2	2	2	2
CRA-147	2	2	2	4	4	4	4,2
CRA-148	2	2	2	2	2	2	2
CRA-149	2	2	2	2	2	2	2
CRA-151	2	2	2	2	2	2	2
CRA-154	2	2	2	2	2	2	2
CRA-158	2	2	2	2	2	2	2
CRA-161	2	2	2	2	2	2	2
CRA-162	2,4	3	3	3	3	3	3,2
CRA-175	2	2	2	4	4	4	4,2
CRA-177	2	2	2	2	2	2	2
CRA-18	2	2	2	2	2	2	2
CRA-180	2	2	2	2	2	2	2
CRA-183	2	2	2	2	2	2	2
CRA-25	3	3	3	3	3	3	3
CRA-30	2	2	2	2	2	2	2
CRA-35	2	2	2	2	2	2	2
CRA-41	2	2	2	3	3	3	3,2
CRA-46	2	2	2	2	2	2	2
CRA-47	2	2	2	2	2	2	2
CRA-51	2	2	2	2	2	2	2
CRA-53	2	2	2	2	2	2	2
CRA-54	2	2	2	2,3	3	3	3,2
CRA-55	2	2	2	2,3	3	3	3,2
CRA-57	2	1	1	1	2	2	2
CRA-58	2	2	2	2	2	2	2
CRA-60	2	2	2	2	2	2	2
CRA-63	2	2	2	2	2	2	2
CRA-64	2	2	2	2	2	2	2
CRA-66	2	2	2	2	2	2	2
CRA-67	2	2	2	2	2	2	2
CRA-72	2	2	2	2	2	2	2
CRA-74	3	3	3	3	3	3	3
CRA-76	2	2	2	2	2	2	2
CRA-78	3	3	3,4	4	4	4	4,2
CRA-90	2	2	2	2	2	2	2
CRA-96	2	2	2	2,1	1	1	2
PAC0002	2,4	4	4	4	4	4	4,2
PAC0016	2,5	5	5	5	5	5	5,2
PAC0025	2	2	2	2	2	2	2
PAC0032	2	2	2	4	4	4	4,2
PAC0087	2	2	2	2	2	2	2
PAC0088	2	2	2	2	2	2	2
PAC0089	2,4	4	4	6	6	6	6,2
PAC0098	2	2	2	2	2	2	2
PAC0128	2	2	2	2	2	2	2
PAC0133	2,3	3	3	5	5	5	5,2

	Haloplex CNVkit calls - Pooled Reference 150bp						
Sample	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B
602	2	2	2	2	2	2	2
699	2,3	3	3	3,2	2	2	3,2
CRA-101	2	2	2	2	2	2	2
CRA-103	2	2	2	2	2	2	2
CRA-106	2,3	3	3	2	2	2	2
CRA-108	2	2	2	1	1	1	2
CRA-11	2	2	2	2	2	2	2
CRA-112	2,3	3	3	3	3	3	3,2
CRA-118	2	2	2	2	2	2	2
CRA-119	2	2	2	3,4	4	4	3,2
CRA-120	2	2	2	2,1	1	1	2
CRA-129	2	2	2	2	2	2	2
CRA-134	2	2	2	2	2	2	2
CRA-138	2	2	2	2	2	2	2
CRA-147	2	2	2	3	3	3	3,2
CRA-148	2	2	2	2	2	2	2
CRA-149	2	2	2	2	2	2	2
CRA-151	2	2	2	2	2	2	2
CRA-154	2,1	1	1	1,2	2	2	2
CRA-158	2	2	2	2	2	2	2
CRA-161	2	2	2	2	2	2	2
CRA-162	2	2	2	2,3	3	3	2
CRA-175	2	2	2	3	3	3	3,2
CRA-177	2	2	2	2	2	2	2
CRA-18	2	2	2	2	2	2	2
CRA-180	2	2	2	2	2	2	2
CRA-183	2	2	2	2	2	2	2
CRA-25	2	2	2	2	2	2	2
CRA-30	2	2	2	2	2	2	2
CRA-35	2	2	2	2	2	2	2
CRA-41	2	2	2	3	3	3	2
CRA-46	2	2	2	2	2	2	2
CRA-47	2	2	2	2	2	2	2
CRA-51	2	2	2	2	2	2	2
CRA-53	2	2	2	1	1	1	2
CRA-54	2	2	2	2,3	3	3	3,2
CRA-55	2	2	2	2	2	2	2
CRA-57	2	1	1	1	2	2	1,2
CRA-58	2	2	2	2	2	2	2
CRA-60	2	2	2	2	2	2	2
CRA-63	2	2	2	2	2	2	2
CRA-64	2	2	2	2,1	1	1	2
CRA-66	2	2	2	2	2	2	2
CRA-67	2	2	2	2	2	2	2
CRA-72	2	2	2	2	2	2	2
CRA-74	2	2	2	2,3	3	3	2
CRA-76	2	2	2	2	2	2	2
CRA-78	2	2	2	3	3	3	3,2
CRA-90	2	2	2	2	2	2	2
CRA-96	2	2	2	1	1	1	2
PAC0002	2	2	2	2	2	2	2
PAC0016	2,4	4	4	4	4	4	4,2
PAC0025	2	2	2	2	2	2	2
PAC0032	2	2	2	3	3	3	3,2
PAC0087	2,1	1	1	1	1	1	1,2
PAC0088	2	2	2	2	2	2	2
PAC0089	2	2	4	5	5	5	5,4,2
PAC0098	2	2	2	2	2	2	2
PAC0128	2	2	2	2	2	2	2
PAC0133	2	2	2	5	5	5	5,2

	Haloplex CNVkit calls - CRA-25 Ref- 150bp						
Sample	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B
602	2	2	2	2	2	2	2
699	2,4	4	4	4,2	2	2	2
CRA-101	2	2	2	2	2	2	2
CRA-103	2	2	2	2	2	2	2
CRA-106	2,4	4	4	4,2	2	2	3,2
CRA-108	2	2	2	2,1	1	1	2
CRA-11	2	2	2	2	2	2	2
CRA-112	2,4	4	4	4	4,2	2	4,2
CRA-118	2	2	2	2	2	2	2
CRA-119	2	2	3	3,5	5	5	5,2
CRA-120	2	2	2	2,1	1	1	2
CRA-129	2	2	2	2	2	2	2
CRA-134	2	2	2	2	2	2	2
CRA-138	2	2	2	2	2	2	2
CRA-147	2	2	2	3	3	3	3,2
CRA-148	2	2	2	2	2	2	2
CRA-149	2	2	2	2	2	2	2
CRA-151	2	2	2	2	2	2	2
CRA-154	2	2	2	2	2	2	2
CRA-158	2	2	2	2	2	2	2
CRA-161	2	2	2	2	2	2	2
CRA-162	2,3	3	3	4	4	4	4,2
CRA-175	2	2	2	4	4	4	4,2
CRA-177	2	2	2	2	2	2	2
CRA-18	2	2	2	2	2	2	2
CRA-180	2	2	2	2	2	2	2
CRA-183	2	2	2	2	2	2	2
CRA-25	2	2	2	2	2	2	2
CRA-30	2	2	2	2	2	2	2
CRA-35	2	2	2	2	2	2	2
CRA-41	2	2	2	3	3	3	3,2
CRA-46	2	2	2	2	2	2	2
CRA-47	2	2	2	2	2	2	2
CRA-51	2	2	2	2	2	2	2
CRA-53	2	2	2	2,0	0	0	2
CRA-54	2,3	3	3	3	3	3	3,2
CRA-55	2	2	2	2,3	3	3	3,2
CRA-57	2	1	1	1,0,2	2	2	2
CRA-58	2	2	2	2	2	2	2
CRA-60	2	2	2	2	2	2	2
CRA-63	2	2	2	2	2	2	2
CRA-64	2	2	2	2,1	1	1	2
CRA-66	2	2	2	2	2	2	2
CRA-67	2	2	2	2	2	2	2
CRA-72	2	2	2	2	2	2	2
CRA-74	2	2	2	2	2	2	2
CRA-76	2	2	2	2	2	2	2
CRA-78	2,3	3	3	3	3	3	3,2
CRA-90	2	2	2	2	2	2	2
CRA-96	2	2	2	2,0,1	1	1	2
PAC0002	2,3	3	3	3	3	3	3,2
PAC0016	2,5	5	5	5	5	5	5,2
PAC0025	2	2	2	2	2	2	2
PAC0032	2	2	2	3	3	3	3,2
PAC0087	2,1	1	1	1	1	1	1,2
PAC0088	2	2	2	2	2	2	2
PAC0089	2,3	3	3	5	5	5	5,2
PAC0098	2	2	2	2	2	2	2
PAC0128	2	2	2	2	2	2	2
PAC0133	2	2	2,4	4	4	4	4,2

Figure 8.1: Comparison of CNVkit calls against MLPA data for different reference genomes in CNVkit. Matches between MLPA and CNVkit were highlighted in green, genes with multiple copy numbers called are highlighted in orange while mismatches remain unshaded. Greatest agreement was achieved using the pooled reference with 150bp bins.

8.4.9 Comparing HaloPlex - CNVkit segmentation algorithms against MLPA CNV calls

	MLPA							Pooled Reference 150bp -CBS							Pooled Reference 150bp -FLASSO							Pooled Reference 150bp -HAAR						
Sample	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B	FCGR2A	HSPA6	FCGR3A	FCGR2C	HSPA7	FCGR3B	FCGR2B
602	G							N	N	N	N	N	N	N	N	N	L	N	N	N	2, 1	L	0, 2	N	N	N	N	N
699		G	G	G	N	N	N	2,3	G	G	3, 2	N	N	3,2	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-101								N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-103	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-106	G	G	G	G	N	N	N	2,3	G	G	N	N	N	N	N	N	N	N	N	N	N	G	G	G	N	N	N	N
CRA-108	N	N	N	L	L	L	N	N	N	N	L	L	L	N	N	N	N	2, 0, 0, 2, 1	L	L	N	N	N	N	N	N	N	N
CRA-11	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-112	G	G	G	G	G	G	N	2,3	G	G	G	G	G	3,2	2,3	G	G	G	G	G	3, 2	N	N	N	N	N	N	N
CRA-118	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-119	N	N	N	G	G	G	N	N	N	N	G	G	G	3,2	N	N	N	3,4,2,4	G	G	3,2,2	N	N	N	N	N	N	N
CRA-120	N	N	N	L	L	L	N	N	N	N	2, 1	L	L	N	N	N	N	2,2,1	L	L	N	N	N	N	N	N	N	N
CRA-129	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-134	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-138	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-147	N	N	N	N	N	N	N	N	N	N	G	G	G	3,2	N	N	N	N	N	G	3, 2	N	N	N	G	G	G	G
CRA-148	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-149	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-151	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-154	N	N	N	L	L	L	N	2,1	L	L	1, 2	N	N	N	2,1	L	1,1,2	N	N	N	N	N	N	L	1,2	N	N	N
CRA-158	N	N	N	L	L	L	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-161	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-162	N	N	N	G	G	G	N	N	N	N	1, 2	G	G	N	N	N	N	2,3	N	N	N	N	N	N	N	N	N	N
CRA-175	N	N	N	N	N	N	N	N	N	N	G	G	G	3,2	N	N	N	2,3	G	G	3, 2	N	N	N	N	N	N	N
CRA-177	N	N	N	G	G	G	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-18	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-180	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-183	N	N	N	G	G	G	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-25	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-30	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-35	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-41	N	N	N	G	G	G	N	N	N	N	G	G	G	N	N	N	N	N	N	G	N	N	N	N	N	N	N	N
CRA-46	N	N	N	L	L	L	N	N	N	N	N	N	N	N	N	N	N	L	L	L	N	N	N	N	N	N	N	N
CRA-47	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-51	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-53	N	N	N	L	L	L	N	N	N	N	L	L	L	N	N	N	N	L	L	L	N	N	N	N	N	N	N	N
CRA-54	N	N	N	G	G	G	N	N	N	N	2, 3	G	G	3,2	N	N	N	2,3	G	G	3, 2	N	N	N	N	N	N	N
CRA-55	N	N	N	G	G	G	N	N	N	N	N	N	N	N	N	N	N	2,0,2,3	G	3,2	N	N	N	N	N	N	N	N
CRA-57	L	L	L	L	N	N	N	N	L	L	L	N	N	1,2	N	L	L	1, 2,0,0,0,2, 1,1	N	N	N	N	N	N	2,0,1,2	N	N	N
CRA-58	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-60	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-63	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-64	N	N	N	L	L	L	N	N	N	N	2, 1	L	L	N	N	N	N	2,1	L	L	N	N	N	N	N	N	N	N
CRA-66	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-67	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-72	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-74	N	N	N	G	G	G	N	N	N	N	2, 3	G	G	N	N	N	N	2, 0,0,2,4	G	G	3, 2	N	N	N	2,0, 2,0,3,2	N	N	N
CRA-76	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-78	N	N	N	G	G	G	N	N	N	N	G	G	G	3,2	N	N	2,3	G	G	N	N	N	N	N	N	N	N	N
CRA-90	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
CRA-96	L	N	N	L	L	L	N	N	N	N	L	L	L	N	N	N	N	2,2, 0,0,0,1, 1	L	1,2,1	N	N	N	N	2,0	N	N	N
PAC0002	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
PAC0016	G	G	G	G	G	G	N	2,4	G	G	G	G	G	4,2	N	N	G	G	G	G	N	N	N	N	N	N	N	N
PAC0025	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
PAC0032	N	N	N	G	G	G	3,2	N	N	N	G	G	G	3,2	N	N	N	2,3	G	G	N	N	N	N	N	N	N	N
PAC0087	L	L	L	L	L	L	1,2	2,1	L	L	L	L	L	1,2	2,1,1	L	L	1,1,0,2,1	L	L	N	L	L	L	1,1,1,0,1,0,2	N	N	N
PAC0088	N	N	N	L	L	?	G	N	N	N	N	N	N	N	N	N	N	N	L	N	N	N	N	N	N	N	N	N
PAC0089	N	N	N	G	G	G	N	N	N	G	G	G	G	5,4,2	N	N	N	G	G	G	N	N	N	N	N	N	N	N
PAC0098	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
PAC0128	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	2,0,3,0,1,2	N	N	N
PAC0133	N	N	N	G	G	G	N	N	N	N	G	G	G	5,2	N	N	N	G	G	G	4, 2	N	N	G	G	G	G	G

Figure 8.2: Comparison of CNVkit calls against MLPA data for different segmentation algorithms fused lasso and Haar against the default circular binary segmentation when using a pooled reference with bin size of 150bp. While some samples and genes are called better using a fused lasso segmentation others are also over-called particularly in FCGR2C making gains lost by more ambigous and increased noise in others. Haar segmentation

8.4.10 HaloPlex - grouping CNVkit calls with known CNRs in the FCGR locus

Sample	Start	End	Copy Number	Region (CNR)
699	161514732	161596980	3	2
CRA-106	161514434	161568118	3	2
CRA-108	161574747	161666938	1	4
CRA-111	161574747	161660321	1	4
CRA-112	161513838	161670658	3	6
CRA-119	161575716	161670809	4	4
CRA-120	161596830	161655372	1	4
CRA-135	161574747	161659701	1	4
CRA-147	161559016	161669128	3	4
CRA-153	161592035	161657486	3	4
CRA-154	161515031	161596680	1	2
CRA-162	161589488	161656278	3	4
CRA-175	161575716	161675797	3	4
CRA-41	161580721	161658090	3	4
CRA-53	161564172	161637469	1	1
CRA-54	161591586	161672925	3	4
CRA-57	161518582	161599992	1	2
CRA-64	161593983	161653106	1	4
CRA-74	161591586	161653106	4	4
CRA-78	161586023	161670809	3	4
CRA-96	161574747	161661673	1	4
PAC0016	161511701	161676704	4	6
PAC0032	161579858	161671111	3	4
PAC0087	161515180	161670658	1	6
PAC0089	161540482	161676855	5	5
PAC0133	161428302	161670809	5	6

Table 8.17: Haloplex samples with copy number variants in the FCGR locus. Shown are the copy number calls with start and end locations combined with the copy number region the variant was grouped under.