

**University of Southampton Research Repository**

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Jed Lye (2022) "Transcriptomics for Investigating Immunodeficiencies of Mendelian and Age-related Origins ", University of Southampton, Faculty of Medicine, PhD Thesis, pagination.

Data: Jed Lye (2021) Title. URI [dataset]

**University of Southampton**

Faculty of Medicine

School of Human Development and Health

**Transcriptomics for Investigating Immunodeficiencies of Mendelian and Age-related**

**Origins**

DOI 22/09/2022

VOLUME 1 OF 1

by

**Jed Lye**

Thesis for the degree of Doctor of Philosophy

# Table of contents

<b>Table of contents</b> .....	<b>i</b>
<b>Table of tables</b> .....	<b>ix</b>
<b>Table of figures</b> .....	<b>xi</b>
<b>List of Abbreviations</b> .....	<b>xviii</b>
<b>Acknowledgements</b> .....	<b>1</b>
<b>Research thesis: declaration of authorship</b> .....	<b>2</b>
<b>Foreword and COVID19 impact statement</b> .....	<b>4</b>
<b>Chapter 1 Introduction</b> .....	<b>1</b>
1.1 The human adaptive immune system.....	1
1.1.1 An introduction to T-cells and humoral and cell-mediated immunity .....	1
1.1.2 The specificity and memory of T-cells in the adaptive immune system .....	5
1.1.3 RNA splicing and its role in the immune system.....	6
1.1.4 Alternative splicing in T-Cells .....	11
1.2 Primary immunodeficiencies – T-cell disorders.....	12
1.2.1 Severe combined immunodeficiency.....	15
1.2.2 X-linked lymphoproliferative disease .....	17
1.2.3 X linked immune deficiency with associated hyper IgM .....	17
1.2.4 The DiGeorge syndrome .....	18
1.2.5 Ataxia Telangectasia.....	19
1.2.6 Other types of genetic variant in PID.....	20
1.2.7 Variants affecting alternative splicing and their role in PID .....	21
1.2.8 Diagnostic challenges in primary immunodeficiencies.....	23
1.2.8.1 DNA in diagnostics .....	24
1.2.8.2 RNA in diagnostics.....	25
1.2.8.3 RNASeq parameters.....	28
1.2.8.4 RNAseq analysis techniques.....	29

1.2.8.5	Considerations in diagnostics .....	30
1.2.9	Hypothesis-free approaches in RNAseq diagnostics.....	32
1.3	Secondary Immunodeficiencies – Immunosenescence .....	33
1.3.1	Epidemiology and Demographics .....	33
1.3.2	A general description of Immune senescence.....	34
1.3.3	The effect of immunosenescence on viral infection and vaccinations.....	36
1.3.4	Cellular and molecular drivers and features of Immunosenescence .....	36
1.3.5	Alternative splicing and immunosenescence .....	38
1.3.6	Using transcriptomics to asses immunosenescence in COVID19 and Influenza39	
1.3.6.1	Ageing Clocks to quantify or diagnose immunosenescence.....	40
1.3.7	Summary .....	41
1.3.8	Aims of research.....	44
1.3.9	Key research questions pertaining to T-cell Primary Immunodeficiencies .....	45
1.3.10	Key research objectives pertaining to Immunosenescence as a Secondary Immunodeficiency.....	45
<b>Chapter 2</b>	<b>Methods .....</b>	<b>47</b>
2.1	Primary immunodeficiency .....	47
2.1.1	Patient enrolment and patient data collection .....	47
2.1.2	PID patient clinical phenotype .....	48
2.1.3	Whole blood and PBMC collection .....	50
2.1.4	Whole blood RNA extraction methods .....	50
2.1.5	RNA extraction QC .....	52
2.1.6	RNA Sequencing by Novogene.....	54
2.1.7	RNA quality control: Novogene .....	54
2.1.8	Library construction and sequencing.....	54
2.1.9	Whole blood control data – The Genome Tissue Expression Consortium .....	57
2.1.10	Whole blood control data - Splicing and Disease Cohort .....	58
2.1.11	Data processing.....	60
2.1.12	Trimming of reads and Quality control .....	61

2.1.13 Alignment of reads.....	62
2.1.14 Read Counts .....	62
2.1.15 Gene panel .....	63
2.1.16 Exploratory Analysis of RNAseq data.....	64
2.1.17 Differential gene expression .....	65
2.1.17.1 Gene expression Z-score calculation .....	65
2.1.17.2 OUTRIDER.....	65
2.1.18 Aberrant splicing - Mendelian RNAseq program .....	68
2.1.18.1 Splice junction discovery, normalisation and junction filtering.....	69
2.2 Immunosenescence investigation.....	71
2.2.1 Patient enrolment and patient data collection .....	71
2.2.2 Data processing.....	72
2.2.3 RNA extraction and sequencing.....	72
2.2.4 Trimming and Alignment with STAR – Performed by Dr. J. Lord.....	73
2.2.5 PcaExplorer .....	73
2.2.6 Pandaomics – Differential Gene expression between infections.....	74
2.2.7 Transcript Counts – Salmon Tool .....	74
2.2.8 DTU use between infections – BANDITS Tool.....	75
2.2.9 GO analysis of results.....	76
2.2.10 Multiple regression performed using Python .....	76
2.2.11 Classification .....	78
2.2.12 Lasso Regression for age prediction – Coding performed by Yaron Strauch. .	79
2.2.13 Analysis of features.....	80
<b>Chapter 3 Results: Whole-blood RNAseq investigation into the transcriptome of patients with primary immunodeficiencies –gene expression .....</b>	<b>81</b>
3.1 Introduction .....	81
3.2 Result from the quality control steps .....	81
3.2.1.1 RNAseq read alignment results.....	83
3.3 Exploratory data analysis .....	84

3.4	Outliers in gene expression: Using FPKM values and Z-scores.....	87
3.5	Splicing and Disease Cohort.....	94
3.5.1	Exploratory data analysis of dataset including ‘Splicing and Disease’ Cohort.....	94
3.5.2	Z-score results after ComBat-seq Batch correction and TPM calculation.....	106
3.5.3	OUTRIDER results.....	110
3.5.4	Pre-processing using OUTRIDER.....	112
3.5.5	Results for OUTRIDER analysis of outlier gene expression.....	115
3.5.5.1	Gene expression outlier results for patient SRB0017.....	117
3.5.5.2	Gene expression outlier results for patient SRB0012.....	118
3.5.5.3	Gene expression outlier results for patient SRB0006.....	118
3.6	Summary Table.....	124
3.7	Discussion.....	125
3.8	Conclusion.....	128
<b>Chapter 4</b>	<b>Results of Splicing Analysis of Primary Immunodeficiency Patients.....</b>	<b>130</b>
4.1	Introduction.....	130
4.2	Validation of the ‘Mendelian RNAseq’ splicing detection program using sample SOT58.....	130
4.3	Testing the Mendelian RNAseq tool on the first patient sample.....	133
4.3.1	Results from SRB003 Splicing analysis.....	133
4.3.1.1	Results of filter 1.....	133
4.3.1.2	Results of filter 2.....	134
4.3.2	Interrogation of events with IGV.....	135
4.4	PID cohort splicing analysis.....	142
4.4.1	Investigation of events in sample SRB0002.....	145
4.4.2	Investigation of events in sample SRB0005.....	148
4.4.3	Investigation of events in sample SRB006.....	151
4.4.4	Investigation of events in sample SRB0013.....	152
4.4.5	Investigation of events in sample SRB0014.....	156

4.4.6	Investigation of events in sample SRB0018 .....	160
4.4.7	Investigation of events in sample SRB0019 .....	163
	Summary of results.....	167
4.5	Discussion.....	169
4.6	Conclusion .....	170
<b>Chapter 5 Understanding transcriptomic differences in COVID-19 and Influenza: Results</b>		
<b>172</b>		
5.1	Introduction .....	172
5.2	Aims.....	172
5.3	Cohort analysis.....	172
5.4	Exploratory Data Analysis with PcaExplorer .....	177
5.5	Comparing and contrasting gene expression between COVID-19 and Influenza..	179
5.6	Comparing and contrasting isoform expression between COVID-19 and Influenza	195
5.7	Analysis of differential gene expression and differential transcript use in Covid19 and Influenza.....	204
5.8	Discussion.....	211
5.8.1	Findings .....	211
<b>Chapter 6 The Effect of Ageing on Host Transcriptomic Profiles during Viral Infection</b>		
<b>216</b>		
6.1	Introduction .....	216
6.1.1	Aims.....	217
6.2	Methods.....	218
6.3	Results.....	221
6.3.1	Exploratory data analysis .....	221
6.3.2	Results from multiple regression performed using Python.....	229
6.3.3	Volcano plots of transcriptome features association with age .....	233
6.3.4	Beta-coefficient distribution with age .....	240
6.3.5	Quantifying the groups of genes associated with ageing in the cohorts .....	243
6.3.6	Analysis of transcriptomic convergence results .....	246
6.3.7	Classification of infection based on gene expression .....	253

6.3.8	Application of classification learner linear support vector machine to split cohort	257
6.4	Discussion.....	258
6.4.1	Significance .....	258
6.4.2	Limitations.....	264
6.5	Conclusion .....	264
6.5.1	'DiCo' occurs at the level of immune response to infection also. ....	265
6.6	Statement of contributions.....	265
<b>Chapter 7 Results: Immune ageing and infection specific ageing clocks – machine learning application .....267</b>		
7.1	Introduction .....	267
7.1.1	Aims.....	268
7.2	Methods .....	268
7.3	Results Iteration 1 .....	270
7.3.1	Results: First iteration .....	270
7.3.2	Results: Second Iteration .....	276
7.4	Discussion.....	285
7.5	Conclusion.....	288
7.6	Statement of contributions.....	288
<b>Chapter 8 289</b>		
<b>Chapter 9 Discussion.....289</b>		
9.1.1	Limitations.....	292
<b>Appendix A 295</b>		
A.1	IUIS PID gene list. ....	295
A.2	Whole blood RNAseq data processing syntax .....	297
A.3	Novogene RNASeq QC methods .....	299
	Novogene next generation RNAseq QC .....	299



A.4 – Genomics England PID gene list .....	307
A.5 HTG T-Cell gene list. ....	309
A.6 Combined IUIS, GeCIP, T-cell panel from HTG EdgeSeq panel .....	312
A.7 OUTRIDER syntax .....	314
A.8 Salmon Script .....	315
A.9 BANDITS R script. ....	316
A.10 OUTRIDER results, in full .....	319
A.11 Differentially Expressed Gene graphs from Covid19 and Influenza cohorts stratified by age from pcaExplorer .....	320
A.12 Machine Learning Classification Performance Plots.....	323
<b>Bibliography.....</b>	<b>343</b>



## Table of tables

Table 2-1 Clinical phenotype of primary immunodeficiency patients.....	48
Table 2-2 Splicing and Disease Cohort.....	59
Table 2-3 - Data processing packages.....	60
Table 3-1 Basic descriptive statistics for read depth. ....	84
Table 3-2 FPKM Z-score outliers from PID cohort .....	91
Table 3-3 - TPM Z-score outliers in gene expression .....	108
Table 3-4 Comparing OUTRIDER results from the Splicing and Disease control cohort with known molecular diagnosis. ....	116
Table 3-5 OUTRIDER results for Primary Immunodeficiency Samples .....	120
Table 3-6 Genes from OUTRIDER results cross referenced with pre-identified gene panels....	121
Table 3-7 Summary table of all results from gene expression outlier detection methods .....	124
Table 4-1 SOT58 splice analysis outputs ranked by normalised read support. ....	132
Table 4-2 Splicing analysis results from SRB0003 :Filter 1.....	134
Table 4-3 Splicing analysis results from SRB0003 : filter 2 .....	135
Table 4-4 - Outcome of IGV investigation.....	141
Table 4-5 Summary Table of Candidate Splicing Events .....	167
Table 4-6 Comparing Results of Methods to Detect Gene Expression Outliers for PID Patients	168
Table 5-1 Covid and Influenza demographic analysis.....	173
Table 5-2 Differentially expressed genes which are more highly expressed in Covid19.....	183
Table 5-3 Differentially expressed genes in Covid and Influenza, ordered by P-value.....	184
Table 5-4 Table of most highly expressed pathways in Covid19 cohort compared with Influenza pathways.....	188

Table 5-5 -Pathways with most upregulation in Influenza cohort when compared with Covid19 cohort. ....	191
Table 6-1 Cohort grouping for machine learning application.....	221
Table 6-2 The distribution of ages for entire Influenza and COVID19 cohorts.....	222
Table 6-3 Top 20 genes with expression associated with age in COVID19 patients. ....	231
Table 6-4 Top 20 genes with expression associated with age in influenza patients. ....	231
Table 6-5 Top 20 Isoform abundance changes associated with age in influenza.....	232
Table 6-6 Top Isoform abundance changes associated with age in COVID19. ....	232
Table 6-7 Cross referencing and comparison of gene list to establish evidence of convergence.	248
Table 6-8 Cross referencing and comparison of isoform list to establish evidence of convergence.	248
Table 6-9 GO analysis for list of 'COVID19 genes' which showed converging expression. ....	249
Table 6-10 GO analysis for list of 'Influenza genes' which showed converging expression. ....	250
Table 6-11 GO analysis: isoforms which were initially higher COVID19.....	251
Table 6-12 GO analysis: isoforms which were initially higher Influenza.....	252
Table 6-13 -Classification Machine Learning Model Performance .....	254
Table 6-14 Performance of classification learning machine learning models on transcriptomic data. .....	257
Table 7-1 Top 10 models for ageing clock in Influenza.....	282
Table 7-2 Top 10 models for ageing clock in COVID-19.....	283
Table 7-3 Top 10 models for ageing clock in combined cohort.....	284
Table 8-1 FastQC output explanation .....	303
Table 8-2 HTG T-Cell gene list.....	309

## Table of figures

Figure 1-1 Development of the T-cell receptor repertoire .....	2
Figure 1-2 Thymic Maturation of T-cells .....	3
Figure 1-3 the mRNA splicing process .....	9
Figure 1-4 Chart showing percentage distribution for groups of primary immunodeficiency....	14
Figure 1-5 The developmental stages at which T-cell primary immunodeficiencies affect function. .....	15
Figure 1-6 The PID patient diagnostic journey.....	26
Figure 2-1 Agilent 2100 electropherogram RNA trace .....	53
Figure 2-2 - Novogene sequencing and QC workflow - received in personal report.....	54
Figure 2-3 Novogene lncRNA Library Preparation Workflow - Image provided by Novogene in email report. (222) .....	56
Figure 2-4 Workflow of GTEx data processing.....	61
Figure 2-5 - Gene list file structure. Columns shown are GeneID, ensemble gene identifier, strand information, chromosome number, start position, end position, and gene type.	68
Figure 2-6 - Splice junction normalization visualised.....	70
Figure 3-1 PID patient MultiQC comparisons; GC content vs Duplication percentage .....	82
Figure 3-2 Reads per sample: GTEx and PID.....	85
Figure 3-3 PCA plot of GTEx and PID RNAseq data .....	86
Figure 3-4 Scree plot for PCA analysis of GTEx and PID RNAseq data .....	87
Figure 3-5 Example image of Z-score tables (excerpt from table 1 shown) with GTEx as controls.	88
Figure 3-6 Number of overexpression outliers obtained from Z-score calculations with the GTEx data included as controls.....	89
Figure 3-7 FPKM and Z score table except: GTEx data not included in calculation.....	90

Figure 3-8 Over expression outliers present with GTEx control data excluded. ....	91
Figure 3-9 Euclidean distance heatmap with data from GTEx, PID cohort and splicing and disease cohort. ....	96
Figure 3-10 Euclidean distance heatmap with data from healthy controls, PID cohort and splicing and disease cohort.....	97
Figure 3-11 Euclidean distance heatmap with data from healthy controls, PID cohort and splicing and disease cohort after batch correction with ComBat-seq.....	98
Figure 3-12 Pearson correlation heatmap with data from healthy controls, PID cohort and splicing and disease cohort.....	99
Figure 3-13 Pearson correlation heatmap with data from healthy controls, PID cohort and splicing and disease cohort after batch correction with ComBat-seq.....	100
Figure 3-14 2D PCA plot of PID, Splicing and Disease and GTEx cohorts.....	101
Figure 3-15 Scree plot from PCA of PID, Splicing and Disease and GTEx cohorts.....	102
Figure 3-16 2D PCAplot of PID, Splicing and Disease cohorts.....	102
Figure 3-17 Scree plot from PCA of PID, Splicing and Disease cohorts.....	103
Figure 3-18 2D PCAplot of PID, Splicing and Disease cohorts after ComBat-seq batch correction.	104
Figure 3-19 Scree plot from PCA of PID, Splicing and Disease cohorts after ComBat-seq batch correction. ....	105
Figure 3-20 3D PCAplot of all GTEx PID, Splicing and Disease cohorts. ....	105
Figure 3-21 3D PCAplot of PID, Splicing and Disease cohorts after ComBat-seq batch correction and GTEx data removed. ....	106
Figure 3-22 PID Z-scores for under expressed Genes.....	107
Figure 3-23 PID Z-scores for Overexpressed Genes.....	107
Figure 3-24 OUTRIDER: Detection of outlier samples based on number of outlier genes. ....	111
Figure 3-25 OUTRIDER FPKM data in a mixed Primary Immunodeficiency and Splicing and Disease Cohort.....	113

Figure 3-26 Heatmap of Primary Immunodeficiency and Splicing and Disease Samples before and after batch correction.....	114
Figure 3-27 Expected vs actual expression of OCLNP1.....	119
Figure 3-28 Predicted Expression vs Actual expression for RRP1B and RRP8 .....	122
Figure 3-29 Rank vs normalised counts for RRP1B and RRP8.....	123
Figure 4-1 IGV alignment and splicing of <i>RAC2</i> .....	137
Figure 4-2 - Sashimi plot for <i>RAC2</i> , exons 3, 4, 5 and 6. ....	138
Figure 4-3 <i>RAC2</i> sequence and alignment .....	139
Figure 4-4 <i>RAC2</i> sequence and alignment 2. ....	140
Figure 4-5 Per-sample unique splicing events in WB RNAseq data .....	143
Figure 4-6 SRB002 Alternative Splicing Event <i>TRIM22</i> .....	146
Figure 4-7 Sashimi plot of SRB002 <i>TRIM22</i> Alternative Splicing Event.....	147
Figure 4-8 Sashimi plot of SRB005 event spanning multiple genes.....	150
Figure 4-9 SRB0013 <i>IL16</i> Sashimi Plot.....	154
Figure 4-10 Variant SRB0013 FOR <i>IL16</i> .....	155
Figure 4-11 Coverage plot for SRB0014 for <i>CD59</i> .....	158
Figure 4-12 Sashimi plot sample SRB0014 <i>CD59</i> splicing. ....	159
Figure 4-13 SRB0018 Variant .....	161
Figure 4-14 Sashimi plot of SRB0018 <i>SIP1R</i> splicing. ....	162
Figure 4-15 Sashimi plot of SRB0019 <i>POLE2</i> splicing.....	165
Figure 4-16 Sashimi plot of SRB0019 <i>IFI44L</i> splicing.....	166
Figure 5-1 Histogram of patient ages in Covid19 cohort .....	175
Figure 5-2 Histogram of patient ages in the Influenza cohort.....	175
Figure 5-3 Bar chart representing the sex distribution of the Covid19 cohort.....	176

Figure 5-4 Bar chart representing the sex distribution of the Influenza cohort.....	176
Figure 5-5 Exploratory data analysis: total aligned reads per samples for Covid19 and Influenza samples .....	178
Figure 5-6 Principal component analysis of Covid19 and Influenza cohort.....	179
Figure 5-7 Differentially expressed genes between Covid 19 and Influenza as determined by Pandaomics software. ....	181
Figure 5-8 - Expression of c-JUN in Covid19 and Influenza cohorts.....	186
Figure 5-9 The most upregulated molecular expression node in the Covid19 cohort .....	190
Figure 5-10 The most highly expressed molecular node in the Influenza cohort when compared to the Covid19 cohort .....	193
Figure 5-11 CD163 expression in Covid19 and Influenza patients.....	194
Figure 5-12 Genes with the greatest change in transcript use between Covid19 and Influenza.....	197
Figure 5-13 Precision of the model in determining differential transcript use. ....	198
Figure 5-14 Genes with top DTU 1-5.....	199
Figure 5-15 Genes with top DTU 6-10.....	200
Figure 5-16 Genes with top DTU 11-15.....	201
Figure 5-17 Genes with top DTU 16-20.....	202
Figure 5-18 Genes affected by changes in expression and splicing.....	206
Figure 5-19 Pathways enriched for differential expression and differential transcript use.....	207
Figure 5-20 Biological processes enriched for differential gene expression and differential transcript use.....	208
Figure 5-21 Biological processes which experience enrichment from DEG between Infectious disease. ....	209
Figure 5-22 Biological processes which experience enrichment from DTU between Infectious disease. ....	210



Figure 6-1 Principal component analysis of Covid19 and Influenza cohort.....	223
Figure 6-2 Principal component analysis of Covid19 and Influenza cohort.....	224
Figure 6-3 Exploratory data analysis: total aligned reads per samples for Covid19 and Influenza samples only patients under 65 years of age .....	225
Figure 6-4 JUN expression in the COVID19 and Influenza cohorts stratified by age. ....	226
Figure 6-5 Differential gene expression for IGHG1 in COVID and Influenza cohorts stratified by decade of life. ....	227
Figure 6-6 CD163 expression in Covid19 and Influenza patients stratified by decade of life. ...	228
Figure 6-7 Volcano plot for ageing gene expression in COVID19 patients .....	234
Figure 6-8 Volcano plot for ageing gene expression in Influenza patients.....	235
Figure 6-9 COVID19 isoform association with age volcano plot.....	236
Figure 6-10 Influenza isoform association with age volcano plot .....	237
Figure 6-11 Volcano plot for ageing splicing factor expression in Covid19 patients.....	238
Figure 6-12 Volcano plot for ageing splicing factor expression in Influenza. ....	239
Figure 6-13 Histogram of Beta-coefficients for genes in Covid19 .....	241
Figure 6-14 Histogram of Beta-coefficients for genes in Covid19 .....	242
Figure 6-15 Histogram of Beta-coefficients for genes in Influenza .....	242
Figure 6-16 Histogram of Beta-coefficients for transcripts in Influenza.....	242
Figure 6-17 Upset plot - Matrix Based Comparison of Gene Expression Panel Data Sets.....	244
Figure 6-18 Machine Learning Classification Performance Using All Models: Decreasing Age.	255
Figure 6-19 Machine Learning Classification Performance Using All Models: Increasing Age ..	256
Figure 7-1 MAE for a range of alpha values in the COVID19 cohort.....	270
Figure 7-2 Number of features for a range of alphas in the COVID19 cohort.....	271
Figure 7-3 R-squared values for a range of alphas in the COVID19 cohort .....	271

Figure 7-4 MAE for a range of alphas in the Influenza cohort.....	272
Figure 7-5 Number of features for a range of alphas in the Influenza cohort.....	272
Figure 7-6 R-square values for a range of alphas in the Influenza cohort .....	273
Figure 7-7 MAE for a range of alphas in the combined cohort.....	273
Figure 7-8 Number of features for a range of alphas in the combined cohort .....	274
Figure 7-9 R-squared valued for a range of alphas in the combined cohort .....	274
Figure 7-10 Feature type for ageing clocks.....	275
Figure 7-11 Features shared between the infection specific and combined clock. ....	275
Figure 7-12 MAE scores for a range of alpha values in Influenza .....	277
Figure 7-13 Number of features for a range of alpha values in Influenza .....	277
Figure 7-14 R-squared values for range of alpha values in Influenza .....	278
Figure 7-15 MAE scores across a range of alpha values for COVID19 .....	278
Figure 7-16 Number of features across a range of alpha values for COVID19 .....	279
Figure 7-17 R-squared scores across a range of alpha values for COVID19 .....	279
Figure 7-18 MAE scores across a range of alpha values for combined infections.....	280
Figure 7-19 Number of features across a range of alpha values for combined infections.....	280
Figure 7-20 R-squared values across a range of alpha values for combined infections.....	281
Figure 9-1 - Novogene Q-score distribution .....	300
Figure 9-2 Error rate distribution: Novogene .....	300
Figure 9-3 GC content distribution: Novogene.....	301
Figure 9-4 Raw read classification QC: Novogene .....	301
Figure 9-5 Differential gene expression for IGHG 1-24 in Covid and Influenza cohorts stratified by decade of life .....	320

Figure 9-6 Differential gene expression for IGLV 3-19 in Covid and Influenza cohorts stratified by decade of life. ....	321
Figure 9-7 Differential gene expression for IGHA1 in Covid and Influenza cohorts stratified by decade of life.....	322
Figure 9-8 Classification matrix for old cohort, using gene expression. ....	323
Figure 9-9 Classification matrix for young cohort, using gene expression. ....	323
Figure 9-10 ROC plot - old test cohort, using gene expression to predict COVID19 .....	324
Figure 9-11 ROC plot - old test cohort, using gene expression to predict Influenza .....	325
Figure 9-12 ROC plot - young test cohort, using gene expression to predict COVID19.....	326
Figure 9-13 ROC plot - young test cohort, using gene expression to predict Influenza .....	327
Figure 9-14 Classification matrix for old cohort, using isoform abundance. ....	328
Figure 9-15 Classification matrix for young cohort, using isoform abundance. ....	328
Figure 9-16 ROC plot - old test cohort, using Isoform expression to predict COVID19.....	329
Figure 9-17 ROC plot - old test cohort, using Isoform expression to predict Influenza .....	330
Figure 9-18 ROC plot - young test cohort, using Isoform expression to predict COVID19.....	331
Figure 9-19 ROC plot - young test cohort, using Isoform expression to predict Influenza.....	332

## List of Abbreviations

CD25	Cluster of differentiation 25
CD28	Cluster of differentiation 28
CD3	Cluster of differentiation 3
CD4	Cluster of differentiation 4
CD44	Cluster of differentiation 44
CD8	Cluster of differentiation 8
CTL	Cytotoxic lymphocytes
DN	Double Negative
ESE	Exon splicing enhancer
ESS	Exon splicing silencer
FPKM	Fragments per kilobase per million
GOF	Gain of function
GTEx	Genome tissue expression consortium
hnRNP	Heterogeneous nuclear ribonucleoprotein
IgA	Immunoglobulin A
IgE	Immunoglobulin E
IgG	Immunoglobulin G
IgM	Immunoglobulin M
ISE	Intron splicing enhancer
ISS	Intron splicing silencer
IUIS	International Union of Immunological Societies
LOF	Loss of function
MHCI	Major histocompatibility complex 1
MHCII	Major histocompatibility complex 2
ML	Machine Learning
mRNA	Messenger ribonucleic acid
PBMC	Peripheral blood mononuclear cells
PID	Primary immunodeficiency
RNA	ribonucleic acid
RNAseq	ribonucleic acid sequencing
SIAD	Selective immunoglobulin A deficiency
snRNP	Small nucleolar ribonucleoproteins
TCR	T-cell receptor
TH1	T helper 1 cells
TH17	T helper 17 cells
TH2	T helper 2 cells
WB	Whole blood

## Acknowledgements

I would like to give thanks to...

The three supervisors of this work, Professor Diana Baralle B.Sc., M.B.B.S, M.D, FRCP, Professor Anthony Williams MBBS, BSc, MSc, MRCP, FRCPath, PhD, and Dr. Andrew Douglas BSc (MedSci), MBChB, DPhil, MRCPE, who provided regular and important advice, critique and creative input to the work,

Dr Jenny Lord and Dr. Htoo Wai, who provided training, to Dr. Lord also who contributed to the work herself including helping develop syntax and strategies for approaching the bioinformatics,

Professor Mark Cragg for his help reviewing and advising on the introduction. Professor Tristen Clark BM, MRCP, MD, DTM&H kindly allowed us access to the infectious disease samples. In addition, the technical team for the IRIDIS HPCC, who provided support and advice when things refused to work,

Yaron Strauch, a fellow PhD student in the same group had a significant contribution to the last two results chapters in which a series of machine learning models were applied to the data and subsequently a tool has been developed from this work as part of a collaborative effort,

The University of Southampton and Southampton General Hospital for accommodating the project,

And lastly my family, for the eternal patience and understanding while I completed this project.

## Research thesis: declaration of authorship

Print name: Jed Lye

**Title of thesis: RNAseq as a diagnostic tool in primary immunodeficiencies**

I declare that this thesis and the work presented in it are 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. Parts of this work have been published as:-

Lye JJ, Williams A, Baralle D. Exploring the RNA Gap for Improving Diagnostic Yield in Primary Immunodeficiencies. *Front Genet.* 2019;10(1204).

Signature:

Date:13/02/2024



## Foreword and COVID19 impact statement

The research in this thesis aims to explore immunodeficiencies using the RNA space as a modality of investigation and or diagnosis. It will focus on two specific types of immunodeficiency, one primary and one secondary. To explore the primary immunodeficiency space, the research aims to characterise the T-cell mediated primary immunodeficiencies from the existing literature first and using a variety of techniques propose and test improvements to the current diagnostic pipeline of T-cell PID using RNA based methods. To explore secondary immunodeficiency space, the research aims to characterise and explore age-related immunosenescence from the existing literature, and then use similar RNA based methods to explore the transcriptomic profile of immunosenescence in cohorts of different models. The investigation then goes on to generate a continuous modelling algorithm which associates the biological age with chronological age of these cohorts.

Of note, it is important to state that the original research project aimed to develop T-cell specific diagnostic pipelines after the broader analysis of whole blood investigation. However due to a number of logistical issues, arising in part because of the 2020 COVID19 global pandemic, this was not possible. The research was significantly impacted by the pandemic, and datasets from COVID19 and Influenza patients were made available for an alternate approach to immunodeficiencies.



# Chapter 1 Introduction

## 1.1 The human adaptive immune system

The human immune system is responsible for the detection and removal of toxins, pathogens, malignant and senescent cells of the host (1, 2). The immune system can be categorised in a number of different ways; humoral and cell mediated (3), innate and adaptive (4), resident and circulating (5)

Whilst all these elements are essential for the functioning of the immune system, the research described in this thesis will focus mainly on the cell mediated adaptive immune response, with the intention of comparing systems of diagnosing immunodeficiencies.

The defining characteristics of the adaptive immune response is the ability to mount a specific response to any of a repertoire of possible threats, to adapt to these and to deliver immunological memory. As such the adaptive immune system is able to generate a more rapid and vigorous response to a repeated exposure to the same threat.

### 1.1.1 An introduction to T-cells and humoral and cell-mediated immunity

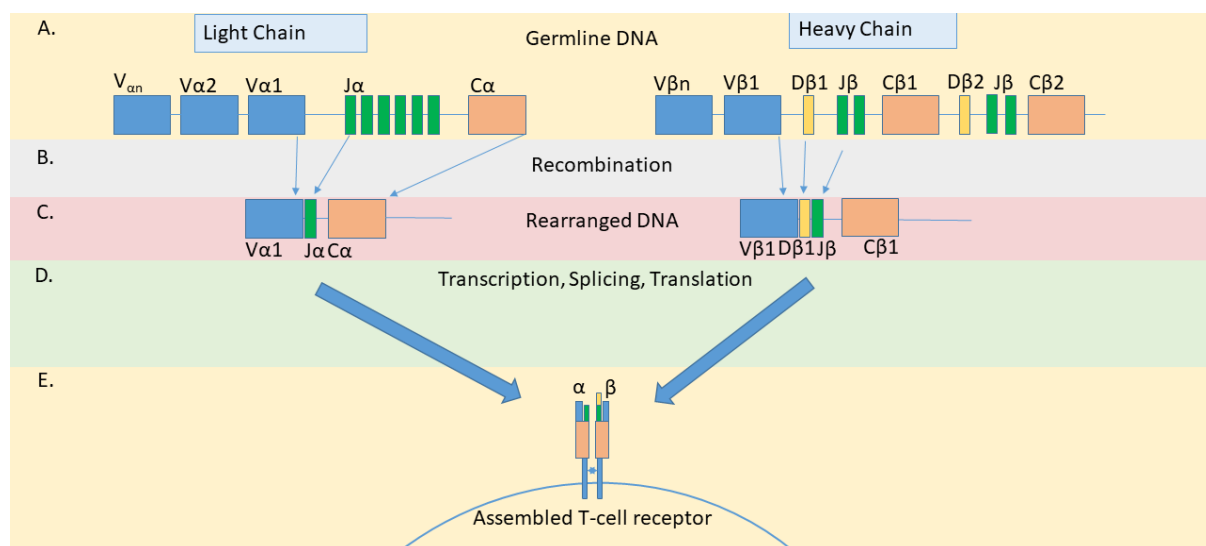
The key cell types responsible for adaptive immunity are B and T-cells. B and T- cells are so called due to the location of their maturation process, the bone-marrow, and the thymus, respectively (6) (7).

Perhaps the most important secreted element of the adaptive immune system are antibodies, produced by plasma cells, which are derived from B-cells and are a key part of humoral adaptive immunity (8). Antibodies are the primary mode of defence against extracellular pathogens and their associated toxins; binding to each of these elements and assisting in their sequestration and elimination (9).

It is the T-cells which primarily deliver cell-mediated adaptive immunity. T-cells are defined by their expression of a surface receptor complex called the T-cell Receptor or TCR. The TCR is comprised of an  $\alpha$  and  $\beta$  chain, which are associated with the CD3 protein complex, common to all T-cell lineages

(10). TCRs bind to cell surface structures known as major histocompatibility complexes (MHCs) (11). MHCs present small sections of a protein or 'antigen' in the form of peptides for recognition by the TCR which are specific to MHC-antigen combination, giving the cells their specificity (11)

T-cell receptors are made up of two polypeptide chains each having a variable and constant section. These receptors are part of the immunoglobulin superfamily. The receptors are comprised of a variable  $\alpha$  (light) chain and  $\beta$  (heavy) chain. Complete beta chains are comprised of three variable sections and one constant, the alpha chains have only two variable sections and one constant. These separate sections are denoted as V, D, and J, standing for 'variable', 'diversity', 'joining' segments. The alpha chain has only the V and J segments. The genes for these chains contain multiple versions of each segment for each of the V, D, and J components (Figure 1-1). These are rearranged during recombination to give any one of  $10^{18}$  possible T-cell receptors.

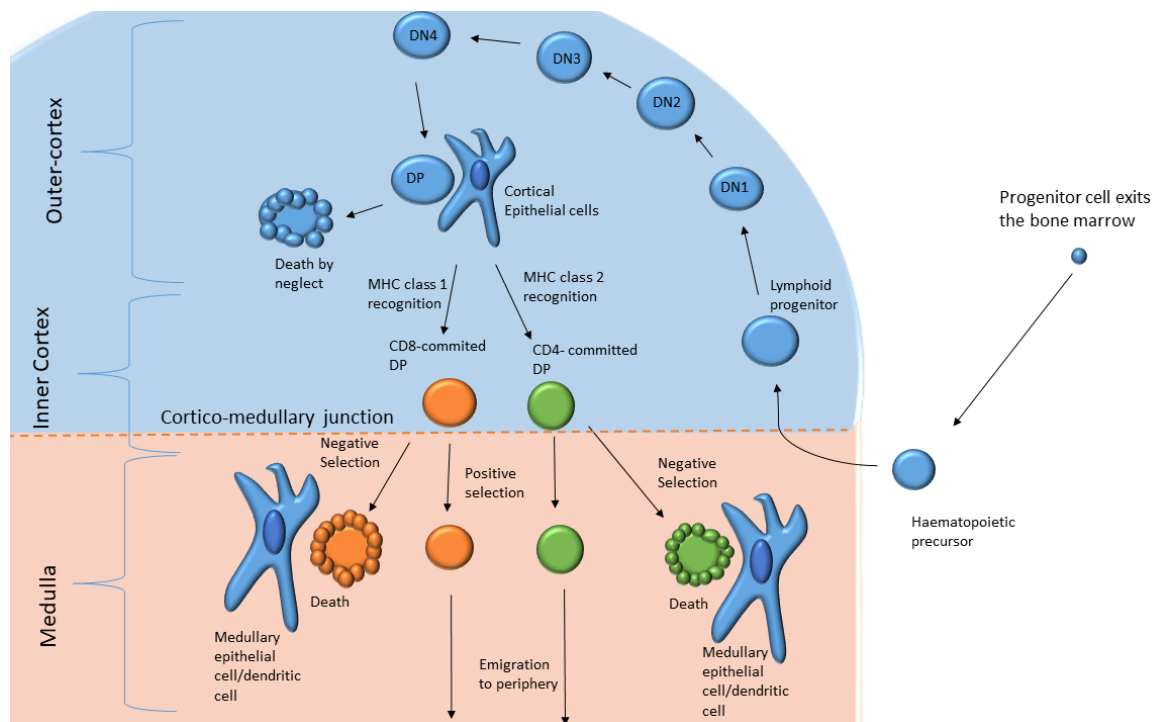


**Figure 1-1 Development of the T-cell receptor repertoire**

**A.** Representation of the arrangement of the genes for heavy ( $\beta$ ) and light ( $\alpha$ ) chains. **B.** Genes undergo recombination to form complete exons with 1 segment from V, one segment from J and one constant region. In heavy chains the D segment is also present. **C.** This produces rearranged DNA sequence, with a small intron between the VJ/VDJ sections and the C sections. **D.** Exons undergo transcription, splicing out of intron and translation. **E.** Receptor complexes are formed and presented on the T-cell surface.

T-cells are first produced from haematopoietic stem cells in bone marrow. Immature, undifferentiated cells from the lymphoid lineage are known as progenitor cells, and progenitor T-cells undergo migration from the bone marrow to the thymus. Once in the thymus, the thymocytes (as they are then called), will mature into T-cells and undergo selection processes based on their individual TCR receptor affinity to self-antigen (12). Initially thymocytes lack the TCR and its co-

receptors CD8 and CD4. Within the thymus they undergo a maturation process shown in Figure 1.2. During this process they sequentially gain key surface receptors, some highlights of which are explained in the figure legend.



### Figure 1-2 Thymic Maturation of T-cells

Lymphoid progenitors produced by haematopoietic stem cells exit bone marrow and migrate to thymus via the blood. These lack the TCR expression, CD4 and CD8 expression so are termed double negative (DN). These go through developmental stages classified by surface protein expression of CD44 and CD25. DN1 =  $CD44^+CD25^-$ , DN2 =  $CD44^+CD25^+$ , DN3 =  $CD44^-CD25^+$ , and DN4 =  $CD44^-CD25^-$ ; DP = dendritic progenitor. Adapted from R.N. Germain, 2002 (12)

The TCR expression begins in DN2-DN4, with a pre-TCR, comprised of a  $\beta$  chain and an  $\alpha$ -chain which has not yet been re-arranged. This early version of the TCR is known as the pre-TCR. The expression of this early version of the TCR in DN4 leads to significant proliferation and production of a mature, expressed TCR- $\alpha$  chain and complete  $\alpha\beta$ -TCR.

The expression of both CD4 and CD8 at this stage alongside the TCR results in the DP ('double positive') notation of these cells. Through interaction with cortical epithelial cells expressing MHCI or MHCII molecules bearing self-peptides the process of selection occurs. If not enough affinity and

thus interaction occurs, these DP cells undergo delayed apoptosis (death by neglect) or pass to the medulla. In the medulla further interaction occurs between the now, lineage committed thymocytes and medullary epithelial cells or haematopoietically derived dendritic cells. If an appropriate intermediate level of thymocyte interaction occurs with MHCI complexes, the thymocytes become CD8<sup>+</sup>T-cells. In the case this interaction is with MHCII complexes, the thymocytes become CD4<sup>+</sup> T-cells. Should the interaction affinity be inappropriately high, the thymocytes undergo negative selection and acute apoptosis.

Once mature, they are considered naïve T-cells and can circulate in the body for almost a decade (13). T-cells circulate in the blood vessels and lymph vessels until they encounter their cognate antigen. When the naïve T-cell encounters its cognate antigen associated with class 1 MHC (CD4 T-cells) or class 2 MHC (CD8 T-cells) on an antigen presenting cell for the very first time, it becomes primed (14). In addition to the professional antigen presenting cells such as dendritic cells, epithelial cells can also act as antigen presenting cells or 'APC's' under pathological conditions (15). As such priming can take place either in the central lymphoid organs, lymphatic circulation or at the site of infection (14) but canonical priming takes place in the T zone of a lymph node (16).

Priming is precipitated by the formation of an immunological synapse between the T-cell and the antigen presenting cells (17). In the immunological synapse, ICAM-1 on the APC binds with LFA-1 on the T-cell, and TCR/antigen interaction triggers LFA-1 conformational change. B7.1 and B7.2 on the APC bind to CD28 on the T-cell and the T-cell is then effectively primed (18). T-cell priming initiates biophysical, biochemical, transcriptomic and proliferative changes, generating expansions of differentiated effector cells, some of which go on to become long lived memory T-cells (14). The subsequent location of T-cells after priming/stimulation by an antigen presenting cell depends on a number of factors which are not yet entirely understood (19). Stimulation by an APC originating at any site can cause the T-cells to migrate to almost every other tissue in the organism, and T-cells primed in secondary lymphoid organs can end up in non-lymphoid tissues by upregulating respective homing receptors (19). Through these methods the adaptive immune system protects the host from pathogens which are found anywhere in the body.

For previously unstimulated naïve T-cells, costimulation from the CD28 cell surface protein is also necessary for the cells to be fully activated (20). Depending on cell type, other co-receptors are also involved in the activation process. CD4 for example in CD4<sup>+</sup> T-cells and CD8 in CD8<sup>+</sup> T-cells (21).

CD4+ and CD8+ are the two major subsets of T-cells. Simplistically, those T-cells expressing the CD4+ surface antigen are often described as helper T-cells or Th cells, and those which express the CD8+ antigen are cytotoxic T lymphocytes or CTL (22). CD4+ Th cells specifically secrete cytokines to help elicit and modulate CTL responses. CD4+ Th cells also help generate and modulate the humoral response through their interactions with B cells. The diverse populations and roles of T-cell are brought about through complex differentiation and regulation processes which continue to be investigated and explored (23).

### **1.1.2 The specificity and memory of T-cells in the adaptive immune system**

CD8+ T-cells can become primed by almost any cell in the body, as expression of HLA—1 is almost ubiquitous (24). A number of cell types are able to prime CD4+ T-cells by presenting antigen on HLA-2. These include dendritic cells (DC), macrophages, follicular dendritic cells (FDC), and a range of epithelial cells, and even sub-categories of fibroblasts (25). Most typically DC's or macrophages phagocytose pathogens or foreign material, and fragments of these (antigens) are then presented on the cell surface via the major histocompatibility complex (25, 26). These dendritic cells or macrophages then return to the local lymphoid tissues; lymph nodes or the spleen (26). When a lymphocyte recognises the antigen presented on the dendritic cell with sufficiently high affinity, it will become activated. The cell adaptive response now begins a process of amplification of response (27).

Naïve CD4+ T-cells interact with the DC and undergo clonal proliferation to produce a greater number of antigen-specific cells which are able to interact with CD8 T-cells and B-cells. This allows germinal centre formation in the lymphatic tissues (28).

Upon encountering its cognate antigen, a T-cell becomes activated and forms a large blast-cell which proliferates and undergoes the process of clonal expansion. The resulting cells either become short-lived effector cells or long-lived memory cells (27). Memory T-cells continue to exist as both circulating central (TCM), effector memory T-cells, or tissue resident memory T-cells and can persist in circulation for 44 to 54 days (29-33). Effector CD4+ T-cells have additional functional subsets which are denoted as Th1, Th2, Th17 (34, 35).

T-cells do not have the ability to hyper-mutate their T-cell receptor variable regions once stimulated. However, evidence has shown that T-cells are able to enhance their antigen responsiveness through a process termed functional avidity maturation (36). Antigen primed T-cells undergo changes in TCR

machinery affecting calcium flux, ERK activation, Vitamin D receptor to increase the responsiveness to cognate antigens by as much as 50% in primed T-cells (36). Furthermore, the requirement for CD28 co-stimulation at the immunological synapse is only required for naïve cells. Similarly, requirement of cytokines as a third activation signal is required for naïve cells but not in primed T-cell (36).

Once the pathogen has been cleared, the majority of these, antigen specific effector cells of both lineages undergo apoptosis in the interest of maintaining homeostasis, allowing other T-cells to expand as required and to limit the chances of autoimmunity (37). The remaining cells which do survive are memory T and B cells, able to respond rapidly to instances of re-infection. These cells are generated in the germinal centres after stimulation (38) (28). Stimulation by antigen alone is not sufficient to instigate the development of memory cells, and factors such as inflammation induced maturation of the APC's is important in determining the cell effector/memory fate (39).

The interaction strength and duration between the TCR and MHC also contribute to the determination of fate. The specifics of factors which induce lymphocytes to become memory cells is still being elucidated and varies between cell types. Both the cells expression of factors (e.g. CD127) in CD8 T-cells and the external signalling molecules (e.g. IL2) are factors in determining memory cell generation (39). These immune cells become either tissue resident or circulating cells. Resident memory cells enhance protective immunity through immediate effector function and rapidly recruiting other immune cells under reinfection (40) .

Differentiation to various lineages of T-cells and their specific functions, maturation, differentiation, activation and stimulation are initiated through distinct transcriptional programs, and are detectable with transcriptomic analysis (41-43). These programs are mediated by changes in gene expression (44), changes in the relative abundance of isoforms through alternative splicing, (45-47) and allele specific expression changes (48, 49). These different transcription variables have important roles in controlling, modulating and tuning many of these processes and as such the RNA signatures provide both quantitative and qualitative data and present a unique opportunity to gain insight into the variance in immune system which can be precipitated by genetic variation (50).

### **1.1.3 RNA splicing and its role in the immune system.**

Processing of the RNA transcripts provides a critical mechanism to amplified and enhance information density of the genome, and this produces a wider range of molecules for greater range

of instructions and responses for the cell (51). RNA processing involves the removal of intronic nucleotides from pre-mRNA and ligating the ends of exons together. This forms a mature transcript for further processing and translation, which is known as canonical splicing (52). There also exists a mechanism for generation of different patterns of exonic nucleotide sequences than is present in the main gene sequences; this is termed alternative splicing (52).

Through this process the cell can produce an array of isoforms from a single gene, or in fact multiple genes spliced together; a lesser-known process termed trans-splicing (53, 54). The chemical process through which introns are spliced out and exons are either ligated together in different combinations is made up of two transesterification reactions. In the first reactions the branchpoint reacts with the splice donor site, then the now free 3' end of the upstream exons reacts with the splice acceptor site. Through this process and other post transcriptional modification steps, the preRNA forms a mature mRNA (53). Importantly, the various species of mRNA brought about by alternative splicing, although originating from the same genomic loci, can have differential and even antagonistic effects (55).

The boundaries of exons and introns are marked by specific consensus sequences known as splice sites, which are recognised by small nuclear ribonuclear proteins (snRNPs). Once bound, these are then joined by further snRNPs which work in a coordinated manner to arrange the intron in a manner which facilitates the transesterification process and the introns subsequent excision. Introns are categorised as either minor or major introns – depending on the sequences which are found at the splice sites and branch point sequences. These classes of introns each have respective spliceosome complexes which are dependent upon different snRNP's (56).

Deep surveying of alternative splicing has shown that 95% of genes which contain multiple exons undergo alternative splicing, and even when only considering moderate to high abundance events, there are reportedly 100,000 individual splicing events in major tissues (57). Alternative splicing occurs both co-transcriptionally and post-transcriptionally, and the action of transcription factors regulates and influences splicing events. This is demonstrated in some of the most crucial mechanisms of the adaptive immune system. (58-60).

Exonic/Intronic splicing enhancers, and exonic/intronic splicing silencers are regions on the transcript which are recognised and bound to, by heterogeneous nuclear ribonucleoproteins and serine arginine rich proteins. These proteins trigger the formation of the spliceosome consisting of U1, U2, U4, U5, U6 and U2af. The intron region is folded back on itself to form a loop (lariat loop)

during the first transesterification, the loop is then excised via the second transesterification process (53) (Figure 1-3).



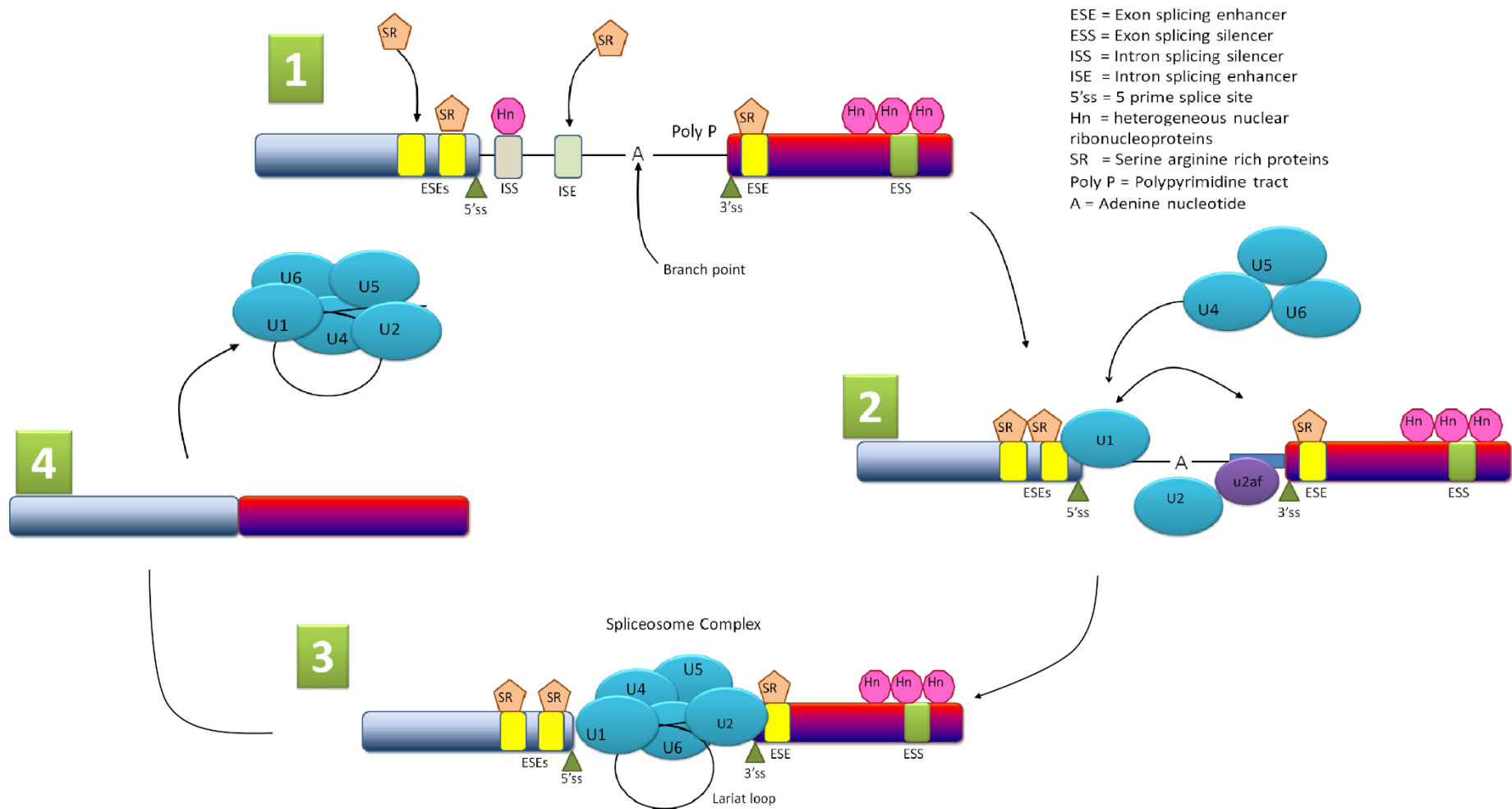


Figure 1-3 the mRNA splicing process

Figure 1-3 the mRNA splicing process- The polypyrimidine tract is a spliceosome assembly promoting element of the intron. Approximately 20 base pairs long and rich in pyrimidine nucleotides, this element acts as a binding site for elements of the spliceosome. The branch point near the 3' end of the intron has an adenine nucleotide and is necessary for lariat loop formation.

1. SR proteins bind to ESE's and ISE to promote splicing, Hn-ribonuclear proteins bind to ESSs and ISSs to inhibit the splicing process.
2. The bound SR proteins promote further binding of snRNPs at splice sites. U2AF binds to the polypyrimidine tract, U1 binds to the 5'ss on the exon, and U2 then binds the branchpoint and subsequently disrupts binding of U2AF. U4, U5 and U6, enter as a complex and bind to the other proteins to form and spliceosome complex.
3. Chemical reactions induce steric conformational changes in the spliceosome complex, inducing formation of the lariat loop.
4. Exon ends are ligated, and the spliceosome complex detaches from the exons, taking the intron for degradation.

Splicing is critical in all tissues, and this is perhaps best demonstrated in the immune system. The ImmGen project found that around 60% of genes in mice are expressed as multiple isoforms in T or B cells, and 70% of these had an impact on the lineage differentiation (61). Important examples include *FOXO1* induced Ikaros splicing which is essential for the recombination of immunoglobulin genes. This allows the immune system to produce its diverse range of antibodies/immunoglobulins (62). In addition, the alternative splicing of *CD45* is necessary for the production of a range of tyrosine phosphatases, imperative for the diverse set of lineage and stage-specific receptor signal transduction thresholds (63). This occurs because the cellular control mechanisms of transcription and splicing are tightly linked, and RNA-polymerase II is a facilitator of splicing factor recruitment (64).

Activation of lymphocytes is a key component of the adaptive immune response to pathogens (65). Part of the central activation of these cells is the degradation of *I $\kappa$ B $\alpha$*  and release of *NF- $\kappa$ B*, which translocates to the nucleus to initiate maturation and activation of the cell. The "CBM" complex that brings about the degradation of *I $\kappa$ B $\alpha$*  is formed by *CARMA1*, *BCL10* and *MALT1* (66). *MALT1*, a crucial component of this complex undergoes alternative splicing of EXON 7. This process produces mRNA isoforms with a differential function, and the activation strength of CD4<sup>+</sup> T-cells is in fact mediated by the relative abundance of the alternatively spliced isoforms of *MALT1*. The splicing of *MALT1* is modulated by the molarity of phosphorylated splicing factor *hnRNPU* in the nucleus (67). These are just some examples which demonstrate the means by which alternative splicing is a key component of the normally functioning immune system and how perturbations can lead to pathology.

In some instances gene expression level studies neglect the importance of precursor mRNA splicing that can precipitate functionally distinct transcriptomes, with differing biological functions (68). These different transcriptome profiles have been observed as a critical component of tissue differences in the human body (69). Recently studies have extended to highlight transcriptional variation through the life course which demonstrated profound change as age progresses (70).

#### **1.1.4 Alternative splicing in T-Cells**

mRNA alternative splicing is known to be inherently utilised in a number of functionally specific genes in T-cells, such as *CD44*, *CD45*, and *CTLA4*, for which multiple RNA isoforms are produced in T-cells (71, 72). These splicing patterns are also known to be altered in response to antigen stimulation, and mediate changes in the complement of functional proteins (73). It is important to note that the alternative splicing events which occur during T-cell activation have been shown to occur in distinct gene sets from those which demonstrate changes in expression (74).

*CD45*, a protein tyrosine phosphatase, elicits control of cell cycle progression and thus proliferation of T-Cells (73). *CTLA-4*, which transduces inhibitor signals in T-Cells has two transcripts of 550 and 650 bp respectively. The shorter isoform, *CTLA-4delITM*, has a deletion of an exon (literature contradictory on exon number 2 or 3) which codes for the transmembrane domain. Anti-*CD3* plus anti-*CD28* stimulation results in the suppression of this isoform, suggesting the shorter, soluble version of the signal transducer may be produced via alternative splicing as a means of regulating immune activity and homeostasis (75, 76).

## 1.2 Primary immunodeficiencies – T-cell disorders

Variations in the genome can cause aberrations in the functions or abundance of the transcript. This can in turn affect a specific facet or pathway of the immune system. On a cellular level, this can result in reduced function rendering the cell inherently unable to react, constantly stimulated, or reacting inappropriately (77) (78). Clinical results of these events include disorders of increased frequency and severity of infection, autoimmunity, aberrant inflammation and malignancy (79). These disorders are termed 'primary immunodeficiencies' (78).

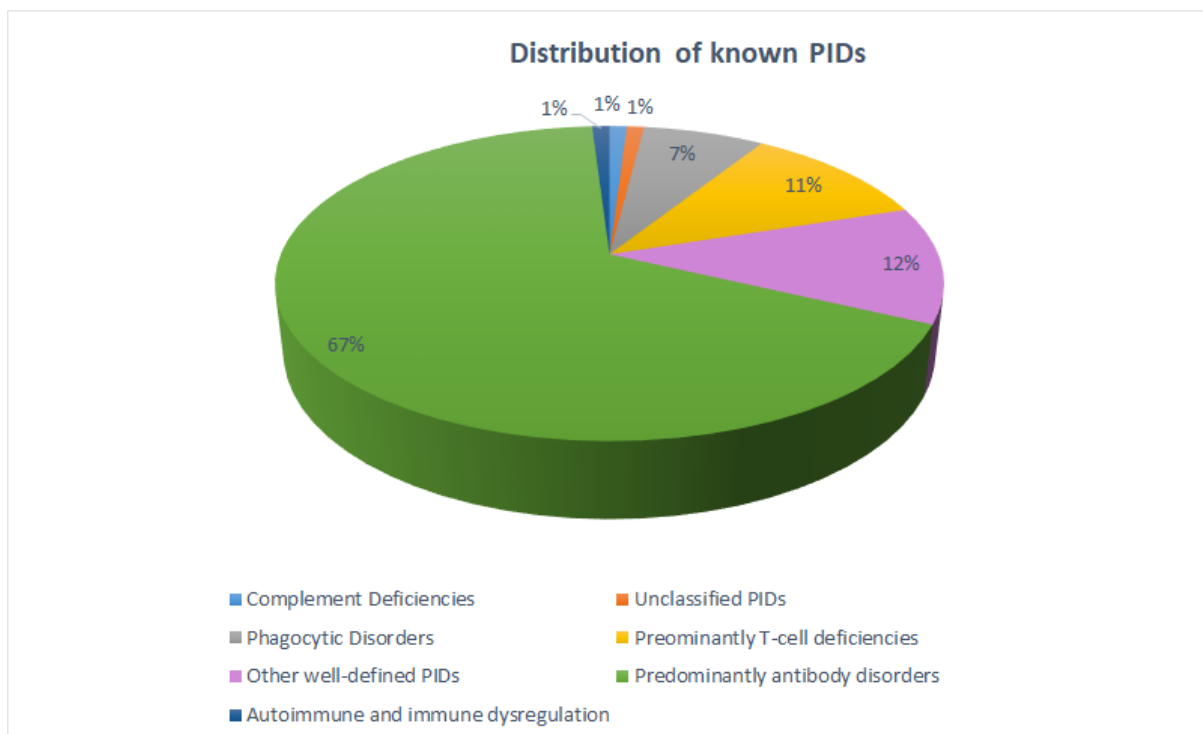
The understanding of the heterogeneity of PIDs has expanded greatly over the last decade, at last count on Jan 2020, the list encompassed over 416 distinct disorders arising from 450 genes, demonstrative of the complexity of the immune system (78, 80, 81). Reports put the average incidence of PID's in the UK at 7.6 per 100,000 (82).

The monogenic variants which precipitate the disorders do so through three primary molecular modalities; loss of function (LOF) of the protein, gain of function (GOF) of the protein, or changes in expression, which can also result from GOF/LOF changes (78) . Monogenic genotypes can manifest as homozygous in which the same mutation is present on both alleles, heterozygous in which only a single allele is mutated, or biallelic which regards mutations occurring in the same gene, although not necessarily the same mutation (compound heterozygotes) (83). Hemizygous patterns of inheritance also exist in which a mutation occurs in a chromosomal segment of which the patient has only one copy (84). Known examples of gene action for heterozygous mutations include dominant gain of function, haploinsufficiency and negative dominance. X-linked recessive traits can be caused by hemizygous pathogenic variants in males or homozygous in females. Rarely, X-linked dominant traits can also manifest (GOF/LOF) (78).

This plethora of disorders constituting PIDs has brought about a need to categorise them for expedited diagnosis and treatment protocols. Some broader methods simply classified the disorders into groups of innate and adaptive immunity (85). The Inborn Errors of Immunity Committee (previously the International Union of Immunological Societies PID expert committee or IUIS) has now devised a precise system, which classifies disorders by the immunological pathway affected. In addition, it now has corresponding phenotypical classification systems for clinicians at the bedside to help identify the disorders. These briefly comprise nine categories; 1. Immunodeficiencies affecting cellular and humoral immunity, 2. CID with associated or syndromic features, 3. Predominantly

antibody deficiencies, 4. Diseases of immune dysregulation, 5. Congenital defects of phagocyte, 6. Defects in intrinsic and innate immunity, 7. Auto-inflammatory disorders, 8. Complement deficiencies, and 9. Phenocopies of PID (80). The most common form of PID is selective immunoglobulin-A deficiency (SIAD), which can manifest with a variety of clinical presentations from Type 1 diabetes mellitus, juvenile arthritis, and ankylosing spondylitis. SIAD has an estimated prevalence of as high as 1 in 143 persons in some countries (86, 87). Whilst these numbers are linked to rates of consanguineous marriage, the discovery of one identical twin with the disorder and other evidence suggests an environmental component (88). Other studies show links with epigenetic factors, microbiome, age, sex and season (89). Some cases may be linked to physiological stress, as seen in elite athletes (90). The genetic origin of SIAD is still unclear, an enrichment of *HLA-B8* was noted in a small scale study (91), and a genomic locus at the proximal end of the *MHC* has been observed to have increased likelihood of being a predisposing factor in disequilibrium tests (92, 93). SIAD is characterised by decreased or absent levels of IgA, which is the most prevalent form of immunoglobulin in luminal secretions. Only when these IgA levels are observed in the presence of normal levels of IgG and IgM, is a diagnosis of SIAD given (94).

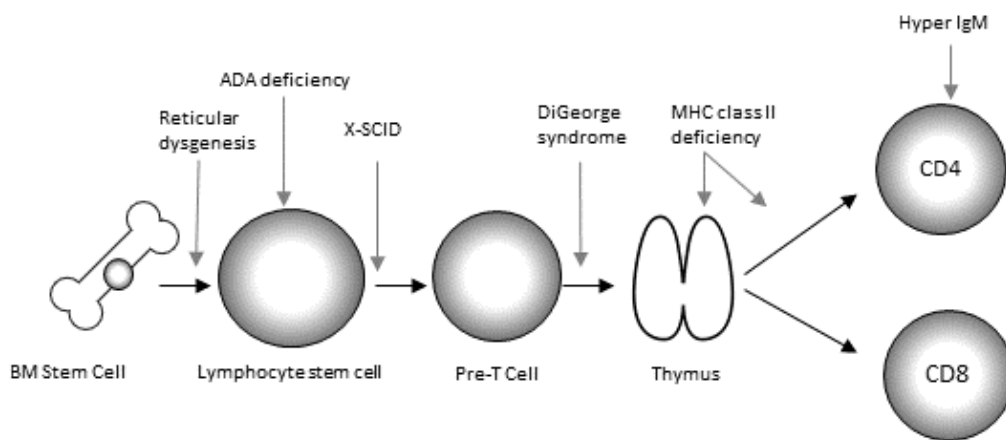
Whilst individually rare, the remaining disorders considered in the wider scope of PID together represent a significant burden on the health and economy of a nation. Current diagnostic levels suggest an incidence of 5.90/100,000 (82), experts suggest however 70%-90% of PID remain undiagnosed and true incidence could be as high as 1:250, as no encompassing screening programs exist (95). A number of types of T-cell disorder exist and usually arise as a result of specific single gene disorder, resulting in maturational or stimulation response abnormality (96). T-cell primary immunodeficiencies make up about 11% of immunodeficiencies, see Figure 1-4 Chart showing percentage distribution for groups of primary



**Figure 1-4 Chart showing percentage distribution for groups of primary immunodeficiency.**

Currently T-cell deficiencies make up around 11% of primary immunodeficiencies (97). By far the largest number of PIDs are antibody disorders. The majority of all PIDs then are from the adaptive immune system.

T-cell disorders are often observed to precipitate or be part of developmental syndromes, regularly affecting cellular repair (98, 99). Presently characterised T-cell specific primary immunodeficiencies consist of severe combined immune deficiency, X linked lymphoproliferative syndrome, X linked immune deficiency with associated hyper IgM, The DiGeorge syndrome, Chronic mucocutaneous candidiasis, Ataxia telangiectasia, Nijmegen breakage syndrome and other rare T-cell deficiencies (96). The various disorders manifest at different stages of the immune systems cytological and histological development (Figure 1-5). The differences observed can occasionally help inform diagnosis. This will be explored in the coming sections.



**Figure 1-5 The developmental stages at which T-cell primary immunodeficiencies affect function.**

T-cell primary immunodeficiencies elicit their affect a variety of immune cell maturation stages from early progenitors to mature T-cells. This diagram is not exhaustive and was adapted from J. Edgar (2008) (96).

### 1.2.1 Severe combined immunodeficiency

Severe Combined Immunodeficiency is a heterogeneous subgroup of immunodeficiencies which present with severe T, B and NK cell aberrations in development and function. This typically leads to a number of opportunistic infections and failure to thrive (96). A number of molecular genetic defects have been identified including those affecting  $IL2-\gamma c$ ,  $JAK3$ ,  $IL-7R\alpha$ ,  $ADA$ ,  $RAG1/2$  (100) The specific mechanism for development of SCID are explored below.

These present with different phenotypic aberrations in lymphocyte number and function, the most common of which is the X-linked  $\gamma c$  form, termed X-SCID. A clinical investigation may point to this disorder when the T-cells and NK cells have low numbers, but B cells appear present at normal levels (100). The unique patterns in ratios of cells can often be used to distinguish between the subtypes of SCID and inform further diagnostic tests (100). SCID is considered a paediatric emergency and immediate priority is to stabilise for bone marrow transplant (96).

The  $\gamma c$  chain (or  $IL-2$  receptor gamma) is a common component of receptors for  $IL-2$ ,  $IL-4$ ,  $IL-7$ ,  $IL-9$ ,  $IL-15$  and  $IL-21$  receptors and has critical roles in the proper functioning for them (101).  $IL-2/IL-2R$  function is essential for functional maturation of regulatory T cells during development of the

thymus (102). Without these signals being interpreted, response to immunological mitogens needed for growth and proliferation is reduced and the normal development of the adaptive immune cells is impaired (103).

The mechanism of action for disease causing mutations on *γc chain* varies; using a cell surface staining-based assay one study has shown that 47% of samples had no staining at all, 32% had trace amount and 21% had normal staining (104). Examples of mutations in IL2-R *γc* include point mutations to produce premature stop codons in exon 3 and exon 7, and resulting in truncation of the transcripts cytoplasmic domain (exon 7 and exon 8) (105).

*JAK3* is an important component of the signal transduction cascade from cytokines, specifically those which contain the *γc* chain. The cytokine receptor proteins are unable to transduce a signal as they have no enzymatic activity, so the pathway is dependent upon the *JAK3*, the cytosolic tyrosine kinase which propagates the signal intracellularly (106). Some cytokine receptors (IL-2, IL-4, IL-7, IL-9, IL-15, and IL-21) rely on intracellular JAK to initiate signalling as they lack enzymatic activity (107). Mutations have been observed in all 7 domains of the gene at various locations on each domain (107). Compound heterozygote mutations were observed in one case. c.81T>G, p.H27Q mutation in exon 2 came from patients' father, while c.665G>A, p. R222H in exon 6 from patient's mother. The exact nature of the mutation's action was not clarified with orthogonal RNA based interpretation. In this instance patient presented with chronic active Epstein-Barr virus, decreased T-cells in peripheral blood (107).

*ADA* (adenosine deaminase) is an enzyme critical in clearing toxic metabolites which accumulate in tissues like the lymphoid tissues, which demonstrate a high rate of metabolism and cell turnover (108). The autosomal recessive condition leads to the accumulation of these metabolites and results in an inability of the cell to conduct normal DNA synthesis and repair (109). As a result, defective thymocyte development and increased levels of apoptosis in the thymus are observed (109). Biochemical testing and genetic testing are usually able to confirm diagnosis.

*RAG1* and *RAG2* are recombinase activating genes. They each produce proteins through which DNA double stranded break are induced and the antigen receptors undergo V(D) J recombination to form their diverse sequences (110). Deficiencies of either *RAG* gene result in deficiencies in T-cell and B-cell number. This occurs due to apoptosis of cells with absent recombination process (111).

In addition to these and *LIG4*, splice site disrupting mutations have been found in *Artemis*, a gene integral to the V(D)J recombination and DNA repair pathways. One mutation caused a 5' splice site



dysfunction, resulting in two novel transcripts. The other caused a novel 3' splice site, which resulted reduced wild-type expression and a transcript with intron inclusion. In each case either low activity or low levels of the transcript were seen (112).

### **1.2.2 X-linked lymphoproliferative disease**

X-linked lymphoproliferative disease or 'Duncan's disease' is a fatal, recessive, lymphoproliferative syndrome with pathophysiology leading to lymphomas, dysgammaglobulinemias, and lymphohistiocytosis. Usually, the syndrome goes undetected until infection with the Epstein-Barr virus (113).

Once carriers of the mutation are infected however, the normally low impact EBV infection can result in severe hepatitis, and bone marrow failure. The disorder occurs due to recessive mutations in *BIRC4*, *XIAP* and *SH2D1A* (96). The majority of the mutations occur in *SH2D1A*, and the majority of these are missense (most frequently appearing in exon2), however, nonsense, frameshift and splice site mutations have all been reported (114). In one example a nonsense mutation was observed in *SH2D1A* (c.300T>A). The mutation was present in exon 3 and converts tyrosine to a stop codon (TAT>TAA). Truncated transcripts were confirmed via PCR, and inheritance was confirmed as X-linked as the patient's mother was identified as a carrier (115).

Patients with deficiencies in SLAM associated protein (SAP), which is encoded by the *SH2D1A* gene, are observed to have impaired ability to form germinal centres for B-cell affinity maturation and memory cell production (116, 117). Consequently they lack the long term antibody responses which form part of the adaptive immune system (116). SAP is now known to bind to a series of receptors from the SLAM family. Several of the SLAM receptors are known to be cytotoxic receptors in both NK and CD8 cells and SAP is most highly expressed in T and NK cells (117). As such cells with deficiencies in SAP also exhibit reduced capacity to adhere to and kill pathogens and EBV transformed cells (118). Through these mechanisms SAP deficiency reduces lymphocyte proliferation and development, whilst also blunting their effects and bringing about immunodeficiency.

### **1.2.3 X linked immune deficiency with associated hyper IgM**

X linked immune deficiency with associated hyper IgM (HIGM) is characterised by abnormally low IgG and high IgM and is often seen with unaffected lymphocyte numbers. Production of IgA and IgE

is also often impaired. This group of disorders also carries with it an increased risk of infection, specifically *Cryptosporidium*, pyogenic infections, or *Pneumocystis jiroveci*, and a susceptibility to lymphoma(96).

These disorders cause a failure of the humoral response to effectively initiate “class switching”; the process of B-cells changing the genomic source of then heavy chain in antibodies whilst maintaining the same variable domain (119). Naïve B cells produce IgM and IgD normally, and will switch to other classes of immunoglobulin in the instance of encountering signalling molecules at their CD40 and cytokine receptors from T-cells (119).

The genetic causes are varied: the most common results from mutations in *CD40LG* gene, which often results in abnormal amounts or absent production of this gene (120). It is therefore likely that the diagnosis will be aided by RNA investigations. Mutations in this gene appear throughout its length, and range from non-sense, missense, insertion deletion and splice site mutations (120). The *CD40* ligand on T-cell surface is required for communication with IgM producing B Cells. Without functioning CD40L class switching cannot take place (121). In one example a patient presented with recurrent infections throughout life and was misdiagnosed as CVID. After follow-up investigation, analysis showed a 6-nucleotide insertion of exon 1 (c.121\_122insCAGCAC). The expression of the CD40L appeared normal until PBMC stimulation with ionomycin, confirmed a diagnosis of X-HIGM.

Additional causes of HIGM syndrome include mutations in the IKK-gamma (*NEMO*) gene which is also X-linked, and *CD40* activation induced cytidine deaminase (*AICDA*), and uracil-DNA glycosylase (*UNG*) also can produce autosomal recessive HIGM syndromes.

#### **1.2.4 The DiGeorge syndrome**

DiGeorge syndrome is a developmental disorder which presents with significant abnormalities of facies, hypoparathyroidism, congenital heart disease, alongside cellular immune deficiency. The severity of the immune deficiency is rarely significant. It may reflect perturbations in T-cell regulatory function and usually improves with time, although thymic hypoplasia is sometimes observed (96). In some cases complete absence of the thymus is present and profound immunodeficiency occurs as a consequence (122). The exact molecular causes remain unclear, but hematopoietic cell defects leading to failure to develop the third and fourth pharyngeal arch is suggested to be a causal factor in thymic development issues (123) (124) (125).

This syndrome is one of a group of disorders related to the same (often *de-novo*) genetic lesion; a deletion at 22q11, with different presentations (122). Affected patients have increased susceptibility to viral or fungal infections, lowered T-cell counts and reduced lymphocyte proliferation in response to challenges (122).

Homozygous deletions in *TBX-1* have been shown to be lethal and produce and mimic the phenotype of DiGeorge syndrome (126). *TBX-1* regulates a number of transcription factors and has wide reaching downstream effects. In addition a number of other genes have been implicated, including *CRKL* and *CHD7*, but, as yet, the whole molecular pathology picture remains unclear (122).

### 1.2.5 Ataxia Telangiectasia

This autosomal recessive disorder is progressive and is both neurological and immunological in nature (127). Presenting with cerebellar originated ataxia, Telangiectasia, and oculomotor dyspraxia, it is caused by mutations in ataxia telangiectasia gene (*ATM*). *ATM* is a serine/threonine protein kinase involved in DNA damage response (specifically double stranded DNA breaks) in response to ionising radiation (128). The *ATM* gene's phosphorylation action is also part of the cell cycle checkpoint, and its deficiency leads to inappropriate stress response and cell cycle progression (129). Molecular modalities for this AR disorder are usually in the form of bi-allelic truncating mutations in *ATM* (130). However, milder forms have been identified in which missense mutations and leaky splice site mutations can precipitate the phenotype (130). In one example, aberrant inclusion of a cryptic exon was established, with a deletion of 4 nucleotides in intron 20. The deletion was found at 12 bp downstream and 53 bp upstream from the 5' and 3' ends of the cryptic exon respectively (131).

Immunologically, depleted T-cell numbers due to thymic hypoplasia can be observed and T-cell toxicity has reduced potency. Usually mild, the immunodeficiency element can ordinarily be treated with immunoglobulin replacement and prophylactic antiviral and fungal agent. Patients regularly (~15%) develop leukaemia or lymphoma in the second and third decades of life (96).

Several other T-cell specific immunodeficiencies do exist including Nijmegen breakage syndrome, NK cell deficiency (although not a true T-cell disorder), chronic mucocutaneous candidiasis, cartilage hair hypoplasia immunodeficiency, centromeric instability and facial anomalies syndrome (ICF).

Metabolic disorders can also present with some level of immune deficiency, and these include orotic aciduria, methionine synthase deficiency and biotin dependant multicarboxylase deficiency. The

various forms of immunodeficiency have a varied and complicated array of phenotypes, due to the many stages of development and function which can be affected, and the distinct genes involved in each stage (46).

### **1.2.6 Other types of genetic variant in PID**

There are a number of other mechanisms through which specific genetic variants manifest their effects to cause PID disease. Gene expression levels are a powerful indicator of pathogenic events in Mendelian disease (132). In the event that a gene expression level is outside of the physiological range, it can be identified as an outlier by using methods based on statistical interpretation of normal ranges. These effects are often correlated with gain or loss of function, splicing, and structural variants (133).

Some of the variants which bring about PIDs have been covered earlier in 1.2, however a recurrent problem is that the causal variant/s are not always obvious. In addition to those variants occurring in non-coding segments, a problem facing those seeking diagnoses are variants with unexpected consequences. Variants may exist in genes not currently understood to be linked to the disease or phenotype, and so can potentially be filtered out in the informatic process (134).

Disorders falling under the PID umbrella can be a result of quantitative differences in gene expression as opposed to qualitative differences in the expressed specific transcripts. Cases of PID have been linked to variants which, although not present in putative PID genes, do affect the expression of genes or networks of genes implicated in the immune response (135).

Expression quantitative loci (eQTL), are genomic loci which are demonstrated to affect the expression of one or more genes (136). These occur throughout the genome, although the majority, surprisingly exist in the noncoding regions of the genome (137, 138). These eQTLs can be tissue specific also, so that one locus will only affect the expression of a gene in a particular tissue, but does not contribute to the expression profile in alternative tissues (139).

eQTLs result from genetic variation at the locus including such as single nucleotide polymorphisms. SNP's at eQTL loci have been demonstrated to affect the transcriptional level of other RNA's, modifying protein expression and causing phenotypic changes to the abilities and behaviours of cells as demonstrated in some immunological cases (140). These eQTLs explain a fraction of the genetic

expression of specific genes, and the vast majority do not exist in the coding regions of genes and are predicted to be involved in gene regulation (141). Important recent investigation results reveal that these eQTLs have a more pronounced effect on immune regulation than the effects of age and sex. Interestingly, immune stimulation exclusive effects have been identified for some of these eQTL variants (142).

### 1.2.7 Variants affecting alternative splicing and their role in PID.

Genomic variants which affect splicing patterns can affect splicing in a qualitative fashion (e.g. splice site is muted or a new splice site used) and quantitative fashion (splice site has increased or decreased affinity and isoform abundance is affected (143). Aberrations in the relative abundance of isoforms has been shown to be linked with disease, and these aberrations have been suggested to be useful as a predictive biomarker for disease and may present therapeutic targets (70, 144-146).

*Cis*-acting mutations pertaining to splicing are those which exist on the RNA molecule and can affect splicing primarily through altering the splice site recognition or altering exon splicing enhancer or silencer sites (53). Their *trans*-acting counterparts are those which affect the part of the molecular machinery responsible for splicing which binds to the RNA in a “*trans*” manner (147). Splice sites usually comprise GT and AG dinucleotides at 5’ and 3’ sites respectively (148, 149). Additionally, mutations in *trans*-acting splice factors – the splicing machinery of the cell, can also bring about disease (53).

The impact of mutations that affect RNA processing/splicing is currently providing a diagnostic revolution. Variants which affect splicing are those which occur in known active splice sites, regulatory elements such as exonic splicing enhancers and intronic splicing enhancers or the activation of cryptic splice sites (150). It is important to note that existing studies looking at variants affecting splicing in PID have determined that the variants which directly influence splice sites are more robustly linked to disease phenotypes than those which effect splicing regulatory elements (150). Interest in the detection of activated cryptic splice site has spurred on the development of a number of *in-silico* tools for prediction of splice site usage (150, 151). These tools are often unable to discern the resulting transcripts exon use patterns ~4/5 of cases (152). Whilst this is enhanced by other orthogonal investigations such as mini-gene assays (150), the multiple facets of splicing control involve more than just the sequence of the splice site in question. These include the activation of other splice sites within the gene, splicing quantitative trait loci, the relative abundance, phosphorylation status and localisation of different and often competing *trans*-acting factors (153).

Further complicating this process, the seemingly benign, synonymous variants which are normally removed by bioinformaticians during the filtering process can disrupt splicing (154). Cummings et al. evidenced this is the *POMGNT1* and *RYR1* genes demonstrated to be causative of Mendelian disease in muscle (155). This type of splice disrupting mechanism has been observed to be causal in cases of PID (156), and in fact, deep learning investigation techniques has shown that between 9%-11% of rare genetic disorders are caused by synonymous or intronic splice-altering mutations (157)

Indeed, much as gene expression can be influenced by multiple loci throughout the genome (eQTLs – section 1.2.6) so too can patterns of splicing. Splicing quantitative trait loci (sQTLs), are points existing throughout the genome which together determine the genomic contribution to the relative usage of splicing events (158). Analysis of sQTLs has been improved by RNAseq methodologies however, it remains a difficult affair as the isoform expression must be estimated using statistical methods (159). Additionally, these sQTLs are not necessarily in close proximity to the splice junction. Characterization of these sites in humans has shown SNPs demonstrating tangible sQTL activity at 100 kb from the relative splice site (160).

PID cases can also be caused by multiple variants which affect splicing in different ways. Compound heterozygous mutations in the *MALT1* were identified as causal in a case of profound combined immunodeficiency (161). The *MALT1* gene is a protease gene implicated in T-cell activation. The causal variants were identified by whole exome sequencing, and consisted of an inherited inactivated splice acceptor site, due to a change from the consensus AG to GG, and a de novo deletion of c.1059C which led to frameshift and premature termination (161).

It is important also, to consider that that genetic variation in non-protein-coding genes can cause disease (162). Examples within the PID research and diagnosis space include a recently discovered variant occurring in coding regions for genes comprising components of the minor spliceosome, which is used for the splicing of at least one exon in ~800 genes (163). Specifically, the noncoding gene *RNU4ATAC* that produces a small nuclear RNA (snRNA) termed U4atac was discovered to cause Roifman syndrome, by preventing normal minor intron splicing in *MAPK1* in B cells, and *DIAPH1* and *HPS1* in megakaryocytes. In the examples of Roifman syndrome seen, the retention of the intron introduces a stop codon and consequentially a truncated protein. Resulting imbalances in the *MAPK1*/*MAPK3* heterodimer leads to complications in cell morphology and a failure of survival and maturation of the cells to naïve B-cell state (164, 165).

Compound heterozygous variants in the *RNU4ATAC* gene, were found to be responsible. They were first discovered in an affected family after traditional filtering methods had not detected viable variants. The link was confirmed by the detection of intron retention during curated splicing analysis of RNAseq data (165).

### **1.2.8 Diagnostic challenges in primary immunodeficiencies**

The importance of early diagnosis in PID cases is high, with relation to both the patient's qualitative experience and the economic cost to healthcare services. Sources vary in cost analysis of undiagnosed PID, some say that whilst a diagnosed US patient costs healthcare services over US\$250,000 per annum, largely due to treatment costs, an early diagnosis of the disorder can save as much as US\$6500 per patient, per annum (166). An alternate source suggests an undiagnosed patient might cost the healthcare system US\$102,552 annually, once diagnosed these costs may drop by as much as \$79,942 (167). In a UK patient survey, 45% of patients reported a diagnostic wait time of between 1-6 years, around 1/6<sup>th</sup> reported waiting 10-20 years. Other key findings of the same survey confirmed undiagnosed patients bring about a dramatically increased burden on NHS resources (168). Identification of the precise molecular origins for each patient's case of PID leads to improved patient care (169), and improved prognosis. The importance of correct genetic cause for a PID phenotype is demonstrated by the different treatment preferences, which exist for conditions that may present with similar clinical phenotypes (170). Precision diagnostics can help to achieve this in part, by allowing targeted intervention to the specific molecular causes (171-173).

Challenges in diagnosing PID are numerous and diverse. Studies which correlate the phenotype and genotype have been useful in diagnostics, developing an understanding of various PID disorders (174). Additionally, these correlation studies have been useful for de-convoluting the pleiotropic nature of the involved genes, and establishing that a single mutation can bring about a variety of clinical phenotypes in people with differing genetic and environmental background although the exact mechanisms are not completely understood (175). However the development of a universal diagnostic pipeline for PID is hindered by the heterogeneity in presentation of disease, even among patients with what appears to be the same pathogenic genetic variant (genetic pleiotropy) (176). Conversely, several genotypes can bring about the same phenotype (genetic heterogeneity) (175). Once a clinical diagnosis of PID is suspected, mainly based upon a compatible phenotype, a family history is usually taken and number of subsequent laboratory tests performed to confirm the type of immune mechanism affected can be performed (177). As more causal genes are identified in cases

of PID, these can be combined into panels for screening suspected cases of PID as part of a diagnostic pipeline (178). With the emergence of targeted sequencing, clinical exomes and complete exomes through short read next generation sequencing technologies, the inclusion of genetic testing within a PID diagnostic work up has become more widespread. This approach to both adult and paediatric onset disease has consolidated the importance of protein based functional immune testing (cytokines, antibodies, etc.) for characterising the nature of the phenotypic presentation, but furthermore to evaluate candidate genetic variants in such pathways that have been identified through parallel germline DNA testing.

### **1.2.8.1 DNA in diagnostics**

DNA Sequencing-based genetic testing is often used where possible, as it provides the best diagnostic capability of existing clinically adopted methods (167). Whole exome sequencing (WES) sequencing currently provides the highest success rate (179, 180), and it achieves this despite the exome comprising only ~2% of the human genome (181). Around 85% of currently annotated variants exist within this portion of the genome (182). It has been hypothesised that this focus has likely led to the underestimation of the contribution to disease of non-coding variants (183). Researchers are calling for universal molecular gene testing for the diagnosis of primary immune deficiencies (184). Evidence from existing literature, however, suggests that even this may be inadequate; currently whole exome sequencing (WES) and whole genome sequencing (WGS) is only able to produce reliable diagnosis in 25-60% of cases (185-190) across a variety of disorders. A recent systematic review showed even greater range for PID, specifically, a diagnostic yield of between 15-79% was observed (191). Whilst these results are encouraging, they also suggest that the methods of variant identification and filtering remain in need of refinement. The development of WGS as a clinically validated routine testing modality is still in its infancy, although many countries have undertaken whole genome sequencing projects to evaluate this approach (192). Within the UK's 100,000 Genomes Project, PID were accepted as an indication for inclusion into the study. Plans to incorporate WGS for PID into routine clinical pathways have been approved following the transition phase of the 100K project to WGS sequencing in routine NHS care across England.

Confirmed formal genetic diagnosis of PID currently relies heavily on clinical interpretation of results. From understanding the phenotypes and prospective pathogenic mechanisms precipitating these features, to understanding of modes of inheritance family history, and understanding consequences of variants in relevant genes (193).



Crucial to this process is the ability of bioinformatic tools and databases to predict the significance of such variants. WES delivers around 20,000-23,000 variants per individual, and WGS produces 3-5 million variants per individual (183). This makes the task of identifying a causal Mendelian disease variant extremely difficult without a series of bioinformatics filters. The diagnostic process makes use of databases which give information, often in the form of scores, about the predicted pathogenicity of the many variants in an individual. These can be filtered further using lists of genes already known and evidenced to be relevant. The gene panels may be devised based on the genes having a known role in the biological system or process, or they may be produced from lists of known molecular diagnosis of genetic disease. By combining results from these different tools with clinical knowledge of the patient's presentation and family history, a causal variant in a known gene, implicated in a biological pathway or linked to the disease can be identified. Problems with the WGS/WES sequencing diagnostic methods also arise when no variant, identified through patient's genome sequencing, can be reliably linked to the clinical presentation and cytological/molecular manifestation of the disorder. Such occurrences include exonic variants of unknown significance, variants in intronic and intergenic non-coding RNA (162), variants in the *cis*-acting regulatory elements of transcription (194) imprinting disorders and repeat expansions (183).

Conventional clinical diagnostics, utilising human phenotype ontology for integration of cases into specific diagnostic groups, and traditional genetic sequencing methods for diagnostics then, are still currently inadequate, and whilst proteomic diagnostic methods are in development – they exist at a relatively early stage of development and can miss the potentially valuable RNA regulatory phenomena.

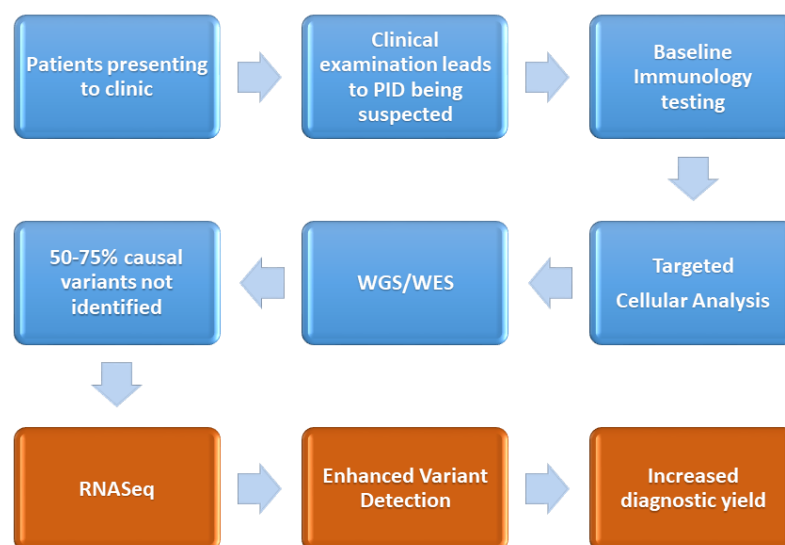
#### **1.2.8.2 RNA in diagnostics**

RNA investigation technology and literature has experienced great leaps forward in recent years in terms of technological advancement and cost reduction (195). RNA sequencing is now the most used quantitative method of mapping gene expression profiles (196). The transcriptome – or RNA expression profile of a given cell, or tissue can give unparalleled insight into the elegant inner workings of the cell. Through capture of all RNA species, it characterises the cytological gene transcription architecture and can deliver an instantaneous picture of environment–cell interaction or response programme (196, 197) .

A range of technologies exist for conducting RNA sequencing, each with its own strengths. Long read sequencing provides reliable structural information but can have sub-optimal reliability in base

calling (198), or is prohibitively expensive for high throughput analysis (199). Short read NGS RNAseq involves sonication or enzymatic degradation of RNA into smaller fragments, selection of fragments using one of a number of methods, cDNA synthesis, the construction of a library and subsequent sequencing (200).

Currently, this technology generates a mixture of both quantitative and qualitative analysis opportunities of RNA species: Qualitative transcriptome profiling outcomes include identification of sequence variants at the level of the genome (201), somatic cell mosaics, non-canonical splice variants, occurring either due to *cis* or *trans*-acting factor aberrations (53). Quantitative outcomes of transcriptional profiling include differentially expressed genes, alternative splicing events and allele specific expression quantification (200). Previous studies have demonstrated that when compared with a large control datasets, identification of expression outliers in peripheral whole blood can contribute to the detection of disease causing variants (202,203), after WGS/WES has been performed (Figure 1-6)



**Figure 1-6 The PID patient diagnostic journey**

Fig. 1 depicts the current diagnostic pathway from patient presenting to the clinic to the results of the whole genome or whole exome sequencing (blue) and where the RNAseq methodology joins the pipeline.

Regulation of alternative splicing of transcripts controls the relative abundances of RNA isoforms of genes. Gene mRNA Isoforms are often required to be kept at specific ratios as the isoforms can have differential function (204), or in some cases, can be antagonistic (55). Through RNAseq or exon junction-spanning probe-based capture, changes in isoform balance can also be resolved. Perturbations in the relative abundance of these isoforms is a driving force in the genesis of many diseases (205, 206). The sensitivity and suitability of RNAseq in transcriptomic investigation was demonstrated in mouse and human models, and has enabled the discovery of ~7600 novel isoforms in mouse Immune Cells (61) and detected 100,000 splicing events with at least moderate abundance (57).

In addition, transcriptome profiling can also give insight into control mechanisms exhibited by the non-coding RNA species, such as lncRNA, and miRNA, the significance of which is continually being elucidated in the molecular pathology of disease (207, 208). Indeed such examples exist in PID; miR-6891-5p accumulation is demonstrated to contribute to selective IgA deficiency, the most common form of PID (209). Thanks to the increasing ability of technology and steady reduction in costs, researchers are also able to cast a wider net.

DiGeorge's syndrome mentioned earlier, is an example of this, in which due to a chromosomal deletion at position 22q11.2, thymus glands are unusually small or in some cases – absent. Alternatively, maturational stages of T-cells can be unaffected but have signalling malfunctions. X-linked immune deficiency is not a result of maturational abnormalities but instead results from abnormal T-cell surface antigen; CD40 ligand. T-cells cannot then signal B-Cells differentiate into IGG producing plasma cells, resulting in a class-switching failure and continued or hyper-production of IGM (96). Such examples are very likely to be able to have a detectable signal in a transcriptomic analysis pipeline through RNAseq.

Through RNAseq based investigation it is possible to examine all of the mRNA species destined for translation in hypothesis free methods and create profiles of normal transcriptomes and pathologic transcriptomes in a tissue specific manner (210). Through applying appropriate bioinformatic pipelines and algorithms, this transcriptomic data can be used for biomarker identification (211). It is also possible to quantify the relative expression of genes coding for the splice factors themselves, which can directly bring about specific pathological processes specific to PID, such as those observed in Roifman's syndrome, mentioned earlier (164, 165).

Micro-fluidic technology adaptations have improved the development of robust, single-cell transcriptomic profiling. In combination with NGS based technologies the single cell technology provides a method for profiling the transcriptomes of individual cells, giving unparalleled insight into the heterogeneity of cell populations and their transcriptional profiles (212). Adaptations such as the SMART-seq2 or Fluidigm C1 library preparation methods also now allow the production of full-length cDNA's, giving transcript isoform level resolution. However these methods do not yet allow multiplexing, massively increasing overall costs and labour in large cohorts (213). The ability to profile the entire transcriptome of a peripheral blood mononuclear cell (PBMC) culture on a single cell basis would give a dramatically increased ability to understand the specific cell populations, sub-types, and cell-cell interactions taking place in an immune challenge *in-vitro*, using existing methods of immune challenge such as those outlined by Martkamcha *et. al.* (2016) (214) This approach could also then be utilised in those patients who are suspected to be genetic mosaics.

### 1.2.8.3 RNAseq parameters

RNA sequencing using next generation sequencing technologies comes with a wide variety of variables in library preparation methods, sequencing parameters and optimisation options. Depending on project questions, material qualities and approaches, the pipeline can be quite distinct and give varying information specificity and sensitivities.

Paired end reads provide double the amount of sequence information, without doubling the cost of the experiment, by sequencing portions at either end of the same fragment. This method is also particularly useful and informative for alternative splicing and novel isoform detection, which are part of the parameters of our investigation. In addition, it has been demonstrated that single end reads can produce a significant number of false positives and false negatives when assessing differential gene expression (215).

Globin is suggested to make up 80-90% of the transcript species in peripheral whole blood (216). Globin depletion methods increase the number of species of detectable transcripts and reduce the number of globin mapped reads from ~80% to ~20 % (216). Poly-A enrichment methodologies have demonstrated increased read mapping to exons (71%) over ribo-depletion alternatives (22%) and have been recommended for alternative splicing investigations (217, 218). This increased read mapping provides more usable reads, improving resolution, which allows for better identification of differential expression incidences. Whilst neither Poly-A selection methods nor rRNA depletion methods are completely free from non-specific effects, poly-A selection methods are preferred for

novel isoform detection in non-degraded samples (217). PAXGene stabilisation tubes provide a robust nucleic acid capture and storage method. However, when using patient's peripheral whole blood, RNA concentration is relatively low, and this can be exacerbated in PID patients, who may be experiencing lymphocytopenia (219).

#### **1.2.8.4 RNAseq analysis techniques.**

At present the whole genome and whole exome technologies employed for molecular diagnostics are an excellent recent addition to the NHS service but have a limited capacity to provide accurate diagnosis in cases of PID. Current literature suggests a diagnostic yield of between 15 -79% (191). This is in part due to the variant filtering process, which is laborious and limited in its reach. In other clinical areas, RNAseq is demonstrated to enhance diagnostic capacity by informing the variant filtering process (220, 221).

RNA Sequencing can provide insight into several aspects of cellular events. It gives good insight in gene expression by assigning and counting all reads to specific loci, strands, genes, alleles, and transcripts (200). A range of tools are available for follow up bioinformatic analysis of the data generated by RNAseq, which have a variety of functions, strengths and weaknesses. These tools utilities are also dependant on the type of question being asked, and the resources available to the researcher (200). This project aims to assess the total gene expression, transcript usage and novel splicing events.

Many informative and useful metrics can be extracted from RNAseq data. As a primary example, the range of RNA transcripts present can be observed, such as mRNA, miRNA, lncRNA, siRNA, snoRNA, piRNA, circRNA, and tRNA (222-224). In addition, expression levels of these RNAs, relative isoform abundance, splicing, and expression of specific alleles can all inform research about pathology or mechanisms (225). Due to time and funding limitations, this study is aiming to look at protein coding genes only. Although data will be retained for future analysis.

The most straightforward and common way to punctuate and decode the result of variants of unknown significance in the genome using RNA, is to look at expression. Gene expression is already used to investigate and model immunodeficiencies and generate profiles for diagnostic and comprehension purposes using hybridisation based methods (226). RNAseq is already evidenced as an extremely valuable tool for generating these profiles in other disorders, and a number of methods for detecting changes in gene expression exist (227, 228). Due to the utilisation of

sequencing rather than hybridisation methods, it also allows for detection of previously unknown splice sites and events (141). As discussed later, the abundance of isoforms and differential alternative splicing is intimately tied to T-cell activation (74).

The use of individual exons provides another useful metric, which has more reliable quantitation in short read RNAseq (229). This is because isoform usage essentially relies on algorithms to infer the isoform to which a mapped read belongs. In some instances, this is based on probability and as such comes with inaccuracies (230). However, the inclusion or not of specific exons is less easily tied to function. Specific isoforms and their respective proteins have had many years investigation through molecular biology, but these the inclusion or not of single exons cannot always be robustly associated with function, as other splicing events within transcripts can cause further functional differences.

#### **1.2.8.5 Considerations in diagnostics**

In order to assess the impact of genomic variation on the unstimulated immune system, the normal immune response and the immune-deficient responses, it is important to experimentally ‘tune out’ the variations in signal arising from environmental factors. Some of these are outlined below.

##### **1.2.8.5.1 Tissue selection**

Untreated whole blood is the simplest patient sample to acquire, and as shown in a review of the literature, can inform diagnosis of PID in patients around 50-60% of the time using WGS/WES (183, 220, 221, 231) . It also has the advantage of not requiring a high degree of preparation or technical training to produce as a sample. However, in the case that differential expression in a specific cell subtype is contributing to the disease progression, this signal may be harder to identify in a mixed cell sample, if for example the expression of this gene or isoform is at normal levels in other cell types.

It has been established that a high degree of the variation in CD8+ cell populations can be attributed to environmental factors. These include metrics such as cell population frequencies and signalling responses to inflammation driving cytokines such as IL-10 and IL-6 as measured by mass cytometry, flow cytometry, and immune signalling experiments (232). This makes them a poor model for genetic variant impact. CD4+ T-cells display a large degree of heritability in these assays and as such should provide a good level of transcriptomic heritability also, allowing for clearer elucidation of the effects of variants on differential gene expression (232).

The immune system's response to pathogen-based challenges is highly dynamic, and observing this dynamic response is therefore more informative regarding identification of impaired response (233). Indeed, it has been shown in innate immune system studies, that the effects on differential expression of some variants can only be observed in a dynamic fashion (140, 234). Similarly, some splicing quantitative trait loci have effects only manifest during an immune challenge (235).

Co-culture of PBMCs provides a greater insight into activation pathways as it allows for the cell – cell communication response programs and produces similar results in terms of ranked gene expression response networks, with a few notable exceptions (233).

Studies of dynamic immune responses to challenges, in concert with machine learning can be used to identify small groups of stimulation pathway-specific genes (236). Comparing the expression profiles of these genes in healthy cohorts with PID patients can potentially be utilised to identify candidate genes that may then harbour a disease-causing variant or indicate some anomaly in the pathway for further investigation.

Individuals' immune responses to pathogenic challenges are exceptionally variable, and the variability in these responses is not easily elucidated. As mentioned earlier, age, sex, seasonality, nutrition and lifestyle all have effects on the specific response profile exhibited by individuals (142). These factors that influence responses have a greater degree of significance in specific cell types. CD8+T-cells for example show a high degree of heterogeneity in the context of temporal changes through the life course of the individual for example, and CD4+T-cells and monocytes are heavily influenced by sex (142). It is therefore useful to be able to discern transcripts from different cell types within a culture. Utilising flow cytometry to separate cell types and then multiplexing the sequencing runs or utilising single cell RNAseq becomes an attractive option. The most appropriate method choice comes down to the number of cell types which need to be analysed, the capacity of the flow cell and the required sequencing depth. The transcriptomic landscape provides an excellent opportunity for advancement of diagnostic yield, and transcriptional profiling has already begun to be utilised across a range of disorders to help build a "molecular fingerprint" of disease. The Immunology community has made a case for PID diagnosis to be supported using transcriptional profiling using whole transcriptome sequencing (190), and these have begun to be answered with examples in primary immunodeficiency cases such as Dock8 CID, GATA2 deficiency, X-linked reticulate pigmentary disorder (XLPDR) (237-239).

### 1.2.9 Hypothesis-free approaches in RNAseq diagnostics

Modern tools such as GWAS and RNAseq allow large scale analysis, without the need for a formal scientific hypothesis (200). This avoids much bias and prevents the investigation missing novel or unelucidated aspects of the topic. In these instances, comparisons of entire genomes, epigenomes and transcriptomes informs novel hypotheses. However reduced sensitivity or specificity can mean signals are harder to detect.

In contrast to the narrower range of proteomic detection mechanisms which also allow for quantitative analysis, RNAseq can easily identify new isoforms of transcripts whilst also giving an indication to their relative abundances, or indeed their absolute abundances in the cases of the Oxford nanopore technology.

Some studies of transcriptomes in other species have begun combining hypothesis free and hypothesis driven analysis, with positive results (240), and the progressive pharmacogenomics field of research has also been to employing these techniques using synergistically to discover 'paradigm shifting' results (241).

Conventionally, large scale hypothesis free RNAseq analysis has been conducted in a manner which compares one set of control data with another set of data which has had some variable altered, and then statistical differences are calculated and investigated (242). Known as "differential" comparison it can be used for expression, exon usage and isoform abundance.

One problem with this approach, is that it needs two homogenous datasets to compare, and in clinical diagnostics this is not a realistic possibility. What is needed is a method to look for aberrant events within a population, or the ability to compare a sample to a sample of healthy controls.

For RNAseq, this has traditionally been difficult as reference ranges will vary depending upon the handling and sequencing method and host of other variables (243). A traditional method for identifying outliers in gene expression is to calculate Z-scores and filter out those below a pre-determined cut-off threshold (210, 244). Brechtmann *et al.* have developed a tool which is able to control for these variables, known and unknown, correct for batch effect and search for outliers in expression using the negative binomial statistical test. The software, known as OUTRIDER, uses a machine learning autoencoder to control for the variables and generate an adjusted count matrix. The model is then applied and the algorithm outputs gene expression outlier tables with log<sub>2</sub> fold change, p-values, p-values adjusted for the false discovery rate, Z-scores, and ranking information.



OUTRIDER is a hypothesis free analysis tool, developed specifically for RNAseq datasets, and compares differences in an intraspecific manner (within a population) as opposed to between groups (132). This is particularly important when considering primary immunodeficiencies, given rarity, and the heterogeneity in both genetic cause and symptoms mentioned earlier. As such this tool presents as an optimal choice for looking for specific changes in rare conditions.

The degree to which hypothesis free methods such as OUTRIDER and EdgeR can succinctly resolve signals, when compared to hypothesis driven methods, is not completely clear, but would be useful knowledge for clinical utility.

### **1.3 Secondary Immunodeficiencies – Immunosenescence**

Secondary Immunodeficiencies are those which are acquired, either through drugs, surgery and trauma, extreme environmental conditions, chronic infections, or malnutrition (245). A relatively newly classified, but ubiquitous example of immune deficiency, is that which occurs as a result of changes in the organism during ageing (246). A feature of the human immune system is that cost of maintaining memory of pathogens over the course of the lifespan depletes the ability of the immune system to respond to novel pathogens in later life, as the cellular compartments numbers shift further from recognition to memory. This is just one feature of a phenomenon termed immunosenescence, which describes the age-related decline or dysregulation in immune function (247, 248).

#### **1.3.1 Epidemiology and Demographics**

Currently, 1 in 11 or 703 million people are above the age of 65 worldwide. By 2050 this is predicted to be 1 in 6 or 1.5 billion people (249). Increasing age is positively associated with both morbidity and mortality from immune mediated inflammatory disease (250), neurodegenerative disease (251), cancer (252) and infectious disease (253). Respiratory tract infections in particular are a major cause of death across the world and two of the top ten global causes of death are a direct result of

infection (254, 255). This is set to increase dramatically as the population of the world continues to age, unless something can be done to improve prognosis of infection in aged individuals.

With advancing age of the individual, immunological competence is decreased (immunosenescence) (248), and as the annual influenza cases and recent COVID-19 pandemic has demonstrated, this leaves the elderly at increased risk of adverse outcomes from infectious disease (256). To further exacerbate the problems that communicable pathogens pose to the aged population, the current preferred prophylactic treatment for infectious disease; vaccinations, have blunted effects in the elderly as a result of immunosenescence (257).

Age-related decreased immune efficacy is currently the focus of much attention (258). Diseases exacerbated by or directly a result of ageing of the immune system are extremely numerous and have outcomes which range from frequent pain to profound shortening of life (259-261). Through developing a comprehensive understanding of the ageing of the immune systems and subsequently restoring its function, the hope is that much of the age-associated morbidity can be compressed into the later stages of life (262). Immune-rejuvenation adjunct therapies might precipitate enhanced vaccine responses and indeed immune resistance to infectious disease and cancer, and this potential is driving investigation into the cellular and molecular features of immunosenescence (248). Drugs which now target the core biological mechanisms underlying ageing (geroprotectors) are being trialled in this immune arena and show promise in preclinical testing (263).

### **1.3.2 A general description of Immune senescence**

Immunosenescence is a broad label which refers to the age related changes occurring in the immune system, and unlike cellular senescence, the term has neither specific functional nor mechanistic designations (262). Whilst the molecular cellular effects of ageing on immune cells and tissues are increasingly elucidated, the precise driving mechanisms driving immunosenescence are not certain as is the case with the fundamental drivers of ageing itself. Some argue the accumulation of somatic cell mutation drives molecular ageing (264). DNA damage is known to be implicated in the induction of senescence, and animal models show that when this occurs in the haematopoietic cells, it can lead to immunosenescence (258). Recent work has disputed this and suggested that epigenetic changes resulting from repair of DNA after mutation is responsible, and not the mutation process itself, however this study has limitations and potential conflicts of interest were apparent (265).

Stem cell exhaustion, cellular senescence, compromised autophagy, dysregulated nutrient sensing, dysregulated RNA splicing, are among 14 currently accepted hallmarks of ageing (266). However, as these factors all overlap and induce each other, and so discerning the most fundamental has been challenging. Most likely, there are multiple converging molecular basis of cellular ageing, and as these affect immune tissues, so too, the immune system ages. The encompassed facets of immunosenescence include chronic low-grade inflammation (inflammaging), reduced capacity to clear infections and cancerous cells, a reduced ability to respond to novel antigen, impaired wound healing and increased autoimmunity incidence (267-269). Inflammaging is connected to various features of immunosenescence, age-related chronic disease and is a known contributor to the decreased cell-mediated response to pathogens observed with advancing age (270, 271). The link between the immune system and neurological function is well established and appropriately immunosenescence is also associated with age related neurodegenerative disorders (272).

Ageing of the hematopoietic system contributes to a decrease in efficacious adaptive immune response. Mouse models demonstrate that this comes as a result of genetic up-regulation of myeloid lineage specific genes, and down regulation of those genes which garner lymphoid specificity in the daughter cells produced (273, 274).

Individuals have a marked difference in immune responses resulting from both genetics, ageing, and environmental factors (268, 275, 276). Immunosenescence is a major axis of variation when considering the heterogeneity of immune response in individuals (275). Some literature suggests that elements of immunosenescence can be causally decoupled from chronological age and indeed, to some degree, the processes of biological ageing too (276). There is a growing body of evidence which demonstrates that some aspects of immunosenescence are linked to antigen exposure history, cell ontogeny programs and intrinsic cellular defects (276, 277). Indeed in a recent landmark study using *Drosophila Melanogaster* as a model, over 70% of gene expression changes previously believed to be linked to advancing age in the immune system, were completely decoupled from chronological age in flies grown in a germ free environments (278).

If these mechanisms can be observed further in humans, it opens up the possibility of immune regeneration possibilities through targeting expression networks of those 70% of genes decoupled from biological ageing and related to antigen exposure. Additionally, this discovery also narrows the focus to a much smaller subset for understanding the fundamental mechanisms of immune ageing.

### **1.3.3 The effect of immunosenescence on viral infection and vaccinations**

The demonstrable effect of immunosenescence on pathogen clearance are a primary factor in the mortality figures from infection (279, 280). As an example, 80-90% of influenza deaths occur in individuals over the age of 65 (281). The effects of immunosenescence are not limited only to infections, and a plethora of evidence shows that vaccinations themselves have extremely limited efficacy on individuals with high levels of immune senescence (282-285). During the early stages of the 2020 global COVID19 global pandemic, data emerging from Wuhan, China indicated the primary risk factor for progression to acute respiratory distress syndrome from COVID19 was indeed, age (286). By May 2021, Centres for Disease Control data was indicative of over 80% of death were in those aged over 65 (287). Whilst co-morbidities are independently associated with age, age itself remains the most significant risk factor for COVID-19 mortality (288).

Impaired vaccination response in both old and young cohort has been linked to factors such as C-reactive protein, which mediate the chronic low-grade inflammation that characterizes immunosenescence (289, 290). Pre-vaccination levels of systemic IL-6, a known indicator of immunosenescence, has been linked with reduced vaccine response in animal models (291). Early work on baseline immune signalling was able to use these approaches to predict vaccine response with relatively good accuracy (289). As mentioned previously, modern literature and drug trials now set their sights on tackling the fundamental causes of immune ageing as a means to enhance the efficacy of vaccinations, and immune response to infections in general. In addition, after the association between immunosenescence and COVID19 morbidity and mortality was established, IL-7 was identified as a potential for adjuvant therapy (292).

### **1.3.4 Cellular and molecular drivers and features of Immunosenescence.**

Whilst the fundamental drivers of immunosenescence are not completely understood, some primary factors are believed to be cellular senescence within the immune tissues, and chronic aseptic inflammation throughout the rest of the body, known as inflammageing, resulting from the senescence associated secretory phenotype (SASP). The SASP varies from tissue to tissue, but interleukins (commonly IL6, IL8,) matrix-metalloproteins and chemotactic proteins are all known to be involved to some degree. Cellular senescence, whilst semantically discrete from immunosenescence, contributes directly to its occurrence by altering production and maturation of immune cells, and reducing the efficacy of their ability to clear other ageing cells (293). The effects of ageing on the innate immune system remain less well characterised than that of the adaptive

immune system, but efforts are increasing in this area (262). Reports around absolute abundance of the various cell types seem to suggest that numbers of cells seem to remain fairly constant; primary differences appear to be in ratios of subsets of cells. Neutrophils and macrophages both show reduced phagocytosis and chemotaxis. Neutrophils have impairments to superoxide production, recruitment of molecules onto the lipid rafts, resulting signal transduction and apoptosis (294). Macrophage cytokine production is also impaired, along with signal transduction and increased PGE2 production (248, 294). Similarly, some populations of dendritic cells suffer an inability to release cytokines (248). They also have decreased IFN I/III production, antigen presentation, TLR mediated signalling and chemotaxis/endocytosis (294).

PBMCs demonstrated a delayed and incongruent release of cytokines and chemokines upon stimulation by pathogen-associated molecular patterns (PAMP's) (248). TLR expression appears to be unchanged with age, indicating pathways which transduce signals may themselves be responsible for the resultant delayed and incongruent cytokine and chemokine release observed (295).

NK cells undergo shifts in relative numbers of high CD56 expressing 'bright' cells which normally constitute around 10% of total population and low CD56 expressing 'dark' cells which make up the remaining 90% (296). NK cells appear to increase in number in individuals who age healthily, whilst concurrently losing some function in both oxidative burst and phagocytosis. As immunosenescence progresses, NK cells have decreased cytotoxic receptor expression of NKP46, NKP30, DNAM1, and increased KIR, NGK2A and PD-1, inhibitor signals (296).

One of the best defined features of immunosenescence is a change in relative population numbers of memory and naive T-cells (259). In some cases absolute accumulation of memory T-cells can occur in people infected with HCMV (297). Additionally reduced numbers of naive T and B cells are seen. This reduction is associated with thymic involution, and a reduction in the IL-7 which stimulates the thymus and the lymphoid lineage cells (298). Increases in P16INK4a, and CD57 expression are indicators of immunosenescence in T-cells (299) and a reduction in T-cell receptor diversity occurs with advancing age (259). CD28 is known marker for cellular proliferation and enhanced stimulation in T-cells, and is observed to have reduced presence on CD4+ and CD8+ T-cell as age progresses (259). In late passage cells in in-vitro this can be as much as 50% (300-302).

Dendritic cells and naive T-cells appear to have the greatest impact on age related immunity changes (303). In studies of vaccination response for yellow fever, reduced numbers of neutralizing antibodies, CD8+ T-cells and new CD4+ T-cells (303).

Treg cells, which are generally identifiable by the expression of CD25 and FOXP3, are heavily implicated in ageing and reviews of available data show they represent an increased proportion of the circulating T-cells as age progresses (304). Treg cells suppress T-cells, NK-cells, dendritic cells and monocytes. Mouse models suggest Tregs limit final numbers of T-cells as opposed to replicative rate or first division time, suggesting the effects are mediated through some concentration gradient. In support of this the effects could be mimicked by IL-2 and CTLA-4 gain and loss of function analysis (305). This together indicates Tregs 'mop up' IL-2 and CTLA-4 prohibiting their promotion of immune function in other cell populations (305). Tregs also produce inhibitory cytokines TGF- $\beta$ , IL-35 (306), and IL10 (307). Through these methods they are able to suppress immune responses and induce apoptosis in effector cells. Aged CD4<sup>+</sup> Treg cells appear to have reduced ability to downregulate IL-17 and IL-2 (304). These changes in Treg cells number and function can lead to both increases in autoimmunity or a suppression of the immune systems clearance of pathogens and mutated cells (304). IL-6, TNF- $\alpha$  exist as more general markers of immunosenescence in peripheral blood (299). This summary is by no means exhaustive, and literature continues to elucidate further changes in immune cells and molecular biomarkers.

### **1.3.5 Alternative splicing and immunosenescence**

The science of RNA biology and the associated technologies are suitably mature to provide highly detailed, quantitative and thorough investigation into a variety of physiological states and developmental stages. Transcriptomics shows us that a huge variety of species of RNA transcript are able to be produced from comparatively small numbers of genes. A central mechanism through which the genome is able to produce such a varied and plastic transcriptome is through the alternative splicing of RNA. Deep sequencing studies have shown that >95% of all human genes are alternatively spliced (57).

Transcriptome studies of gene expression and splicing have started to give unparalleled insight into mechanisms of human development, health and disease (155, 308, 309). The association of ageing, age related disease and changes in the transcriptome is becoming increasingly well documented and maps of changes in gene expression and splicing are being produced to help understand the association (70, 310, 311). RNAseq methodologies been used in conjunction with statistical methods to quantify the association of expression and splicing changes in each gene with advancing age in a variety of tissues (311). One such study found that relative splicing levels were more reliably able to predict biological age than gene expression or absolute isoform abundance (70).

Attempts at discerning the causality or regulatory mechanisms of these splicing changes has revealed promising therapeutic opportunities, and targeting splicing has proven to be a valid method of ameliorating the pathogenic changes to cells (312). Using RNAi screens, Georgilis *et al.* (313), were able to show that splicing factors are able to directly mediate the senescence associated secretory phenotype; one of the cellular aspects of advancing age. This cellular phenomena is one which has been linked to number of age related diseases, including through mechanisms associated with dysregulated alternative splicing (314). Splicing factors are expressed at lower levels with advancing age, proteomics studies show the spliceosome itself is disrupted (315, 316). Research has shown that alternative splicing is highly correlated to ageing and age-related disease across tissues and can directly affect aged cell function and phenotype (313, 314, 317).

Genes essential to immune function and response are alternatively spliced in pathological conditions and during inflammation (235, 318). Analysis of splicing in single cell types has also shown the stimulation causes remodelling of the transcriptome via splicing (235, 318, 319). These include adaptor proteins, (e.g., CD3, CD28, CD8, CTLA-4, MAP4K2, MAP3K7, MAP2K7, CD45, VAV1) transcription factors (such as NFKB1 and STAT2) chromatin modifying enzymes and RNA binding proteins (235, 318, 319). Disruptions to normal splicing patterns occur in inflammatory disorders and autoimmunity (320), both of which are features of age related immunosenescence (289, 290). Complicating the picture, viruses themselves also hijack host machinery, which they rely on to create proteins from their own genome, and disrupt the hosts cellular processes (321). Studies using cell lines infected with Influenza A virus showed a broad program of changes to host gene alternative splicing (322). Sars-cov-2 proteins bind to U1 and U2 splicing RNAs and suppress global splicing activity (323). Alternative splicing patterns then, are affected by individual host differences in genetics, age, and environment. Splicing, like gene expression, provides a mechanism for direct response to immune challenge but also are disrupted as a result of viral manipulation. It is therefore potentially important to explore and map the effects of immunosenescence in various infections to identify shared and discrete features. The stimulated immune system is yet to be investigated in a manner which demonstrates the effects of immunosenescence in a human cohort.

### **1.3.6 Using transcriptomics to assess immunosenescence in COVID19 and Influenza**

Studies of transcriptomics and ageing have produced hugely useful biomarkers of ageing in specific tissues which not only enhance greatly the current understanding of the biological ageing process, but also provide an *a-posteriori* list of potential therapeutic intervention targets.

Observing challenged immune system of an age a stratified cohort, is likely to be highly informative and give different indications than baseline immunosenescence studies, eliciting the identification of pathways and processes impaired in the immune response (233). Indeed, it has been shown in innate immune system studies, that the effects on differential gene expression of some primary immunodeficiencies can only be observed in a dynamic fashion (140, 234). In ageing studies conducted on human monocytes, only when the cells were stimulated by TLR4 and TLR7 /8 and retinoic acid inducible gene 1 agonists were significant differences seen in expressed genes; decreases were seen in IFN- $\alpha$ , IFN- $\gamma$ , IL-1 $\beta$ , CCL20, and CCL8 production whilst increases were seen in CX3CR1 (324). It may well be that immunosenescence related changes in other genes may also be obscured from view until challenged.

Because of the age-related mortality profile with infections such as Influenza, they present an important and useful opportunity to gain some insight into the effects of immunosenescence during an immune challenge. In accordance with these efforts, the present study aims to characterise the effects of ageing on the challenged immune system using the suitably mature RNA space and technologies.

Using SARS-COV-2 and Influenza cohorts we aim to map the changes in whole blood transcriptomes of patients and look at the contribution made by changes in gene expression and alternative splicing.

#### **1.3.6.1 Ageing Clocks to quantify or diagnose immunosenescence.**

Immunosenescence is organism wide and encompasses changes across many tissues and cells (262). Investigations must make use of a variety of different techniques and orthogonal approaches. These include comparing genomic, metabolomics epigenomics, and transcriptomics (325).

The ability to tackle age-related causes of disease necessitates the characterisation of ageing in tissues as it manifests, particularly because ageing occurs in a heterogeneous manner in different individuals, whereby different organs and systems decline at different rates. As with producing strategies to tackle disease, identification of biomarkers of ageing has been highlighted as one of the most important tasks facing geroscience research and has enormous potential in both fundamental and translational biomedical research (326-328).

In an effort to systematically investigate the molecular changes occurring with age, modern studies use machine learning in the form of penalized regression methodologies to identify molecular changes most robustly associated with advancing age (329).



These models, now termed 'ageing clocks', take advantage of the extremely large datasets now being created in the areas of epigenomics, transcriptomics, proteomics and metabolomics (330). Transcriptome based ageing clocks have been demonstrated to be particularly effective (331) and are now approaching the theoretical limit of accuracy, whereby random noise between individuals prevents statistical accuracy (332). As such, transcriptomics is a useful modality for mapping and understanding mechanisms of ageing, despite the heterogeneity of ageing process.

Ageing clocks can be used to accurately predict the chronological age of individuals. Consequently, they can also be used to quantify the rate of ageing of that individual compared with peers, and this can be done in a tissue specific manner. The applications of this are far reaching, from understanding and informing biological research, to pharmaceutical and therapeutic developments to comparing the physiological effects of lifestyle and environmental factors as disparate as smoking or population migration of ageing in specific systems (333).

While some broader, pan-tissue clocks have great utility, disease specific clocks have been noted to be of enormous value as disease specific monitors and risk calculators in the clinic (334).

The first steps towards immune ageing clocks were made with the development of a multi-omic iAGE clock which quantifies systemic age-related inflammation. This ageing clock correlated to multimorbidity, immunosenescence, frailty and cardiovascular ageing (335).

The immune systems primary functions extend beyond its interactions with host tissues however, and one of the features most associated with immunosenescence is effective pathogen clearance. This feature is of particular interest currently due to the global pandemic and the robust associations between advanced age and COVID-19 mortality (336).

### **1.3.7 Summary**

The early and accurate diagnosis of Primary Immuno-deficiencies is important to ensure that positive patient outcome is attained, and economic cost is minimised. Diagnosis of the disorders remains difficult due to clinical challenges in identifying the presence of a primary immune system disorder, stratifying the phenotype to a myriad of overlapping candidate genes and then the laborious task of variant filtering, interpretation and lack of knowledge pertaining to variants, especially those

residing in the non-coding segments of the DNA. Functional validation of a candidate variant is currently undertaken with protein-based *ex vivo* tests which are difficult to standardise and mostly available in research laboratories.

Primary Immunodeficiency provides an interesting model for several reasons; the heterogeneity of presentation even in patients with the same variants, and similar presentation in patients with different variants presents an interesting diagnostic challenge. It is hoped that utilisation of RNAseq alongside WGS/WES can help to avoid diagnostic odyssey type eventualities.

In addition to these interesting challenges, the primary tissue of investigation for PID is usually whole blood. This is readily available, easy to isolate, and comes without a great deal of risk, in contrast to some other contemporary examples of RNAseq diagnostics.

RNAseq is an emerging molecular profiling technology which, when combined with WGS/WES provides unprecedented insight into differential gene expression, splicing activity, allele specific expression and may contribute a further insight into candidate's variants derived from proband or family-based WGS/WES sequencing results.

Hypothesis free transcriptomic analysis hold promise with relation to both disease mechanism elucidation and diagnostics i.e., variant filtering (183, 221). However, RNAseq remains relatively novel as a diagnostic testing tool in rare diseases and the control datasets and cellular contributions to complex tissue profiles (i.e., whole blood) will require further dissection.

Utilising large control datasets for comparison enhances the power of the transcriptional profiling through RNAseq and improves resolution for differential gene expression. Existing projects have developed these datasets for whole blood and immune cells, which provide a starting point for the interrogation of clinical samples for diagnostic research.

Over the coming years, an extended diagnostic approach to PID testing may develop that builds on a clinical module of phenotype, family history and baseline immunological testing. This will be complimented by a DNA module of coding and non-coding variant analysis, utilising sophisticated bioinformatic pipelines to prioritise candidate genetic variants of new loci that would be consistent the clinical phenotype and family segregation. These effects of these candidate variants for monogenic disease may then be functionally interrogated via RNAseq.

In parallel, functional testing of candidate genes through protein-based assays may be undertaken to characterise the impact of a putative monogenic pathogenic variant within a reductionist model at

the protein level. The sharing of these modular assessments across the international community will incrementally improve the standardised analysis of novel variants that will continue to grow over the next few years.

There is still much to be done to improve the current diagnostic yield in PID and RNAseq combined with WES/WGS with a large control dataset will expedite the discovery of PID causing variants, in tissues with high heritability. The presence of candidate gene lists will further enhance the diagnostic capabilities, as filtering and processing of large datasets remains challenging.

Immunosenescence is a broad process which contributes to many of the leading causes of morbidity and mortality. Reviews of the literature show that whilst retrospective analysis of the outcomes of immunosenescence has been studied, and baseline biological analysis of healthy individuals of a range of ages is now a mainstay approach to this topic, there are very few examples of cohort studies of immunosenescence occurring in challenged immune system. Two particular cases of immune challenges with age specific outcomes are Influenza and more recently COVID19 (sars-cov-2). Transcriptomics is a suitably mature modality to use to investigate these models and gene expression and alternative splicing has been shown to provide comprehensive insight in other examples. Given the competing forces of ageing, immune response and viral infection from RNA based pathogens, transcription and splicing are of particular interest as metrics to evaluate the effects. It has already been established that an immune challenge is required to properly capture and quantify problems in immune function (140, 234), and we hypothesised that an ageing clock for infectious diseases that have an age related morbidity and mortality profile might provide previously undiscovered insight into pathogenicity of such diseases. Moreover, less specific multi-infection ageing clocks, derived from mixed cohorts of infected patients might be informative to the more general aspects of immunosenescence pertaining specifically to immune challenges as opposed to baseline immune activity.

Reviews of the literature showed also that many transcriptomic ageing clocks were neglecting to account for the effect of alternative splicing in the ageing transcriptome. Splicing factors are expressed at lower levels with advancing age, proteomics studies show the spliceosome itself is disrupted (315, 316). Research has shown that alternative splicing is highly correlated to ageing and age related disease across tissues and can directly affect aged cell function and phenotype (313, 314, 317) and early work on multi tissue age prediction showed that alternative splicing was more robustly able to predict age than gene expression or isoform expression level changes (70).

### 1.3.8 Aims of research

In this section, a brief discussion of the general aims are presented followed by a concise list of formal research questions.

There exists a gap in PID diagnostics for research specialising in transcriptomic and genomic based diagnostic methods for T-cell disorders. This project aims to discern the potential of RNAseq technology to enhance this diagnostic ability by informing the variant filtering process. An important consideration is the translational capacity of the research – can the diagnostic methods succinctly be applied in a clinical setting, and what is the minimum viable complexity of assay implementation? Alternatively, what is an acceptable trade-off in sensitivity, when this complexity is reduced? To address these questions the project aims to assess the clinical utility of a range of tissues and preparation methods with a view to optimise the ability of the technology to achieve diagnosis whilst keeping clinical preparation time and skills requisites to a minimum. It is hoped this research can provide valuable resource management and efficiency information and contribute to clinical pipeline development.

An additional aim of this project is to use a similar informatics processing pipeline to investigate secondary immunodeficiencies. To do this, a large amount of transcriptomic data will be analysed in depth. This data captures the whole blood transcriptome of patients across a range of ages, who have presented to the clinic with respiratory tract infections. The research will compare the features of these datasets to identify key differences and similarities in the transcriptome between infections, and to see if these differences remain with advancing age and the onset of immunosenescence and associate immunodeficiency.

The association with age of all transcriptomic features will be quantified to establish which are the greatest, and this information will be used to develop the understanding of the biology of immunosenescence and potentially identify novel therapeutic targets.

Lastly, the project aims to develop disease specific ageing clocks. It is hoped that these will deliver further insights regarding the most important alterations in immune response which manifest with age. In addition, industrial applications of these tools include clinical trials management and admissions, personalised medicine, and assessing the efficacy of therapeutic interventions which target immunosenescence.

**1.3.9 Key research questions pertaining to T-cell Primary Immunodeficiencies**

1. Can RNAseq be used to enhance diagnostic capability by using gene expression profiles, dysregulated alternative splicing event frequency, allele and isoform expression to inform variant filtering and identification?
2. Can unstimulated whole blood act as reliable source of patient sample for identifying immunodeficiency expression profiles from the above mentioned 3 metrics?
3. Do PBMCs provide a more robust sample from which to identify changes in these metrics? If so, to what degree, and is this clinically important in terms of diagnostic pipeline development?
4. Does immune challenge provide a more robust sample from which to identify changes in these metrics? If so, to what degree and is this clinically important in terms of diagnostic pipeline development?
5. Can these combined methods identify disease causing mutations and provide specific diagnosis for currently undiagnosed patients?

**1.3.10 Key research objectives pertaining to Immunosenescence as a Secondary Immunodeficiency**

6. Using the existing transcriptomic dataset, set up an informatics pipeline to generate transcriptomics metrics to quantify expression, splicing, and validate the tool/s on other datasets.
7. Compare the transcriptomic profiles in whole blood of hospitalised patients with COVID-19 and Influenza infections at a gene level and isoform level using the tools.

8. Using the metrics above, combined with statistical machine learning methods, identify changes in transcriptomic signatures of patients with these infections which are associated with increasing age. Use this information to draw conclusions about the nature of disease pathology with age.
9. Find Transcriptomic biomarkers of immunosenescence from current literature and investigate their prevalence in the cohort's whole blood samples to conclude their presence in ageing cohort.
10. Use lasso regression to identify a core set of biomarkers without collinearity which produce a COVID-19 ageing clock and an INFLUENZA ageing clock.

## Chapter 2    Methods

### 2.1    Primary immunodeficiency

#### 2.1.1    Patient enrolment and patient data collection

Ethical approval for the project was granted by the University of Southampton Research Ethics Committee (UREC). Recruitment to the study was opportunistic at the University Hospital Southampton; informed consent for recruitment was obtained when known or suspected PID patients were having routine bloods taken on a given day, dependant on presence of participating clinicians. The cohort is a mixture of patients and family members of patients from the ‘Deep immune phenotyping’ study (study number SRB0014; REC reference 12/NW/0794; HTA license no 12009) ongoing at Southampton General Hospital. The aim of the study is to characterise the molecular fingerprints of primary immunodeficiencies and, where possible, identify precision medicine targets in patients within the immunodeficiency cohort. The patients that were assessed via RNAseq presented with a clinical phenotype of primary immunodeficiency and most did not yet have a clinical molecular diagnosis, despite having either WES or WGS performed. Some of the patients had been recruited and had genomes sequenced via the 100,000 genomes project. The primary researcher of this study had no knowledge regarding which participants had a confirmed genetic diagnosis, nor how many this pertains to.

### 2.1.2 PID patient clinical phenotype

At presentation to the clinic, participating clinicians assessed the patients' phenotypes the details of which are recorded below in Table 2-1. These details were not made available to the investigator until after analysis was completed.

**Table 2-1 Clinical phenotype of primary immunodeficiency patients**

Participant ID	Phenotype	Molecular Diagnosis	IUIS Category
SRB001	Panhypogammaglobulinaemia Recurrent Bacterial Infections Recurrent Viral infections Recurrent Fungal Infections Crohn's Disease	No Diagnosis in GECIP	Predominantly Antibody Deficiency
SRB002	Asthma Allergy Recurrent Pneumonia Recurrent Viral skin infections Recurrent Bacterial skin Infections	No information	Undefined Immune deficiency
SRB003	Panhypogammaglobulinaemia Recurrent Bacterial Infections Enterocolitis	No Diagnosis in GECIP	Predominantly Antibody Deficiency
SRB004	Panhypogammaglobulinaemia Recurrent Bacterial Infections ITP	No information	Predominantly Antibody Deficiency
SRB005	Asthma Allergy Recurrent Pneumonia Recurrent Viral skin infections Recurrent Bacterial skin Infections Specific Polysaccharide Antibody Deficiency (SPAD) Impaired T cell Function	CARD 11 A-C @2987250	Immunodeficiency affecting Cellular and Humoral Immunity
SRB006	Chronic Mucocutaneous Candidiasis	STAT1	Defects in Intrinsic and Innate Immunity
SRB007	Chronic Mucocutaneous Candidiasis	STAT1	Defects in Intrinsic and Innate Immunity
SRB008	Chronic Mucocutaneous Candidiasis	STAT1	Defects in Intrinsic and Innate Immunity



SRB009	Panhypogammglobulinaemia Recurrent Bacterial Infections Autoimmune Haemolytic anaemia Autoimmune Neutropenia Autoimmune Thrombocytopaenia	NKKB1 A>AT @102582929	Predominantly Antibody Deficiency
SRB010	Panhypogammglobulinaemia Recurrent Bacterial Infections Autoimmune Haemolytic anaemia Autoimmune Neutropenia Autoimmune Thrombocytopaenia	NKKB1 A>AT @102582930	Predominantly Antibody Deficiency
SRB011	Panhypogammaglobulinaemia Alopecia Severe Viral Infection	ILRG;CXorf65;FOXO4	Predominantly Antibody Deficiency
SRB012	Recurrent Fungal Infection Recurrent Viral infection Decrease in T cell count Autoimmunity	ILRG;CXorf65;FOXO4	Immunodeficiency affecting Cellular and Humoral Immunity
SRB013	Panhypogammaglobulinaemia Recurrent Bacterial Infections Splenomegaly Lymphoid Interstitial Pneumonia	No Diagnosis in GECIP	Predominantly Antibody Deficiency
SRB014	Panhypogammaglobulinaemia Recurrent Bacterial Infections Splenomegaly Lymphoid Interstitial Pneumonia Autoimmune haemolytic anaemia, Enteropathy	No information	Predominantly Antibody Deficiency
SRB015	Hydroa Vaccineforme	No Diagnosis in GECIP	Defects in Intrinsic and Innate Immunity
SRB016	Panhypogammaglobulinaemia Recurrent Bacterial Infection Impaired T Cell proliferation Eczema Enteropathy	No information	Combined Immunodeficiency
SRB017	Panhypogammaglobulinaemia Recurrent Bacterial Infection Absent B cells	No information	Predominantly Antibody Deficiency
SRB018	Panhypogammaglobulinaemia Recurrent Bacterial Infection Bronchiectasis Type 2 DM	No information	Predominantly Antibody Deficiency
SRB019	Recurrent Abscess with Pseudomonas Candidal Discitis Nephrectomy with Klebsiella abscess SLE	No information	Disorder of Phagocytes

SRB020	Panhypogammaglobulinaemia Recurrent Bacterial Infection Neutropaenia	No information	Predominantly Antibody Deficiency
SRB021	Panhypogammaglobulinaemia Recurrent Bacterial Infection Bronchiectasis	No information	Predominantly Antibody Deficiency

### 2.1.3 Whole blood and PBMC collection

Venepuncture methods and blood collection methods were performed as per manufacturer's guidelines (337). PAXgene™ RNA tubes were utilised for RNA isolation and, and BD Vacutainer® Heparin tubes were used for PBMC collection. Tubes were equilibrated to room temperature, and 'mid-flow' peripheral blood samples were taken from participants. The RNA tube vacuum is designed specifically to draw 2.5ml blood into the RNA tube, the Vacutainer draws 3ml. The PAXgene tubes were inverted 8 to 10 times and left to incubate at room temperature for a maximum of 4 hours before transferring to a -80°C freezer. All PBMCs were extracted and frozen within 4 hours from venepuncture.

### 2.1.4 Whole blood RNA extraction methods

Whole blood was sourced from two healthy donors and 21 primary immunodeficiency patients into PAXgene tubes as described above. PAXgene tubes were removed from -80°C storage and allowed to equilibrate to room temperature for 2 hours before RNA isolation RNA was extracted using PreAnalytiX Blood RNA kit by following manufacturer's instructions, described below.

1. Samples contained in PAXGene Blood RNA tubes were the centrifuged in the tubes at max speed (3229 x g) for 10 minutes.
2. Supernatant was decanted off, and rim dried with a clean paper towel. 4ml of RNAase free water was added, and tubes were closed with new BD Hemogaurd.
3. Tubes were then vortexed to dissolve pellet and centrifuged again at 3229 x g for 10 minutes. Supernatant was discarded.
4. 350µl of resuspension buffer was added. Tubes were then re-vortexed until all sample was dissolved.

5. Sample was transferred to a 1.5ml microcentrifuge tube (MCT), and 300µl of binding buffer 2 was added, before 40 µl of proteinase K. This was mixed by vortexing for 5 seconds and incubated at 55°C in a shaker at 250rpm for 10 minutes.
6. Lysate was then pipetted directly onto PAXGene 'Shredder' spin column in 2ml processing tube and centrifuged at 1700 x g, for 3 minutes.
7. Flow through was then transferred to a fresh 1.5ml MCT, without disturbing the pellet.
8. 350µl ethanol was added to the 1.5ml and sample was then vortexed and briefly centrifuged with microfuge for 1-2 seconds to concentrate all fluid in bottom of the tube.
9. 700µl was pipetted in RNA spin column, and centrifuged for 1min at 17000 x g. Place column in new 2ml collection tube and discard the old tube containing flow through,
10. This step was repeated until all remaining samples was concentrated into RNA spin column. The RNA spin column is transferred to a new collection tube.
11. A wash step was performed by adding 350µl wash buffer to the column and performing centrifugation at 17000 x g for 1 minute. Flow through was completely discarded.
12. 10ul DNase and 70 DNA digestion buffer was mixed in an MCT, mixed by gently flicking and centrifuged to collect residual liquid.
13. Sample was then transferred to a new processing tube, and pre-mixed 10µl I:70µl per sample DNase/DNA buffer mix was pipetted directly onto the membrane of the RNA spin column and this was allowed to sit for 15 minutes on the desktop to facilitate enzymatic digestion.
14. This solution was washed off using 350µl wash buffer 1, pipetted into the RNA spin column, which was then centrifuged at 17000 x g for 1 minute. The spin column was replaced with a new and the flow-through was discarded.
15. 500µl of wash buffer 2 was then added to the PAXgene RNA spin column, which was then centrifuged at 17000 x g for 1 minute. The spin column was replaced with a new and the flow-through was discarded.
16. Step 16 was then repeated with a longer, 3-minute centrifugation time.
17. A Drying step was then performed in which the spin column is transferred to a new collection tube, centrifuged for 1 minute at 17000 x g without further substance added.
18. The processing tube was discarded along with the flow-through. The spin column was transferred to a new 1.5ml MCT. To elute the RNA, 40ul of elution solution was pipetted directly onto the RNA column membrane and centrifuged for 1 minute at 17000 x g,
19. This step was then repeated with the same collection tube.

20. RNA samples were incubated for 5 mins at 65°C to denature the RNA, optimising the molecular conformation for downstream sequencing application.
21. 4ul was removed for QC, and the remaining samples were immediately frozen at -80°C.

J. Lye performed all extractions.

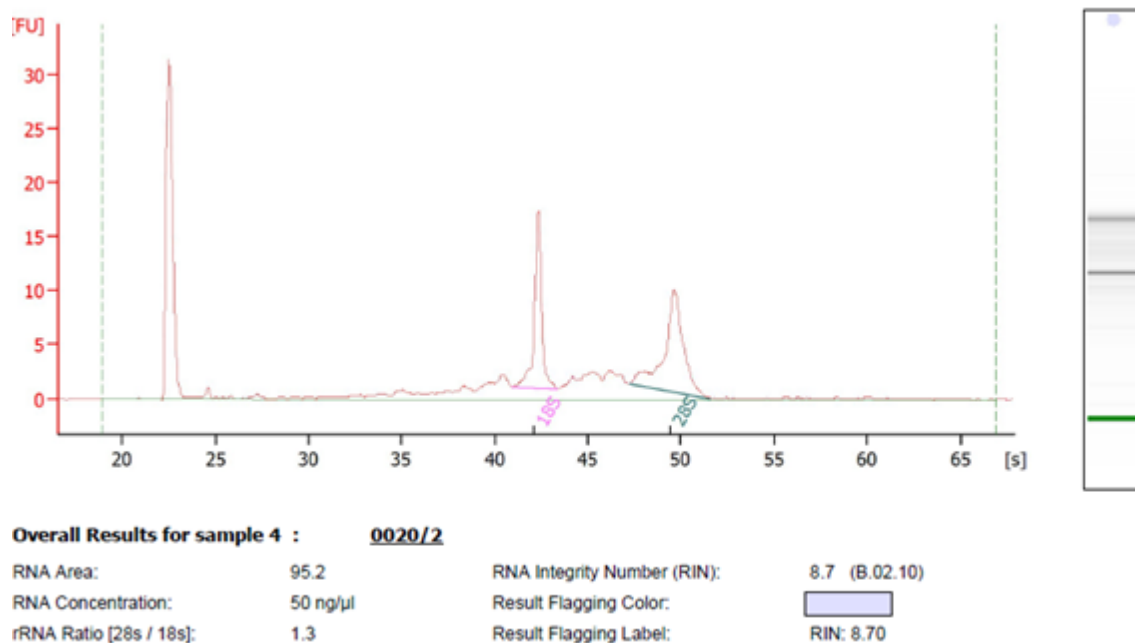
#### **2.1.5 RNA extraction QC**

Quality control of RNA extractions was performed in two ways. In the first instance, using a Nanodrop ND-1000 to check RNA concentration and yields. The Nanodrop uses spectrophotometry to assess purity and concentration of nucleic acids. Briefly, the ratio of the amount of light absorbed at specific wavelengths (260/280 and 230/280) is recorded. This ratio is then compared to the optimal ratios to discern quality.

RNA was subsequently assessed for quality and degradation using the Bioanalyser 2100 (Agilent, Santa Clara, CA. USA). This is a service offered by within the Human Development and Health laboratories at Southampton General Hospital and was performed by Dr Melissa Doherty.

Using micro-capillary electrophoresis, the fragment sizes of RNA are observed, recorded and interpreted to generate an RNA integrity number (RIN score). This score gives an indication of level of RNA degradation.

Key aspects of the trace from the Bioanalyser include the relative abundance of two peaks (18S, 28S) and the noise between the peaks, shown on the trace below in Figure 2-1 Agilent 2100 electropherogram RNA trace.



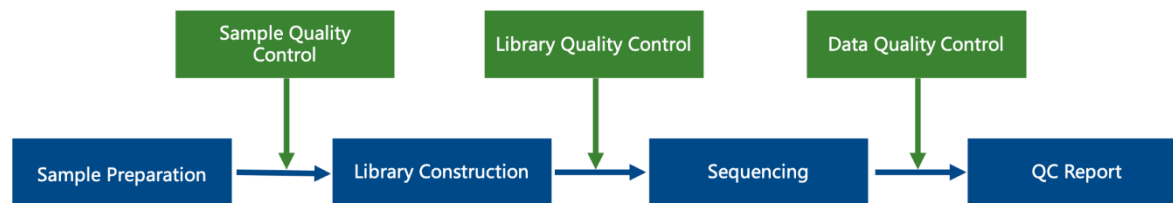
**Figure 2-1 Agilent 2100 electropherogram RNA trace**

Observable in the electropherogram is a small amount of noise, displayed as an increase in the jagged lined between the two peaks and to the left of the first 18s peak, indicative of partial RNA digestion. Also demonstrative of this is decrease in overall size of the peaks and the 28S subunit of the ribosomal RNA being smaller, as this often degrades first. In the event that the first peak at around 23seconds is accompanied by other peaks either side, and the 18S and 28S peaks have shifted to the left, the result would be indicative of heavy degradation and probably not be useable.

We considered the following to be threshold for RNA extraction RIN <7, 280/260 <1.8 or >2.3, RNA conc. Or RNA concentration <20 ng/μl across both platforms. No threshold for 230/260 ratios was applied, as literature did not suggest a lower limit.

### 2.1.6 RNA Sequencing by Novogene

Frozen RNA was sent to Novogene™ Ltd (Hong Kong) on dry ice for RNA sequencing. Reports from Novogene indicate that sample processing was completed as per Figure 2-2.



**Figure 2-2 - Novogene sequencing and QC workflow - received in personal report.**

Quality control steps are represented in green, with procedural steps in blue.

### 2.1.7 RNA quality control: Novogene

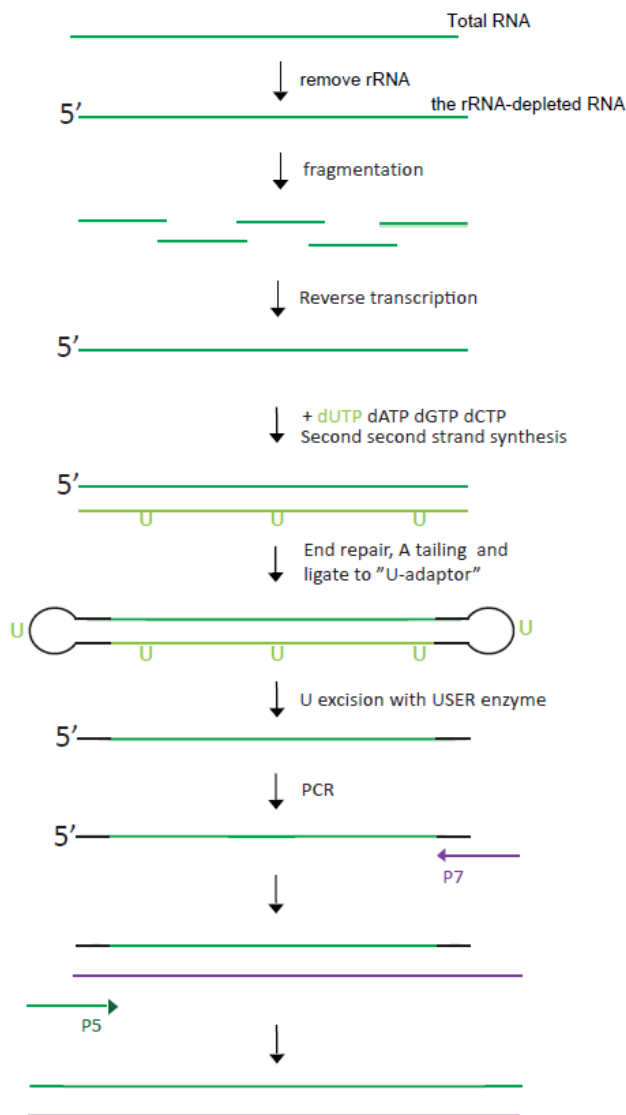
Samples were checked for quality control by Novogene by an array of methods at different stages of the sequencing pipeline (Figure 2-2). Sample quality control was performed by agarose gel electrophoresis and providing samples passed further QC was performed as follows. Sample quantitation and purity assessment was performed via Nanodrop. Sample integrity was checked with Agilent Bioanalyser 2100. Providing samples passed the quality control, the sample library preparation and sequencing was then carried out.

### 2.1.8 Library construction and sequencing

Initially, concerns about the ability to recognise differentially expressed genes because of read depth, combined with financial constraints meant that library prep strategies were employed to maximise coverage over the known, protein coding PID genes mRNA. This included 8 PBMC samples and three whole blood RNA samples (one in duplicate). These samples underwent library preparation using Novogene's mRNA library prep, which captures only the RNA species with a poly-A tail and is performed using oligo(dT) beads. These samples were also treated with globin-zero, which depletes globin transcripts, reducing the reads mapped to these very highly expressed genes. These samples were patient SRB0003, and controls SRBC0001 and SRBC0002. These were sequenced at minimum of 40M reads per sample using strand specific paired end sequencing, with 150bp read

lengths. All PBMC RNA, from each batch were also prepped using the poly-A mRNA isolation method.

However, after discussion, the importance of non-coding RNA was highlighted as an area of emerging interest, given how lack of diagnosis using existing clinical approaches was a factor. Moving forward the remaining whole blood samples were treated with ribo-zero kit leaving total RNA, and then underwent Novogene lncRNA library preparation (Figure 2-3) and were sequenced using strand specific paired end sequencing, with 150bp read lengths as before, but to the greater depth of 50M reads, to compensate for the lack of targeted sequencing and resulting reduction in mapped read to protein coding genes.



**Figure 2-3 Novogene IncRNA Library Preparation Workflow - Image provided by Novogene in email report. (222)**

Once the RNA has been selected in one of the above-mentioned methods, it is fragmented using a fragmentation buffer, and cDNA is synthesised using RNA template and random hexamer primers. A custom, second strand synthesis buffer is added, which is provided by New England Biolabs Ltd. (Ipswich, Mass, USA). To initiate second strand synthesis, dNTPs, RNase H, and DNA polymerase 1 are added to this mix. After terminal repair, a ligation and sequencing adaptor ligation step is performed. Finally, the dsDNA library undergoes size selection and PCR enrichment. Quality control is then carried out at this stage by Novogene using Qubit 2.0 for preliminary concentration, the Agilent Bioanalyser 2100 which tests insert size, and qPCR which quantifies the library concentration.



The quality-controlled libraries are then pooled and fed into Illumina sequencers to effective concentration and data volume expectations.

### **2.1.9 Whole blood control data – The Genome Tissue Expression Consortium**

Acknowledgement Statement: The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this thesis were obtained from dbGaP accession number phs000424.v7.p2 on 28/06/2019.

The Genotype-Tissue Expression (GTEx) project is an established resource database harbouring hundreds of volunteers' samples from a range of tissues, including whole blood (338). Originally established to study the effects of genetic variation on gene expression, the transcriptomic data also serves as a valuable resource for other studies looking at expression. RNA is isolated and sequenced from deceased donors' tissue with a low-post-mortem interval time. Whole blood transcriptomic data from this database served as the control data for novel splicing event discovery.

To ensure all data was treated in the same manner, the original raw RNAseq data files were downloaded from the control data source (GTEx) as described below.

Once access was formally granted, appropriate licences were configured for use by the SRAconfig tool, part of the SRA toolkit available from NCBI. To avoid downloading unhelpful data, records of participants' samples were collated into Excel 365 (Microsoft, Redmond, USA) file, high priority samples were identified by being filtered by age (18-49), a Hardy Death Score which indicates the duration, if any, of morbidity preceding mortality (339) and tissue source. The scores deemed to be applicable for analysis were 1 (violent and fast deaths due to accident) and 2 (fast death of natural causes).

Subsequently, parameters were selected to filter out those samples which:

- Came from people who had complex and chronic diseases, which were also more likely to have disruption in immune system transcriptomes.
- Came from other tissues than whole blood.

113 GTEx samples remained after filtering. The results were then amalgamated in a single file. This was submitted as a request detailing “unique SRR identifiers” for instructing SRA file downloads. The very large quantities of data needed to be downloaded from NCBI and collating the requisite SRA files in the “cart function” expedited this process. The access permission only allows downloaded files to be worked on at the pre-designated file location, set when configuring the download option in SRAConfig. Files were converted to a fastq.gz, achieved by “fastq dump”, an executable file also found in the SRA toolkit. This converter component also split paired end reads into separate ‘fastq’ format files.

The SRA toolkit commands were executed using Windows Powershell – an intrinsic, command line-driven shell, available in Microsoft Windows Desktop.

This download/filter step was carried out by Dr J Lord.

#### **2.1.10 Whole blood control data - Splicing and Disease Cohort**

The Splicing and Disease cohort are a group of patients which are part of an ongoing study investigating suspected Mendelian disorders and variants of unknown significance. The VUSs were identified through the Splicing and Disease research study at the University of Southampton. Informed consent was obtained for all patients enrolled onto the splicing studies. Ethical approval was granted by Health Research Authority (IRAS Project ID 49685, REC 11/SC/0269) and by the University of Southampton (ERGO ID 23056).

At the time of writing the cohort numbered nearly 196 patients. Filtering needed to be applied to ensure controls samples and PID patients were of the same age. All samples selected were over the age of 18 and under the age of 65 and had already had their peripheral whole blood sequenced for efficiency. The samples were also filtered to remove those which had a known VUS occurring in the 2020 list of known PID genes (See Appendix A.1). This included SOT063 which had a *TP53* VUS, and SOT158 which had two *TERT* variants. 23 samples remained. All RNAseq data were processed using the same tools with identical syntax to the PID cohort described in section 2.1.11

Table 2-2 Splicing and Disease Cohort

ID	AGE	Variant	Sequencing Batch	Processed by
<b>SOT010</b>	22	NF1 c.1158A>C, p.(=), c.1168_1179del12, p.Asn390_His393del	3	Jenny Lord
<b>SOT017</b>	62	NF1 c.7832A>G, p.Asp2611Gly	3	Jenny Lord
<b>SOT018</b>	20	NF1 c.5489C>G, p.Pro1830Arg	3	Jenny Lord
<b>SOT020</b>	18	NF1 c.4122G>T, p.Gln1374His	3	Jenny Lord
<b>SOT027</b>	46	BRCA2 c.10249 T>C p. (Tyr3417 His)	3	Jenny Lord
<b>SOT029</b>	54	BRCA2 c.6935 A>T p.(Asp 2312 Val)	3	Jenny Lord
<b>SOT033</b>	62	SMAD3 c.802C>T, p. (Arg268Cys)	4	
<b>SOT037</b>	45	BRCA1 c.4987-11T>C	3	Jenny Lord
<b>SOT040</b>	50	BRCA2 c1127T>G heterozygote	3	Jenny Lord
<b>SOT043</b>	48	BRCA1 c. 1731 A>G p. (=)	3	Jenny Lord
<b>SOT045</b>	52	BRCA1 c.4676-8C>G	3	Jenny Lord
<b>SOT049</b>	64	BRCA2 c. 9502-13 C>G	3	Jenny Lord
<b>SOT058</b>	19	MED13L: c.2570-4_2574del	4	Jenny Lord
<b>SOT070</b>	40	DKC1 c.915+10 G>A	4	Jenny Lord
<b>SOT082</b>	22	FOXP1 c.583C>T	5	Jed Lye
<b>SOT104</b>	62	TMEM127 c.411T>A, AIP c.317G>A_Arg106His and WT1 c.871A>T p.(Ser291Cys)	5	Jed Lye
<b>SOT117</b>	51	Clinical diagnosis MEN1 but not confirmed molecularly	5	Jed Lye
<b>SOT123</b>	40	Negative for NF2, SMARCB1, LZTR1	5	Jed Lye
<b>SOT130</b>	18	No diagnosis - negative 100KGP	5	Jed Lye
<b>SOT140</b>	21	No diagnosis - negative 100KGP	5	Jed Lye
<b>SOT152</b>	46	Polyposis but NAD from 100KG.	5	Jed Lye
<b>SOT175</b>	23	Suspected connective tissue disease. No specific variant.	5	Jed Lye
<b>SOT189</b>	30	No diagnosis - negative 100KGP	5	Jed Lye

Error! Reference source not found. provides details for the control samples used including ID, age, variant (if known) processing batch number and operator who processed the data.

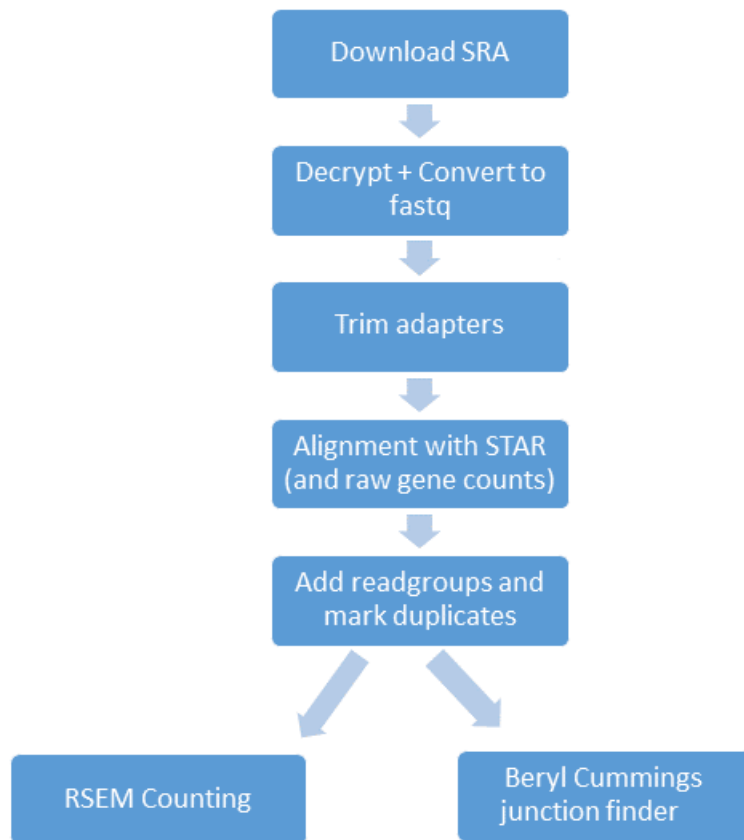
### 2.1.11 Data processing

**Table 2-3 - Data processing packages**

Software Name	Task	Reference
<b>FastQC v.0.11.3</b>	Pre and post trimming quality control	(340)
<b>MultiQC v1.5</b>	Comparison of quality control metrics to identify outliers	(341)
<b>Trimmomatic v0.3.6</b>	Trimming adapters from reads	(342)
<b>STAR aligner software v2.6.1</b>	Aligning to genome/transcriptome and performing counts	(343)
<b>Samtools v1.3.2</b>	Sorting and indexing of Bam files	(344)
<b>Picard v2.8.3</b>	Read groups added and duplicates marked	(345)
<b>RSEM v1.3.1</b>	Generate gene/transcript counts, isoform percentage metrics	(346)
<b>PCAExplorer</b>	Exploratory data analysis	(347)
<b>ComBat-Seq</b>	ComBat-seq	(348)
<b>OUTRIDER</b>	Gene expression outlier analysis	(132)
<b>Mendelian RNASeq</b>	RNA splicing analysis	(155)

All data was stored on a secured shared drive for research or on a password protected external hard drive. Most of the files were uploaded to a scratch folder on the University of Southampton's high-performance computing cluster, IRIDIS 4. At each stage of file transfer, MD5 sums were checked in Linux using 'md5sum' command to ensure complete and successful file transfer.

The data processing pipeline for the GTEx raw data, which had previously been obtained and processed by Dr J. Lord served as the initial model for processing the PID cohort data. The present study replicates these processes from the point at which fastq files are obtained.



**Figure 2-4 Workflow of GTEx data processing.**

GTEx data was downloaded from the sequence read archive and processed as described above with quality control steps performed with Fastqc, before and after adapter trimming (see 2.1.12). Fragments per kilobase per million, (FPKM) and transcripts per million (TPM) reads counts were obtained using the RSEM package, and STAR was set up to give absolute read counts. Details of all of these steps are covered in section 2.1.14. These steps were performed by Dr. Jenny Lord

### 2.1.12 Trimming of reads and Quality control.

The fastq files were loaded into Trimmomatic v0.3.6 which was obtained via biobuilds/2017.11 for adapter trimming. Adapter sequences were obtained from the Novogene reports and loaded into a separate .txt file on the IRIDIS 4 workspace. These were then used as templates to search and trim in the fastq files in an attempt to remove the any untrimmed adapters. Trimmomatic shell script can be found in the complete processing syntax in Appendix A.2. Sequence data exists in fastq or 'fq' file format. This consists of read identifiers, nucleotide base calls, and quality scores delivered in a per-

base metric. Quality “Q” scores are denoted by Phred scores, which represents the likelihood that the base has been incorrectly called. This is achieved with the formula.

$$Q = 10 \times \log_{10} p$$

Once this has been calculated, reads with many low-quality base calls can be excluded or the true identity of a base can be determined by comparing many reads spanning the same regions. In-house QC was performed using FastQC v.0.11.3 on the unprocessed fastq files stored on the IRIDIS 4 high performance computer cluster. Explanations of parameters and example of the analysis of FastQC can be found in Appendix A.3. The outputs from FASTQC were then compiled into a single report using MultiQC and compared.

#### **2.1.13 Alignment of reads.**

Unpaired reads were discarded, fastq files containing 150bp paired-end reads were loaded into FastQC v.0.11.3 once again for QC. Reads were then aligned using STAR aligner software (343) V2.6.1, in line with the processing performed by Dr. J Lord on the GTEx data. Reads were aligned to GRCh38 reference genome and annotated using Gencode GRCH38.p12 Genome v30 annotation file in GTF format. STAR alignment was set to ‘two-pass’ mode for allowing for unmapped reads to be mapped to the junctions found on the first ‘pass’. Overall, the parameters for the alignment were selected to match those of the pre-processed whole blood RNAseq data obtained from the GTEx, performed by Dr. J Lord, which also followed the recommendations for the ENCODE project (349) found in the manual for the STAR aligner software (350). The specific parameters can be found in the syntax provided in the appendix A.2. The data from output log.txt files of these aligned libraries were compared using Microsoft Excel for differences in number of aligned reads and splice junctions detected.

#### **2.1.14 Read Counts**

Raw read counts for each gene were obtained from STAR during alignment with the quant mode setting (see appendix A.2). Aligned bam files were then sorted and indexed using Samtools v1.3.2 (344), read groups were added and duplicates marked using Picard v2.8.3 and finally RSEM v1.3.1

was used to generate gene counts, transcript counts, and isoform percentage metrics for analysis. RSEM is an accurate and fast read counter, which can act with or without a reference genome and was selected over other tools as it provides more options for types metrics for analysis: Gene and Transcript level FPKM, TPM and isoform percentage (60). The specifics of these parameters can also be found in the syntax provided in the appendix A.2.

#### **2.1.15 Gene panel**

Gene panels are necessary to reduce data to a manageable size and prioritise research time based on known gene associations. Two PID specific gene panels were obtained. The first was comprised of a panel of genes, curated by the IUIS, known to have variants which have been a demonstrable cause of primary immune deficiency (appendix A.1). This gene panel is regularly updated and curated and was acquired from the IUIS (78). This panel was obtained to aid in the variant filtering process, and also to filter the RNA events which are linked to the PID symptoms specifically during

The number of known causal genes which produce PID in patients is increasing annually. Methods based purely on filtering events and variants using an existing gene panel are likely to miss new not-yet-linked genes, which reduces the ability of the investigation to generate diagnosis in the cohort. To address this and expand the gene panel, two additional gene lists were combined with the IUIS panel. These were firstly the Genomics England list of PID associated genes from the PanelApp (351), found in the Genomics England portal (Appendix A.4) and secondly, the HTG EdgeSeq library prep platform's list of T-Cell specific genes from the Immuno-oncology panel. Eventually a final gene panel, submitted by Professor Anthony Williams was included for splicing analysis, known henceforth as panel 'AW'.

The HTG EdgeSeq Immuno-oncology assay panel was originally developed by HTG™ and designed to include 549 human RNA transcripts, known or believed to be involved in the innate and adaptive response to cancer. It covers the spectrum of activity from infiltrates to activation to checkpoints. In the current research this has been repurposed by isolating a single group of protein coding genes on this list which, comprise those genes involved with T-cell activity specifically (Appendix A.5).

These panels were collated into a list in Microsoft Excel. All entries for “unknown origins and large multi gene deletions” were removed and 294 duplicates were removed. The final list of 501 genes can be found in appendix A.6 In the event that a diagnosis was not reached with the original IUIS PID panel, filtering could be expanded to include genes present in this larger, curated list.

### 2.1.16 Exploratory Analysis of RNAseq data

Initial exploratory data analysis was conducted using the `pcaExplorer` (347) package in the R environment. Raw read counts and meta data were loaded into `pcaExplorer`. Metadata for the samples consisted of origin (GTEx/UHS), disease state (Control/PID), and batch number (1, 2, 3). Information pertaining to the donor's age and sex were not included in initial exploration as this data for the patients had not been provided. Count data was  $\log_2$  transformed, and the `pcaExplorer` program amalgamated the log transformed data and meta-data into a DESeq data set – the parameters of which are highlighted below.

General information, count data and count data statistics were calculated using the `pcaExplorer` functions. The data was colour-coded on all subsequent images by batch number. 135 samples were included in total, which comprised 20 PID patient samples, 2 control samples and 113 selected and curated GTEx samples. 58825 transcripts were included which included non-coding genes, microRNA genes, and pseudo-genes in addition to the protein coding genes.

Filters were then applied to data as follows; the threshold on the row sums of the counts was set to a value of 1, and a threshold row mean value was set as 1. This was to prevent completely unexpressed genes from being included, and genes which expression was so low that statistical interpretation would have been inaccurate.

A two-dimensional principal component plot was then produced using the `pcaExplorer` function, and 95% confidence interval ellipses were included to demonstrate clustering based on principal components.

A scree plot for the top 8 principal components was produced to demonstrate the contribution of each principal component to the overall variance between the samples, and another for the cumulative variance explained by the principal components.

Next the data was explored using the principal component by gene function. A 2D plot showing projections of gene abundances onto pairs of components, with samples as biplot variables. This allows the identification of genes or groups of genes which have particular impact on the top and bottom loading of the principal components investigated. Some of those genes contributing in the greatest way to the variability were investigated. Violin or boxplots were produced to visualise the differences between batches and help explain the variability between groups.



## 2.1.17 Differential gene expression

### 2.1.17.1 Gene expression Z-score calculation

All patient and GTEx FPKM values were calculated from reads obtained from RSEM (as described above). These were compiled into Microsoft excel spreadsheets and then filtered into data for IUIS tables of known causal PID genes. The mean, standard deviation and Z-scores were calculated for each patient. Z-scores were calculated using the Excel function 'STANDARDIZE(X, mean, standard\_dev)'.

$$Z = \frac{X - \mu}{\sigma}$$

Where Z = standard score, X = observed score,  $\mu$  = mean of sample  $\sigma$  = standard deviation of sample.

Subsequent to this Z-score calculation, the raw reads were corrected using ComBat-seq. The batch corrected data was re-visualised using PCaExplorer. TPM values were used to replace the FPKM values, as literature suggested these may provide more robust results. Z-scores were calculated using data for only PID patients and two control samples (SRBC0001 and SRBC0002) with GTEx data excluded.

To visualise results, colour coded conditional formatting of red – green for low to high, was applied to the table rows (gene wise) in the case of the TPM values, and sample wise in the case of the Z-scores. This was to show the spread of the data for each gene and identify gene expression outlier candidate genes per sample respectively.

The resulting tables of Z-scores were collated into a single table, and a conditional formatting threshold of  $Z > 3$  was applied, to highlight overexpression outliers for each sample.

These were tallied using the Excel count function and plotted onto a bar chart to show the frequency of outliers in each sample, in each scenario: With GTEx or without GTEx as controls.

### 2.1.17.2 OUTRIDER

OUTRIDER provides an outlier gene expression detection platform (352). The OUTRIDER environment for the R package was installed on a locally hosted version of R/4.0.0 in the author's local ./scratch drive on the IRIDIS HPCC.

To validate the proper functioning of OUTRIDER, the developers test dataset with positive controls was used to run the program. The analysis and results from this dataset are demonstrated in the user walkthrough. The dataset itself is contained within the program as a system file. The OUTRIDER program was run using the standard syntax which can be found in Appendix A.6 . The output was then cross-referenced with the output found in the manual as positive controls to validate the functionality.

Raw RNAseq count data was combined from the Splicing and Disease Cohort and the Primary Immunodeficiency Cohort in the form of a text file. This was uploaded to the remote scratch server on Iridis4 HPC at the University of Southampton. Using the programming language R (version 4.0.0-cairo) The OUTRIDER software was used to process the data to identify aberrant expression outliers. First an OUTRIDER data set (ODS) was created using the count data and meta data which included cohort information (Splicing and Disease or PID) and the batch number, of which there were two for the PID cohort and 3 for the Splicing and Disease cohort. Using the Gencode v30 annotation dataset, FPKM values were calculated, and non-expressed genes were marked in OUTRIDER. The distribution of these counts was plotted, and statistics for expressed genes were visualised in a graph using the option for this in the OUTRIDER program. Heatmaps were produced to show the level of batch effect within the dataset, with the metadata groups marked by colours. Using OUTRIDERS noise correcting autoencoder, confounders were then controlled for using the default values of  $q=20$  and iterations  $=15$ . After this process the heatmap was re-plotted to demonstrate successful batch correction. The negative binomial model is then fitted to the data to identify outliers, p-values are calculated, and for reference the program was also instructed to compute Z-scores. The package gives the option to rank samples by the number of outliers or aberrantly expressed genes. Outlier samples were identified from the dataset and removed. Using the options in OUTRIDER, results were printed to a csv file, and Volcano plots were produced for each PID sample. In addition, graphs for expected vs observed counts, QQ plots, and ranked expression were produced for genes of interest, based on results from the analysis or clinical diagnostic information given to the researcher. Results were cross referenced using the IUIS panel of genes known to cause PID, the GECIP panel of genes known to cause PID, HTG EdgeSeq Immuno-oncology panel (T-cell specific and in its entirety) and also a curated list of genes with statistically significant changes in naïve CD4+ T-cell stimulation, acquired from the DICE database and filtered for significance by the researcher. For reference and to give context to fold change values extracted from the DICE database, the values of the full curated list of these CD4+ T-cell genes were also plotted into a graph using Microsoft Excel.



### 2.1.18 Aberrant splicing - Mendelian RNAseq program

Cummings et al. developed a novel method for the detection and comparison of alternative splicing events using RNAseq data. A description of the code, use case, syntax and file structure were all found on the MacArthur laboratory blog (353) and GitHub page (354), although these have since been removed. The syntax for the tool requires an input of RNAseq reads in the format of sorted and indexed bam files, with read groups added and duplicates marked. Appropriately processed bam files were uploaded to the IRIDIS 4 high performance computing cluster at the University of Southampton.

These steps were performed by Dr Jenny Lord on the GTEx data, who kindly forwarded the outputs and the necessary syntax to join these with the outputs of the discovery steps from the samples in the current study. The Cummings' syntax also requires a gene list for input which contains the gene symbols, gene IDs, strand information, location coordinates, chromosome number and gene type. The exact syntax used by Cummings et al. for developing these lists was not made publicly available, however results from the control dataset were replicated exactly using an equivalent Python script developed by Dr J. Lord. This script stripped the relevant columns from the Gencodev30 annotation .gtf file and merged them into the specified format required by the syntax.

1	OR4F5	ENSG00000186092.6	+	chr1	65419	71585	protein_coding
2	OR4F29	ENSG00000284733.1	-	chr1	450703	451697	protein_coding
3	OR4F16	ENSG00000284662.1	-	chr1	685679	686673	protein_coding
4	SAMD11	ENSG00000187634.12	+	chr1	923928	944581	protein_coding
5	NOC2L	ENSG00000188976.11	-	chr1	944203	959309	protein_coding
6	KLHL17	ENSG00000187961.14	+	chr1	960584	965719	protein_coding
7	PLEKHN1	ENSG00000187583.10	+	chr1	966497	975865	protein_coding
8	PERM1	ENSG00000187642.9	-	chr1	975204	982093	protein_coding
9	HES4	ENSG00000188290.10	-	chr1	998962	1000172	protein_coding
10	ISG15	ENSG00000187608.9	+	chr1	1001138	1014540	protein_coding
11	AGRN	ENSG00000188157.15	+	chr1	1020120	1056118	protein_coding
12	RNF223	ENSG00000237330.3	-	chr1	1070967	1074306	protein_coding
13	C1orf159	ENSG00000131591.17	-	chr1	1081818	1116361	protein_coding
14	TTL10	ENSG00000162571.13	+	chr1	1173884	1197935	protein_coding
15	TNFRSF18	ENSG00000186891.14	-	chr1	1203508	1206592	protein_coding
16	TNFRSF4	ENSG00000186827.11	-	chr1	1211340	1214153	protein_coding

**Figure 2-5 - Gene list file structure. Columns shown are GeneID, ensemble gene identifier, strand information, chromosome number, start position, end position, and gene type.**

### 2.1.18.1 Splice junction discovery, normalisation and junction filtering

#### Discovery

Reads obtained from the process of RNA sequencing are aligned to a reference genome using alignment tools. One such tool, STAR, is a 'splice-aware' aligner. This tool compares the continuous sequence of a read with that of the genome and matches them based on quality control metrics determined by the user. When this tool finds a sequence for which two sections align to two locations in the genome, the aligner identifies this as a splicing event, (with particular quality control parameters).

The Mendelian RNAseq splicing tool acquires and combines the splicing data contained in inputted bam files, such as genomic coordinates, gene ID, number of samples in which it was observed, total number of reads supporting the junction in all control and patient samples combined and per sample, and compiles this into a multi-sample splicing dataset for interrogation.

#### Normalisation

The normalisation step of the workflow is performed by executing a section of syntax comparing the read support (number of reads spanning the splicing event) in each sample with the read support for the highest shared wild-type annotated junction within the whole dataset. Figure 2-6 demonstrates this process. The wild-type splice events are exon junctions A-B, and C-D. An exon skipping event can be seen to take place between A-D. The relatively small number of supporting reads for A-E may be indicative or this event likely being due to mapping noise, although other explanations exist. The exon-exon junction event is normalised by the maximum read support of a shared exon-intron junction annotated in Gencode v30. In this example, 200 reads support the exon skipping event, and the exon-intron junction, which is annotated and shared, have 100 and 300 support. As the maximum is 300,  $200/300$  gives the normalised value (0.66) which supports the event. The normalised read value of A-E =  $3/300 = 0.01$ .

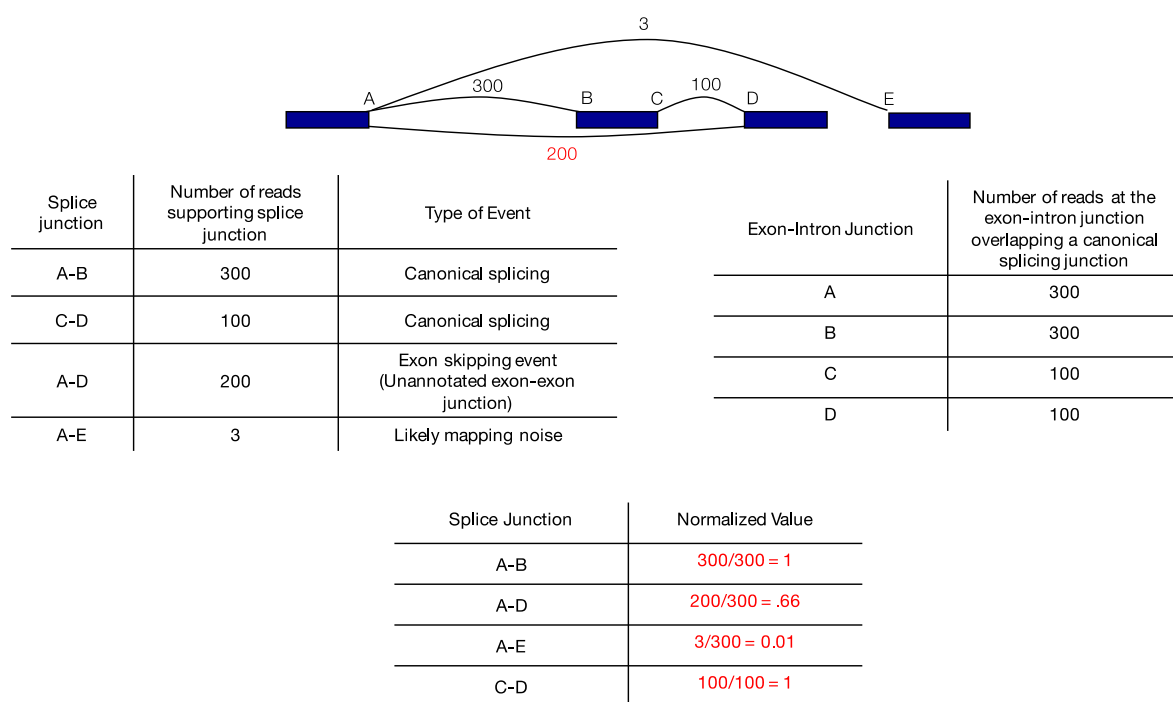
#### Filtering

Due to the extremely high number of events and the low likelihood of pathogenicity, multiple filtering steps of those events were conducted based on different strategies. This was performed by running the Mendelian RNAseq program filtering syntax (155). The stringency of the various strategies could be relatively low, as the events explored would be later filtered by the gene panel.

The aim of the filtering process was to reduce the events to a manageable number for follow-up analysis using IGV (15 events or less), to identify events which were unique or seen almost nowhere else, to identify those that were likely to be pathogenic.

Filtering was performed and repeated using varying sets of parameters within the program which made use of metrics including absolute read support, normalised read support, the number of samples in which the events occurred, events which only occurred in a specific sample, read support/ normalised read support being highest in a specific sample, and the ratio of normalised read support between the highest samples and the next highest sample.

Results from these filtering steps were downloaded as .txt files, compiled in Microsoft excel and cross-referenced with identified gen panels to identify splicing events in gene which may be causative of the phenotypes observed. Finally, these events were investigated through manual inspection of the aligned transcriptome reads using interactive genome viewer (IGV)(355).



**Figure 2-6 - Splice junction normalization visualised.**

Figure depicting the normalisation process in the case specific to heterozygous exon skipping. Figure appears in Cummings et al. 2018 (155)

## **2.2 Immunosenescence investigation**

### **2.2.1 Patient enrolment and patient data collection**

Ethical approval was granted by South Central Hampshire Research Ethics Committee. COVID-19 point of care study REC reference was 20/sc/0138, granted March 16<sup>th</sup>, 2020. Influenza point of care study REC reference was 17/sc/0368, granted September 7<sup>th</sup>, 2017. Written informed consent was obtained from patients. Demographic and clinical data were collected at enrolment, outcome data was collected from case notes and electronic systems. ALEA and BC platforms were used for data capture and management.

Studies were registered with ISRCTN trial registry. COV-19POC registration took place on 18<sup>th</sup> March 2020; ISRCTN14966673, and the FluPOC registration took place on November 13<sup>th</sup>, 2017; ISRCTN17197293. The COV-19POC study was non-randomised interventional trial in adult patients presenting to hospital with suspected COVID-19. The study was to evaluate clinical impact of molecular point of care testing using QIAGEN QIAstat-Dx Respiratory SARS-COV-2 Panel on the QIAstat-Dc PCR testing platform (356). The COV-19POC molecular point of care trial took place from 20<sup>th</sup> Marc to 29<sup>th</sup> April 2020. All patients were recruited to acute areas of Southampton General Hospital. The FluPOC study was a randomised controlled trial to evaluate clinical impact of molecular point of care testing for adults hospitalised with acute respiratory illness, during influenza season. Testing was performed with Biofire FilmArray platform with Respiratory panel 2.1 (357). The trial was conducted during 2017/18 and 2018/19 during the influenza seasons. The multicentre trial was conducted at Southampton General Hospital and Hampshire County Hospital. All participants were recruited within the first 24 hours of admission to hospital, and prior to any treatments.

## 2.2.2 Data processing

**Table 2-4 Software packages**

Software Name	Task	Reference
<b>FastQC v.0.11.3</b>	Pre and post trimming quality control	(340)
<b>MultiQC v1.5</b>	Comparison of quality control metrics to identify outliers	(341)
<b>Trimmomatic v0.3.6</b>	Trimming adapters from reads	(342)
<b>STAR aligner software v2.6.1</b>	Aligning to genome/transcriptome and performing counts	(343)
<b>Samtools v1.3.2</b>	Sorting and indexing of Bam files	(344)
<b>Picard v2.8.3</b>	Read groups added and duplicates marked	(345)
<b>RSEM v1.3.1</b>	Generate gene/transcript counts, isoform percentage metrics	(346)
<b>PCAExplorer</b>	Exploratory data analysis	(347)
<b>PandaOmics</b>	Gene expression outlier analysis	(132)
<b>Salmon</b>	RNA splicing analysis	(155)
<b>BANDITS</b>	Relative abundance calculation	(360)
<b>ToppGene</b>	Pathway and processes	(361)
<b>scikit_learn</b>	Data engineering and regression	(362)
<b>Matlab™ R2020b</b>	Classification machine learning	(363)
<b>Statsmodels 0.14.0</b>	p-values were determination for regression	(364)
<b>EnhancedVolcano</b>	Volcano plot production	(365)
<b>UpSet (366)</b>	Upset plot production	(366)

## 2.2.3 RNA extraction and sequencing

Peripheral whole blood was collected in PAXgene tubes and total RNA was isolated using the PreAnalytix PAXgene Blood RNA kit according to manufacturer's protocol. Extractions were performed in containment level 3 Tripass Class 1 hood, and RNA isolates were stored at -80°C. Cytoplasmic, mitochondrial rRNA and globin mRNA were depleted using QIAseq FastSelect-RNA/Globin kit from Qiagen, with fragmentation interval of 7 or 15 minutes. Library preparation was



performed using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (New England Biolabs), with 11 or 13 amplification cycles using AMPure XP beads. Quality control steps were performed using the Qubit and 2100 Bioanalyser from Agilent.

Libraries were then pooled to obtain equimolar ratios. Sequencing was performed using 150bp paired end reads on a Illumina® NovaSeq 6000 (Illumina®, San Diego, USA). This step was performed by Centre for Genomic research at the University of Liverpool

#### **2.2.4 Trimming and Alignment with STAR – Performed by Dr. J. Lord**

Illumina adapter sequences were trimmed using Cutadapt v1.2.1 (358) and reads which matched the adapter sequence by 4 or more bp was trimmed off and read sections with minimum window quality of 20, any reads left with less than 10bp total were also removed. Read data was then processed using FastQC (v0.11.9) and compiled into MultiQC (v1.5)(341). A lower limit on total sample read depth filter of 20 million was applied, all samples under this threshold were excluded from further analysis. Reads were aligned to the human genome version GRCh38, with Gencode annotation v34, using STAR aligner v2.7.6a. Alignment was performed in two pass mode, and according the Encode standard setup options in the STAR manual (Dobin and Gingeras, 2015). Resulting Bam files were sorted and indexed using Samtools (v1.8) (344). 5 influenza patients did not pass minimum sequencing depth thresholds of 20 million reads, 2 COVID-19 patients were excluded due to having seasonal Coronavirus and two were excluded due to having underlying chronic lymphocytic leukaemia. In total 78 patients with COVID19, and 83 patients with Influenza passed filtering for further analysis. Statistical analysis of baseline clinical characteristics was performed, and no statistically significant differences exist between the cohorts for age or sex.

#### **2.2.5 PcaExplorer**

Exploratory data analysis was performed to compare the cohorts' transcriptomes using PCAexplorer. Raw gene expression counts were loaded into pcaExplorer and  $\log_2$  transformation was selected before analysis was started. PCAplots with 95% confidence interval circles produced, along with scree plots, showing the representative contribution of each principal component. This process was the repeated, excluding all participants over the age of 65 years. After top differentially expressed

genes were determined using Pandaomics (Paragraph 2.2.6), some of these genes were then stratified by age and plotted to show difference in expression between the cohorts, with advancing age.

#### **2.2.6 Pandaomics – Differential Gene expression between infections**

The expression patterns of the host's genes in the blood were assessed to given insight into differences in pathogenicity. Pandaomics software was selected for its easy-to-use end to end application, and molecular topology aware algorithms which prioritise upstream changes in gene expression, as well as those of greater magnitude (359). The tool uses the "limma" statistical software which incorporates linear modelling and empirical bayes moderation (360) The software however is incompatible with ensemble gene id's with version numbers, and also cannot be used with transformed data, and instead requires raw gene count data. From the STAR outputs 60669 genes were output, after removing all version numbers from the data, 45 duplicates were also identified by the software which were removed before proceeding. Raw counts for 60624 genes were loaded into Pandaomics software with associated metadata consisting of sample ID, Infection type (Flu/Covid), sex and age. No further data engineering was performed with the Pandaomics software. Differential gene expression and associated pathway analysis was conducted automatically on the data using PandaOmic tool, and the most differentially expressed genes and associated pathways data were downloaded as CSV files.

#### **2.2.7 Transcript Counts – Salmon Tool**

Transcript counts were generated using the tool, Salmon as this program was isoform aware and not computationally intensive (159). Genome build38, and Gencode v.34 annotation. Selective alignment method was used for assigning and reads which gives high accuracy without sacrificing computational speed. This involves the creation of a combinational genome/transcriptome index file referred to as a 'Gentrome'. This was conducted using the Alevin tool protocol developed by the Combine lab (361). Quantification was performed using the script found in appendix A.8. Transcripts equivalence classes were used to estimate transcript abundances. Equivalence classes represent the number of different transcripts to which fragments in a specific class map. In many cases this allows accurate determination of the relative abundance of specific isoform, however with increasing isoform complexity, the ability to accurately discern the transcript of origin decreases.

This step was subsequently repeated and optimised by Yaron Strauch. Specifically, the results henceforth pertain to those obtained using the same methods with Gencode annotation version 39 which was used whilst packaging the tool as a product for commercial application.

### **2.2.8 DTU use between infections – BANDITS Tool**

Salmon transcript counts were loaded into the BANDITS software (362) to compare the relative abundance of transcripts between infections. Pre-analysis filtering of data was performed using the default BANDITS configuration. Transcripts were excluded if they; a) represented a proportion less than 0.01 of total transcripts, b) had less than 10 counts, c) were from a gene which had less than 20 counts total.

BANDITS produces two sets of results, those at gene level and those at transcript level. P-values are produced and adjusted using Benjamini-Hochberg correction. In addition, inverted p-values are provided which vary only when the dominant transcript remains the same in both groups despite abundance changes. This is calculated by taking the square root of p-value which results in an inflated value. This is performed to give priority ranking in results to those results in which differential splicing results in the change of a dominant transcript. The BANDITS software also produces a novel metric denoted as 'DTU measure'. This is intended to measure the intensity of the DTU change, similarly, to fold change quantification in differential expression analysis. DTU measure represents the sum of absolute difference between two most expressed transcripts between the groups. A value of 0 indicates proportions are identical whereas 2 represents different transcripts always being used. BANDITS further gives precision information for mean and sd, higher precision parameters indicate lower sample to sample variability in these values.

The results metric `Max_Gene_Tr.p-value` and `Max_Gene_Tr.Adj.p-value` are conservative hybrid p-values. These are the maximum between gene and transcript level p-values, adjusted respectively. Transcripts are only considered significant if the corresponding gene is also significant. Means and standard deviations are also delivered. The full R script for BANDITS appears in appendix A.9.

Results were filtered by adjusted inverted p-values  $<0.05$  and these values were plotted against the BANDITS unique DTU measure using Microsoft Excel.

The isoform ratios of the top 20 differentially spliced genes as ranked by DTU measure were then plotted in stacked column charts to demonstrate the relative abundance of each isoform irrespective of gene expression changes.

### 2.2.9 GO analysis of results

To understand which pathways were differentially expressed and which were alternatively spliced, the list of isoforms which experienced differential transcript use was converted to lists of genes from which they originate. ENSG id's from this list and the differentially expressed genes list were pasted into ToppGene enrichment analysis online tool (363) with the following parameters. 'Probability density function' was selected for P-value method, Bonferroni correction was selected, 0.05 p-value cut-off and gene limits were 1 and 2000 for lower and upper respectively.

Go terms were retrieved from the ToppGene output and pasted into Revigo tool (364) version 1.8.1 to produce graphs of the enriched biological processes. The option for 'large lists' (includes a higher percentage of results), was selected when using genes, but 'small lists' when using isoforms, as some filtering was needed on the larger lists of isoforms to make the data manageable. The option to remove obsolete terms was set to yes. Semantic similarity for the graph was set to default 'SimRel'. The Revigo tool only allows for a maximum of 2000 GO terms to be input, which was exceeded by both sets of isoforms' associations. Only the top 2000 GO terms were included in the analysis.

This was repeated for both gene specific pathways and isoforms specific pathways for each infection.

### 2.2.10 Multiple regression performed using Python.

#### Feature engineering

Both gene expression and relative isoform abundance were regressed from age. Initially gene expression TPM's were calculated from raw reads-per-gene values, originally obtained during the STAR output files during mapping process. This was performed in Microsoft excel using the formula.

$$RPK \text{ of } i = \frac{\text{read counts of } i}{\text{gene length in kilobases of } i}$$

$$TPM \text{ of gene } i = \frac{RPK \text{ of gene } i}{(\sum ALL RPKS)/1,000,000}$$

(Where  $i$  denotes any gene of interest)

The relative isoform abundances were calculated by stripping transcript level counts for all transcripts from each gene, combining them into a .txt file, and deriving relative abundances in Microsoft Excel, by summing all the transcript reads from a gene and dividing this by the value assigned to each transcript.

These two steps were subsequently optimised for commercialisation by Yaron Strauch during packaging for commercial applications and now occurs automatically during feature engineering using a python script. The TPM values are instead calculated from summing all transcripts reads from various equivalence classes from produced by Salmon instead of counting directly to genes using Star. All results henceforth are derived using this optimised method.

Using the python package `scikit_learn`, data engineering and regression-based ML was performed, and p-values were determined using the `statsmodels` package. Missing CRP-values for 7 patients were filled using clinical phenotypes-based regression which included age, white blood cell count, neutrophil count, and lymphocyte count trained on the remainder of the influenza cohort. 10-Fold cross-validation was used to validate this procedure.

Filtering was applied for any features with a single value across the dataset. One-hot vector encoding was used to represent categorical data. A standardisation step was performed for all transcript features by subtracting mean and dividing by standard deviation.

To identify transcriptome changes associated with advancing age in the two cohorts, multiple linear regressions were performed to regress engineered transcriptomic metrics (isoform abundance or gene expression TPM) from the confounding factors age, sex, white blood cell count, neutrophil count, lymphocyte count, C-reactive protein levels, Diabetes status, immunosuppression status, smoking status, presence of cardiovascular disease, and presence of respiratory disease. A total of 60669 gene and 207749 transcripts were regressed.

The alpha intercepts and beta coefficients were extracted using scikit-learn, and ordinary least square regression provided by statsmodels was used for P-value quantification.

Beta coefficients values describe the nature of the relationships between dependent and response variables and allow comparison between experiments. The beta coefficient values for gene expression and isoform abundance were extracted and printed to .csv.

The linear regression steps were performed by Yaron Strauch.

For each infection, features with a beta coefficient value of zero were removed and histograms were plotted to show the distribution of beta-coefficients representing association between age and both gene expression and relative transcript use. Significant associations were determined and filtered for with p-value threshold of 0.05 using Microsoft Excel.

Beta coefficient values for significant changes in isoform abundance were plotted against the  $-\log_{10}$  P-value, in modified volcano plots. Using the R package EnhancedVolcano (365). This process was then repeated for splicing factors. The list of splicing factors was taken from the previous work conducted on splicing abundance by Lye et al. 2019 (314). For each infection, those genes which showed statistically significant changes in expression with age were compared with those which underwent changes in isoform abundance with age and the results were represented in Venn diagrams.

### **2.2.11 Classification**

To understand if transcriptomic signatures in peripheral whole blood could be used to automatically distinguish between infections, machine learning classification algorithms were employed. The top 100 most differentially expressed genes between the infections, as identified by Pandaomics, were selected as features and classification machine learning was performed using Matlab™ R2020b, with the infection type as a response. Multiple types of classification models were used and compared. Classification was conducted for the cohort using the follow models: fine tree, medium tree, Coarse Tree, Linear discriminant, Logistic Regression, Gaussian Naïve Bayes, Kernel Naïve Bayes, Linear SVM, Quadratic SVM, Cubic SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian, Fine KNN, Medium KNN, Coarse KNN, Cosine KNN, Cubic KNN, Weighted KNN, Boosted Trees, Bagged Trees, Subspace discriminant, Subspace KNN, RUSBoosted Trees. To ascertain if the age of the cohort, and so the ageing of the immune system, affected the ability of these algorithms to distinguish between the transcriptomic profile of those infected a series of repeats were conducted. The classification

was conducted for the entire cohort, and then for each of the following age restricted subsets; <81 years old, <71 years old, <61 years old, less than 51 years old, >30 years old, >40 years old, > 50 years old and > 60 years old. The peak performance of each model, for each respective cohort was recorded. Cumulative peak performance values and mean peak performance values for all models for each age restricted cohort was then calculated to show the overall performance of machine learning classifier models for each cohort subset to show how infection classification prediction performance varies with age of cohort.

### **2.2.12 Lasso Regression for age prediction – Coding performed by Yaron Strauch.**

To produce disease transcriptomic ageing clocks, which are not splicing agnostic but incorporate both types of features (gene expression and relative isoform abundance) the gene expression TPM data and relative isoform abundance proportion data were used as independent features to predict age.

Further data engineering and Machine Learning was performed with the python package scikit-learn (366). Features with one distinct value across the entire dataset were dropped. 200,000 features were pre-selected by performing  $f$ \_regression (f-test) of features and ranking by score.

All 200,000 features were standardised by subtracting mean and dividing by standard deviation. 15,000 lasso regressions were fitted for evenly distributed  $\alpha$  values between 0 and 4 inclusive; low  $\alpha$  values select for more features, high  $\alpha$  values drive more beta coefficients towards zero. For each  $\alpha$ , a lasso model was fitted and non-zero values are extracted. 4-fold cross validation was performed whereby the model is trained one  $\frac{3}{4}$  of the data and tested on the remaining  $\frac{1}{4}$ . The mean and standard deviation of the training split was recorded, and the data is standardised using these parameters. A model is then fitted, and the mean average error (MAE) root mean squared error (RMSE) and  $R^2$  is recorded. The lowest MAE from each set of features was recorded and a linear model was then fitted on the corresponding feature set for all data.

To assess the potential overfitting of the best performing model, the mean absolute error values of a subset of all the models were plotted. This subset consisted of the best performing algorithm for

each set of features. To produce this subset, the alpha with lowest RMSE for each subset of features was written to CSV. Graphs were produced to show the numbers of isoforms abundance changes and genes expression changes which were most predictive of biological age in the disease cohorts.

### **2.2.13 Analysis of features**

We hypothesised features with each individual disease-specific ageing clock might be similar, indicative of non-disease related ageing biomarkers, the features of each algorithm were compared to identify any similarities and examined to understand contribution of expression changes and of splicing changes in the feature list.



## **Chapter 3 Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene expression**

### **3.1 Introduction**

Gene expression outlier identification was investigated as a modality for identifying causal variants in the Primary immunodeficiency cohort. Control datasets included the pre-selected, GTEx dataset of 113 samples, the Splicing and Disease cohort, and the two healthy controls included in the sequencing of the PID cohort. Quality control steps were performed as every stage to ensure data integrity and reliability as described in the methods chapter. Z-score calculation and subsequently the OUTRIDER tool were identified in the literature as potential methods for outlier detection, and these were employed to interrogate the datasets as described in the methods sections. This chapter presents the results of quality control steps, the exploratory data analysis, and the testing of the different methods, with some follow up investigation of results.

### **3.2 Result from the quality control steps.**

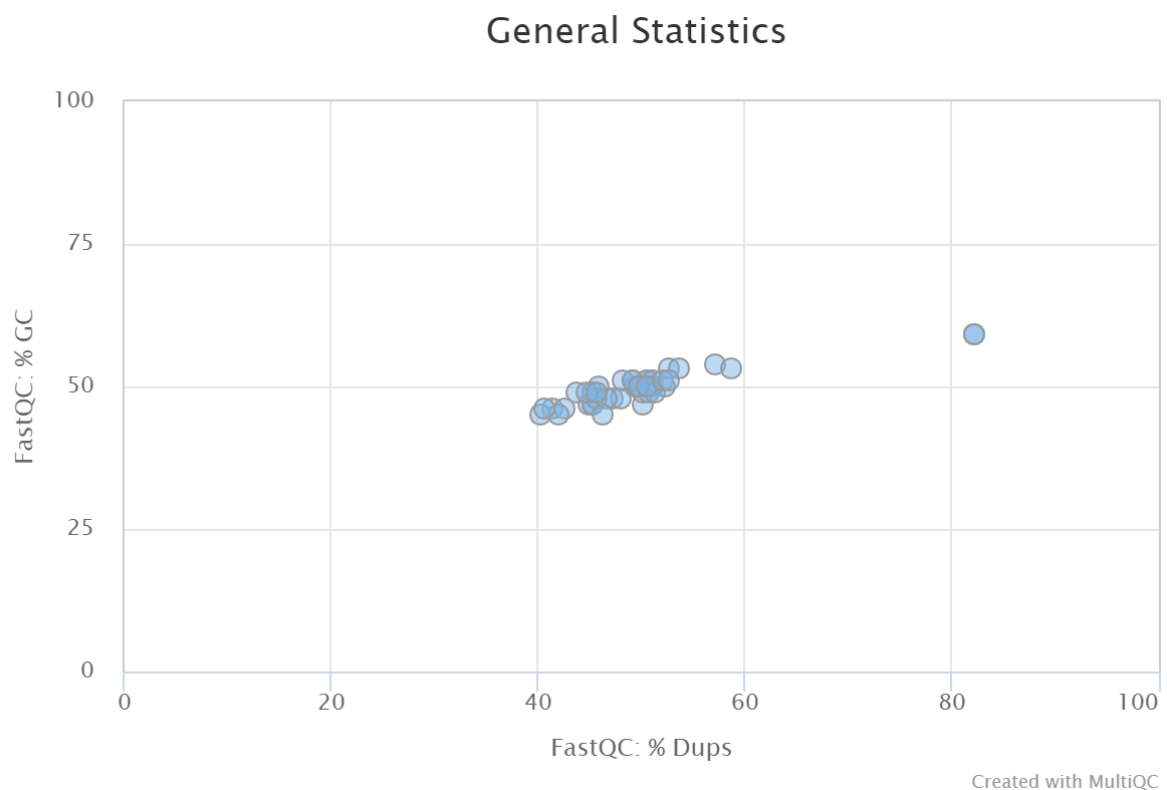
Quality control steps were performed as every stage to ensure data integrity and reliability as described in the methods chapter.

The quality control steps performed at Southampton General Hospital indicated that the RNA yield for sample SBR00001 ( $19.4 \text{ ng. dl}^{-1}$ ) appeared to be just below our pre-defined threshold ( $20 \text{ ng. dl}^{-1}$ ). However, this sample was still progressed to sequencing with the caveat that it could be removed from analysis steps if it appeared to be an outlier or skewed/affected the latter results. Sample SBR0004 failed quality control steps twice by a large amount and was removed from the study. Sample SRB0001, along with SRB0013 and SRB0014, were also 'QC flagged' for RNA quality and amount by Novogene during quality control, as they may produce low quality data due to low yields

and RIN values. It was decided the project should proceed with analysis of these samples, but they could be excluded from statistical models if necessary.

MultiQC results from the prepared RNAseq libraries reflected low number of reads in two healthy control samples and SRB0003 which is expected due to lower sequencing depth for these first samples. However, SRB00014 also flags as having less than 10M reads. It was noted that the earlier RIN value for this sample was below acceptable threshold and so this low sample quality probably contributed to the low read numbers.

After adapter trimming, MultiQC analysis highlighted further outlier characteristics with SRB0001, which had an 82.3% sequencing duplication percentage, 24% higher than the next highest sample, which was at 58%, and was also an outlier for GC content (Figure 3-1). High GC content combined with duplication levels is often indicative of adapter contamination (A.3).



**Figure 3-1 PID patient MultiQC comparisons; GC content vs Duplication percentage**  
Duplication percentage along the X axis and GC percentage along the Y axis both shows SRB0001 to be an outlier when compared the rest of the samples.

Inspection of combined results in MultiQC after trimming of residual adapter presence for all samples, however this is not uncommon. The GC content trace for sample SRB00001 had an irregular distribution compared with the other samples; specifically, it appeared to have lower percentages of samples with 40% GC content, SRB001 also had the highest peak in adapter content.

### **3.2.1.1 RNAseq read alignment results.**

Post alignment QC was performed by comparative analysis of log.out.final files created by the STAR aligner program. Sample SRB0001 displayed severe lack of quality alignment via a number of metrics. Sample SRB0001 had only 27% uniquely mapped reads, about 66% less than the mean, compared to ~ 50% for the next lowest sample. SRB0001 also had 10x as many reads mapped to multiple loci as some other samples, this equated to 58%. There were no other significant outliers in any category. Sample SRB00013 and SRB00020 were flagged as having greater than 1% reads mapped to too many loci, however they were only just above threshold for flagging (1%) and did not warrant discontinuation of investigation.

### 3.3 Exploratory data analysis

PcaExplorer is a software package for the programming language R (367). This package was used for the exploratory analysis of the RNAseq data and as a first pass to assess the utility of the control datasets. Through the quality control and alignment steps, the number of uniquely mapped reads in the PID cohort appeared to be between 31M and 63M. After pcaExplorer filtering which removes low expressed genes based on thresholds described in the methods section, the total number of reads per samples was dramatically reduced. A graphical representation of post-filtering read counts can be seen below in Figure 3-2. There were less than half of the original number of reads in some cases. The PID group for example appears to have 15-25M reads per sample once lowly expressed genes are filtered out. The overall variability in aligned read counts is much higher in the GTEx group of 113 samples. Basic statistics for all samples were automatically generated from read count data (Table 3-1).

**Table 3-1 Basic descriptive statistics for read depth.**

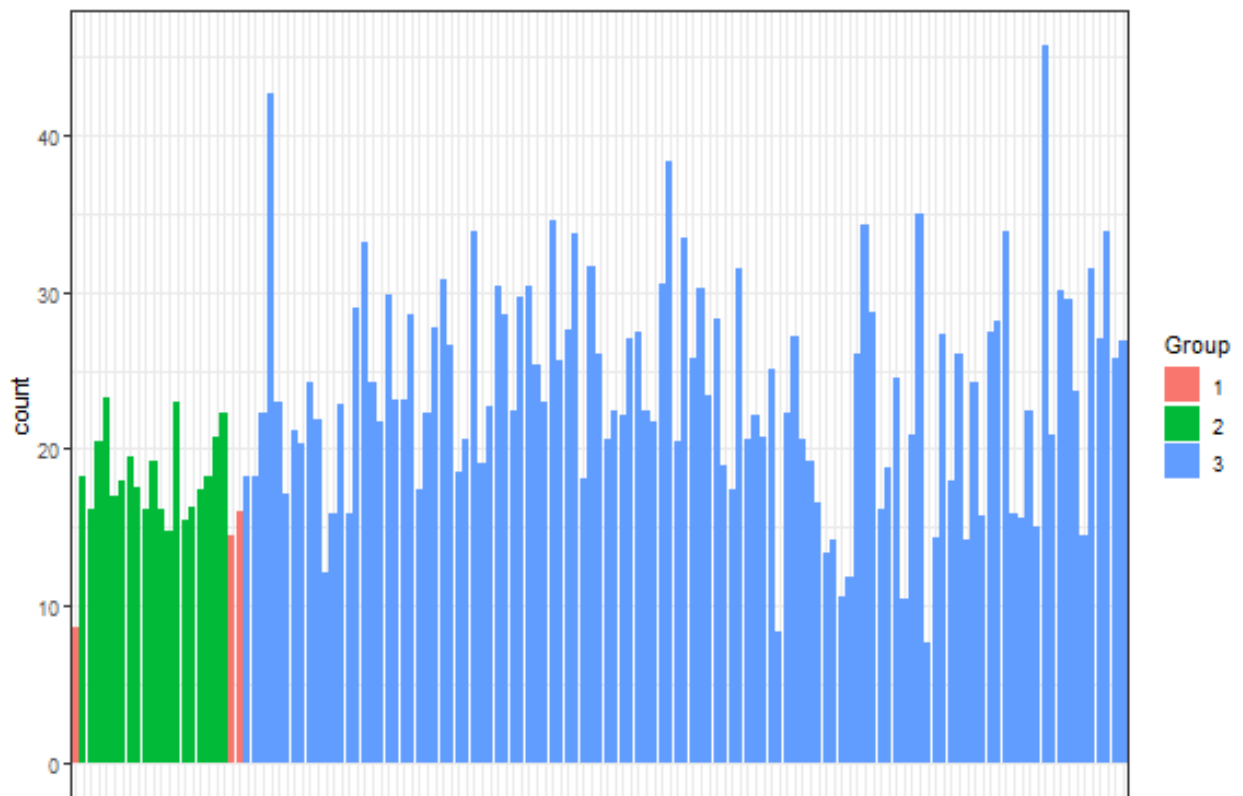
<b>Minimum</b>	<b>1<sup>st</sup> quartile</b>	<b>Median</b>	<b>Mean</b>	<b>3<sup>rd</sup> Quartile</b>	<b>Maximum</b>
7.618	17.729	22.304	22.746	27.405	45.661

The majority of the PID samples were around or below the mean and median values for the reads per sample across all groups. There is enormous variability of the GTEx dataset seeming to range from <10M to >45M reads.

Principal component analysis (Figure 3-3) presents the GTEx samples as clustering completely separately from the PID or healthy control samples, with no 95% confidence interval overlap between datasets. The separation of the data was by a very large margin in terms of principal component 1, and there was greater in-group variability in the GTEx data than between the PID data and controls, regardless of batch or initial sequencing depth. Principal component 2 displayed greater differences within the group in GTEx data than between GTEx and PID/control data. Scree

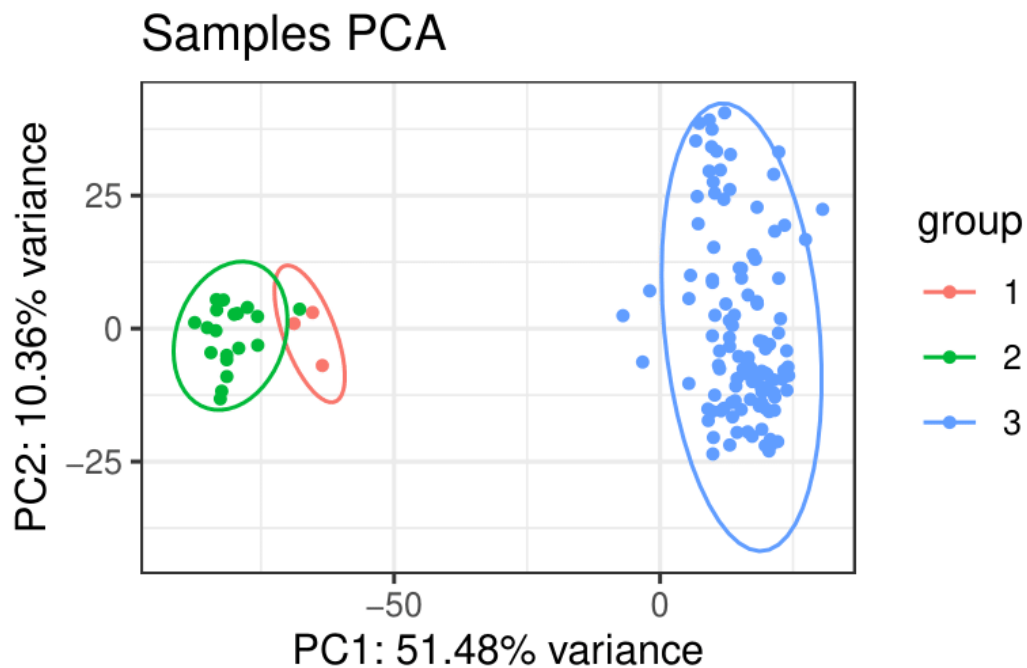
plots (Figure 3-4) show that the variance caused by PC1 is almost 10 times as large as the variance caused by PC3.

## Number of million of reads per sample



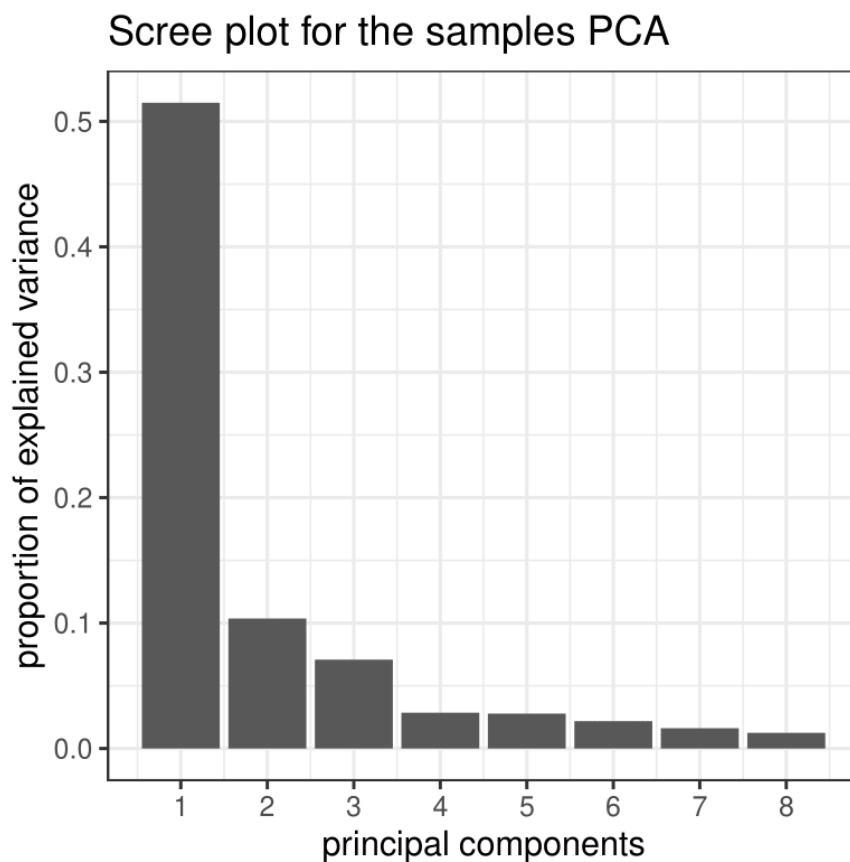
**Figure 3-2 Reads per sample: GTEX and PID**

Sequencing depth of the samples, Group 1 (red) comprises the two control samples which were produced and extracted internally and patient SRB0003, group 2 (green) includes all of the remaining 19 PID patients and group 3 comprises the 113 GTEX samples. Few PID samples have more than 20M reads per sample.



**Figure 3-3 PCA plot of GTEx and PID RNAseq data**

A principal component analysis of filtered and transformed RNAseq data. Group 1 (red) comprises the two control samples which were produced and extracted internally and patient SRB0003 prepared using Poly-A enrichment library preparation methods, group 2 (green) includes all of the remaining 19 PID patients using lncRNA library prep methods and group 3 comprises the 113 GTEx samples. 95% confidence interval ellipses were included to demonstrate clustering based on principal components.



**Figure 3-4 Scree plot for PCA analysis of GTEx and PID RNAseq data**

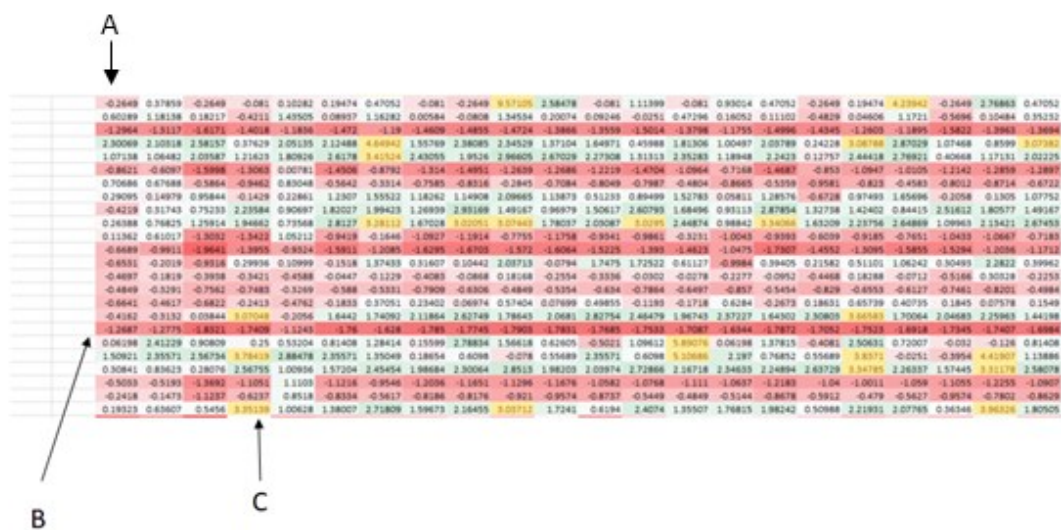
The first principal component accounts for over half of all of the variability between samples and is almost five times as much as the next largest principal component. PC1 also contributes double the total variance which is contributed by the next seven principal components combined.

### 3.4 Outliers in gene expression: Using FPKM values and Z-scores.

Quantifying gene expression for mendelian disease has previously been successfully performed using RPKM and Z-score calculation (210). The present analysis uses FPKM (fragments per kilobase per million) values, an almost identical metric which takes into account the paired end read sequencing method. Outliers in gene expression for FPKM PID data were calculated using Z-scores in Microsoft Excel. Conditional formatting was applied to Excel spreadsheets containing the Z-scores for each gene, in each PID panel, for each patient. In this intermediate processing stage, each sample was represented by a vertical columns, and conditional formatting was applied to each vertical column,

Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene expression

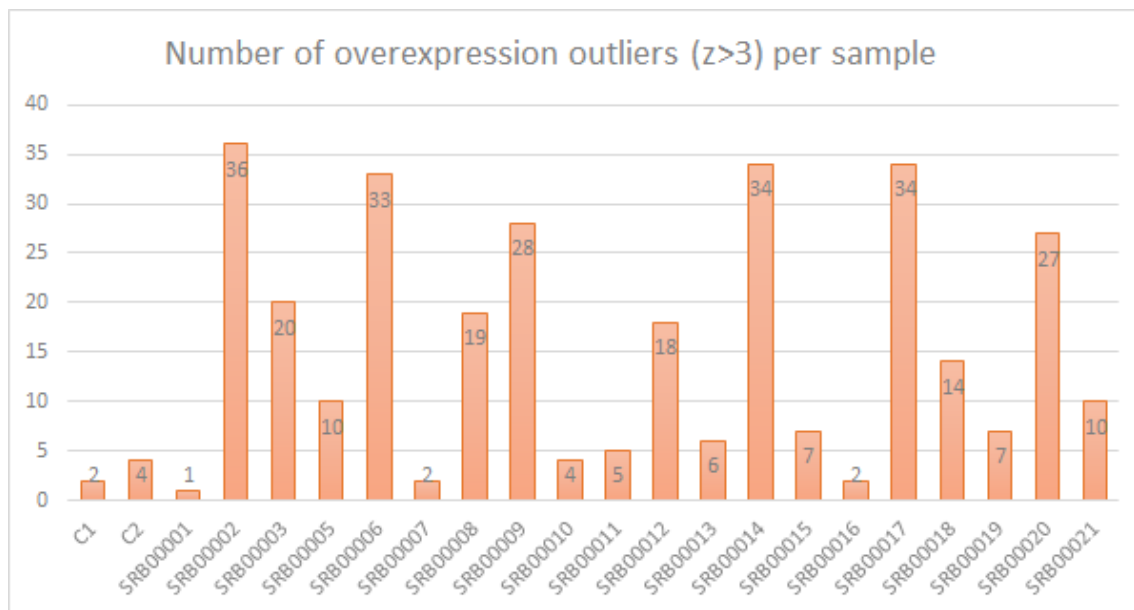
where the highest values were green and the lowest values were red. Visual inspection of the PID table spreadsheets at this interim stage showed clear and strong systematic effects on Z-scores; for some genes all PID samples were statistical outliers (Figure 3-5). The effect of this skewing could also be visualised by the FPKM heat-map (before Z-scores were calculated) which included the mean values for the GTEx. These horizontal correlations were almost exclusively occurring when the GTEx mean was a clear outlier compared with the samples plus controls. In addition, there were a high number of outliers ( $z > 3$ ) per sample (Figure 3-6).



**Figure 3-5** Example image of Z-score tables (excerpt from table 1 shown) with GTEx as controls.

A) Each column represents a sample. B) Each row represents a gene, the gene which arrow B points to shows an example of the horizontal correlations in Z-scores, as all samples are highlighted red (low expression) indicative of skewing of PID Z-scores by the GTEx. data C) Yellow highlighted boxes represent expression outliers (overexpression  $Z > 3$ ).



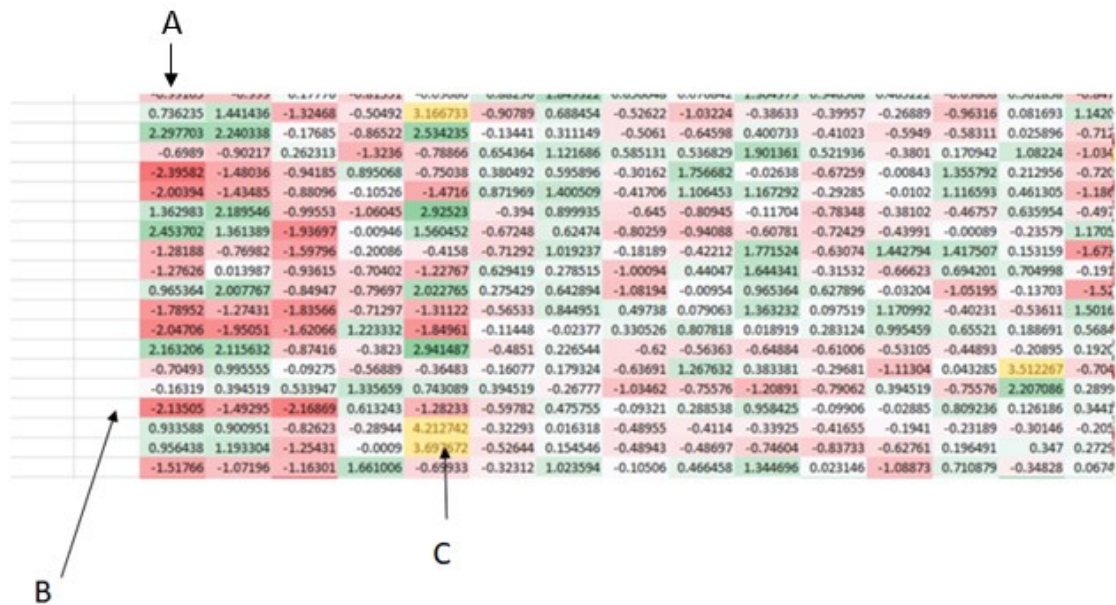


**Figure 3-6 Number of overexpression outliers obtained from Z-score calculations with the GTEx data included as controls.**

The graph shows the number of outliers (overexpression,  $z > 3$ ) on a sample wise base, with all GTEx values included in the calculations but not included in the chart. These are calculated from the PID IUIS panel, made up of 413 genes. This indicates that in some cases (SRB00002) the outliers make up around 8.7% of genes.

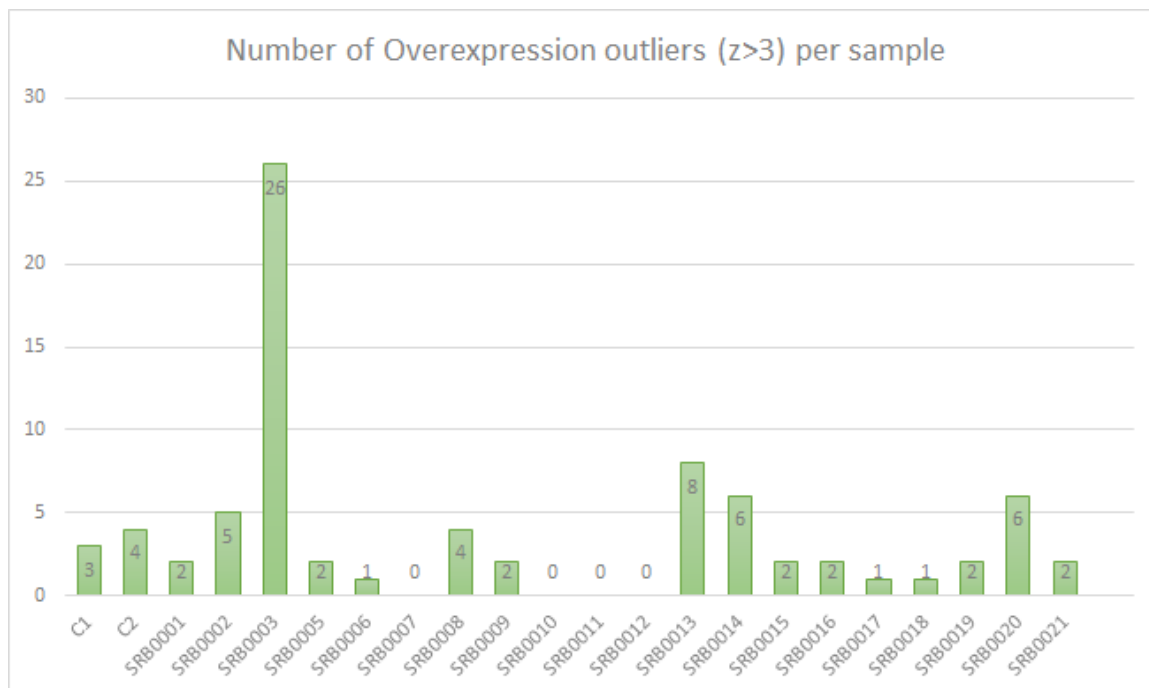
Due to the clear effects on outlier status caused by inclusion of GTEx data, this control data set was removed from subsequent Z-score calculation, and the spreadsheets were re-calculated (Figure 3-7). The skewing of data was absent, and correcting the systematic bias also resulted in a greatly reduced number of gene expression outliers across the PID samples (Figure 3-8), these were now at levels which could be investigated individually. SRB003 had unusually high numbers of expression outliers ( $n=26$ ) compared to the other samples. Under-expression outliers were not readily detected in the PID datasets, however; Only 1 under-expression outlier was present. This occurred in sample SRB0001 (not shown). The Z-score outliers found using this method were presented to the clinician responsible, Professor Tony Williams. The results (Table 3-2) did not appear to show changes in expected genes in patients with known diagnosis, nor did they reflect any known associations to phenotypes in patients with only clinical phenotype data. It was suggested that alternative methods should be devised, and an alternative control group should be established if possible.

Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene expression



**Figure 3-7 FPKM and Z score table except: GTEx data not included in calculation.**

- A) Each column represents a sample. B) Each row represents a gene, the gene which arrow B points to shows an example of the horizontal correlations in Z-scores, as all samples are highlighted red (low expression) indicative of skewing of PID Z-scores by the GTEx. data C) Yellow highlighted boxes represent expression outliers (overexpression  $Z > 3$ ).



**Figure 3-8 Over expression outliers present with GTEx control data excluded.**

This graph shows a greatly reduced number of expression outliers per sample, as would be expected from a heterogenous population. This is the case with the sole exception of SRB003 which has an increased number of expression outliers.

**Table 3-2 FPKM Z-score outliers from PID cohort**

Sample	Genes with TPM based expression outliers	Z-scores for genes	Diagnosis
SRB0001	SEMA3E	3.023668	No Diagnosis in GECIP
	CFTR	3.518489	
	C7	3.319817	
SRB0002	FOXN1	4.274695	No information
	PSENN	3.870322	
	ALPI	3.435364	
	IL36RN	4.18131	
SRB0003	Too numerous		No Diagnosis in GECIP
SRB0004	Degraded sample		No information
SRB0005	CCBE1	4.364673	CARD 11 A-C @2987250
	IL2RA	4.118184	

Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene  
expression

	CEBPE	3.91707	
SRB0006	RNF31	3.105485	STAT1
	NFKB2	3.342633	
	TNFSF12	3.486897	
	C9	4.541959	
SRB0007			STAT1
SRB0008	GIN51	3.138757	STAT1
	IL23R	3.027007	
	CFH	3.079085	
	CFHR5	4.582576	
SRB0009	IL21	3.91511	NKKB1 A>AT @102582929
	CFI	3.162278	
	FANCI	3	
SRB0010			NKKB1 A>AT @102582930
SRB0011			ILRG;CXorf65;FOXO4
SRB0012	0		ILRG;CXorf65;FOXO4
SRB0013	RAG1	3.512267	No Diagnosis in GECIP
	FAT4	3.178482	
	IL17F	3.360506	
	C4A	3.258921	
	C4B	3.208775	
	CFHR2	3.524924	
	CFHR3	3.41927	
	BRIP1	3.097605	
SRB0014	TAP1	3.305794	No information
	NCF1	3.473473	
	APOL1	3.319696	
	IL17RA	3.206766	
	SERPING1	3.167219	
SRB0015	FERMT1	3.592095	No Diagnosis in GECIP
	CFHR1	4.582576	
SRB0016	C8A	3.775805	
SRB0017	C6	4.582576	No information
SRB0018	CFI	3.162278	No information
SRB0019	IL12B	3.66049	No information
	THBD	3.031026	
SRB0020	CD40	3.003646	No information
	TTC7A	3.178153	
	CARD9	3.310548	
	C1QA	3.260571	
	C1QB	3.197129	

Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene  
expression

	FCN3	3.338513	
SRB0021	XRCC2	3.008934	No information

### **3.5 Splicing and Disease Cohort**

Due to the need for alternative control groups, the ‘Splicing and Disease’ cohort data, described in chapter 2.1.10, which had been sequenced with the same parameters, was processed identically and included in the rest of the analysis, replacing the GTEx data as a control. This was primarily because PID patient samples were limited in number potentially reducing statistical power.

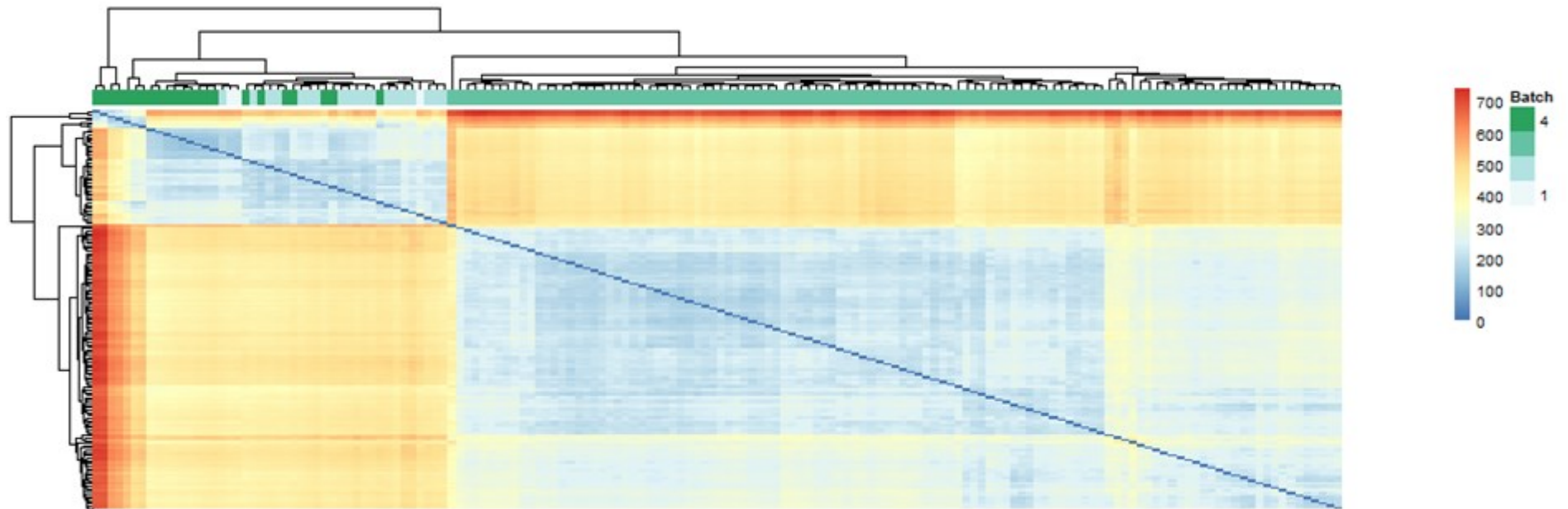
#### **3.5.1 Exploratory data analysis of dataset including ‘Splicing and Disease’ Cohort.**

Exploratory data analysis was conducted using the `pcaExplorer` tool. The Euclidian distance heatmaps shows clear batch difference between GTEx and other groups of data (Figure 3-9). After removal of GTEx data, the Euclidean distance heatmaps showed overall reduction in batches in the data, although some clustering was still present (Figure 3-10). This was comparatively small however, as indicated by the reduced intensity blue colour. The presence of outlier samples was observable. Once ComBat-seq was implemented, the batches appeared to partially dissolve, and increased interspersed of the samples was apparent (Figure 3-11). Healthy controls remained close together, although these has been sequenced together at lower depth. As a further indication of the converging of the data after batch correction, the automatically generated scale of the Euclidean distance colour key reduces across the three figures from maximum value of 700 with GTEx, to 600 without the GTEx, and finally 150 after ComBat-seq implementation (arbitrary units).

Pearson correlation heatmap produced before ComBat-seq application demonstrate, through both colouration patterns and dendrogram structure, the presence of a strong, unseen clustering effect (Figure 3-12). Investigation showed this to not be resulting from batch, sex, or disease state. After the application of ComBat-seq this effect was removed, there was a higher degree of correlation overall, demonstrated by the increased red hue overall. Importantly there is a change in the scale after batch correction also, which indicates improvement of overall correlation (Figure 3-13). No Pearson correlation heatmap could be produced for the combined dataset, the `pcaExplorer` package was unable to display the figure due to the large number of samples.

2D PCA plots initially show no overlap between GTEx data and other data in terms of confidence intervals, distances between the GTEx and other datasets is approximately 3x size of the variation within any group across this axis (Figure 3-14). The scree plot for this pca plot shows principal component 1 to be responsible for over 70% of the variation in samples. There is overlapping of confidence intervals for the PID and Splicing and Disease cohorts indicative of improved utility as a control dataset when compared with the GTEx, as evidenced by the pca plot with the GTEx data removed, which showed the PID cohort to exist entirely within the 95%CI for the splicing and disease cohort, although the 95%CI for the PID cohort did extend beyond the splicing and disease boundaries, indicative datasets not being completely analogous (Figure 3-16). After ComBat-seq was implemented the 2D pca plots showed the 95% CI circle to be completely contained within the 95% CI boundaries for the splicing and disease patients (Figure 3-18). The number of points not completely inside the PID cohort reduced from appx. 13 to appx 5. Scree plots produced from the data before and after ComBat-seq was applied demonstrate a more balanced effect from the various principal components, indicative of a reduced impact from any batch effect or technical variation (Figure 3-15)(Figure 3-17)(Figure 3-19).

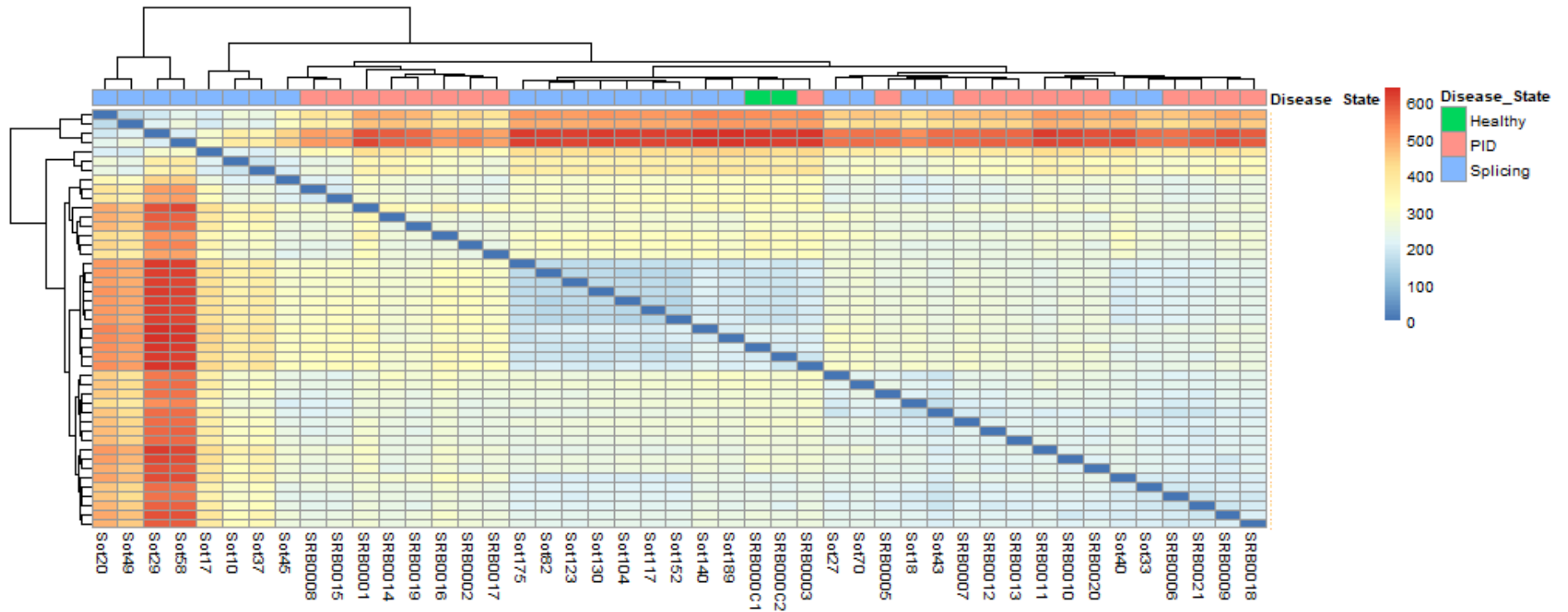
SRB0001 appears to consistently be an outlier sample, likely resulting from the low quality of the sample. 3D pca plots were also produced for the combined datasets and the final batch corrected dataset to visualise the relationships when considering the top 3 principal components. The clear interspersed increase after application of batch correction gives confidence to the utility of the splicing and disease dataset after ComBat-seq is implemented to account for surrogate variables.



**Figure 3-9 Euclidean distance heatmap with data from GTEx, PID cohort and splicing and disease cohort.**

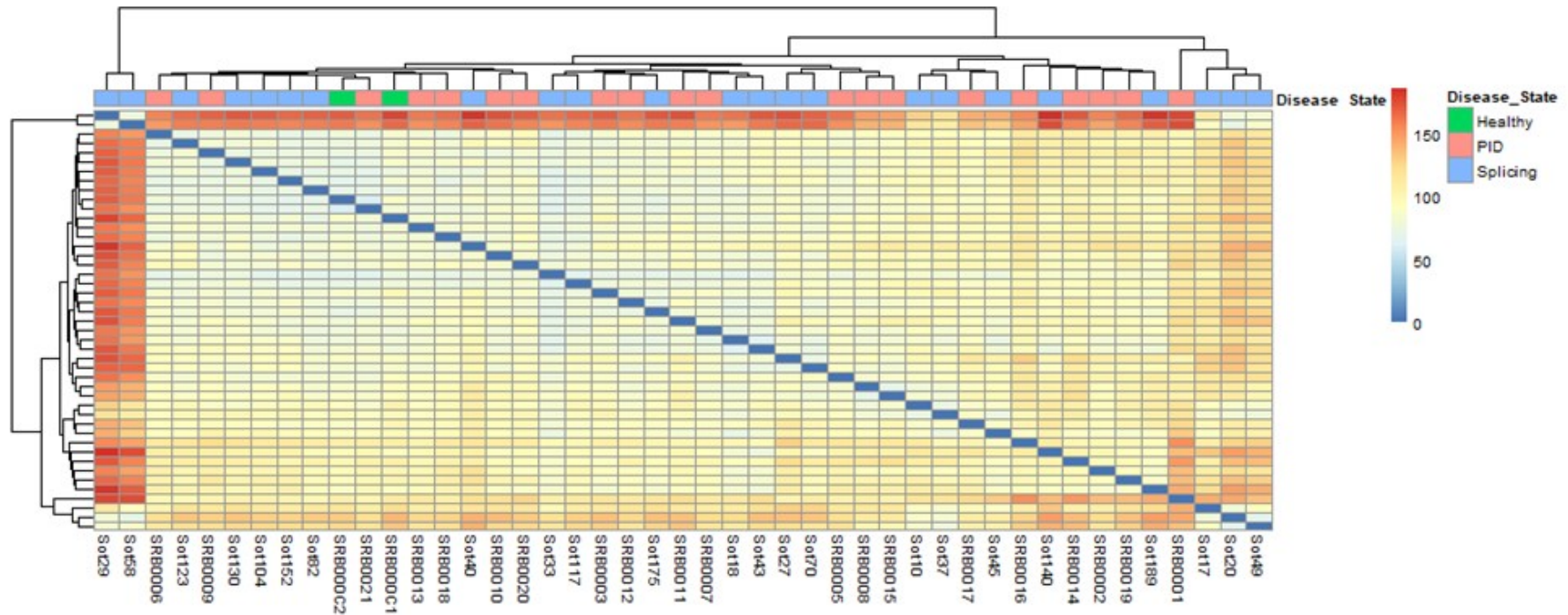
Group 1 represents the first batch of RNAseq including SRBC001, SRBC002, and SRB0003. Group 2 represents the PID cohort data, group 3 represents the large GTEx data set, group 4 represents the Splicing and Disease cohort data. Dendrogram represents hierarchical clustering based on Euclidean distance between samples.





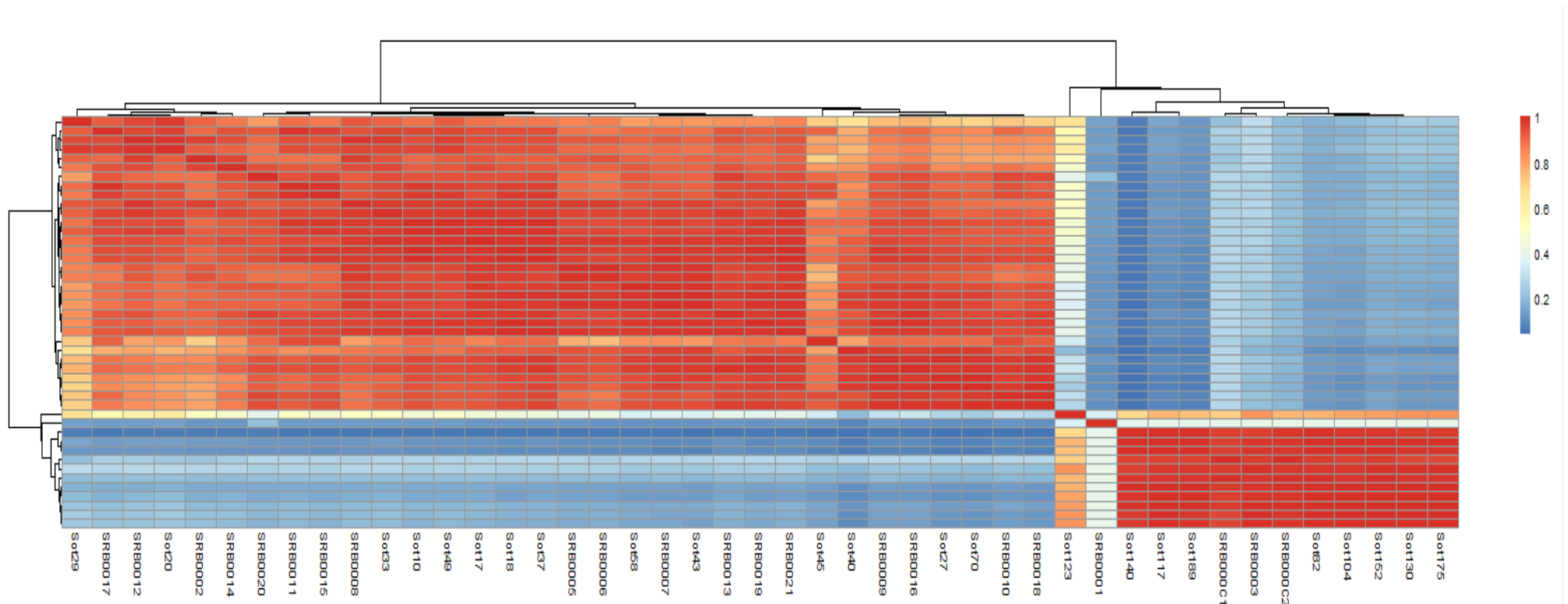
**Figure 3-10** Euclidean distance heatmap with data from healthy controls, PID cohort and splicing and disease cohort.

Group 1 (green) represents the healthy controls SRBC001, SRBC002. Group 2 (pink) represents the PID cohort data, group 3 (blue) represents the large GTex data set, group 4 represents the Splicing and Disease cohort data. Dendrogram represents hierarchical clustering based on Euclidean distance between samples.



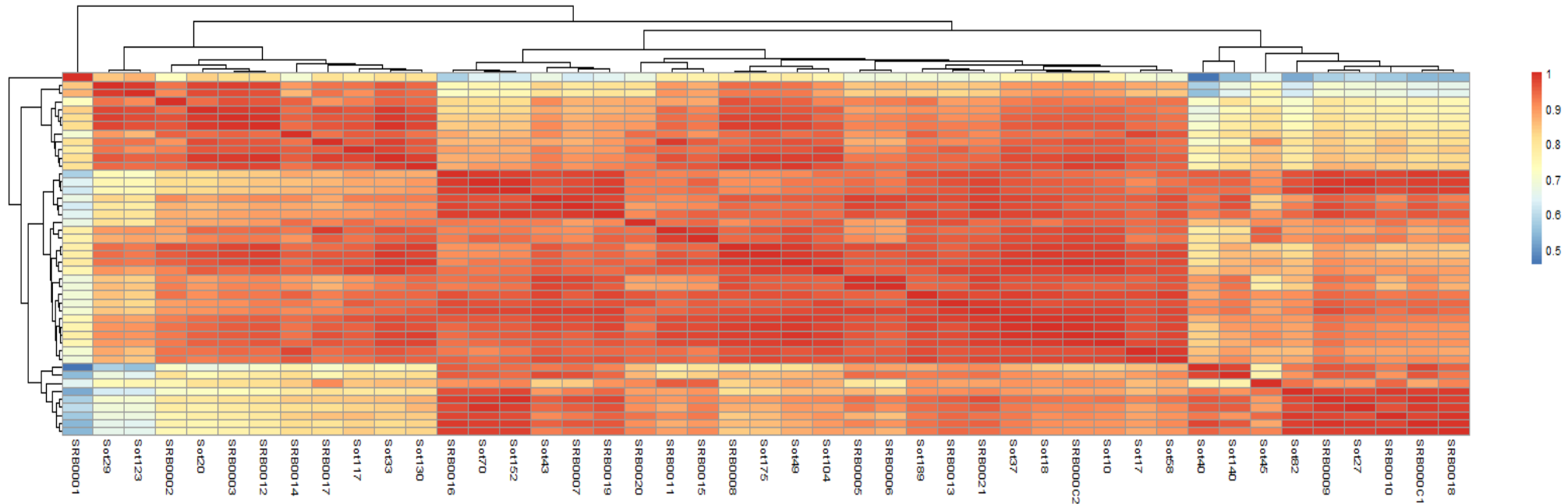
**Figure 3-11 Euclidean distance heatmap with data from healthy controls, PID cohort and splicing and disease cohort after batch correction with ComBat-seq.**

Group 1 (green) represents the healthy controls SRBC001, SRBC002. Group 2 (pink) represents the PID cohort data, group 3 (blue) represents the large GTEx data set, and group 4 represents the Splicing and Disease cohort data. Dendrogram represents hierarchical clustering based on Euclidean distance between samples.



**Figure 3-12 Pearson correlation heatmap with data from healthy controls, PID cohort and splicing and disease cohort.**

Pearson correlation coefficient map shows the degree of similarity between samples based on gene expression. Scale runs from 0-1. 1 (red) is indicative of identical expression and zero is indicative of no similarity.



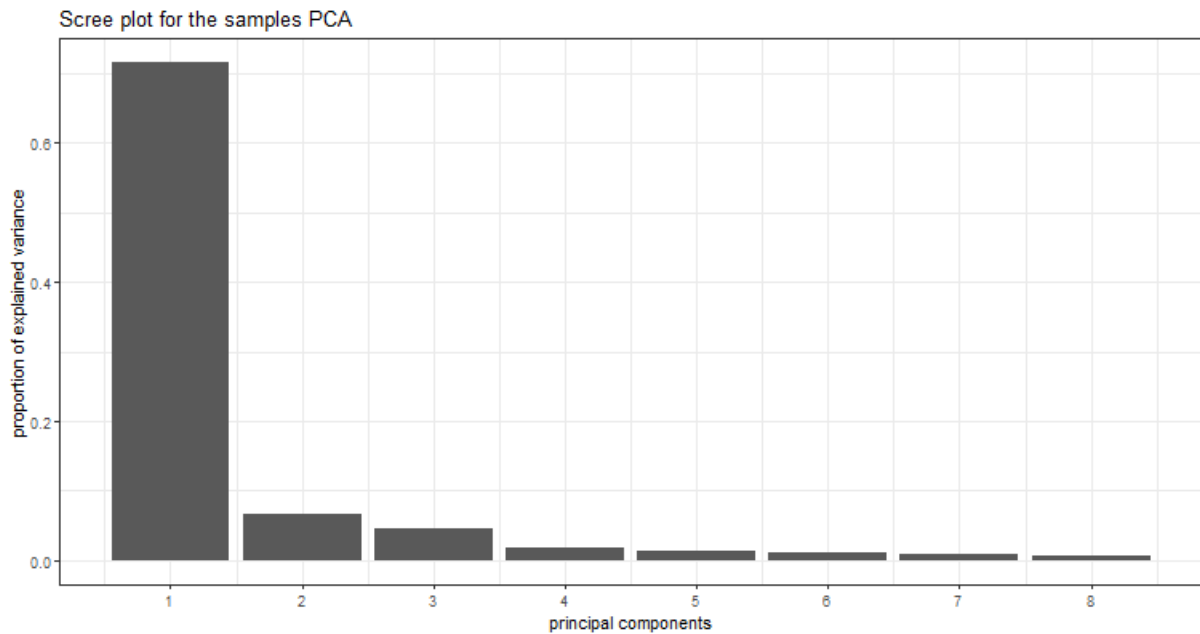
**Figure 3-13 Pearson correlation heatmap with data from healthy controls, PID cohort and splicing and disease cohort after batch correction with ComBat-seq.**

Pearson correlation coefficient map shows the degree of similarity between samples based on gene expression. Scale runs from 0-1. 1 (red) is indicative of identical expression and zero is indicative of no similarity.



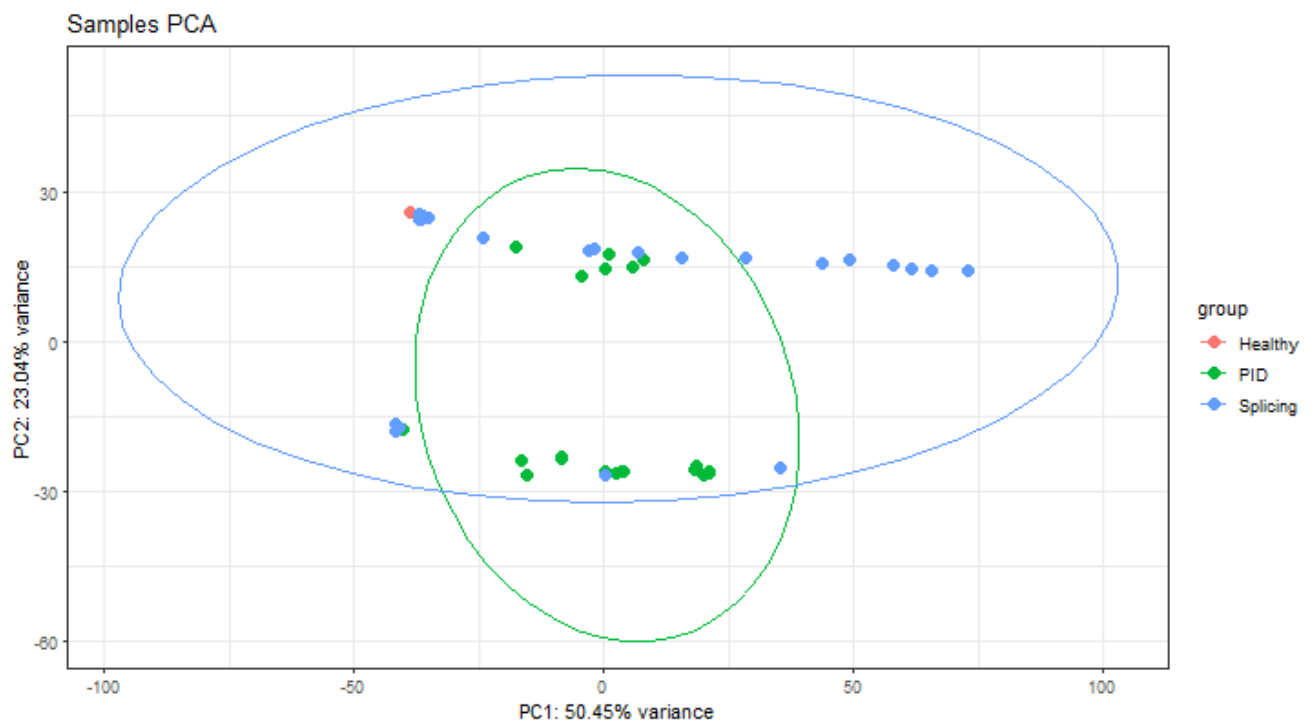
**Figure 3-14 2D PCA plot of PID, Splicing and Disease and GTEx cohorts.**

Red (group 1) represents healthy control samples, green represents PID samples, red represents 'Splicing and Disease' samples, and aqua represents the GTEx samples. Separation around PC2 (Y axis) was determined to be a result of expression of sex-specific genes. 95% confidence intervals represented by ellipses.



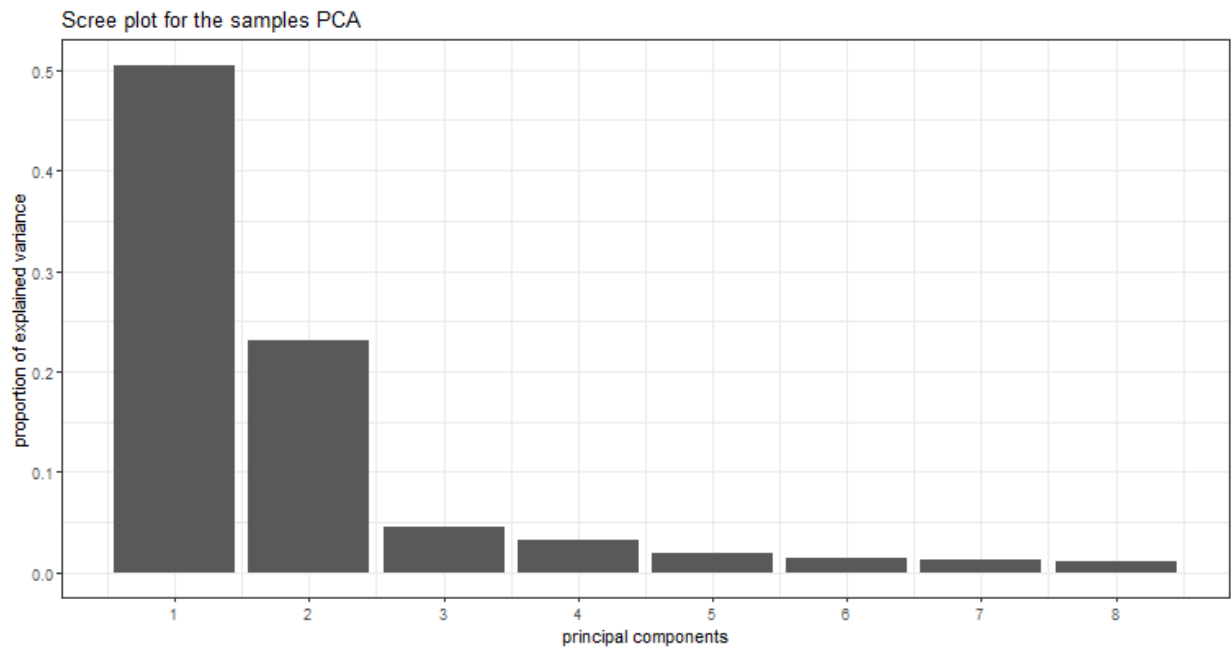
**Figure 3-15 Scree plot from PCA of PID, Splicing and Disease and GTEx cohorts.**

Each column represents the total proportion of variance which is explained by the principal component on the x axis. Principle component 1 is responsible for around 72% of the total variance.



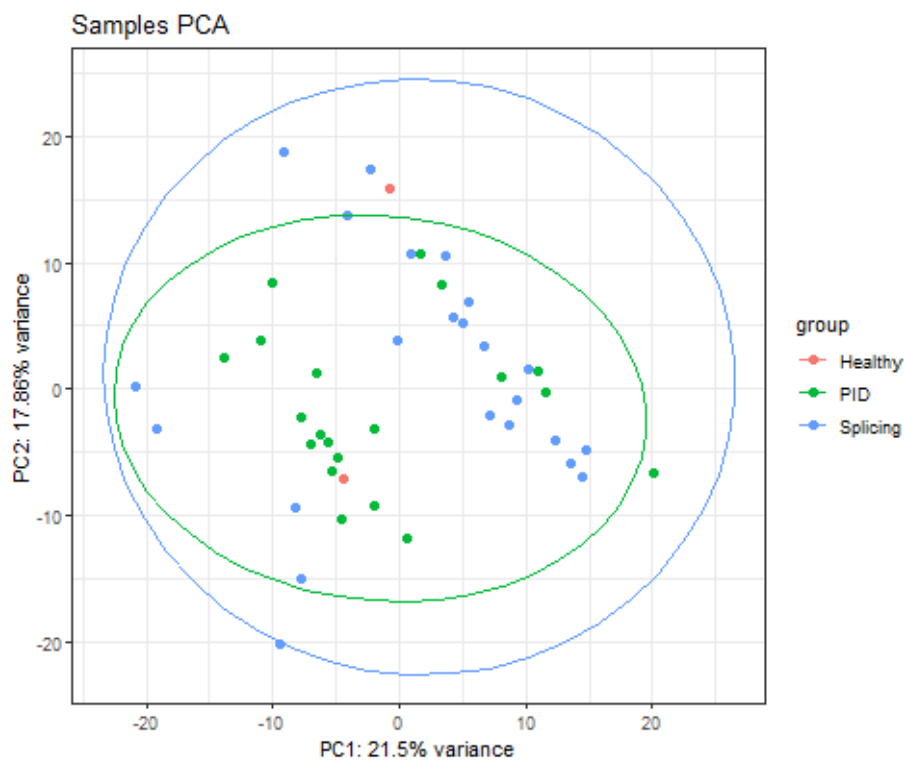
**Figure 3-16 2D PCAplot of PID, Splicing and Disease cohorts.**

Red (group 1) represents healthy control samples, green represents PID samples, red represents 'Splicing and Disease' samples, and aqua represents the GTEx samples. Separation around PC2 (Y axis) was determined to be a result of expression of sex-specific genes. 95% confidence intervals represented by ellipses.



**Figure 3-17 Scree plot from PCA of PID, Splicing and Disease cohorts.**

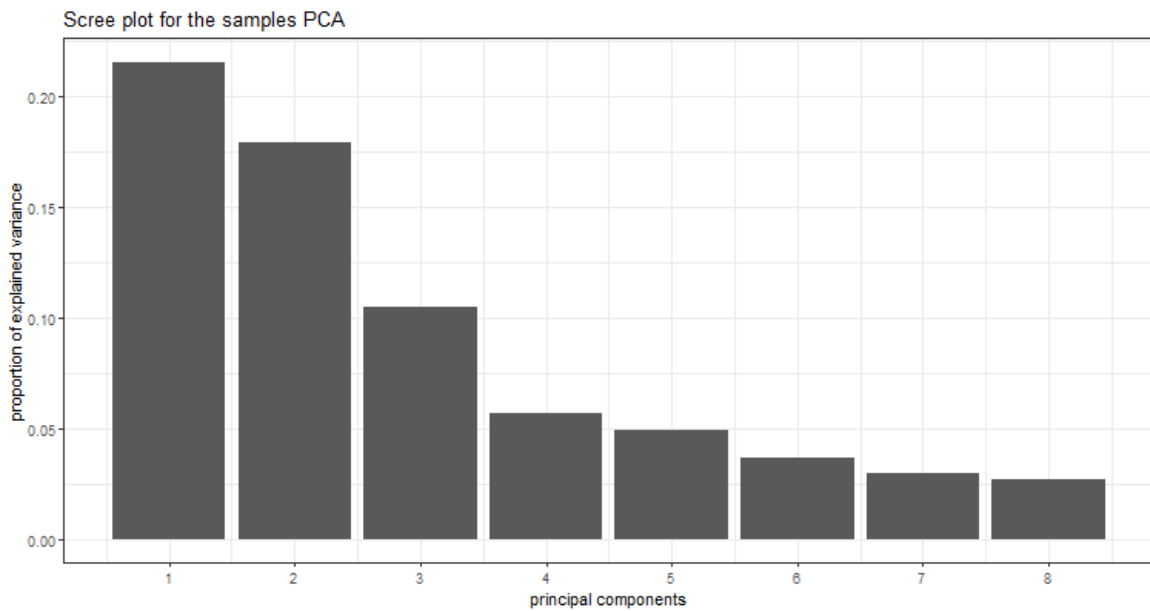
Each column represents the total proportion of variance which is explained by the principal component on the x axis. Principle component 1 is responsible for around 51% of the total variance. Principle component 2, around 23%



**Figure 3-18 2D PCAplot of PID, Splicing and Disease cohorts after ComBat-seq batch correction.**

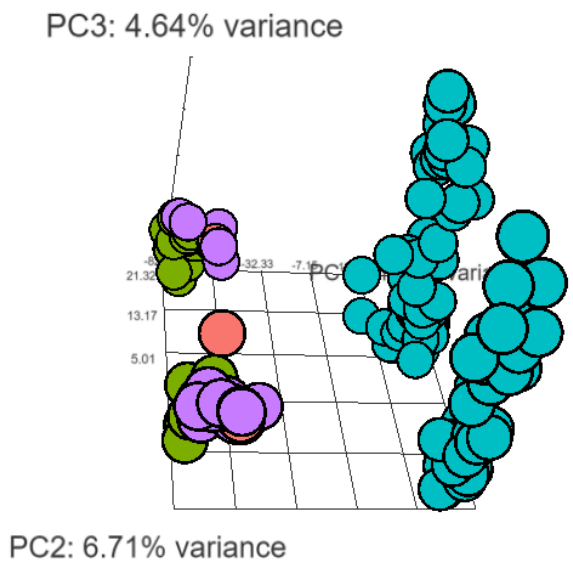
Red (group 1) represents healthy control samples, green represents PID samples, blue represents 'Splicing and Disease' samples. 95% confidence intervals represented by ellipses.





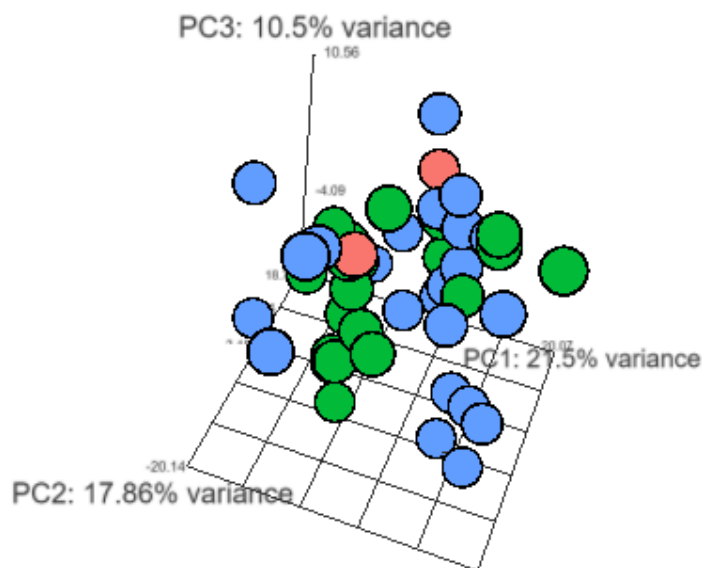
**Figure 3-19 Scree plot from PCA of PID, Splicing and Disease cohorts after ComBat-seq batch correction.**

Each column represents the total proportion of variance which is explained by the principal component on the x axis. Principle component 1 is responsible for around 21% of the total variance, principal component 2, around 17%



**Figure 3-20 3D PCAplot of all GTEx PID, Splicing and Disease cohorts.**

GTEx data in teal, healthy controls in red, Splicing and Disease in green, PID data in red. Divides within group are a result of sex differences in the transcriptome. The GTEx data here is obviously distinct in transcriptome from the other samples.

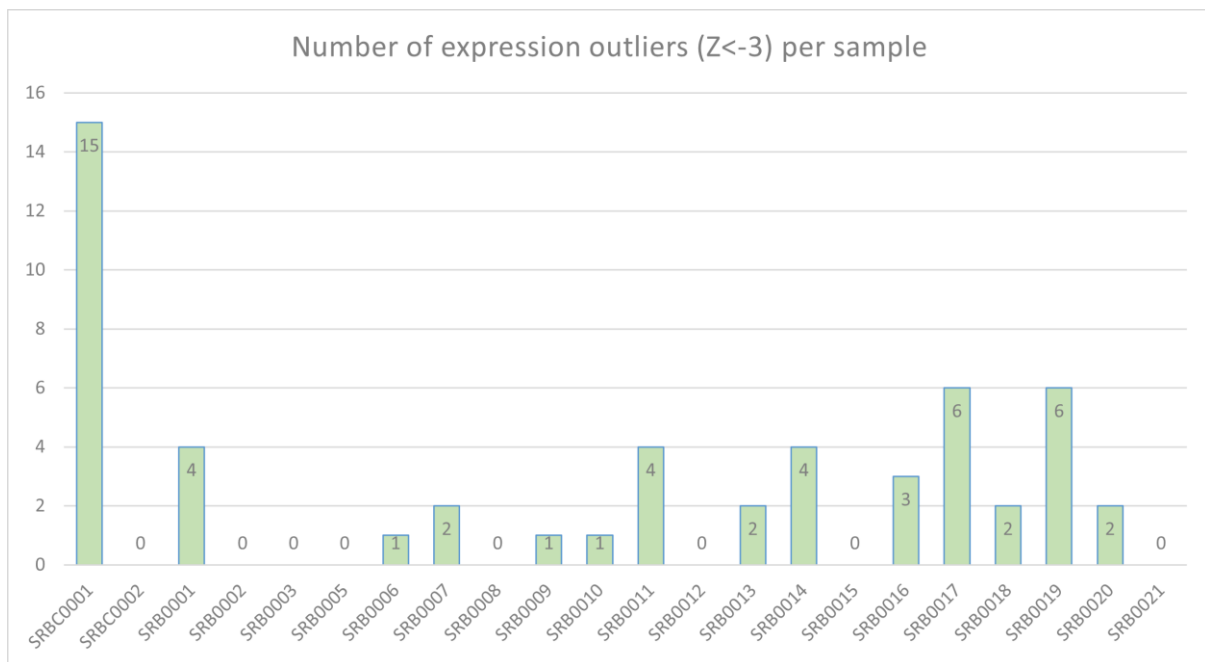


**Figure 3-21 3D PCAplot of PID, Splicing and Disease cohorts after ComBat-seq batch correction and GTEx data removed.**

Healthy controls in red, Splicing and Disease in green, PID data in blue. The GTEx data here is obviously distinct in transcriptome from the other samples. Data shows no clear distinction in PCA three-dimensional clustering.

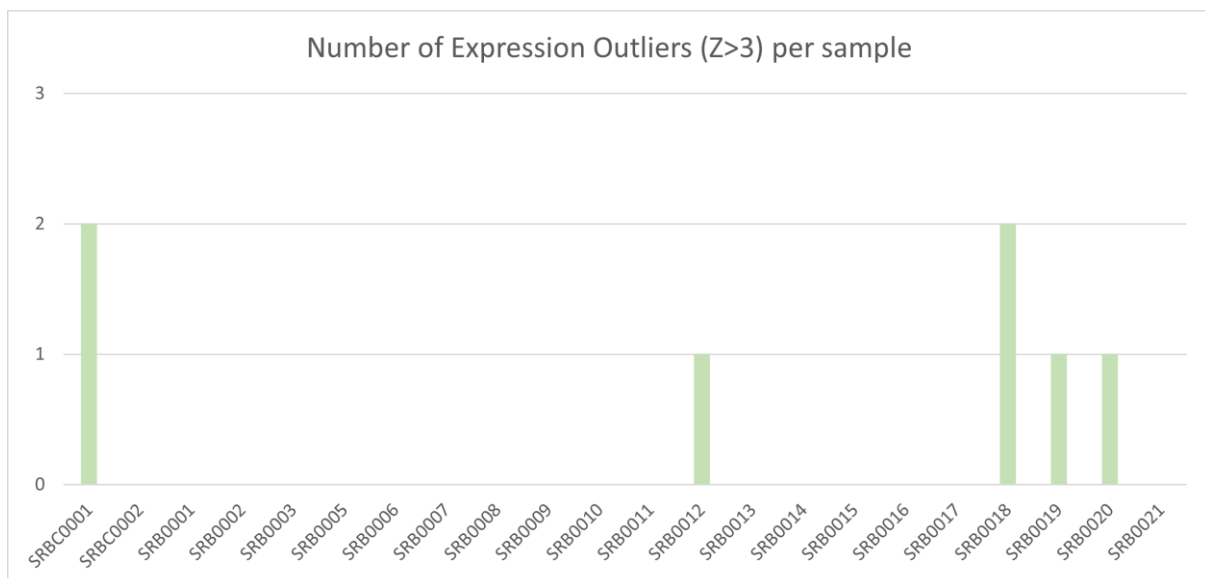
### 3.5.2 Z-score results after ComBat-seq Batch correction and TPM calculation.

Using batch corrected samples with Splicing and Disease samples as controls, with GTEx data excluded, the number of Z-scores remained manageable for downstream analysis and interpretation in most cases, and also for the first time showed expression outliers in both directions (i.e., over-expression and under-expression). The presence of outliers in the control sample SRBC001 can also be observed, however. Two patients with a known deleterious variation in NFKB1/NKKB1 were present in the cohort. After batch correction and TPM calculation, one of these patients' results was above threshold, and another was approaching threshold for Z-score outlier ( $Z=+/-3$ ). None of the other known diagnosis were captured using this technique. Very large Z-scores were obtained for SRB0017 for the IGKC and IGHM genes, both present on B cells, which has some synergy with the patient's clinical phenotype, (Panhypogammaglobulinaemia, Recurrent Bacterial Infection, Absent B cells). It is not possible to discern a gene of interest from this data, as the lack of RNA resulting from absent B cell expression may be the effect of a variant elsewhere in the genome as opposed to the specific genes with reduced expression. Other findings have been examined with the responsible clinician (Professor A. Williams) and the evidence to support a clinical diagnosis is too low in this case. As such further assessment using alternative methods is recommended.



**Figure 3-22 PID Z-scores for under expressed Genes**

This graph shows a number of samples with moderate amount (3-6) of under expression outliers, some with a low number (0-2) and one control sample with a high number (15) of expression outliers.



**Figure 3-23 PID Z-scores for Overexpressed Genes**

This graph demonstrates very few samples having overexpression outliers. Only four samples from the PID cohort have overexpression outliers and a single control sample also have two examples.

**Table 3-3 - TPM Z-score outliers in gene expression**

Sample	Genes with TPM based expression outliers	Z-scores for genes	Diagnosis
SRB0001	G6PC3	-3.3506	No Diagnosis in GECIP
	MVK	-3.22697	
	TRAF3IP2	-4.7455	
	BRIP1	3.36824	
SRB0002	0		No information
SRB0003	0		No Diagnosis in GECIP
SRB0004	Degraded sample		No information
SRB0005	0		CARD 11 A-C @2987250
SRB0006	IL2RG	-3.40762	STAT1
SRB0007	IL10	-3.61385	STAT1
	TP53	-5.17661	
SRB0008	0		STAT1
SRB0009	NFKB1	-3.05206	NKKB1 A>AT @102582929
SRB0010	G6PC3	-3.16447	NKKB1 A>AT @102582930
	<b>NFKB1*</b>	<b>-2.6*</b>	
SRB550011	RFX5	-3.70562	ILRG;CXorf65;FOXO4
	FAAP24	-4.02424	
	FASLG	-3.07656	
	THBD	-3.12753	
SRB0012	CD3E	3.022289	ILRG;CXorf65;FOXO4
SRB0013	NFE2L2	-3.12145	No Diagnosis in GECIP
	C8B	-3.00155	
SRB0014	BLNK	-3.61168	No information
	CD19	-3.3991	
SRB0015	0		No Diagnosis in GECIP
	CD3G	-3.03867	
	CD81	-3.2711	
	CD70	-3.2516	
SRB0017	CD40	-3.30509	No information
	BLNK	-3.31889	
	CD19	-3.74433	
	CD79A	-3.87160	
	IGHM	-6.53492	
	IGKC	-6.31391	

Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene  
expression

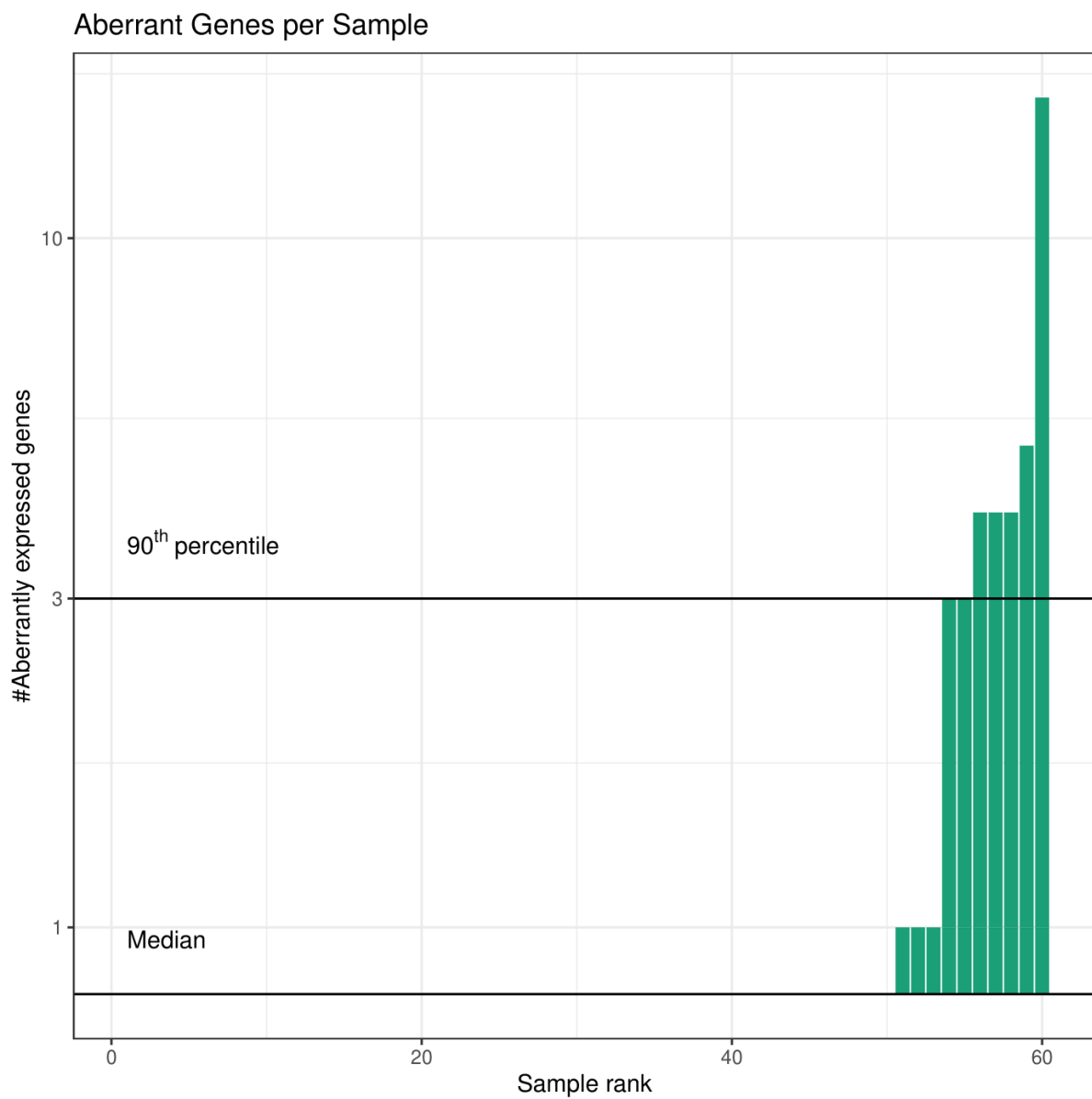
SRB0018	PLEKHM1	-3.49300	No information
	C1QB	-3.28583	
	LYST	3.041579	
	TNFRSF9	3.117149	
SRB0019	HELLS	-3.21443	No information
	TNFRSF13C	-4.13524	
	TOP2B	-3.10538	
	IL18BP	-4.07737	
	CD46	-3.39481	
	DKC1	-3.34669	
	NLRP3	3.19988	
SRB0020	XIAP	-3.08174	No information
	ACD	-3.13444	
	TNFAIP3	3.00717	
SRB0021	0		No information

Table shows patient ID's and the detected Z-score outliers calculated from TPM data.  
\*=sub threshold findings. Clinical diagnosis is also included for reference.

### 3.5.3 OTRIDER results

The OTRIDER program was first validated by downloading the test data used in the OTRIDER manual (352), to determine utility and ensure proper functioning. The results of the test data set were slightly incongruent with those from the manual. P-values deviated in some cases by approximately  $1 \times 10^{-13}$  and  $4 \times 10^{-13}$ . In discussions with the developer, it was clarified that these very small deviations were due to using the program on different operating systems and CPU cores and to be expected to some degree.

OTRIDER observed that several samples had multiple outliers (SOT104=3, SRB0017=4, SOT102=4, SOT117=4, SRB0006=5, SOT120=16), and SOT120 had over 300% as many outliers as the previous sample (Figure 3-24). As such this sample was excluded from the model to prevent statistical bias.



**Figure 3-24 OUTRIDER: Detection of outlier samples based on number of outlier genes.**

Figure shows the sample rank derived from the number of outliers in the sample. SOT120 was the highest ranked sample with 16 outliers, over 300% more than the previous sample.

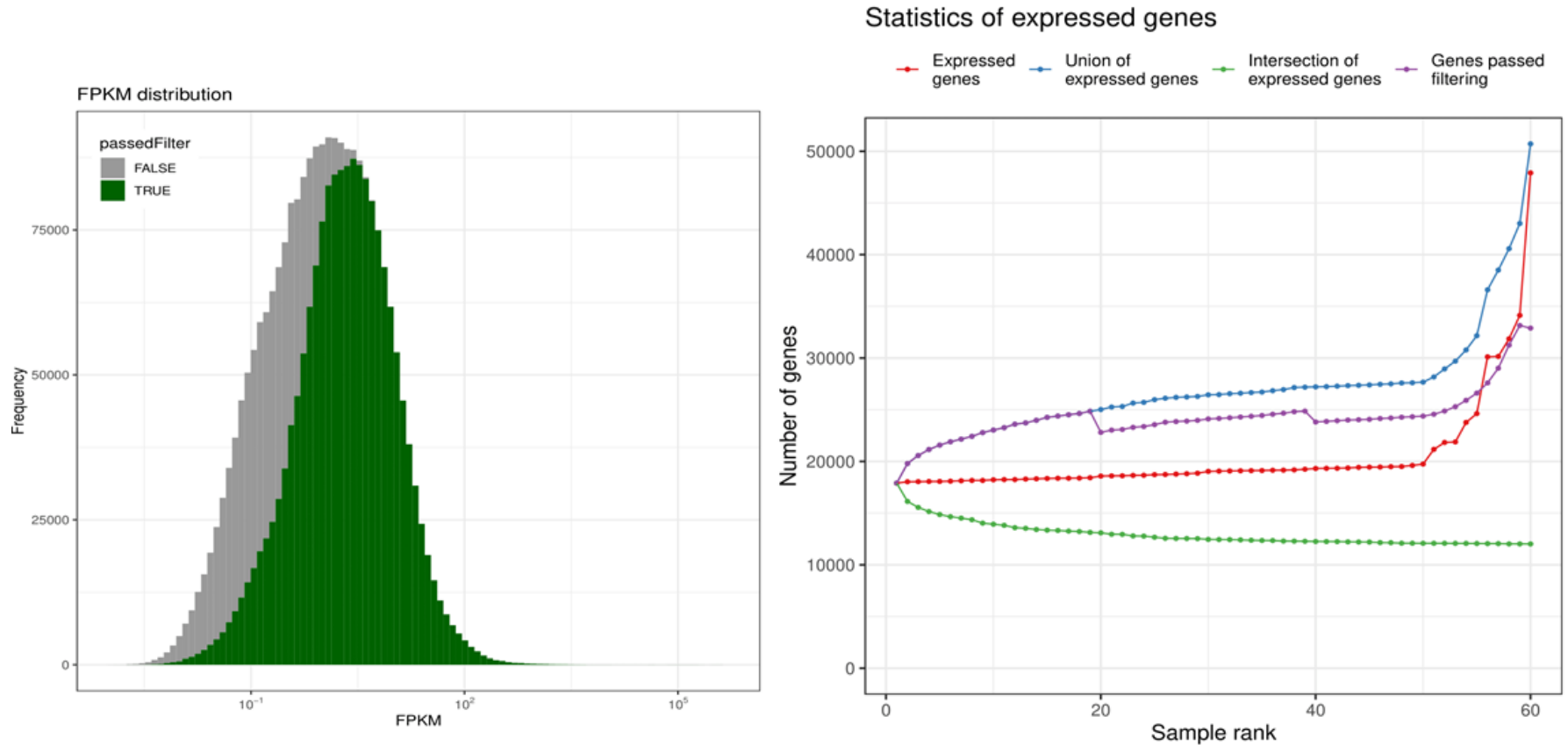
### 3.5.4 Pre-processing using OTRIDER.

OTRIDER uses a neural network for normalisations. Combining two batch correction methods may have unintended consequences and therefore the continued use of the ComBat-seq batch-corrected reads were not appropriate and raw reads from STAR aligner program were used for OTRIDER.

OTRIDER's stringent filtering process removed 25643 genes or 43.6% of the total genes. A graph representing the filtering process and the statistics derived from gene expression values was produced (Figure 3-25). The graphs show a generally linear and regular increase in number of genes until the last 10 samples which seem to have improved overall gene expression detection, likely due to some unseen batch effect. Also visible is the union of expressed genes, which shows the cumulative total of expressed genes with each additional ranked sample. The variance between these two trends is indicative of variance in specific genes detected. The number of genes which pass filtering as each additional sample is a clear increasing trend, is representative of total number of mean expressions being above threshold increasing, a feature of the ranking system. This data has two 'steps', indicative of drops in number of genes passing threshold, suggesting that some batches, whilst having increased total detection, actually have lower expression of groups of genes, thus reducing the mean below expression threshold for inclusion.

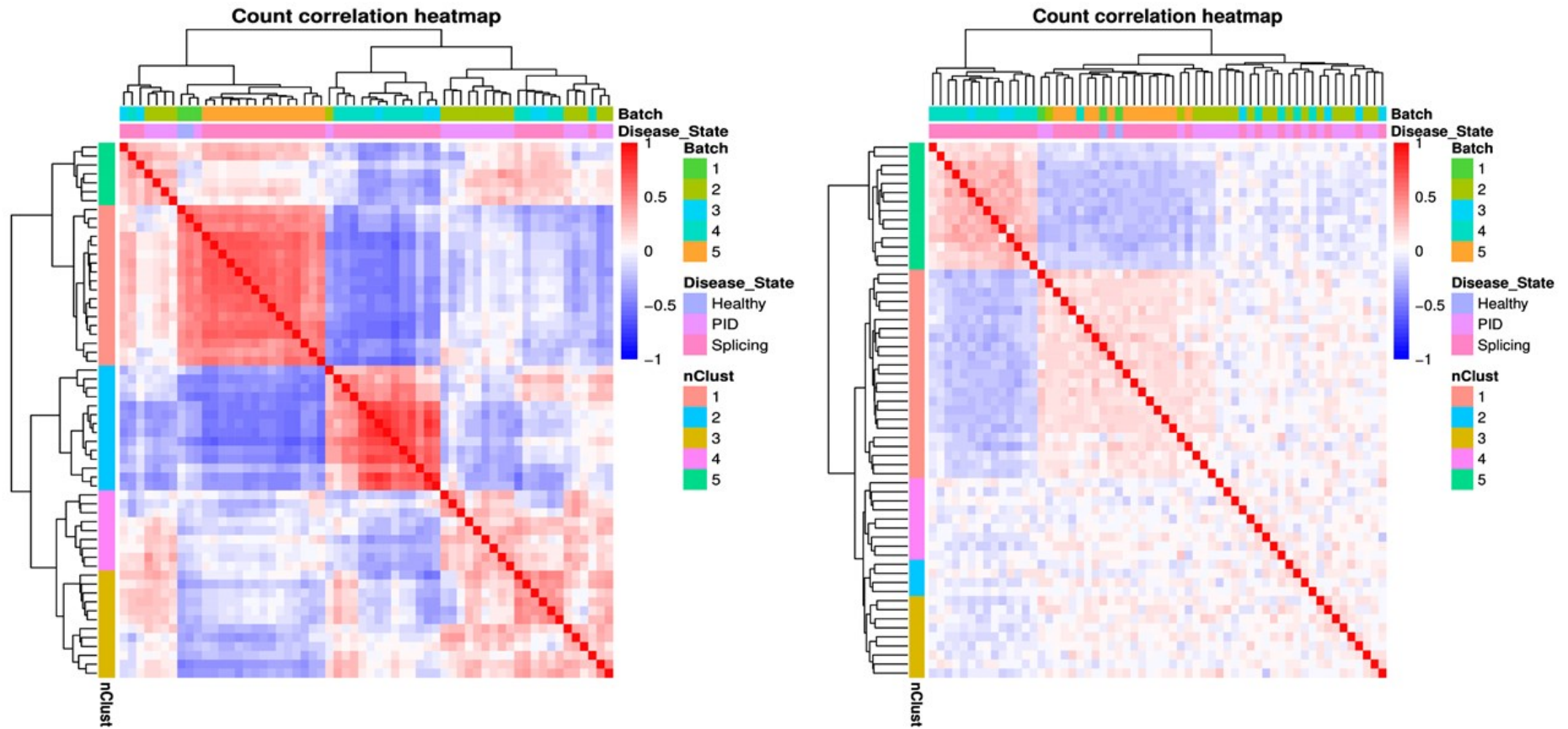
To observe the effect of seen variables including batch effect, OTRIDER produces a count correlation heatmap. Clusters are apparent for batches and disease states and the overall differences in gene expression are apparent from the large amount of red-blue hue in the figure overall. By comparing the heatmaps before and after the normalisation process (Figure 3-26) it can be seen that some of the batches have dissolved and the impact of the new/remaining batches is very much reduced, although not eliminated. Whilst the separation between the two cohorts is gradual there does still appear to be a gradient. The sequencing batches (1-5) are also reduced in this manner. Overall, there is less hue, indicative of decreased expression changes overall.





**Figure 3-25 OUTRIDER FPKM data in a mixed Primary Immunodeficiency and Splicing and Disease Cohort**

**FPKM distribution** (left) represent gene-sample combinations. Those which will be removed in subsequent steps are represented by the grey shaded area. **Statistics of expressed genes** (right) represents the samples, ranked by number total number of detected genes expressed in the samples. The red points number of expressed genes in each sample. The blue points represent the cumulative total of expressed genes with each additional ranked sample. The purple points indicate the number of genes which pass filtering as each additional sample is added.



**Figure 3-26 Heatmap of Primary Immunodeficiency and Splicing and Disease Samples before and after batch correction.**

Figure shows the sample relationships based on count correlation and hierarchical clustering of the samples via dendrogram on the left Y axis. Differences in expression are indicated by the blue to red colouring, with solid red diagonal line indicating samples are the same.

### 3.5.5 Results for OUTRIDER analysis of outlier gene expression

SOT120 was removed from the analysis due to it being an outlier sample (Figure 3-24). A total of 40 gene expression outliers were present from the total dataset of 59 samples (appendix A.10). 28 of these were discovered in the 39 Splicing and Disease cohort samples. 12 were found in the PID cohort (Table 3-5). *OCLNP1*, an unprocessed pseudogene had aberrant expression detected in both direction across 11 of the Splicing and Disease cohort control samples, and 3 samples in the PID cohort, SRB0006, SRB0011, and SRB0013. Notably expression was completely ablated in SRB006. However, the “aberrant by gene” column indicates *OCLNP1* is an outlier in 14 samples. A look at the complete results dataset (appendix A.10) shows us that in total, three samples have completely ablated or very low expression of *OCLNP1*. Further exploring this by plotting the raw counts vs the expected counts, shows many samples had undetectable expression of the pseudogene, whilst others seemed to follow a clear trend (Figure 3-27), This followed no discernible phenotype or characteristic, from age, sex, disease state or batch. Therefore, this result was excluded. Comparing the OUTRIDER results for the Splicing and Disease control group cohort and their known molecular diagnosis (Table 3-4), demonstrated that no diagnosis was resolved using this algorithm. This removes credibility from any results obtained from the OUTRIDER tool, and additional follow-up using alternative methods is likely to be necessary for validating findings.

**Table 3-4 Comparing OUTRIDER results from the Splicing and Disease control cohort with known molecular diagnosis.**

<b>OUTRIDER RESULT</b>	<b>SAMPLE ID</b>	<b>P-VALUE</b>	<b>MOLECULAR DIAGNOSIS</b>
<b>OCLNP1</b>	SOT058	2.02E-20	MED13L: c.2570-4_2574del
<b>OCLNP1</b>	SOT019	1.11E-17	No diagnosis
<b>OCLNP1</b>	SOT017	8.19E-14	NF1 c.7832A>G, p.Asp2611Gly
<b>OCLNP1</b>	SOT045	1.78E-13	BRCA1 c.4676-8C>G
<b>OCLNP1</b>	SOT049	5.61E-10	BRCA2 c. 9502-13 C>G
<b>OCLNP1</b>	SOT018	7.29E-10	NF1 c.5489C>G, p.Pro1830Arg
<b>OCLNP1</b>	SOT020	1.21E-09	NF1 c.4122G>T, p.Gln1374His
<b>NIPA2</b>	SOT117	2.07E-09	Clinical diagnosis MEN1 but not confirmed molecularly
<b>SFT2D1</b>	SOT102	2.18E-09	UBF4 c.8488+3A>G Class 3 VUS
<b>KIDINS220</b>	SOT018	8.1E-09	NF1 c.5489C>G, p.Pro1830Arg
<b>PALM</b>	SOT104	1.33E-08	TMEM127 c.411T>A, AIP c.317G>A_Arg106His and WT1 c.871A>T p.(Ser291Cys)
<b>PRDM16</b>	SOT104	1.09E-08	TMEM127 c.411T>A, AIP c.317G>A_Arg106His and WT1 c.871A>T p.(Ser291Cys)
<b>OCLNP1</b>	SOT069	7.62E-09	No diagnosis
<b>OCLNP1</b>	SOT152	7.9E-09	polyposis but NAD from 100KG.
<b>USP9Y</b>	SOT033	1.88E-08	SMAD3 c.802C>T, p. (Arg268Cys)
<b>OCLNP1</b>	SOT130	2.52E-	No diagnosis - negative 100KGP

		08	
<b>ESYT2</b>	SOT019	6.7E-08	No diagnosis
<b>PRDM13</b>	SOT104	1.26E-07	TMEM127 c.411T>A, AIP c.317G>A_Arg106His and WT1 c.871A>T p.(Ser291Cys)
<b>CYFIP1</b>	SOT117	1.08E-07	Clinical diagnosis MEN1 but not confirmed molecularly
<b>TUBGCP5</b>	SOT117	1.69E-07	Clinical diagnosis MEN1 but not confirmed molecularly
<b>AGBL5</b>	SOT102	1.24E-07	UBF4 c.8488+3A>G Class 3 VUS
<b>SHROOM1</b>	SOT102	1.75E-07	UBF4 c.8488+3A>G Class 3 VUS
<b>SLC39A11</b>	SOT018	2.31E-07	NF1 c.5489C>G, p.Pro1830Arg
<b>OCLNP1</b>	SOT038	8.22E-08	BRCA2 c.1480 G>A heterozygote
<b>NIPA1</b>	SOT117	3.59E-07	Clinical diagnosis MEN1 but not confirmed molecularly
<b>SNX17</b>	SOT010	9.21E-08	NF1 c.1158A>C, p.(=), c.1168_1179del12, p.Asn390_His393del
<b>SCTR</b>	SOT140	1.09E-07	No diagnosis - negative 100KGP
<b>GPATCH2</b>	SOT038	2.51E-07	BRCA2 c.1480 G>A heterozygote

OUTRIDER results column represents expression outliers. Known molecular diagnosis are also included for comparison. In the Splicing and Disease cohort, none of the known molecular diagnosis could be resolved using the OUTRIDER program.

### 3.5.5.1 Gene expression outlier results for patient SRB0017

SRB0017 has a collection of genes which are aberrantly expressed, with large Z-scores, large log2 fold changes, and completely ablated *IGKC* expression. *IGKC* is known to be a causative PID gene in both GeCIP and IUIS gene panels, specifically for Immunoglobulin Kappa light chain deficiency. *PAX5*

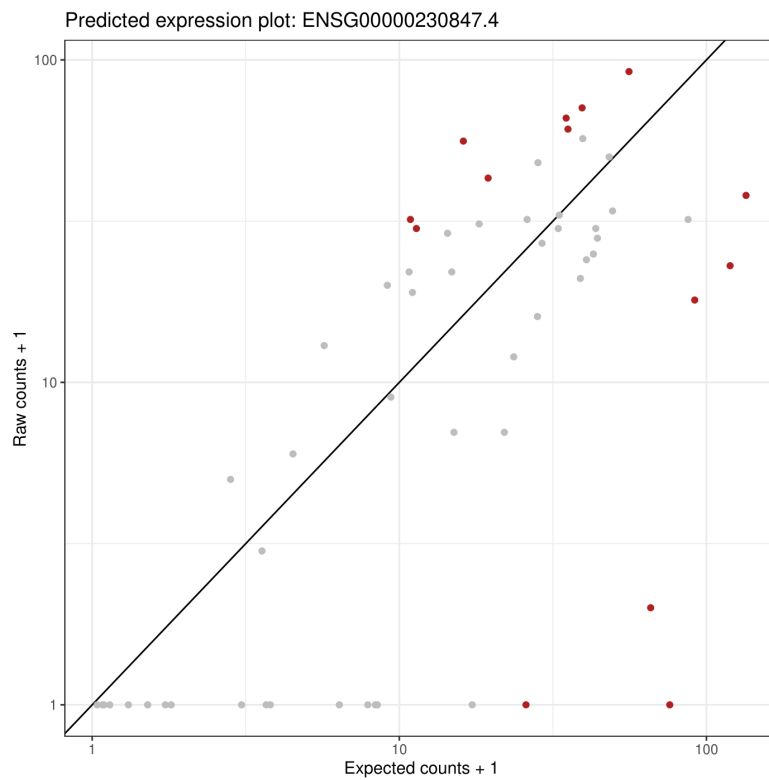
did not appear on the IUIS or GECIP panels but was present in the T-cell panel of the HTG-Edgeseq. SRB0017 Also has reduced expression of *IGHM*, *CD22*, *MS4A1*, *FCRL2*.

### **3.5.5.2 Gene expression outlier results for patient SRB0012**

*ACOT9* was an extreme outlier in SRB0012, demonstrating less than 10% the mean expression of the group. *ACOT9*, a mitochondrial acyl-CoA thioesterase, did not appear on any PID gene panels, and is only linked to syndromic X-linked intellectual disability turner type (368), and Mental Retardation, X-Linked, Syndromic, Claes-Jensen Type (369). There was little evidence linking the gene to immune disorders.

### **3.5.5.3 Gene expression outlier results for patient SRB0006**

SRB0006 also harboured aberrant under-expression of *RRP8* (*Chr* 11p15.4) and *RRP1B* (*Chr* 21q22.3), both of which are ribosomal processing proteins. Expression of both was lower than expected (Figure 3-28, Figure 3-29) these genes were not present in the IUIS, GeCIP or HTG gene panels (Table 3-6).



**Figure 3-27 Expected vs actual expression of OCLNP1.**

Points in grey are non-significant, points in red are significant  $P < 0.05$ . All counts are given +1 to mitigate effects of zero values on statistical models.

Table 3-5 OUTRIDER results for Primary Immunodeficiency Samples

	sampleID	p-value	padjust	zScore	l2fc	Raw counts	Norm counts	Mean Corrected	theta	Aberrant By Sample	Aberrant By Gene	Padj rank	
<b>1</b>	OCLNP1	SRB0006	1.86E-25	6.8E-20	-3.27	-6.25	0	0	13.41	1539.82	3	14	1
<b>3</b>	OCLNP1	SRB0011	2.53E-19	9.25E-14	1.44	1.79	55	41.56	13.41	1539.82	1	14	1
<b>7</b>	ACOT9	SRB0012	1E-11	3.66E-06	-7.09	-4.09	33	62.34	1071.99	13.63	1	1	1
<b>9</b>	OCLNP1	SRB0013	5.74E-10	0.00021	1.3	1.56	31	35.99	13.41	1539.82	1	14	1
<b>11</b>	PAX5	SRB0017	8.41E-10	0.000307	-6.51	-3	5	52.29	475.58	49.53	6	1	1
<b>13</b>	IGHM	SRB0017	2.84E-09	0.000519	-6.53	-3.67	19	227.34	3109.13	12.89	6	1	2
<b>21</b>	RRP8	SRB0006	3.29E-08	0.006004	-5.93	-1.95	141	174.26	672.04	27.65	3	1	2
<b>23</b>	RRP1B	SRB0006	5.85E-08	0.007118	-5.77	-1.47	241	240.76	647.03	42.98	3	1	3
<b>24</b>	CD22	SRB0017	7.47E-08	0.009097	-6.12	-2.26	11	124.93	639.68	45.43	6	1	3
<b>27</b>	MS4A1	SRB0017	1.44E-07	0.013149	-5.92	-2.76	26	252.66	1735.64	16.04	6	1	4
<b>38</b>	FCRL2	SRB0017	6.02E-07	0.043928	-5.79	-2.38	12	103.59	567.25	25.66	6	1	5
<b>40</b>	IGKC	SRB0017	8.15E-07	0.049605	-6.31	-5.98	0	0	1226.38	6.08	6	1	6

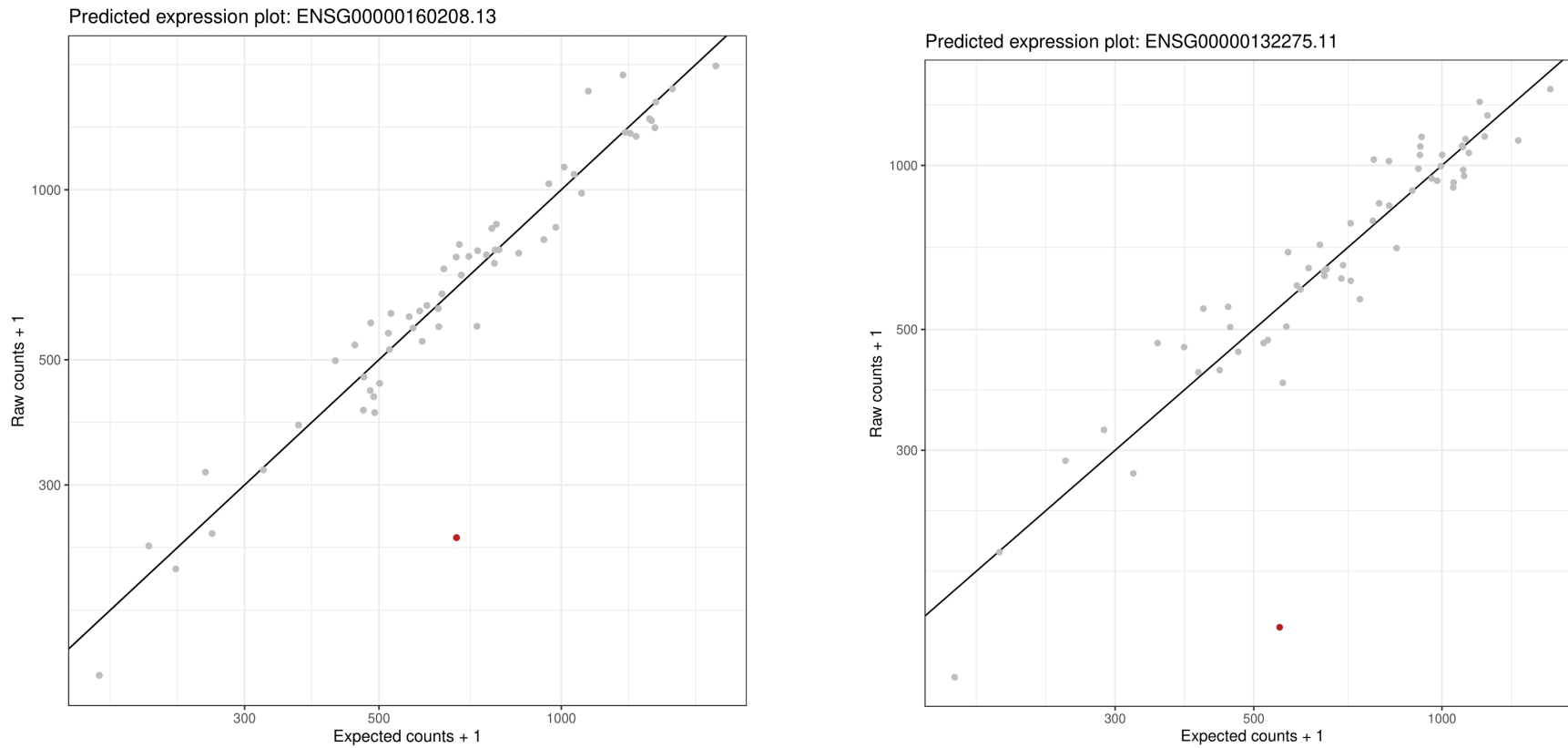
Table shows OUTRIDER results from the PID samples. Adjusted P-value (padjust) cut off was 0.05, l2fc represents the log to base2 fold change, Norm counts indicates the counts for that sample after the de-noising auto-encoder (Ae) removes the outliers. Mean corrected represents the mean value for that gene across the dataset after Ae normalisation. Theta value represents the distribution. Aberrant by sample indicates how many aberrant genes the sample in question has, aberrant by gene represents the number of times a specific gene is seen to be aberrantly expressed. Padj rank is a significance associated ranking system for the detected outliers.



**Table 3-6 Genes from OUTRIDER results cross referenced with pre-identified gene panels.**

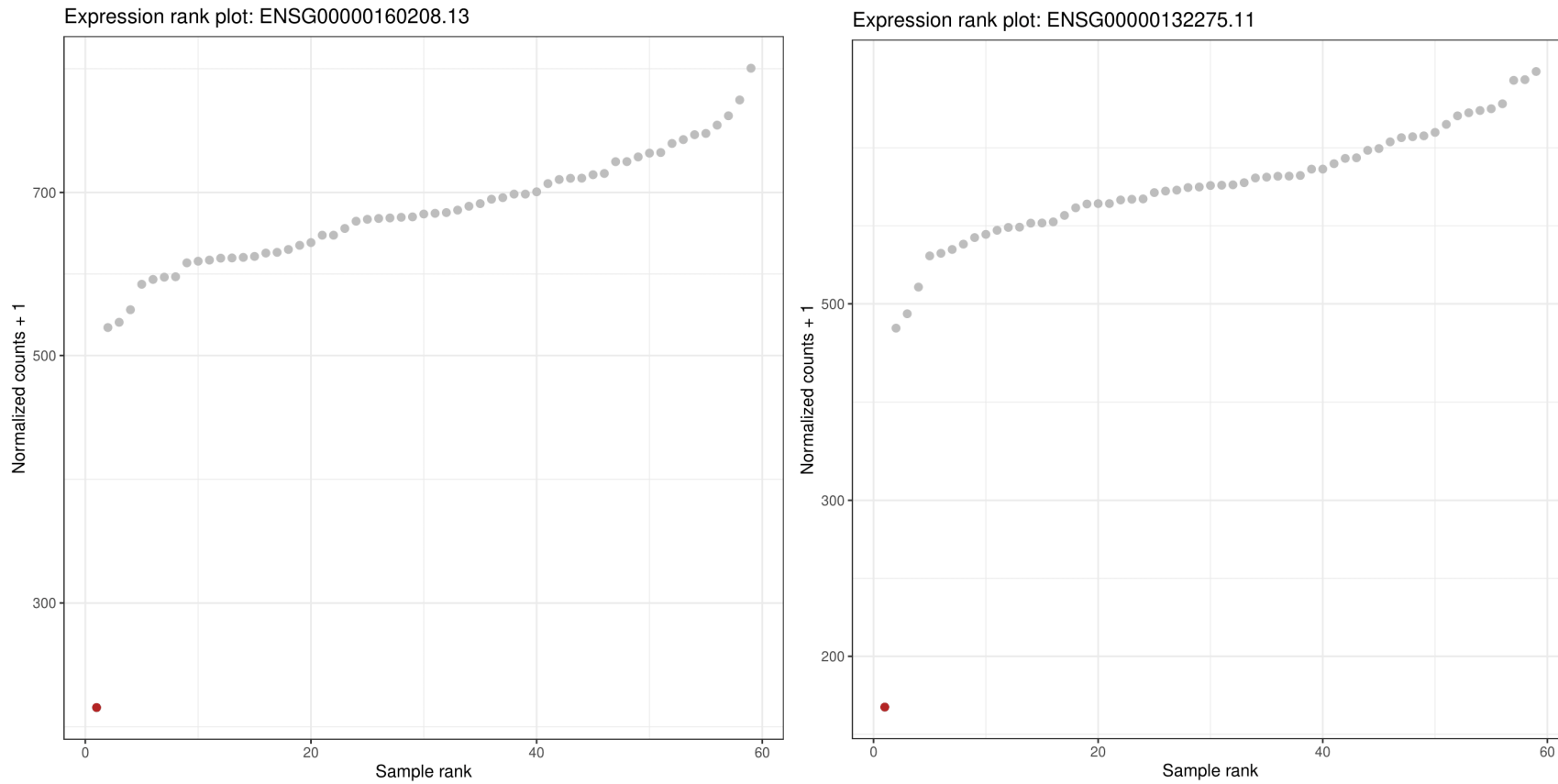
	<b>GECIP</b>	<b>IUIS</b>	<b>HTG -T-cell panel</b>	<b>HTG Immunooncology panel</b>
<b>OCLNP1</b>	--	--	--	--
<b>ACOT9</b>	--	--	--	--
<b>PAX5</b>	--	--	--	✓ <b>(B-cell function subset)</b>
<b>IGHM</b>	✓	--	--	--
<b>RRP8</b>	--	--	--	--
<b>RRP1B</b>	--	--	--	--
<b>CD22</b>	--	--	--	--
<b>MS4A1</b>	✓	--	--	--
<b>FCRL2</b>	--	--	--	--
<b>IGKC</b>	✓	✓	--	--

Table shows the presence of the OUTRIDER results on existing panels of genes, including the GECIP 100,000 genomes PID panel, the IUIS PID panel, the HTG T-cell panel, and the HTG Immunooncology panel.



**Figure 3-28 Predicted Expression vs Actual expression for RRP1B and RRP8**

Predicted expression of RRP1B (ENSG00000160208.13) and RRP8 (ENSG000000132275.11) respectively. Expected counts and raw counts both have pseudo count of 1 added to account for any genes lacking expression completely. SRB0006 is highlighted in red.



**Figure 3-29 Rank vs normalised counts for RRP1B and RRP8**

Rank based on gene expression versus the value of expression with plus 1 pseudo-count of RRP1B (ENSG00000160208.13) and RRP8 (ENSG00000132275.11) respectively. SRB0006 is highlighted in red.

### 3.6 Summary Table

**Table 3-7 Summary table of all results from gene expression outlier detection methods**

Sample	FPKM Z-Score	TPM Z-score	OUTRIDER
SRB0001	SEMA3E, CFTR, C7	G6PC3, TRAF3IP2, MVK, BRIP1	
SRB0002	FOXN1, PSENN, ALPI, IL36RN, SAMD9L		
SRB0003			
SRB0004			
SRB0005	CCBE1, IL2RA, CEBPE		
SRB0006	RNF31, NRKB2, TNFSF12, C9,	IL2RG	OCLNP1, RRP8, RRP1B
SRB0007		IL10, TP53	
SRB0008	GIN51, IL23R, CFH, CFHR5,		
SRB0009	IL21, CFI, FANCI	NFKB1	OCLNP1
SRB0010		G6PC3, <b>NFKB1*</b>	
SRB0011		RFX5, FAPP24, FASLG, THBD	OCLNP1,
SRB0012		CD3E	ACOT9
SRB0013	RAG1, FAT4, IL17F, C4A, C4B, CFHR2, CFHR3, BRIP1	NFE2L2, C8B	
SRB0014	TAP1, NCF1, APOL1, IL17RA, SERPING1	BLNK, CD19, CD79A, POLR3F,	
SRB0015	FERMT1, CFHR1,		

SRB0016	C8A,	CD3G, CD81, CDE70	
SRB0017	C6	CD40, BLNK, CD19, CD79A, IGHM, IGKC	PAX5, IGHM, CD22, MS4A1, FCRL2, IGKC
SRB0018	CFI	PLEKHM1, C1QB, LYST, TNFRSF9,	
SRB0019	IL12B, THBD	HELLS, TNFRSF13C, TOP2B, IL18BP, CD46, DKC1, NLRP3	
SRB0020	CD40, TTC7A, CARD9, C1QA, C1QB FCN3,	XIAP, ACD, TNFAIP3	
SRB0021	XRCC2		

### 3.7 Discussion

The aims of this chapter were to conduct quality control of the data, use exploratory data analysis to identify appropriate control sets, and finally generate a bioinformatic pipeline for processing and interrogation of the transcriptome of PID patients using gene expression. This was to be conducted with the aim of enhancing diagnostic ability and variant filtering. We aimed to utilise multiple tissues to find novel signals which inform diagnosis.

The project progress in the primary immunodeficiency investigation was hindered by several factors. In the first instance that assumption that an externally produced dataset could serve as a control group was incorrect, and this proved to be a significant obstacle. Even utilising purpose-built tools for batch correction could not mitigate the large differences which resulted from different cohort capture (deceased in the case of GTEx), different sample acquisition methods and sequencing methods. Based on PCA, the differences in expression between the GTEx group and other samples were larger even than the differences between the most different samples within the non-GTEx group. Whilst some degree of batch correction is possible using tools like ComBat-seq, the greater these differences are at the outset, the more correction has to be done, and therefore the more

blunted any biological outlier signals become. It was therefore decided that the results from correction of such large differences would render data unreliable afterwards. Indeed, even after the batch correction has taken place within samples which are more analogous, there is still concern that some outliers will become sub-threshold after algorithmic batch correction.

It may be that the batch effect is made worse by using whole blood, which is subject to a number of other pressures, and that by reducing the number of cell types and thus the effects of other variables, signals from immune disorders may be able to be resolved using RNAseq approaches with standard batch-correction tools. In addition, alternative methods of quantification, which compare RNA abundance to housekeeping genes, or relative amounts of immune gene expression, might make RNAseq quantitative approaches more viable.

It was also discovered that methods used by other groups, such as the Z-score methods to detect outliers, were now being replaced by alternatives which are designed for outlier detection in such datasets (352, 370). Using differential expression tools such as EdgeR, which are designed for comparing groups of samples with each other, as opposed to outlier detection, were also not suitable. The identification of appropriate methods, development of a modern and suitable bioinformatic pipeline, learning and training around implementing such a pipeline was not a trivial task, and significant delays occurred.

The most suitable tool for outlier detection of gene expression (OUTRIDER) required a large (>50) sample number for its statistical analysis, which was not met by the PID cohort alone. This meant a second dataset needed to be sourced, processed and batch-corrected. In addition to these challenges, the OUTRIDER tool itself proved challenging to set up and validate leading to further resources being spent on training and familiarisation.

Once operational the OUTRIDER tool did demonstrate comprehensive analysis features, and quickly identified the loss of expression of IGKC, and nominal expression of IGHM in patient SRB0017. These are both predominantly expressed in B-cells. However, a genome sequence for the patient was not present in this instance and so no causal variant could be identified. Clinical consultation with Professor A. Williams was sought, and it was advised that patient phenotype was not completely typical for IGKC knockout, but that this lack of expression could have caused the observed phenotype. Confirmation of this finding using targeted PCR and whole exome sequencing is necessary for diagnosis confirmation, which has since been scheduled by the clinician.

\*\*Clinical information for this patient has since been obtained and it is noted the patients has no B-cells and as such the expression loss can be explained without IGKC involvement.

IGKC appeared first in the 2017 IUIS table of genes which can cause primary immunodeficiencies, although it was highlighted as often asymptomatic. At the project outset IGKC was not a gene which appeared on the GECIP gene panel, although it has since been added. This is an important consideration and evidence of the requirement to include hypothesis-free approaches to outlier detection in Mendelian disease molecular diagnostics. Whilst panel-based approaches are extremely useful for first pass diagnostic investigation, whole transcriptome approaches are where many new discoveries are likely to be found and, in previously undiagnosed patients, it is likely to provide a rich reservoir of data which has been overlooked in early clinical investigation.

In the case of SRB007, the effective total loss of expression in this patient was conspicuous, and as such detection should have been possible using less-complex methods. Given either a suitably sized healthy control group or alternative more homogenous tissues (i.e., PBMC/T-cells), this tool may have been able to identify more potential causal genes via expression changes. Since this sample showed complete loss of expression of IGKC, a critical component of immunoglobulin, sequence information was not available to determine any potential causal variants within the gene.

No discernible RNA signal linked to known causative variants in PID positive controls was able to be identified, and no other significant events which could be linked clinically to the symptoms were found. The project was not able to compare multiple tissues types due to technical and logistical challenges in the early stages. Nor was the project able isolate T-cells and stimulate these specifically for diagnosis uplift. This meant that questions around the optimal complexity / sensitivity ratio could not be explored and developing a robust and effective bioinformatic pipeline even at the minimum viable complexity did not achieve a level of success comparable to other examples from the literature.

In addition, the known molecular diagnosis for the 'Splicing and Disease' cohort data was compared with the result from this control data obtained using the OTRIDER program, in an attempt to validate the OTRIDER program. None of the genes known to have causative variants appeared in the OTRIDER results from the Splicing and Disease control data.

It cannot be ruled out that the same approaches applied to alternative tissues, (PBMC's or stimulated T-cells for example) might produce stronger signals in the transcriptome, leading to diagnosis. In a sample which had immune cells only, more reads would be applied to the cells of interest potentially raising the chances of detecting aberrant expression of genes or splicing. Other examples of RNAseq in Mendelian diagnostics have used a much more homogenous samples than whole blood or smaller gene panels due to the simpler system (210). It is not possible however to rule out that improved bioinformatic pipeline design might produce higher rates of diagnostic uplift in whole blood.

### **3.8 Conclusion**

RNA sequencing as a diagnostic tool has had some success and been demonstrated to validate existing molecular diagnosis and inform new diagnosis where previously one has not been present; 25-40% success rate is found in literature. In the current study, one family (SRB10 and SRB11) had the gene of interest NFKB identified in one member with Z-scores above threshold and another just below, and a third patient appeared to have a novel diagnosis, although this has not yet been validated. This represents a diagnostic yield of around 9% although no uplift itself was had as these patients were already found to have molecular diagnosis. The current study has had limited success, and this is most likely due to experimental design, primarily inadequate controls, and use of whole blood as a tissue.

Inherent limitations in this project included significant batch effect and the use of 'unhealthy' controls. The current approach is does not have adequate sensitivity, as patients with a known diagnosis which affects expression levels were not consistently outside of confidence thresholds for Z-scores. Although wider gene panels devised, these were not used due to time constraints around planning and implementations. The method also does not capture non-coding and intronic regions which can affect expression levels. There exists potential for causative variants outside of the panels to be missed, and therefore this hypothesis-based approach is inherently limiting.



Results: Investigation into the transcriptome of patients with primary immunodeficiencies –gene  
expression

## Chapter 4 Results of Splicing Analysis of Primary Immunodeficiency Patients

### 4.1 Introduction

Alternative splicing outlier identification was investigated as a modality for identifying causal variants in the Primary immunodeficiency cohort. Control datasets included the pre-selected, GTEx dataset of 113 samples, the Splicing and Disease cohort, and the two healthy controls included in the sequencing of the PID cohort. Alternative splicing was mapped using STAR aligner first and compared with 'Mendelian RNAseq' tool. This chapter presents the results of validation steps, PID investigation, and follow-up analysis with Interactive Genome Viewer.

### 4.2 Validation of the 'Mendelian RNAseq' splicing detection program using sample SOT58

As a positive control step for the Mendelian RNAseq splice detection tool developed by Dr. Beryl Cummings, a pathogenic event already seen and validated in a previous publication was used to ensure the syntax and data structure was properly operational. RNAseq data obtained from the 'Splicing and Disease' cohort patient SOT58 and was used as a positive control. The patient has previously had the disease causing mutation event identified; a small, 8 nucleotide deletion in the *MED13L* gene which exists on chromosome 12 (NM\_015335.4:c.2570-4\_2574del) and the subsequent use of an alternate 3' splice site (152). This variant has been successfully identified using the proposed 'Mendelian RNAseq' splicing assessment method in another study from our group and so serves as a useful positive control.

Fastq files were processed in the same manner as described in section 2.1.11 and resulting bam files were merged with bam files from the GTEx data set. These all underwent splice event discovery and normalisation. The events found were then filtered specifically for those which only appeared in the sample SOT58 and had a read support greater than 5. This returned 1270 events in a text file. Further filtering was necessary to make this number manageable, so the data was then transferred

to Microsoft Excel, and a further filter was applied which retained only events which had at least one junction already annotated as these were more likely to be “real events” events and not artefacts from the programs algorithm, or a result of inappropriate read splitting.

Inappropriate read splitting is a common complication and occurs when the STAR aligner program incorrectly aligns part of a read to one location on the genome, and the other part to a different location which has a short, similar sequence to the gene in questions. The read has then been split, and indicates splicing has taken place, when in fact it has not. This returned 72 junctions. These were sorted by the highest amount of normalised read support, and only those with a normalised read support value greater than 1 was retained to make the number manageable and increase likelihood of significant events being kept. The known pathogenic variant was identified in the ranked output as number 10 of 15 events which then remained (Table 4-1). This indicated that the tool was sensitive to identify positive controls but lacked specificity to resolve this from other outliers without clinical information. It also demonstrated that multiple filtering steps would need to be applied in an iterative fashion, often ad-hoc as the results were different each time and changes in filtering strategy may have been required based on the number of events found.

**Table 4-1 SOT58 splice analysis outputs ranked by normalised read support.**

<b>Gene</b>	<b>Locus</b>	<b>Anno status</b>	<b>sample support</b>	<b>Normalised read support</b>
<b>NBPF26</b>	chr1:120834546-120836569	One annotated	8:SOT058	8.0:SOT058
<b>MYL12A</b>	chr18:3255838-3277858	One annotated	57:SOT058	57*:SOT058
<b>GSE1</b>	chr16:85634132-86561671	One annotated	6:SOT058	3.0:SOT058
<b>LPAR1</b>	chr9:110973558-111005465	One annotated	6:SOT058	2.0:SOT058
<b>TMEM64</b>	chr8:90684030-90685005	One annotated	12:SOT058	12*:SOT058
<b>AC093668.3</b>	chr7:102541662-102640204	Both annotated	20:SOT058	1.538:SOT058
<b>POLR2J3</b>	chr7:102541662-102640204	Both annotated	20:SOT058	1.538:SOT058
<b>AC093668.2</b>	chr7:102541662-102640204	Both annotated	20:SOT058	1.538:SOT058
<b>AC093668.1</b>	chr7:102541662-102640204	Both annotated	20:SOT058	1.538:SOT058
<b>MED13L</b>	chr12:115997221-116003003	One annotated	41:SOT058	1.281:SOT058
<b>HIRA</b>	chr22:19447545-19447665	Both annotated	20:SOT058	1.0:SOT058
<b>TTN</b>	chr2:178779118-178779229	Both annotated	12:SOT058	1.0:SOT058
<b>PDE4B</b>	chr1:65793248-65913245	Both annotated	8:SOT058	1.0:SOT058
<b>CCDC175</b>	chr14:59561228-59563737	Both annotated	6:SOT058	1.0:SOT058

Table displaying the novel splicing events in SOT58. Anno status indicates which if any of the junctions were annotated, sample support indicates the total number of reads in each sample which span the event. Normalised read support is the ratio of reads supporting this event, compared with the other splicing patterns using the exons.

### 4.3 Testing the Mendelian RNAseq tool on the first patient sample

To further test the developed splicing analysis informatics pipeline, patient SRB0003 which was sequenced in a first, separate batch to the other PID patients underwent whole blood RNA splicing analysis. This step was performed using the ‘Mendelian RNAseq’ program developed by Cummings et al., (155). SRBC0001, SRBC0002 and GTEx data served as controls for this analysis step. The use of GTEx data was appropriate in this instance as controls served only to identify existing splice sites, and not as a quantitative measure to draw conclusion from. Bam files of aligned reads were sorted and indexed, with read groups assigned and duplicates marked were loaded into the Mendelian RNAseq splice discovery file architecture on IRIDIS4, the high-performance computer cluster at the University of Southampton.

The program was executed, and the SRB003 patient output file which contained 1384287 detected splice junctions in the first instance, underwent normalisation by comparing the read support for each splice event to read support for surrounding splice patterns as described in the methods section. The list of junctions with their normalised read support values was then written to a new text file by the program. This file was subsequently filtered in two separate manners, the syntax for which is contained in the following sections.

#### 4.3.1 Results from SRB003 Splicing analysis.

##### 4.3.1.1 Results of filter 1

```
“python "/scratch/jl5e18/RNA_SEQ/Cummings/analysis/FilterSpliceJunctions.py" -splice_file
"/scratch/jl5e18/RNA_SEQ/Cummings/analysis/All.genelist.normalized.splicingJLL.txt" -
include_normalized -sample_with_highest_normalized_read_support ID003 -n_samples 3 -
print_simple”
```

This syntax version for the filter script selects all events in which SBR0003 had the highest normalised number of reads supporting it, providing the event is not seen in more than 3 samples. This returned 744 hits. Further filtering steps were performed manually as follows. Junctions must have greater read support than 10 – to help distinguish from background noise. This left 11 genes

remaining. Junctions should not be present in only the patient and the non-GTEX controls, as this may be an artefact of the different sample processing and sequencing methods.

These were then cross referenced with the genes from the immune panel, giving only one candidate. The normalised read support of this was too low to be considered “real” or potentially pathogenic, and as such it was not investigated further.

**Table 4-2 Splicing analysis results from SRB0003 :Filter 1**

Gene	Location	Total reads supporting event	Junctions known	Samples observed in	Reads in samples	Normalised read support.
<i>ATM</i>	chr11: 108256340- 108257500	13	One annotated	2	4:IDC001, 9:ID003	0.038:IDC001, 0.138:ID003

#### 4.3.1.2 Results of filter 2

```
"python "/scratch/jl5e18/RNA_SEQ/Cummings/analysis/FilterSpliceJunctions.py" -splice_file
"/scratch/jl5e18/RNA_SEQ/Cummings/analysis/All.genelist.normalized.splicingJLL.txt" -
include_normalized -n_read_support 20 -n_samples 3 -print_simple"
```

This broader filter selected events which has read support greater than 20 overall and was seen in no more than 3 samples. This returned 11,686 splice junctions, these were then re-filtered using the term “ID003” in Microsoft Excel; this was the term given for SRB0003. This returned 138 splice junctions. These were then again negatively filtered for all those events only seen in the three non-GTEX groups and not in the GTEX, as these had a high probability of being an artefact of the processing steps. These were then negatively selected for all those which had less than 10 read support in ID003 specifically. In addition, globin genes were also filtered out, due to the highly repetitive nature and the inherent unreliability of the outcomes. 121 events remained. These were cross referenced using the gene panel, after which 7 events remained.

**Table 4-3 Splicing analysis results from SRB0003 : filter 2**

Gene	Location	Total Read support	Annotation status	Number of samples	Read support: Sample name
<i>RAC2</i>	chr22:37232702-37232819	113	Neither annotated	2	27:SRR810945_MD.bam, 86:ID003
<i>NCF1</i>	chr7:74778233-74778371	48	Neither annotated	1	48:ID003
<i>ADAR</i>	chr1:154583095-154583140	42	Neither annotated	3	8:SRR661553_MD.bam, 12:ID003, 22:SRR656445_MD.bam
<i>SH3BP2</i>	chr4:2794407-2794432	38	Neither annotated	3	4:SRR1328407_MD.bam, 11:ID003,23:IDC001
<i>RIPK1</i>	chr6:3092443-3093219	31	Neither annotated	3	3:IDC001,6:IDC002,22:ID003
<i>B2M</i>	chr15:44711578-44711604	28	Neither annotated	3	6:IDC002,11:ID003, 11:SRR1414559_MD.bam
<i>STAT1</i>	chr2:190966551-190966576	24	Neither annotated	2	4:SRR657468_MD.bam,20:ID003

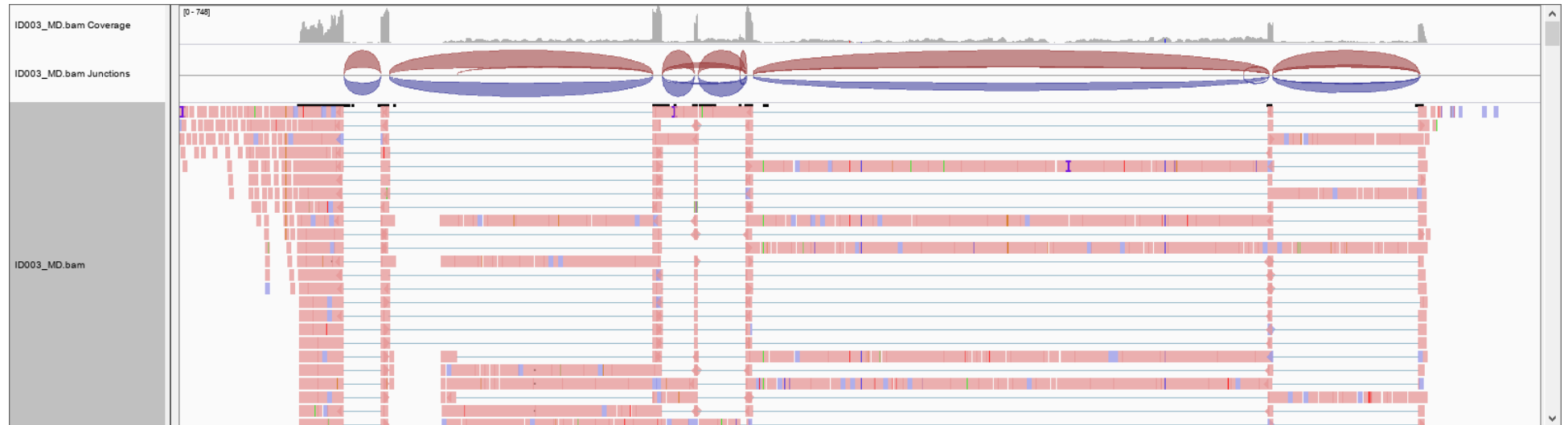
#### 4.3.2 Interrogation of events with IGV

The patients' sorted and indexed bam file was imported to the IGV (Interactive Genome Viewer) software for the visual inspection of the junction's results. *RAC2* was first investigated, and indeed there appears to be a splicing anomaly in the mapping occurring just after the 3<sup>rd</sup> exon (Figure 4-1) (gene is in reverse orientation).

From the third exon in the reverse orientation a small splicing event can be visualised which begins at the end of the third exon and ends part way into the intron. Generation of a sashimi plot shows an alternate view of this event. A small arch, representing the end of one aligned part of the read and the beginning of the next part of the aligned read can be seen, with the number 90 representing the overall read support for the small splicing event (Figure 4-2). Under close inspection, this event appears to be an alignment artefact.; specifically, the reads have been inappropriately split where

sequence homology exists in two areas, one of which being the mis-aligned site. This event can therefore be discounted. The sequence homology is highlighted Figure 4-3 and Figure 4-4





**Figure 4-1 IGV alignment and splicing of RAC2.**

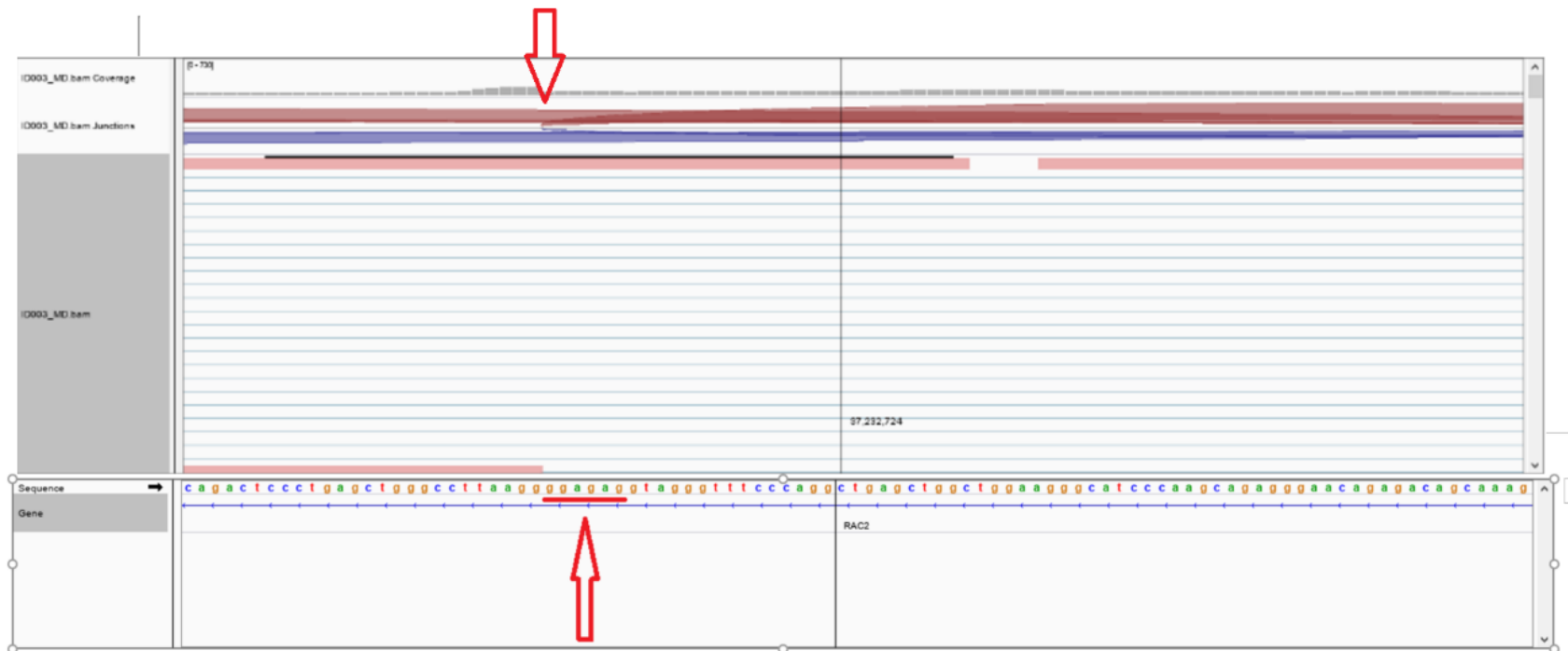
A screenshot from IGV. The grey peaks represent read coverage, the splicing track consisting of blue and red arches. The rows of red bands underneath demonstrate the aligned location of the reads on the genomes.



**Figure 4-2 - Sashimi plot for *RAC2*, exons 3, 4, 5 and 6.**

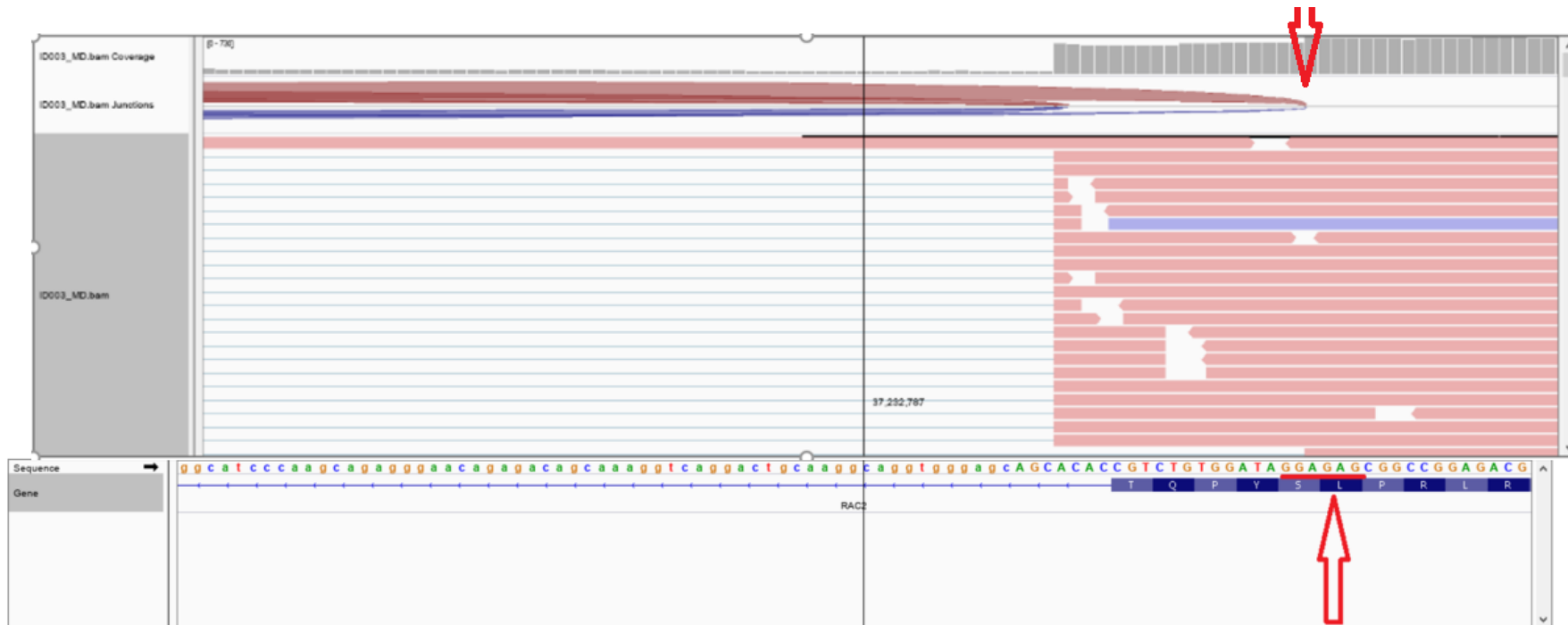
The event in question can be seen in the sashimi plot above the blue exon (track along the bottom) which is first from the left. At this stage it looks like a real event with good read support.

Evidence of reads being inappropriately split. *RAC2* sequence and alignment.



**Figure 4-3** *RAC2* sequence and alignment

Red arrows indicate the section of intronic region which has a sequence matching that of the end of the previous intron. Some shorter reads have mis-aligned this exon tail, causing an apparent aberration. Underlined sequence matches that in Figure 4-4.



**Figure 4-4 RAC2 sequence and alignment 2.**

Red arrow indicates the exonic sequence which 90 reads should have continued their mapping to.

Each event remaining after filtering in SRB003 was investigated with IGV using the same methods as described in 4.3.2, and the results are collated in Table 4-4 - Outcome of IGV investigation.

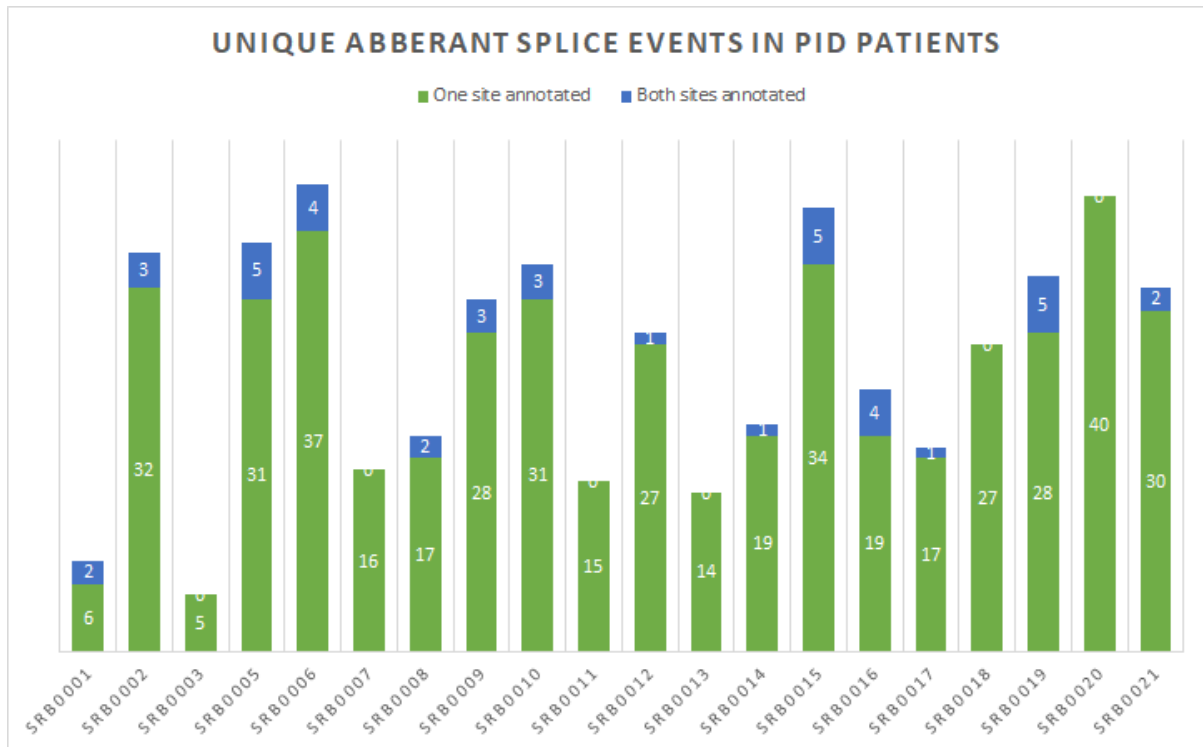
**Table 4-4 - Outcome of IGV investigation**

<i>Gene Splice event</i>	<i>Investigation outcome summary</i>
<i>RIPK1 chr6:3092443-3093219</i>	This event appears to just be the result of background noise as no exons are mapped to this location.
<i>RAC2 chr22:37232702-37232819</i>	Inappropriate read splitting.
<i>NCF1 chr7:74778233-74778371</i>	No reads linked the sites which were spliced to the rest of the transcripts, and there existed a very high degree of sequence homology between the sites immediately after the splicing. CGCCCTCC – TACCCTCT – Likely inappropriate read splitting
<i>ADAR chr1:154583095-154583140</i>	Read support too small to be significant considering surrounding read support. 12 reads in downstream untranslated region. 315 reads between the previous exon and UTR.
<i>SH3BP2 chr4:2794407-2794432</i>	Inappropriate reads splitting
<i>B2M chr15:44711578-44711604</i>	Read support too low. 9 reads for this junction vs 5400 for the nearest other junction
<i>STAT1 chr2:190966551-190966576</i>	Inappropriate read splitting.

#### 4.4 PID cohort splicing analysis.

Following these investigatory analysis steps and further review of the literature, the 20 PID samples RNAseq data underwent splicing discovery, normalisation and finally filtering using the methods described in section **2.1.18**.

Filters were developed to that event had to be unique, have read support greater than 5 and at least one half of the junction needed be annotated. The total number of unique-to-sample events which had read support greater than 5 and at least one annotated splice site varied from 5 in SRB0003 to 41 in SRB0006 (Figure 4-5). Very low numbers were obtained in samples SRB0001, which was noted to be of low quality in QC stages, and SRB0003 which alongside negative controls SRBC001 and SRBC002 was sequenced to a lower depth than the other samples. Events where both splice sites are annotated are those which do not involve the creation or activation of cryptic splice sites, this again adds a further amount of confidence that they might be real, and the result of events such as exon skipping. There are fewer events which can occur uniquely in a sample whilst still utilising two annotated splice sites. This is demonstrated by there being fewer events of this nature recorded (blue sections of split bars). Total number of events range from 5 to 41; a maximum difference of 273%



**Figure 4-5 Per-sample unique splicing events in WB RNAseq data**

The range of per sample unique events spans 5 (SRB00003) to 41 (SRB00006). Regarding events with two annotated splice sites, many samples have no junctions which have both splice sites annotated. The greatest number is 5, appearing in SBR0015 and SRB0005 SRB0019. These make up a maximum fraction of the total event of  $\frac{1}{4}$  in SBR0001.

At this stage the number of events is still too high to investigate manually through IGV and as such alternative filtering is needed. To reduce the number of events to a manageable amount for investigation, the following parameters were employed for event filtering:

- Novel splicing events in globin genes are not included, due to the extremely complex splicing patterns and lack of association with PID.
- At least one of the junctions for the splice site must be annotated.
- Any events which have less than 0.1 normalised read support are filtered; this value was chosen as it produced a manageable number of events for manual inspection.
- Events which appear in the table in green have read RNA support  $> 1$  meaning they became the dominant splicing pattern involving the neighbouring exons.
- The events were cross referenced with the three PID gene panels.

This filtering strategy will not capture some possible events in which the transcripts produce non-functional transcripts which are subsequently broken down in a mechanism known as nonsense mediated decay (NMD). However, reducing the number of events being investigated manually to manageable numbers is important, as such the filtering method is designed to strategically prioritise those events which are more likely to be real and not a result of aberrant alignment. It will also not capture events in non-protein coding genes. Finally, the method is blind to events where both junctions are unannotated. Even in the case of cryptic exons the final reads should have one known exon binding to one cryptic exon so these should not be missed. In the event of complex splicing changes, for example where one exon has a new end point and the following exon is an activated cryptic exon, the event will likely be missed. Additionally, the program cannot detect events where splicing canonically takes place, but does not, such as with abnormal intron retention.

Colour coding of results is systematic based on normalised read support: Green = Dominant, Yellow = read support (RS) >0.25, Orange = read support (RS) >.1 Clinical features of the patients were only present for some of the samples. It was not possible therefore to always determine if the results seen could be linked to the patient. Clinical features are stated when known.



#### 4.4.1 Investigation of events in sample SRB0002

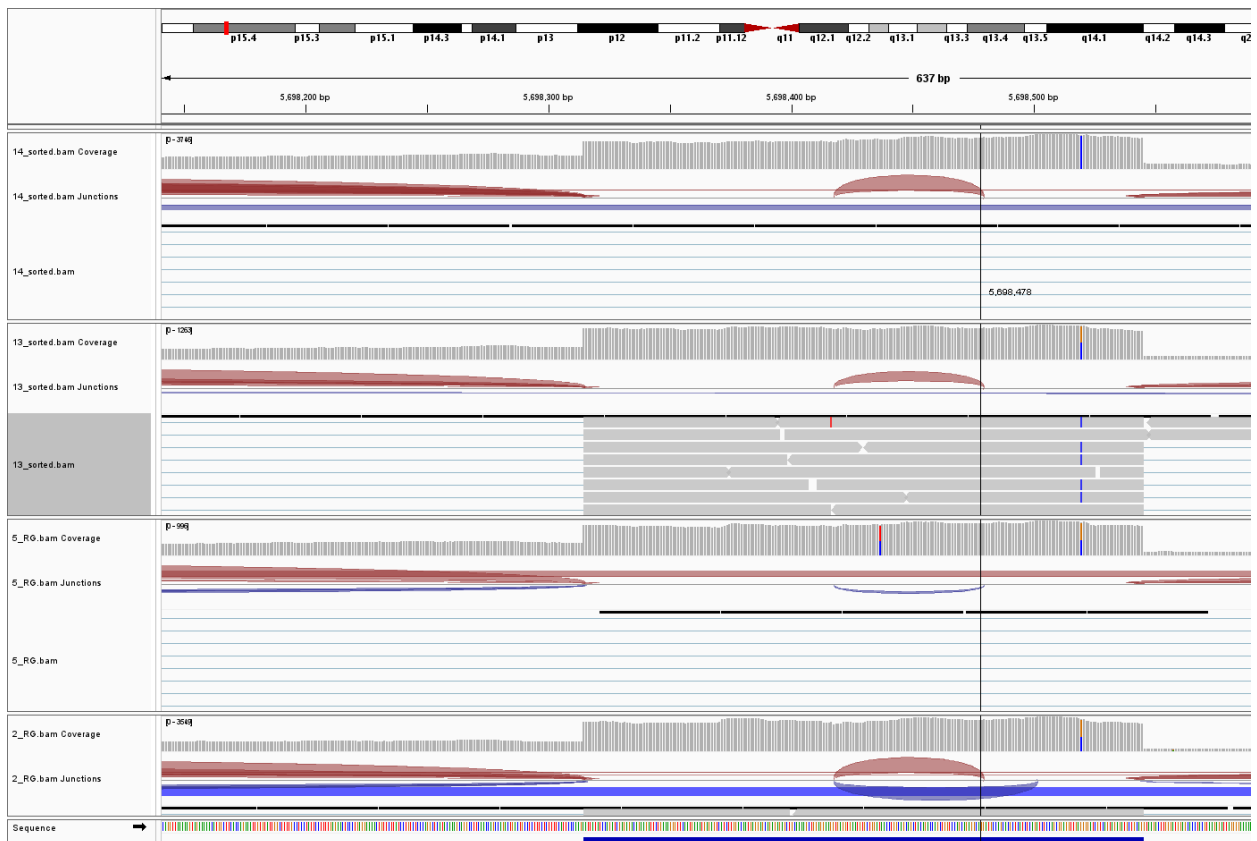
Patient SRB0002 presented with Asthma, Allergy, Recurrent Pneumonia, Recurrent Viral skin infections, Recurrent Bacterial skin Infections in what was considered and undefined immune deficiency.

Gene	Location	RS	Annotation	Norm.R.S	Panel
<i>TRIM22</i>	chr11:5698417-5698502	21	One annotated	0.636:SRB0002	GECIP

SRB0002 has 334 total events before filtering for annotation status. After filtering for annotation status, read support and globin genes 10 events remained. Of these only *TRIM22* appeared in the GECIP PID panel. *TRIM22* had 21 total reads supporting and normalised read support of 0.636 suggesting a high number of the total reads used this splicing pattern. However, it was noted that this statistic was derived from the other intra-exon splicing event within exon 4, and not the total number of reads of the exon, highlighting a weakness in the algorithm used in the tool. The IUIS gene list describes *TRIM22* related PID as manifesting with Granulomatous colitis.

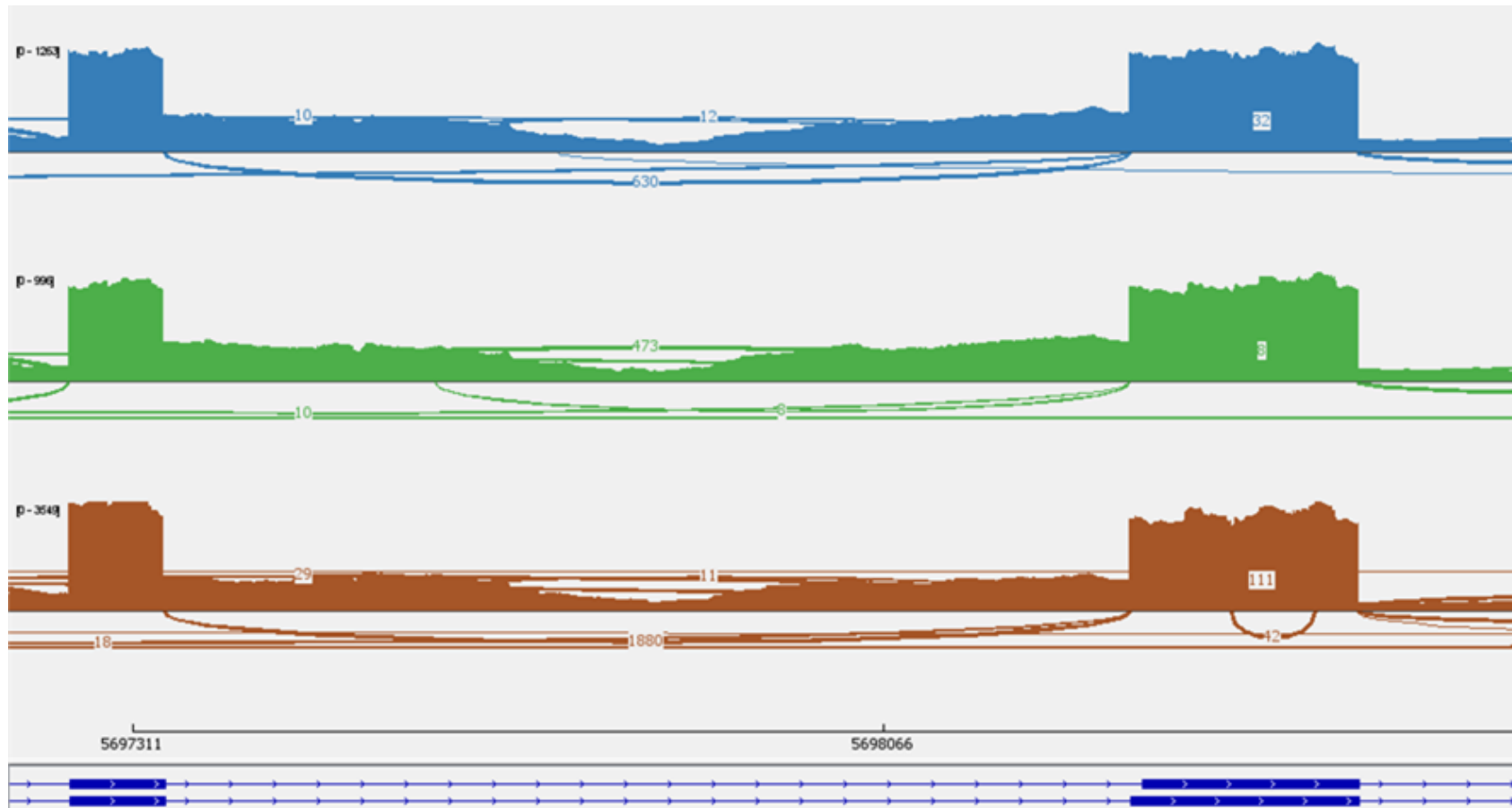
RNAseq alignment was investigated using Integrative genome viewer. An event could be seen where a portion of exon 4 of 8 is spliced out at some level for all samples (Figure 4-6). The read support for this event is low across all samples – around 5% of the total reads. In addition to this, for SRB002 only there is an alternative splicing pattern in which a larger portion of the exon is spliced out. This has a read support of 42, or about 2% of the total reads. This event is 727 base pairs in length, indicating the induction of a frameshift and consequently indication of nonsense mediated decay, which could explain the low read support for the event. While this is significant within the context of intra-exon splicing event seen across all samples, it is extremely low in read support compared to the reads across surrounding junctions, which number around 1880 (Figure 4-7). *TRIM22* associated primary immunodeficiencies have been primarily associated with gastrointestinal symptoms manifesting

## Results of Splicing Analysis of Primary Immunodeficiency Patients



**Figure 4-6 SRB002 Alternative Splicing Event *TRIM22***

Image shows the read coverage track for 4 samples of gene *TRIM22*, the bottom of which is SRB002. The blue arc upside down represents the novel splicing event which is not present in any other samples.



**Figure 4-7 Sashimi plot of SRB002 *TRIM22* Alternative Splicing Event**

Four tracks represent coverage and splicing within four samples. Brown sample represents SRB002, and the half loop represents the novel splicing event, 42 represents the number of reads on the event.

## 4.4.2 Investigation of events in sample SRB0005

Clinical Features	Cellular Molecular Features	Sample
<b>Asthma</b> <b>Allergy</b> <b>Recurrent Pneumonia</b> <b>Recurrent Viral skin infections</b> <b>Recurrent Bacterial skin Infections</b> <b>Specific Polysaccharide Antibody Deficiency (SPAD)</b> <b>Impaired T cell Function</b>	Immunodeficiency affecting Cellular and Humoral Immunity	005

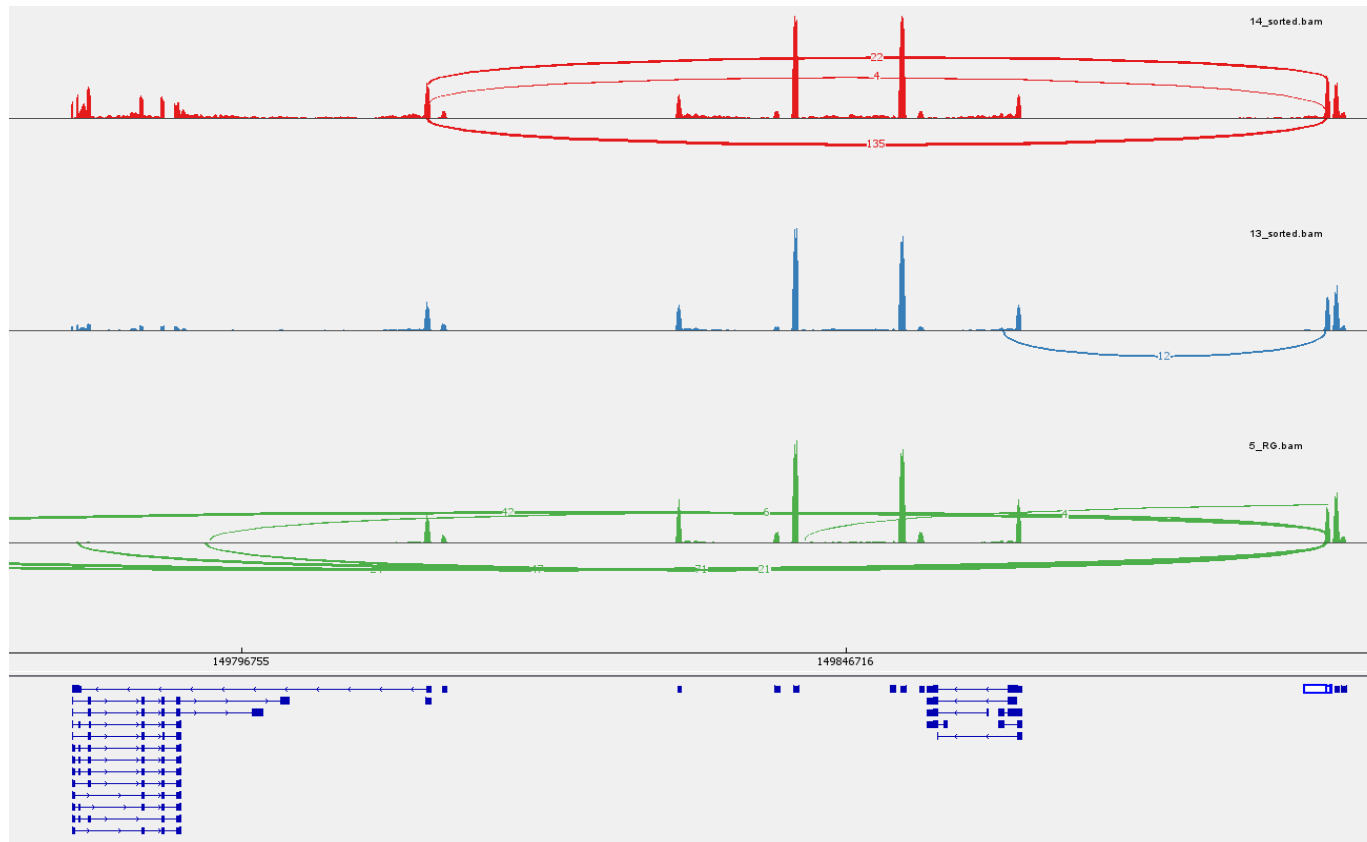
Gene	Location	RS	Annotation	Norm.R.S	Panel
<i>FCGR1A</i>	chr1:149783260-149886264	21	One annotated	7.0:SRB0005	A.W.

In total 300 events were present in SRB0005 before filtering for annotation status and normalised read support. 10 remained after second round filtering was performed. Of these, one appeared in known PID panels when cross referenced *FCGR1A*. Read support for this event was moderate at 21.

Investigation using IGV revealed that this event spanned a from a portion of the *FCGR1A* gene which overlapped the *H2BC1A* histone gene and another high correlated sequence in another histone gene a considerable distance away. It therefore appears that this event is an artefact of alignment and not related to the *FCGR1A* gene. No further investigation of this event was conducted.

Addendum note: GECIP has since updated their Panelapp database to include *FCGR1A* in PID associated genes, although no details on phenotype of pathology is available. From the OMIM database, patients with deleterious mutations in the *FCGR1A* gene appear to be unable to activate anti-CD3 induced T-cell mitogenesis. This occurs through lack of *CD64* expression on monocytes and

dendritic cells, although linked literature suggests patients can appear otherwise healthy (371). It is therefore unlikely to be the causal variant, but further investigation is warranted.



**Figure 4-8 Sashimi plot of SRB005 event spanning multiple genes**

The figure above highlights splicing events and respective coverage. The green sample represents SRB005. The second blue cluster on the bottom map of exons represents both *FCGR1A* (arrows pointing right) and *H2BC1A* (arrows pointing left). Other clusters represent other histone genes. One highlighted gene *H2BC21* is selected so only splicing events related to that gene are present as the event spans the distance from the *FCGR1A* gene and the *H2BC21* gene with a total of 21 reads.

Clinical Features	Cellular	Molecular Features	Sample
<b>Chronic</b>	<b>Mucocutaneous</b>	Defects in Intrinsic and Innate	SRB006/7/8
<b>Candidiasis</b>		Immunity	

#### 4.4.3 Investigation of events in sample SRB006

Gene	Location	RS	Annotation	Norm.R.S	Panel
<i>TFRC</i>	chr3:196075360-196075728	16	One annotated	0.128:SRB0006	IUIS Table 1

SRB0006 had 404 events before second round filtering. 14 remained after filtering for both annotation status and normalised read support > 0.1. Of these events, only one appeared in a gene which also featured in cross referencing panels; *TFRC*. The features of *TFRC* related PID were recurrent infections, thrombocytopenia, and low neutrophil count (372). However, as this event was not present in the other members of the family trio, it therefore was ruled out as the causative variant. No follow up required.

#### 4.4.4 Investigation of events in sample SRB0013

Clinical Features	Cellular Molecular Features	Sample
<p><b>Predominantly Antibody Deficiency</b></p> <p><b>Recurrent Bacterial Infections</b></p> <p><b>Splenomegaly</b></p> <p><b>Lymphoid Interstitial Pneumonia</b></p>	<p>Predominantly Deficiency</p> <p>Antibody</p>	013

Gene	Location	RS	Annotation	Norm.R.S	Panel
<i>IL16</i>	chr15:81306545-81306611	22	One annotated	0.129:SRB0013	A.W.

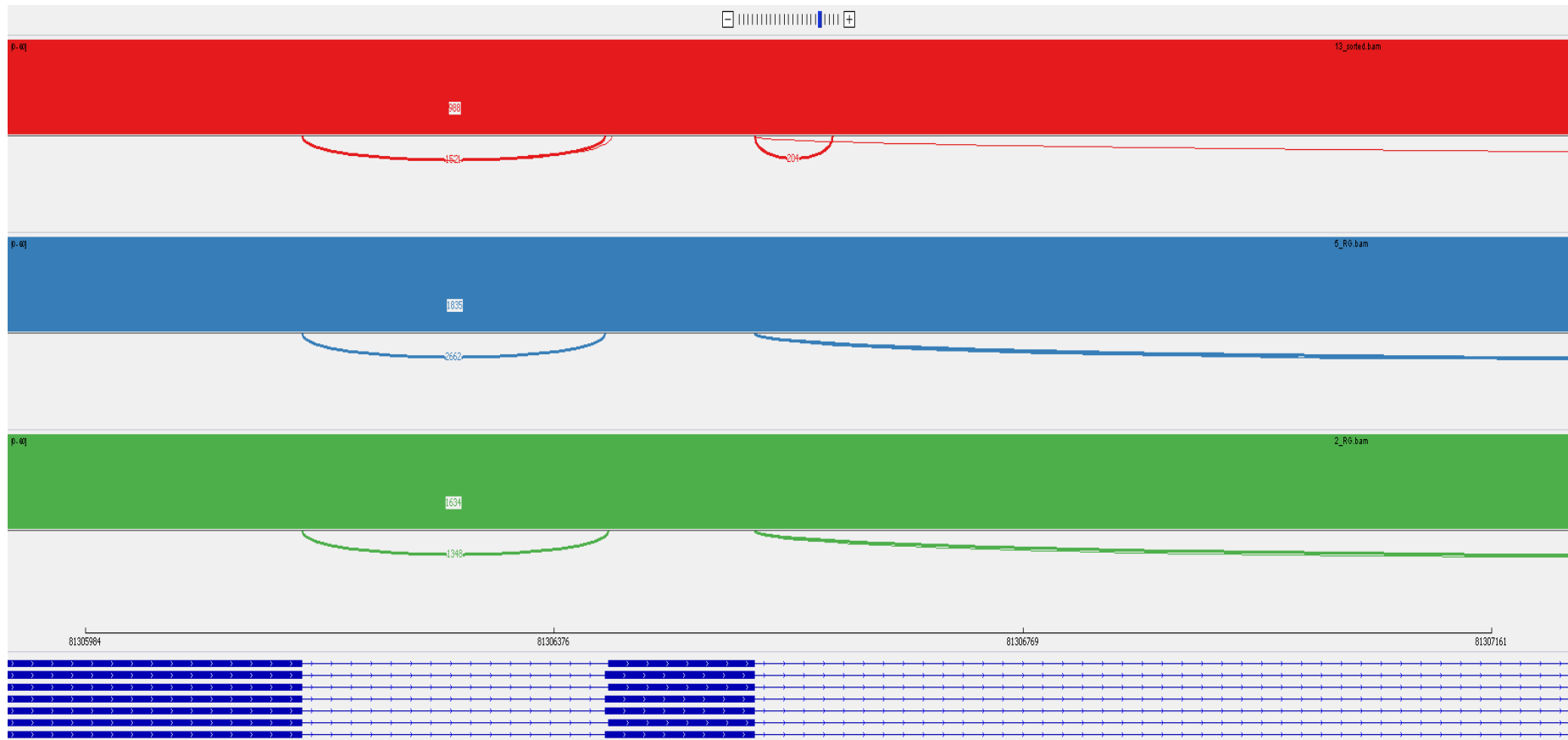
In total 183 events were present in SRB00013 before filtering for annotation status and normalised read support. 5 remained after second round filtering was performed. One of the events appeared in a gene (*IL16*) found in the PID cross-referencing panels. Read support was low at 22, and this comprised only .13 of the normalised read support around this junction.

Genetic disease involving *IL16* is not recognised as a primary immunodeficiency. A number of phenotypes which *IL16* is associated with include, Graves' disease – an auto immune disorder. Sarcoidosis, susceptibility to common cold lymphocyte count and monocyte count. Other non-immune phenotypes include: height, pro-interleukin-16 measurement, blood protein measurement, hair colour measurement, obesity coronary artery disease, drug use measurement, aspirin use measurement, NSAID use measurement, colorectal cancer, primary biliary cirrhosis, body mass index, myocardial infarction, ocular sarcoidosis, peripheral arterial disease, behaviour (373). Graves' disease does not follow a clear inheritance pattern, and is instead a multifactorial condition as demonstrated by twin studies (374). Some examples of Graves' disease have presented with splenomegaly (375) and also can mimic B-cell lymphoproliferative disorders, which manifest with symptoms such as this patient is experiencing (375).

Under inspection with IGV the event is appears to be a case of partial intron retention between exon 18 and 19 (Figure 4-9). In about 10% of the transcripts a small subsection of the intron is spliced out, but the majority remains. Interestingly, a great deal more read support is present in the IGV analysis that that produced by the Mendelian RNAseq tool. There also appears to be a heterozygous A-G mutation in the following terminal exon of *IL16*, at Ch15:81,308,981 which produces a sequence of



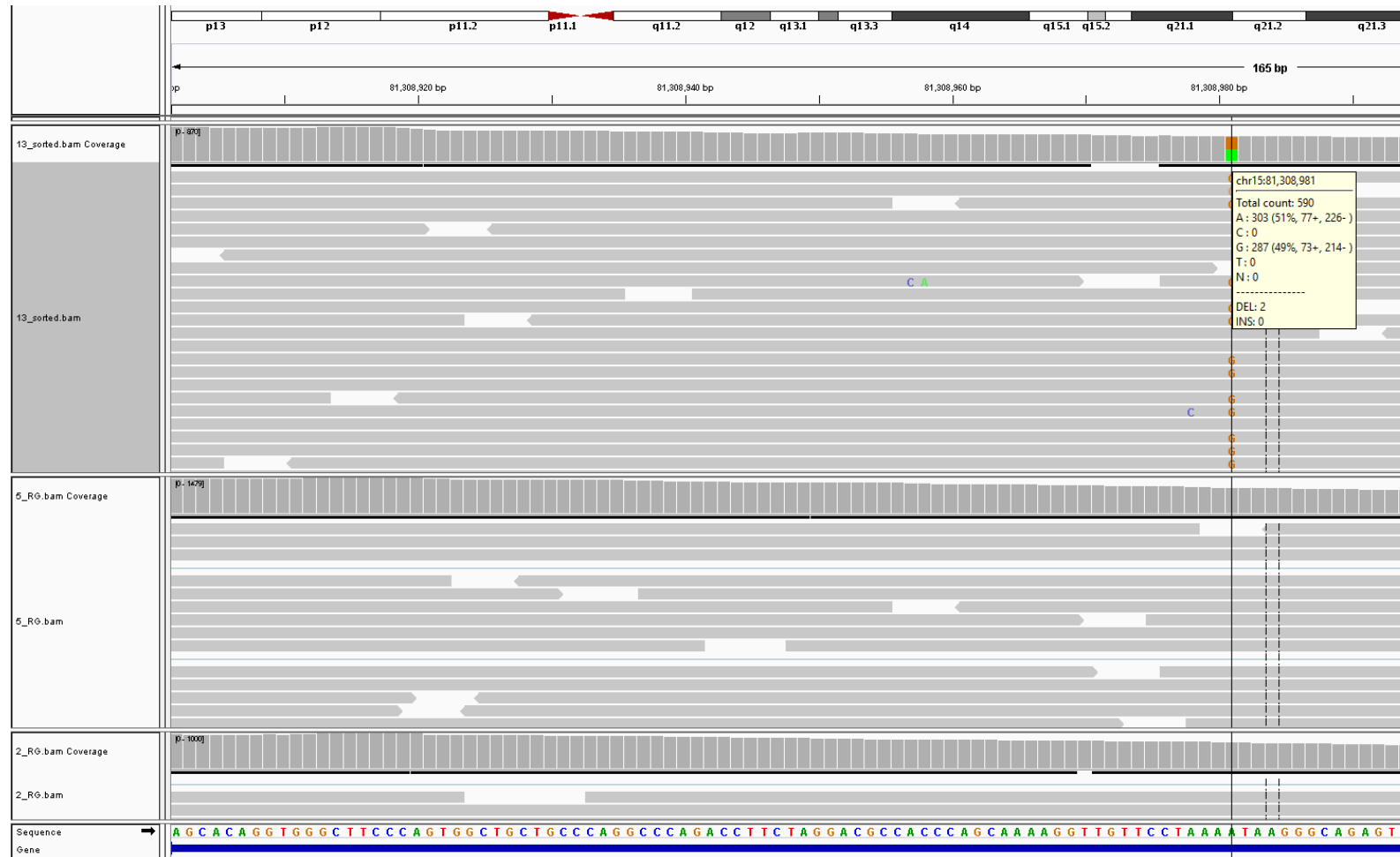
AGTA, potentially creating a new splice site (Figure 4-10). There is some overlap or prospective links between conditions associated with *IL16* and those seen in the patient; biliary cirrhosis complications often include splenomegaly. Moreover, elevated liver enzymes can be present in cases of infection which might mean liver involvement in this patient's condition could have been overlooked and not reported. It is possible that this variant may be directly related to the patient symptoms; further clinical investigation will be required.



**Figure 4-9 SRB0013 *IL16* Sashimi Plot**

Sashimi plot shows the splicing events around exon 19 of *IL16*. The red sample represents SRB0013, and the extra event can clearly be seen. With a read coverage of around 10% of that of the neighbouring junctions.

Results of Splicing Analysis of Primary Immunodeficiency Patients



#### 4.4.5 Investigation of events in sample SRB0014

Clinical features	Cellular Molecular features
<b>Panhypogammaglobulinaemia</b> <b>Recurrent Bacterial Infections</b> <b>Splenomegaly</b> <b>Lymphoid Interstitial Pneumonia</b> <b>AIHA Enteropathy</b>	Predominantly Antibody Deficiency

#### Results from Mendelian RNAseq tool for SRB0014

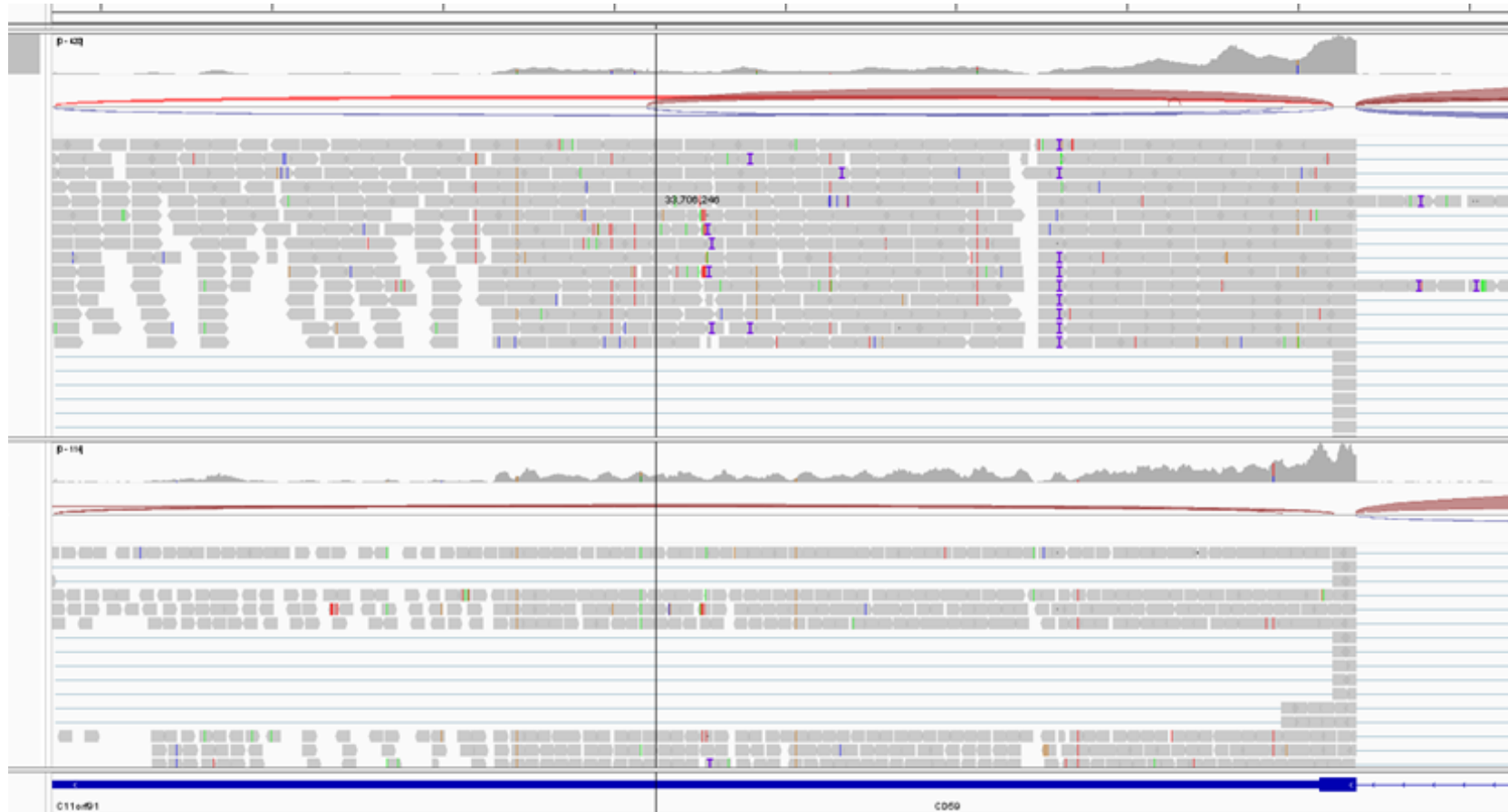
Gene	Location	RS	Annotation	Norm.R.S	Panel
<i>CD59</i>	chr11:33706201-33710206	14	One annotated	14*:SRB0014	IUIS Table 8

In total 224 events were present in SRB00014 before filtering for annotation status and normalised read support. 5 remained after second round filtering was performed. One of the events appeared in a gene (*CD59*) found in the PID cross-referencing panels. IUIS Table 8, describes *CD59* associated PID features as: Haemolytic anaemia, polyneuropathy, thrombosis.

Read support was low at 14. Normalised read support values suggest that this became the only splicing pattern around this junction. IGV investigation was carried out on this sample. Read support appeared higher at 32 for this event, and this represents about 10% of the reads which cover the canonical junctions across neighbouring exons, contrary to the results indicated by the Mendelian RNAseq tool.

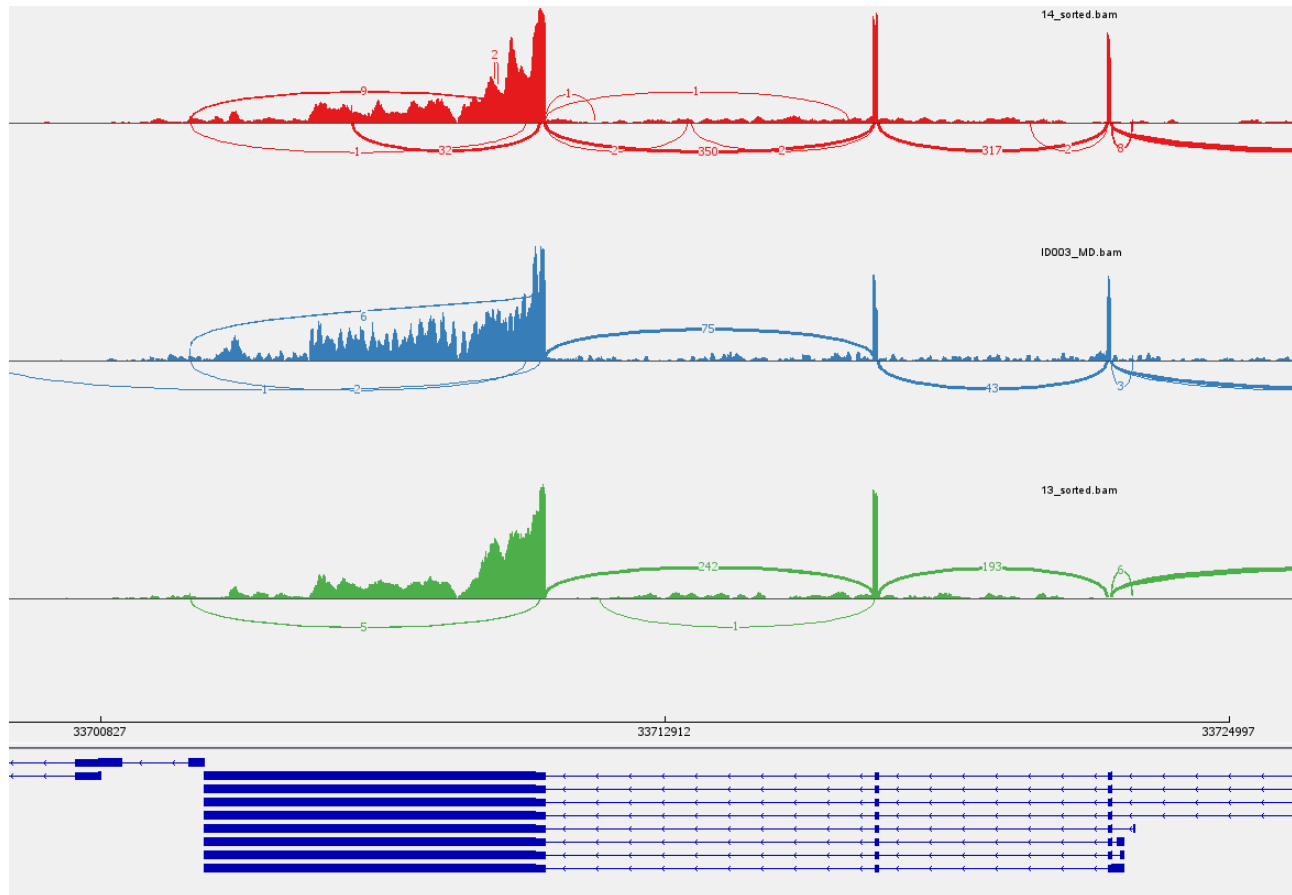
This region of the genome appears to be 'noisy' in all the samples which were investigated, with variants appearing in this first exon for all samples, including *CD59* (Figure 4-11). The splicing event covers a fairly substantial section of 4005 bases (Figure 4-12). As this value is a multiple of 3, a frameshift does not occur and instead a mis-sense outcome is expected.

CD59 is a cell surface molecule which inhibits the formation of the terminal attack complex (376), and leads to complement overactivation . The molecule is expressed on tissues which come into contact with the immune system to prevent autoimmunity, including those of the nervous system (376). In the event that this molecule is not expressed, or functional autoimmunity can ensue. Common effects include autoimmune haemolytic anaemia, and neuropathy (377). Examples of CD59 related pathology usually involve both copies of the gene (377). With the high number of variants around CD59, it may be that two of these are working in parallel to produce compound heterozygous mutations, and as such DNA sequencing and analysis may be necessary.



**Figure 4-11 Coverage plot for SRB0014 for CD59**

Coverage plot shows some noisy data for patient SRB0014 around CD59. This substitutions and deletions at various points in the CD59 first exon.



**Figure 4-12 Sashimi plot sample SRB0014 CD59 splicing.**

Image represents the splicing patterns around exon 1 of *CD59*, with SRB0014 highlighted in red. The upside-down arch with bisects the first exon with 32 reads is the event in question which spans a significant 4005 bases in the patient, accounting for approximately 10% of all reads across neighbouring exon junctions.

## 4.4.6 Investigation of events in sample SRB0018

Clinical features	Cellular Molecular features
<b>Panhypogammaglobulinaemia</b>	Predominantly Antibody Deficiency
<b>Recurrent Bacterial Infection</b>	
<b>Bronchiectasis</b>	
<b>Type 2 Diabetes mellitus</b>	

Gene	Location	RS	Annotation	Norm.R. S	Panel
<i>S1PR1</i>	chr1:101236991-101236993	23	One annotated	1.769: SRB0018	A.W.

252 events were present in SRB00018 before filtering for annotation status and normalised read support. 7 remained after second round filtering was performed. One of the events appeared in a gene (*S1PR1*) found in the PID cross-referencing panels. Read support was low at 23. This transcript appears to have become the dominant transcript around the neighbouring junctions with a normalised read support of 1.77. This event was investigated using IGV. It was quickly established that this was not a true splicing event, but rather a heterozygous point deletion in the patient's genome. Indeed, the event table shows the splicing only spanning a single base (Figure 4-13) (Figure 4-14). The variant in question is at the exact point of a splice site, and so would result in a non-sense mutation. Of the two canonical splicing patterns around the site, both use alternative starting exons. Whilst the use of the exon which does not rely on the splice site which was affected has read support levels which are approximately average when considering the other samples, the read support for the splicing pattern which does rely on the affected exon is 50% that of the next sample. The isoform which uses this splicing pattern is lower across all samples and represent around 1/5 of the reads, in the affected sample SRB0018 this splicing pattern represents around 1/20 of the reads from this sample, supporting the notion of nonsense mediated decay occurring as a result of this frameshift inducing mutation.

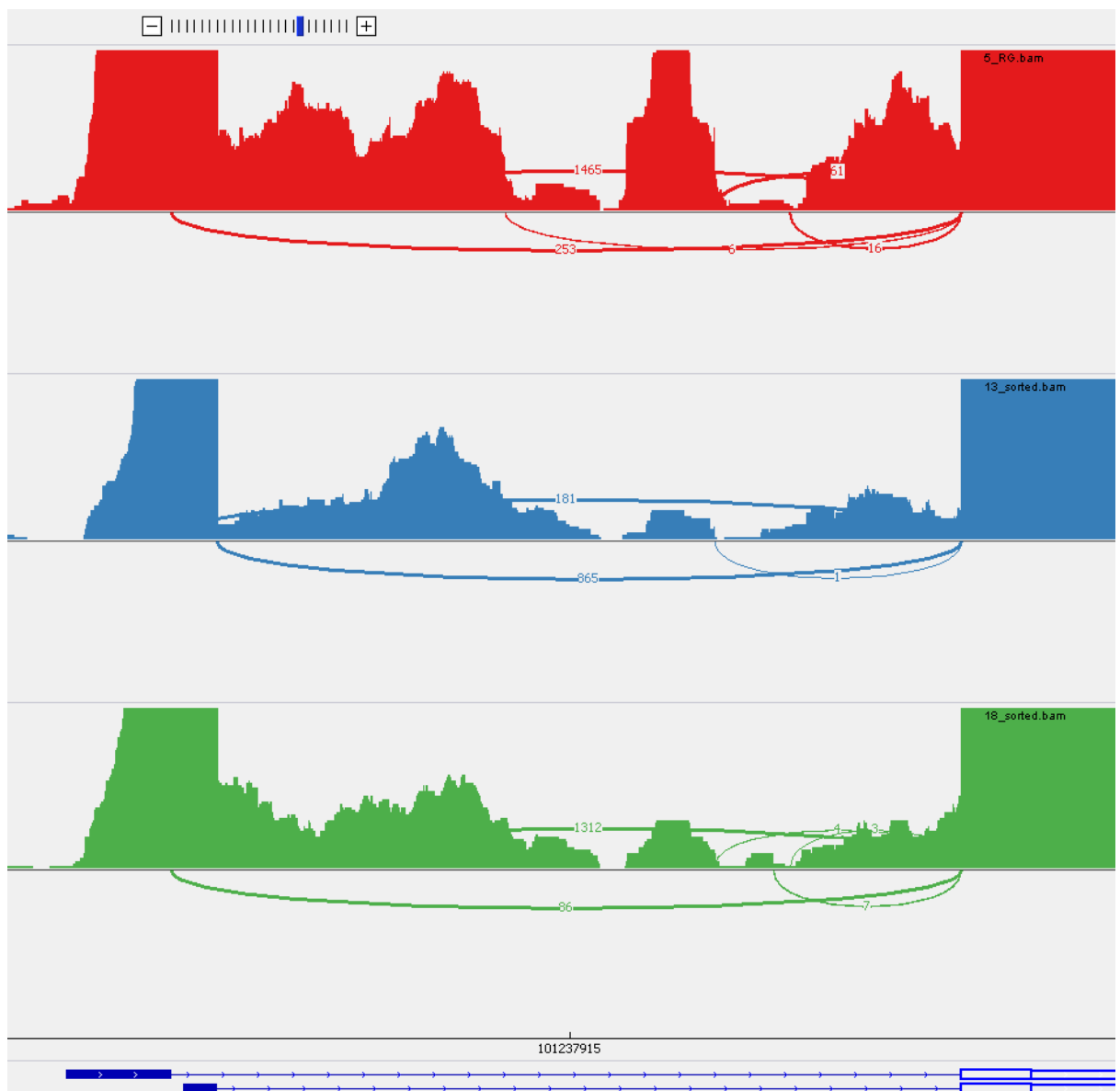


S1P is protective against Type 2 Diabetes Mellitus, indicating a link between a variant in the receptor and the patients' symptoms (378). S1P is expressed on lymphocytes and promotes egress from lymph nodes. S1P1R is also responsible for egress of mature T-cells from the thymus (379) and ultimately their phenotypes also. A body of literature now suggests that S1P1R may also be responsible for the fate and function of T regulatory cells T helper type 17 cells and memory T-cells (380). Similarly, this signalling pathway also plays an important role in mature B-cell egress from the bone marrow (381). The variant in question (rs3835397) has previously been discovered, in a study looking at variants which much be linked to asthma (382), which is elsewhere linked to bronchiectasis (383). Taken together, it is probable this this variant is causative.



**Figure 4-13 SRB0018 Variant**

Imagine shows the deletion of a G from the splice site of exon 1 in *S1P1R* gene at Chr1:101,236,992.



**Figure 4-14 Sashimi plot of SRB0018 *SIP1R* splicing.**

Figure shows splicing patterns of *SIP1R* gene around exon 1, and alternative exon 1. SRB0018 is shown in green and the reduced ratio of use of exon 1 and exon 2 as start sites is reduced, 253:1465 and 181:865 in controls vs 86:1312 (arcs flipped) in SRB0018.

#### 4.4.7 Investigation of events in sample SRB0019

Clinical features	Cellular Molecular features
<b>Recurrent Abscess with Pseudomonas</b> <b>Candidal Discitis</b> <b>Nephrectomy with Klebsiella abscess</b> <b>Systemic lupus erythematosus</b>	Disorder of Phagocytes

Gene	Location	RS	Annotation	Norm.R.S	Panel
<i>TLR1</i>	chr4:38791231-38845830	9	One annotated	0.409:SRB0019	A.W.
<i>POLE2</i>	chr14:49668445-49669524	7	One annotated	0.233:SRB0019	IUIS Table 2
<i>IFI44L</i>	chr1:78628393-78629809	11	One annotated	0.229:SRB0019	A.W.

274 events were present in SRB00019 before filtering for annotation status and normalised read support. 13 remained after second round filtering was performed. Three of the events appeared in genes (*TLR1*, *POLE2*, *IFI44L*) found in the PID cross-referencing panels, one of which is a known causal gene in PID.

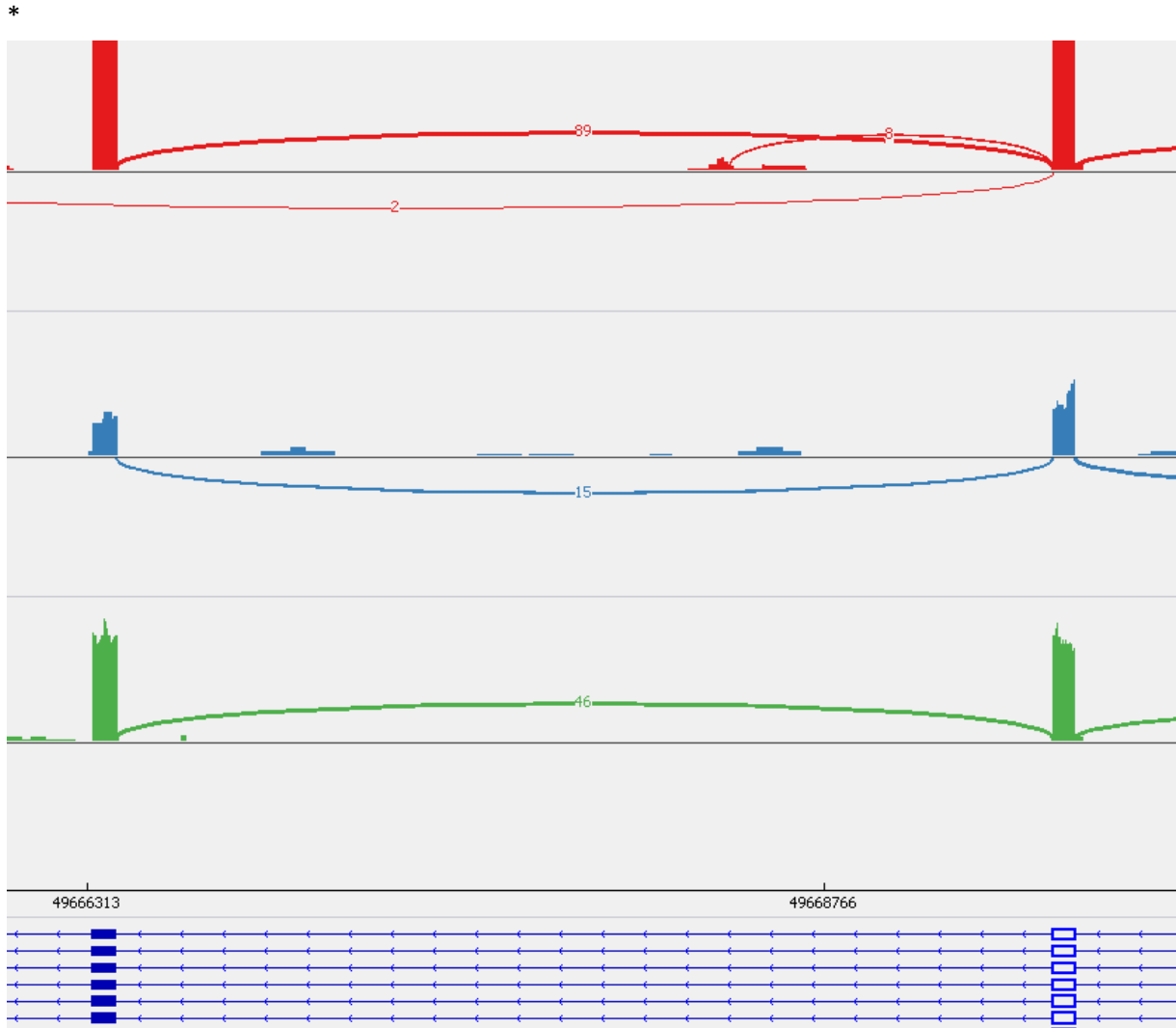
Read support was low in all cases, at the event in *TLR1* =9. Normalised read support of 0.409 indicating a high relative use of the novel splicing pattern. *POLE2* has very low read support (7). The gene is found in the IUIS gene panel, contained in Table 2; Combined immunodeficiencies with associated or syndromic features. The phenotypes associated in the IUIS table include: Recurrent infections, disseminated BCG infections, autoimmunity, type 1 diabetes, hypothyroidism, facial dysmorphism (372). The event in *IFI44L*, another gene from the list provided by Professor A. Williams, has read support 11.

These events were investigated with IGV. The first event, which occurred in *TLR1* was found to be a case of inappropriate read splitting in which the latter part of the read was mapped to another similar part of the genome as where it normally would have been and so this dismissed and not investigated further.

The second event was found to be a case of partial intron retention, in which exon 6 of *POLE2* gains 1079 extra bases of the subsequent intron (Figure 4-15). There was some evidence that there were variants nearby in the intron, but the read support was very low (1-2 reads). There is possibility that if the patient undergoes WGS, a variant may be discovered in the intron, which activated a cryptic exon. Read support for this splicing event was about 20% of that which supported the normal canonical splicing junction. The retained intron is of length 1079, which would cause frameshift and nonsense mediated decay.

*POLE2* is a subunit of a multi-subunit DNA polymerase (384). Deleterious variants in *POLE2* lead to autoimmunity, combined immunodeficiency, reduced proliferative capacity of lymphocytes and resulting recurrent infections. In addition, facial dysmorphism and short stature are recorded effects (384). Whilst some of the symptoms meet the description of *POLE2* related pathology, it's unclear without further clinical investigation if this can be the molecular cause of all the symptoms.

*IFI44L* is a type I interferon-stimulated gene involved in antiviral and antibacterial activity. It promotes macrophage differentiation and inflammatory cytokine secretion (385). And importantly is regarded as a biomarker for systemic lupus erythematosus (386) .

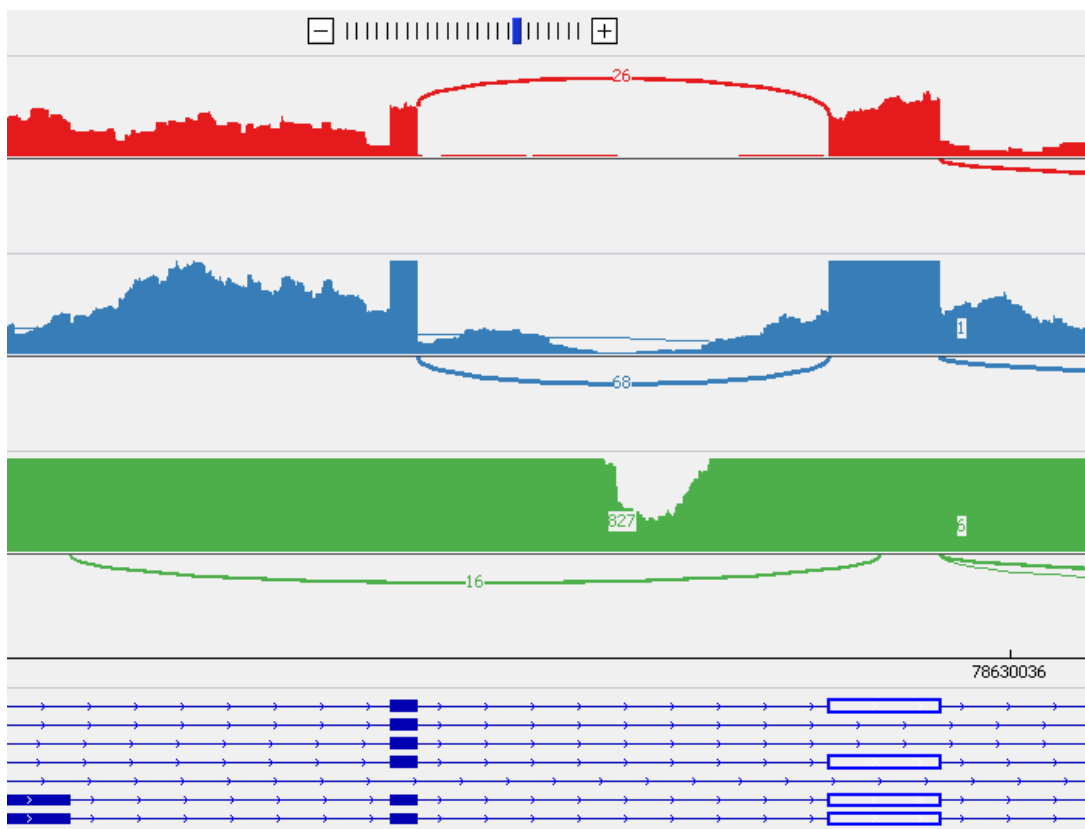


**Figure 4-15 Sashimi plot of SRB0019 *POLE2* splicing.**

Plot shows splicing patterns present in *POLE2* of patient SRB0019 shown in red, versus controls, shown in blue and green. 8 reads, or around 10% of the total read around the nearest exons indicate significant intron retention.

The final event was a case of exon skipping in the *IFI44L* gene. In canonical expression, exon 3 is only included if exon 4 is present also, also the converse is not always true. However, in patient SRB0019 a splicing pattern suggests exon 3 is present, but exon 4 and half of exon 5 is spliced out (Figure 4-16). A great many splicing patterns are present for the *IFI44L* gene in the samples examined and read support for canonical junction in the SRB0019 sample can be as high as 1500. The dominant splicing pattern in the most proximal exons have a read support of 123, and the event in questions

has a read support of 16, or around 10% of the closest alternative. However, the 123 reads supporting this event, are only around 10% of the reads across other exons which are present in the dominant transcript. Therefore, within the wider context of all the isoforms of the gene, this is a very low abundance transcript and unlikely to be causative. In addition, no variant could be observed in the sequence for this gene, although the low coverage of RNAseq does not form reliable variant identification data, and intronic variants may be missed entirely.



**Figure 4-16 Sashimi plot of SRB0019 *IFI44L* splicing.**

Figure shows the splicing of *IFI44L* gene of exons 3, 4 and 5 from right to left with patient (green) versus controls (blue and red).

## 4.5 Summary tables of results

**Table 4-5 Summary Table of Candidate Splicing Events**

Patient	Gene	Event	Known molecular diagnosis	Causal
SRB0002	TRIM22	chr11:5698417-5698502	No information	Unlikely
SRB0005	<i>FCGR1A</i>	chr1:149783260-149886264	CARD-11 A-C @2987250	Unlikely
SRB0006	<i>TFRC</i>	chr3:196075360-196075728	STAT1	Unlikely
SRB0013	<i>IL16</i>	chr15:81306545-81306611	No Diagnosis in GECIP	Unlikely
SRB0014	<i>CD59</i>		No information	Possible
SRB0018	<i>S1PR1</i>	chr1:101236991-101236993	No Diagnosis in GECIP	Likely
SRB0019	<i>TLR1</i>	chr4:38791231-38845830	No information	Unlikely
SRB0019	<i>POLE2</i>	chr14:49668445-49669524	No information	Unlikely
SRB0019	<i>IFI44L</i>	chr1:78628393-78629809	No information	Unlikely

Table shows the gene, event and the diagnosis and the determination on causality of phenotype for the patients for which splicing events were found using the filtering strategy adopted and after being cross-referenced with the panels.

**Table 4-6 Comparing Results of Methods to Detect Gene Expression Outliers for PID Patients**

Sample	Splicing	FPKM Z-Score	TPM Z-score	OUTRIDER	Overlap
SRB0001		SEMA3E, CFTR, C7	G6PC3, TRAF3IP2, MVK, BRIP1		
SRB0002	TRIM22	FOXN1, PSENN, ALPI, IL36RN, SAMD9L			
SRB0003	ATM				
SRB0004					
SRB0005	FCGR1A	CCBE1, IL2RA, CEBPE			
SRB0006	TFRC	RNF31, NRKB2, TNFSF12, C9,	IL2RG	OCLNP1, RRP8, RRP1B	
SRB0007			IL10, TP53		
SRB0008		GINS1, IL23R, CFH, CFHR5,			
SRB0009		IL21, CFI, FANCI	NFKB1	OCLNP1	
SRB0010			G6PC3		
SRB0011			RFX5, FAPP24, FASLG, THBD	OCLNP1,	
SRB0012				ACOT9	
SRB0013	IL16	RAG1, FAT4, IL17F, C4A, C4B, CFHR2, CFHR3, BRIP1	NFE2L2, C8B		
SRB0014	CD59	TAP1, NCF1, APOL1, IL17RA, SERPING1	BLNK, CD19, CD79A, POLR3F,		
SRB0015		FERMT1, CFHR1,			
SRB0016		C8A,	CD3G, CD81, CDE70		
SRB0017		C6	CD40, BLNK, CD19, CD79A, IGHM, IGKC	PAX5, IGHM, CD22, MS4A1, FCRL2, IGKC	IGKC, IGHM
SRB0018	S1PR1	CFI	PLEKHM1, C1QB		
SRB0019	TLR1,	IL12B, THBD	HELLS, TNFRSF13C,		



	POLE2, IFI44L		TOP2B, IL18BP, CD46, DKC1		
SRB0020		CD40, TTC7A, CARD9, C1QA, C1QB FCN3,	XIAP, ACD		
SRB0021		XRCC2			

Table represents findings from the various methods of outlier detection, with the final column showing any overlaps using these methods. Only overlap found was for the B-cell genes IGKC and IGHM in SRB0017, a patient without B-cells. This is expected due to the strong signal such a loss of expression would produce.

## 4.6 Discussion

The aims of this piece of research were to generate a bioinformatic pipeline for processing and interrogation of splicing of the transcriptome of PID patients to enhance diagnostic ability and variant filtering. We aimed to utilise multiple tissues to find novel signals which inform diagnosis. This section discusses the results of the assessment of alternative splicing within the cohort.

The approach to resolving aberrant splicing occurring in individuals employed (Mendelian RNAseq tool) was extremely sensitive, and able to detect positive controls. It was not however able to specifically identify the positive control as causal and clinical interpretation and follow up alternative investigation is still critical in interpreting the findings from this tool. As a result, further disease causing aberrant splicing events may have been detected, but lost in the lengthy, suboptimal filtering process. This component was a significant constraint, and filters had to be designed and optimised multiple times, with event types in mind, adding bias. Yet many event types would not have been identified. A number of events have been identified which will need further clinical interpretation and orthogonal experiments but due to time constraints, this has not occurred. Access to the genomes of these patients is not available currently and they will need to be recalled for WES/WGS. It is possible that as many as 6 causal splicing events may have been discovered. Based on the strong sequencing evidence, and the biological data surrounding the patterns on inheritance and the link between the S1P1R gene and the clinical phenotype, it seems likely that the deletion event found for patient SRB0018 in gene S1P1R, which represents a deletion of a G nucleotide from the splice donor site of exon 1 in SIPR1 gene (g.128delG, Chr1:101236992) is directly responsible for the phenotype. The CD59 splicing abnormality, could be contributing to the patients' phenotype, although a causative variant isn't clear. If another is present. Further investigation is needed for this, including DNA sequencing and variant interpretation.

Only whole blood, bulk RNAseq was conducted. PBMC and activated T-cell analysis was not conducted and furthermore, incorporation of allele specific expression as a metric into the analytic pipeline was not completed. By concentrating the work on PBMC's or T-cells, it is likely further diagnostic uplift will be achieved, due to the reduced noise and enhanced read depth on tissues or importance.

Other bioinformatic tools such as FRASER (387) now exist for assessing alternative splicing outliers. Although we were not able to test these alternative methods, the single package approach will likely expedite the process, although each package will have different benefits and drawbacks. It's likely that combination of multiple tools is likely to be more reliable in producing robust findings. Implementation of end-to-end tool which combines RNAseq data and genomic data to identify causal variants should be prioritised as a first pass in future work.

Assessment of alternative splicing using the syntax generated by Beryl Cummings proved to be sensitive enough to detect alternative splicing aberrations which cause disease. It however was not specific enough to be able to accurately determine which event was causal when using an example candidate with known causal variant. Furthermore, the filtering steps mean that it was likely that disease causing events would be lost in the process, and certain events such as intron inclusion, would not be identified at all.

This is due to the nature of the tool; the utilised algorithm can detect when a splice event takes place, but not when an annotated event does not take place. Due to this limitation and others, the authors put the overall diagnostic rate of disease in their own cohort, for splice altering variants at 35% (155).

## **4.7 Conclusion**

The area of transcriptomics in Mendelian disease is evolving and tools are being generated which take into account many of the nuances of this process. However significant challenges remain, and at best 25-40% success rate is found in literature for uplift of undiagnosed patients using RNASeq. Diagnostic uplift was had for a single sample using this splicing specific approach and experimental design, and critical information gained about other samples. This work also highlighted the importance of experimental design, process optimisation and tissue selection. It is likely that the

success rate cited in the literature was improved by selection of candidates with likely chance of success and the use of tissues with lower heterogeneity of cell type than whole blood.

Experimental design is critical and this process, including generation of an adequate control group, should be conducted with the limitations of the informatic tools in mind. The tool utilised is extremely sensitive, but lacks specificity, and is subject to false positives for inappropriate read splitting. A less sensitive approach or further fine tuning of filtering and alignment parameters is needed. In the case of PID, alternate tissue selection, and even immune challenge may be necessary to identify the causative variant, as splicing changes are an important component of immune response, and not always discernible at baseline status.

Inherent limitations in this project included significant batch effect and the use of 'unhealthy' controls. The tools are also unable to resolve variants which might not affect expression levels, splicing or allelic imbalance, but instead have deleterious effects on the gene product. Moreover, the limitations of the statistical model must be considered. If genes are low expressed, the changes may not be statistically significant. The approaches used for splicing also rely on complicated filtering strategies, none of which can capture all types of effect which have the potential to cause disease. This process is also then subject to cross-referencing of known PID genes or panels and as such, is limited in efficacy.

## **Chapter 5      Understanding transcriptomic differences in COVID-19 and Influenza: Results**

### **5.1      Introduction**

Infectious diseases elicit distinct and specific responses from the host immune system. The ability to respond to various pathogens relies on the ability of the cell to manage the production of RNA and proteins. To this end, the cell can produce a huge variety of proteins and control them within strict ranges. This is mediated in part by changes in gene expression and isoform abundance. To explore this further, the transcriptomic differences and congruence between the whole blood patients with COVID19 and Influenza was investigated. Exploratory data analysis was carried out initially, followed by investigation of differences in gene expression and alternative splicing. Tools for assessing gene expression are more mature and have a greater range of utilities as evidenced. However, differences in splicing may be more reliable, and robust, as well as being relatively untapped for therapeutic and diagnostic potential.

### **5.2      Aims**

The aims of this work were to quantitate these differences in transcriptome through both gene expression and differential isoform abundance metrics. This was performed in order to create a dataset for further analysis of age-related changes in host response between infections.

### **5.3      Cohort analysis**

We aimed to first assess the transcriptomic differences between the cohorts of patients admitted to hospital with either Influenza or Covid19. This analysis does not benefit from the presence of a control data set of patients without infection. Basic statistics were calculated for each cohort using R. This included the mean, median, variance, standard deviation, Shapiro-Wilk test for normality and t-test to compare the means (Table 5-1, Figure 5-1, Figure 5-2).

Both groups were similar in terms of the age of the patients, Covid19 patients were slightly older on average, but had less variance, standard deviation about the mean indicating the cohort followed a slightly more centralised distribution. Shapiro-Wilk W test for normality showed good evidence that both were normally distributed. Covid19 patients age may follow a slightly more normal distribution and the cohort is slightly older in general. Both groups have a slight preponderance of males compared with females (Figure 5-3, Figure 5-4), this increased proportion was also more prevalent in the Covid19 cohort compared with the influenza cohort. Other clinical differences include a slightly higher proportion of white British participants in influenza patients, (p-value  $1.12 \times 10^{-05}$ ), COVID19 patients had a higher prevalence of hypertension (p-value  $1.42 \times 10^{-02}$ ) and liver disease (p-value  $3.63 \times 10^{-02}$ ). COVID19 patients also had a longer symptom duration, higher respiration rate on admission (p-value  $2.79 \times 10^{-02}$ ), administration of supplementary oxygen (p-value  $6.81 \times 10^{-03}$ ), CRP levels (p-value  $1.73 \times 10^{-03}$ ), and lymphocyte numbers (p-value  $2.76 \times 10^{-02}$ ), they also had a longer duration of stay (p-value  $5.51 \times 10^{-10}$ ), and increased 30 day mortality (p-value  $4.42 \times 10^{-05}$ ) (388).

**Table 5-1 Covid and Influenza demographic analysis**

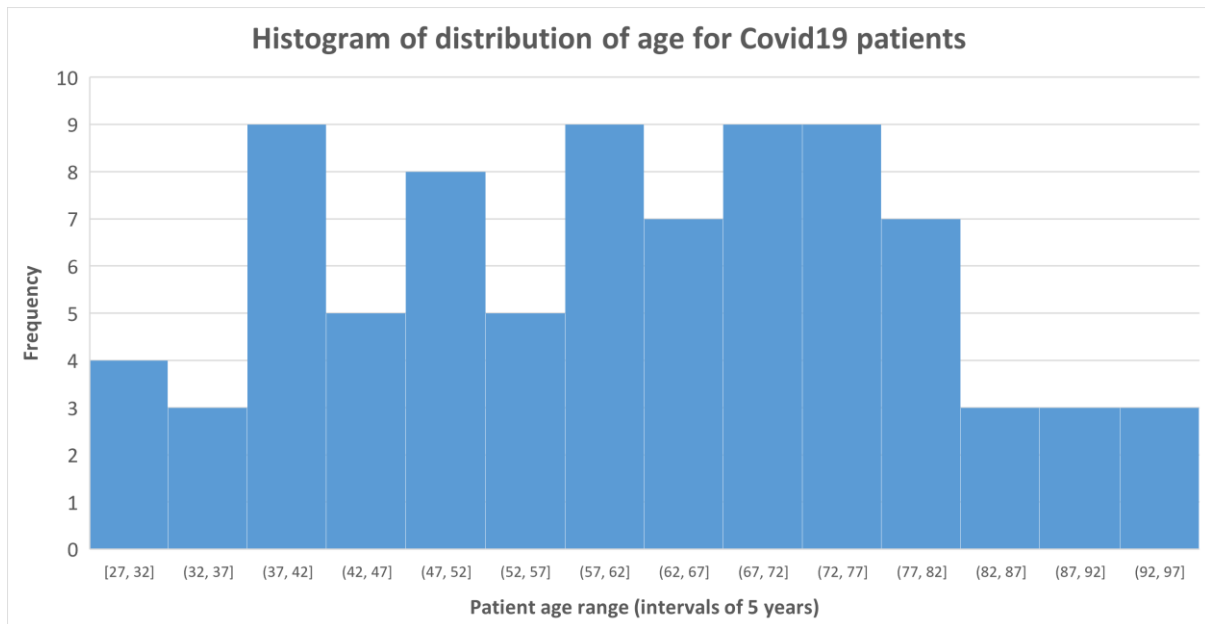
<i>Statistic</i>	<i>Covid19</i>	<i>Influenza</i>
<i>Mean age, trimmed mean (0.1)</i>	61.381 (61.265)	57.843 (57.970)
<i>Median age</i>	61.5	59
<i>Variance in age</i>	308.1905	337.890
<i>Standard deviation of age</i>	17.555	18.382
<i>Mean absolute deviation in age</i>	20.239	22.239
<i>Age range</i>	97-27 (70)	93-19 (74)
<i>Age quartiles</i>	47.75, 61.50, 73.25	42, 59, 73
<i>Shapiro-Wilk W test for normality</i>	(0.9793, 0.01949)	(0.97052, 0.0497)

*(W, p-val,  $\alpha=0.05$ )*

*T-test*

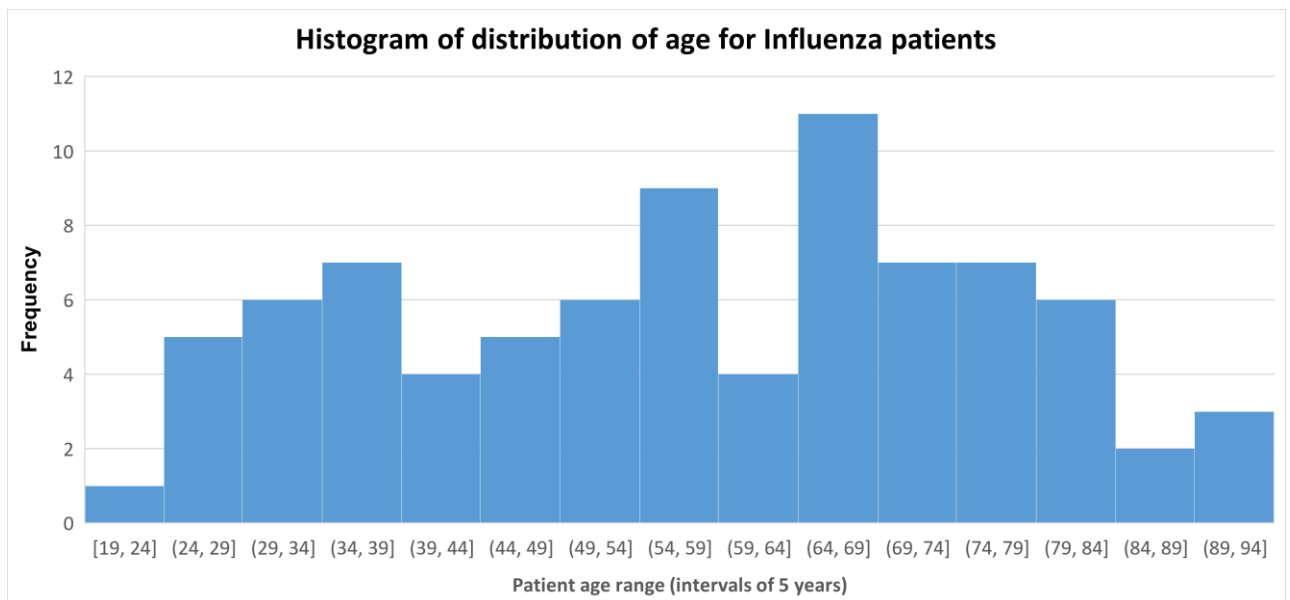
T=2716, df=164.45, p-value=0.2053,

Accept alternate hypothesis: true difference in means of age is not equal to 0.



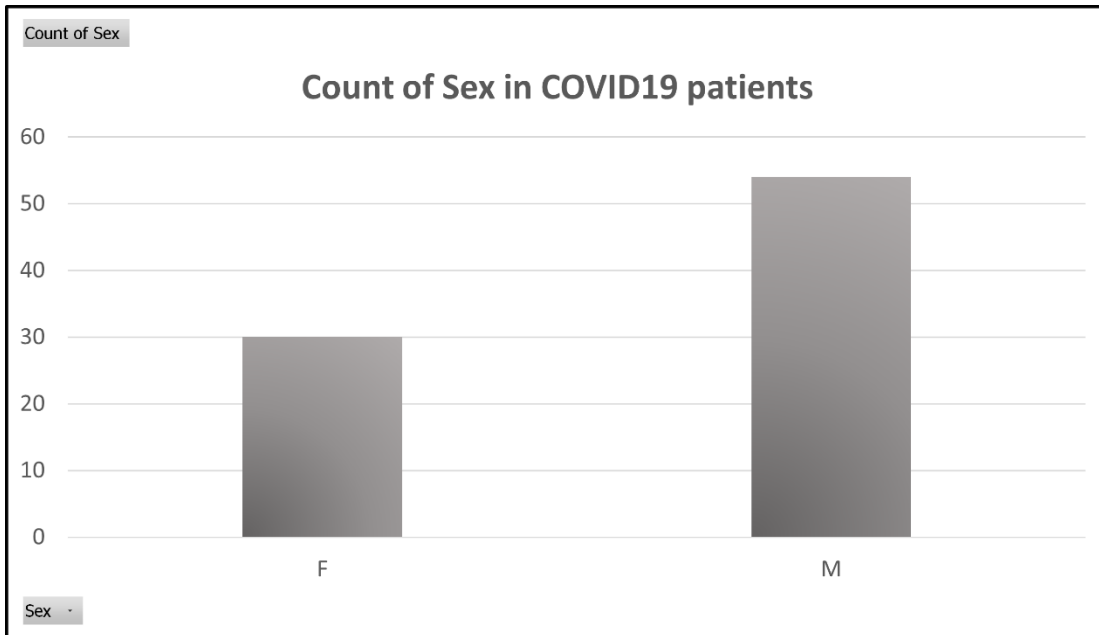
**Figure 5-1 Histogram of patient ages in Covid19 cohort**

Histogram of distribution of age for Covid19 patients, intervals for the histogram are 5-year sections. Peak frequency is N=9 for found in 4 demographic age categories and are concentrated in the central 50% of the distribution curve. The lowest frequencies exist at either of the approximate tails of the distributions.

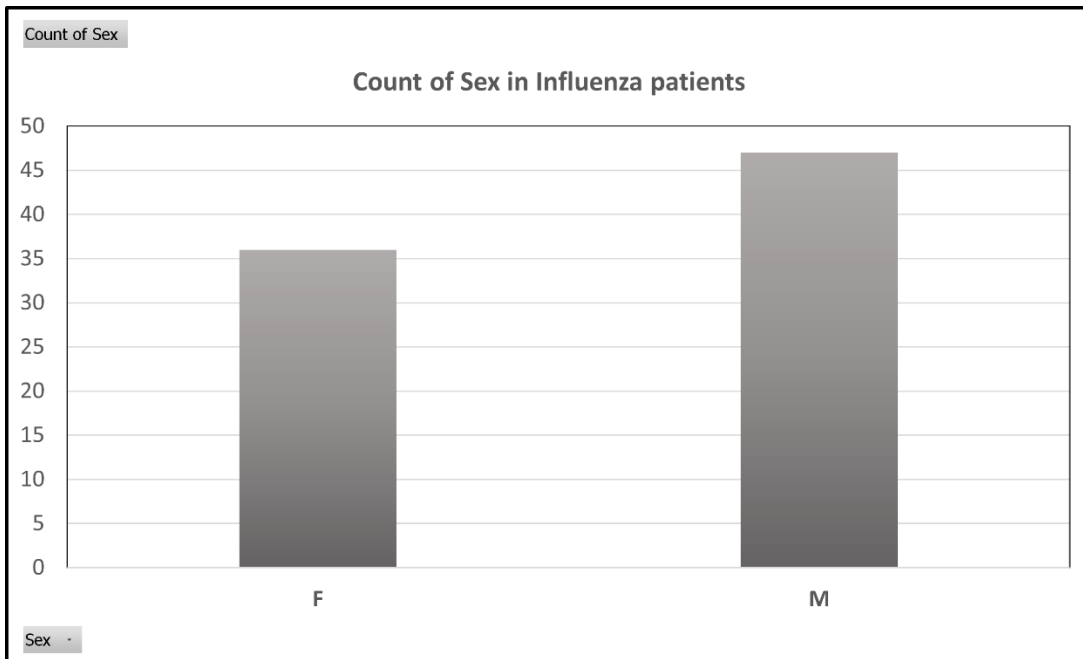


**Figure 5-2 Histogram of patient ages in the Influenza cohort**

Histogram of distribution of age for Influenza patients, intervals for the histogram are 5-year sections. Peak frequency is N=11 for found in 1 demographic age category (64-69), which is near the centre of the distribution curve. The lowest frequencies exist at either of the approximate tails of the distributions.



**Figure 5-3 Bar chart representing the sex distribution of the Covid19 cohort.**  
The cohort is comprised of a minority of females and a majority of males with males almost representing twice as much of the cohort as females.



**Figure 5-4 Bar chart representing the sex distribution of the Influenza cohort.**  
The cohort is comprised of a minority of females and a majority of males. In this cohort, the majority which males represent is comparatively small.

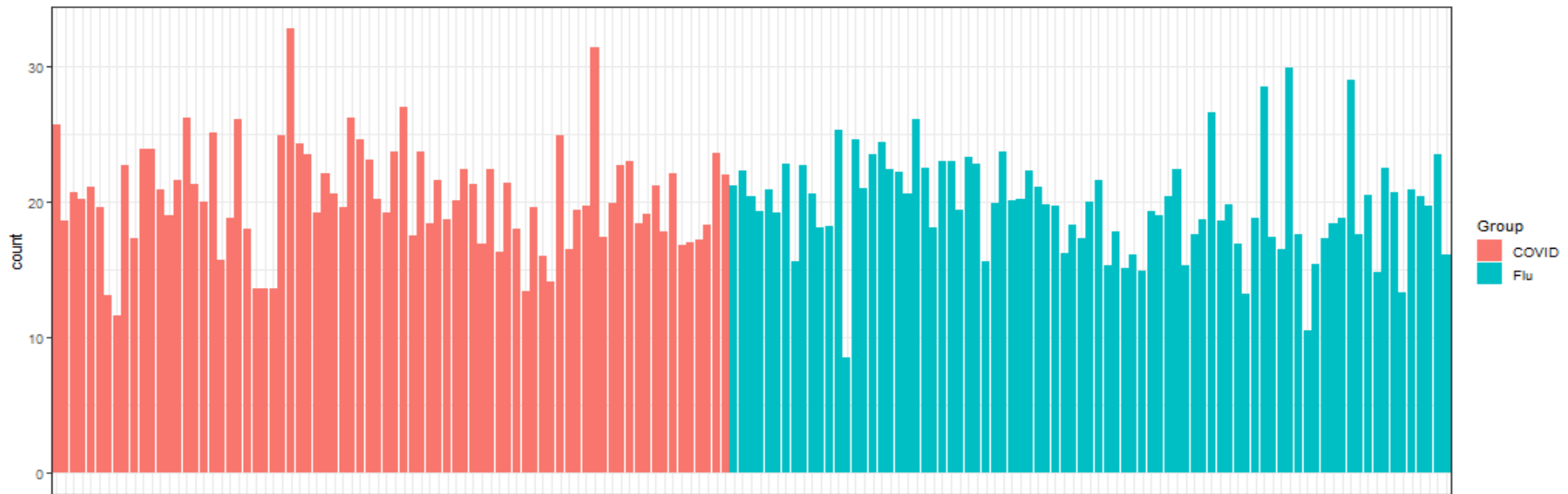


## 5.4 Exploratory Data Analysis with PcaExplorer

Raw gene expression reads were used for exploratory data analysis. This analysis revealed similar distributions of numbers of aligned reads between the COVID19 and Influenza infected patients. The mean reads were around 19M after low abundance reads were filtered, although there were outliers in both groups, in either direction, with lows of around 10M reads and highs of around 30M reads. Principal component analysis showed shows clear grouping of patients about principal component 1 (PC1). After investigation in the genes which most heavily weighted PC1 this was determined to be a result of sex specific differences in gene expression, and so the two groups, separated by horizontal space on the X axis, (PC1) were understood to represent the males and females of the groups.

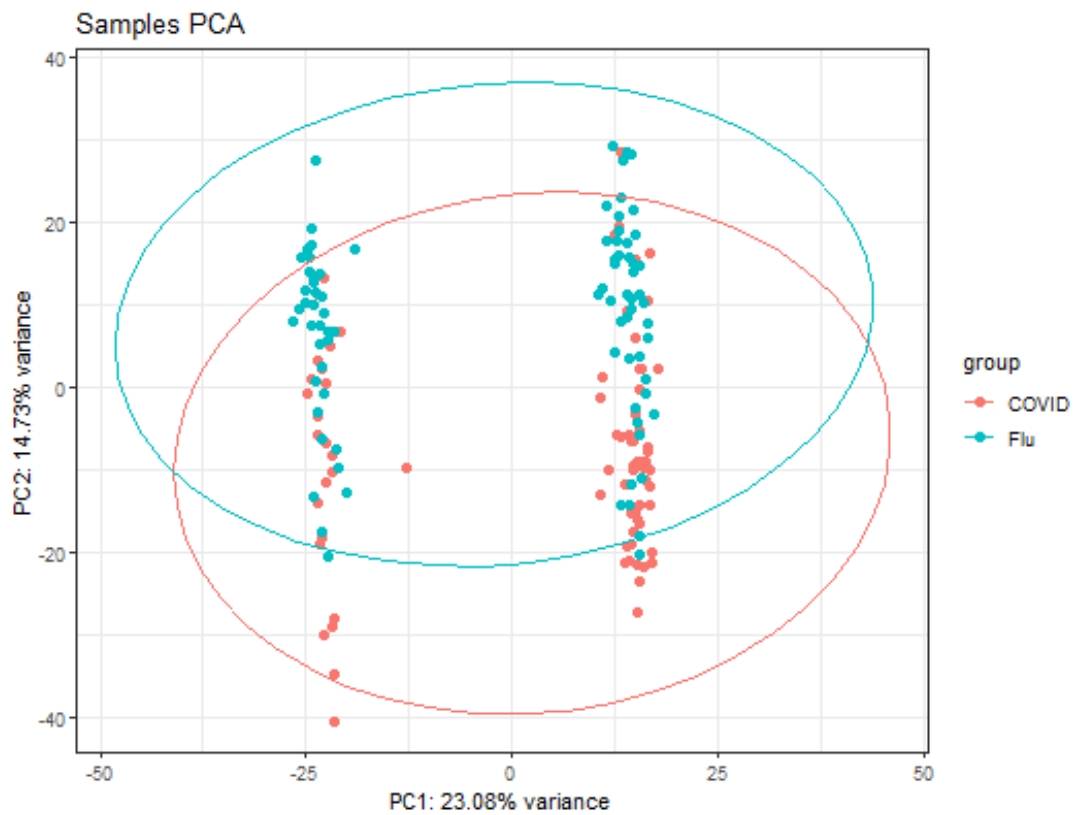
There also exists clustering of the two cohorts, with an incomplete separation of the groups whereby Influenza patients are primarily above Y axis point 0, and the Covid19 patients are primarily below Y axis point 0. 95% confidence intervals have a significant overlap between the groups. In addition, whilst there were general trends of separation about principal component 2, there was significant overlap of 95% C.I. circles indicating that statistically discerning between the two groups using principal components was not possible. There was however observed a general trend in the datapoints around PC2 which showed the Influenza patients tended to cluster higher of PC2 and Covid19 datapoints tended to cluster lower on the axis of principle component 2.

Number of million of reads per sample



**Figure 5-5 Exploratory data analysis: total aligned reads per samples for Covid19 and Influenza samples**

The table displays total aligned reads per sample after the data was cleaned using the pcaExplorer tool as per the parameters in methods section, the data appears to be consistent across both samples.

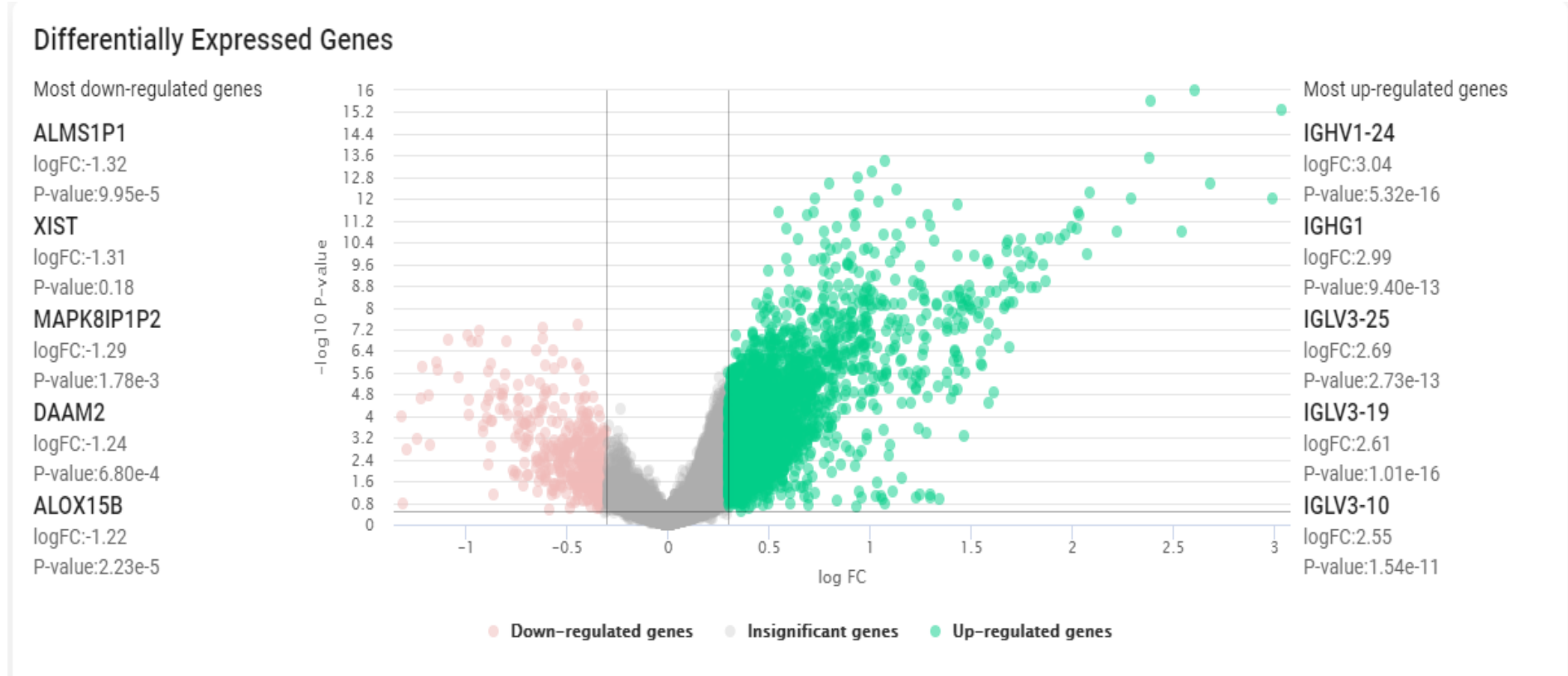


**Figure 5-6 Principal component analysis of Covid19 and Influenza cohort**  
 Principle component plot. Red points are COVID19 patients, blue points are Influenza patients. Circles indicate 95% confidence intervals. X axis is principal component 1, y axis is principal component 2.

## 5.5 Comparing and contrasting gene expression between COVID-19 and Influenza

Differential gene expression between the two cohorts was conducted using Pandaomics software. This software was chosen because of the comprehensive downstream analysis for target ID, and ease of use. The results were plotted in the diagram below (Figure 5-7). When considering genes which are differentially expressed, Pandaomics thresholds for magnitude are automatically assigned as 0.3-fold change, and an FDR-adjusted statistical significance level of 0.05. Top results in terms of magnitude were shown in the diagram on the left and right, for Influenza and Covid19 respectively.

Of the differentially expressed genes, the majority were higher in the COVID19 patients, moreover, those which had the greatest differences in expression were also expressed at higher levels in the COVID19 patients. It cannot be determined from these results alone if these expression changes were as a result of upregulation in one infection, or downregulation in another, only that they were significantly different and in which cohort expression was higher.



**Figure 5-7 Differentially expressed genes between Covid 19 and Influenza as determined by Pandaomics software.**

Genes which were not significantly different between the two infections are represented in grey. Those genes for which expression was significantly higher in COVID19 patients are shown in green. Those genes for which expression was significantly higher in Influenza patients are represented in red. At either side of the diagram, there exists the list of those genes which had the highest fold change, in some cases these were not statistically significant

The most differentially expressed genes were those related to immunoglobulins and therefore related to the adaptive immune response. It was noted that some of those genes which had the greatest magnitude of change, were not reaching statistical significance.

To enhance stringency, Bonferroni correction test was performed, where error rate (E) x test number (N) = p-value threshold. The acceptable error rate was set as 0.05, the number of genes which remained after Pandaomics filtering for low expression was 37794, and the calculated Bonferroni critical Value was determined to be 1.32296E-06. The results from Pandaomics were printed to a CSV file and then filtered for p-values less than 1.32296E-06 to give a list of the most differentially expressed genes.

The top differentially expressed genes from this list, ordered by magnitude after Bonferroni correction for both infections are displayed in the table below (Table 5-2). This process was then repeated using the P-value significance as the metric for ordering. The top genes from this method for both Covid19 and Influenza are found in the second table. Of note, when ordering by P-values, the top 220 genes all had logFC values which were positive, indicated that these were all genes which were expressed at higher levels in Covid19 patients (Table 5-3). To find the genes differentially expressed, which were higher in Influenza samples, and which were statistically significant the genes list was cleared of all positive values and ordered by P-value inversely.

**Table 5-2 Differentially expressed genes which are more highly expressed in Covid19.**

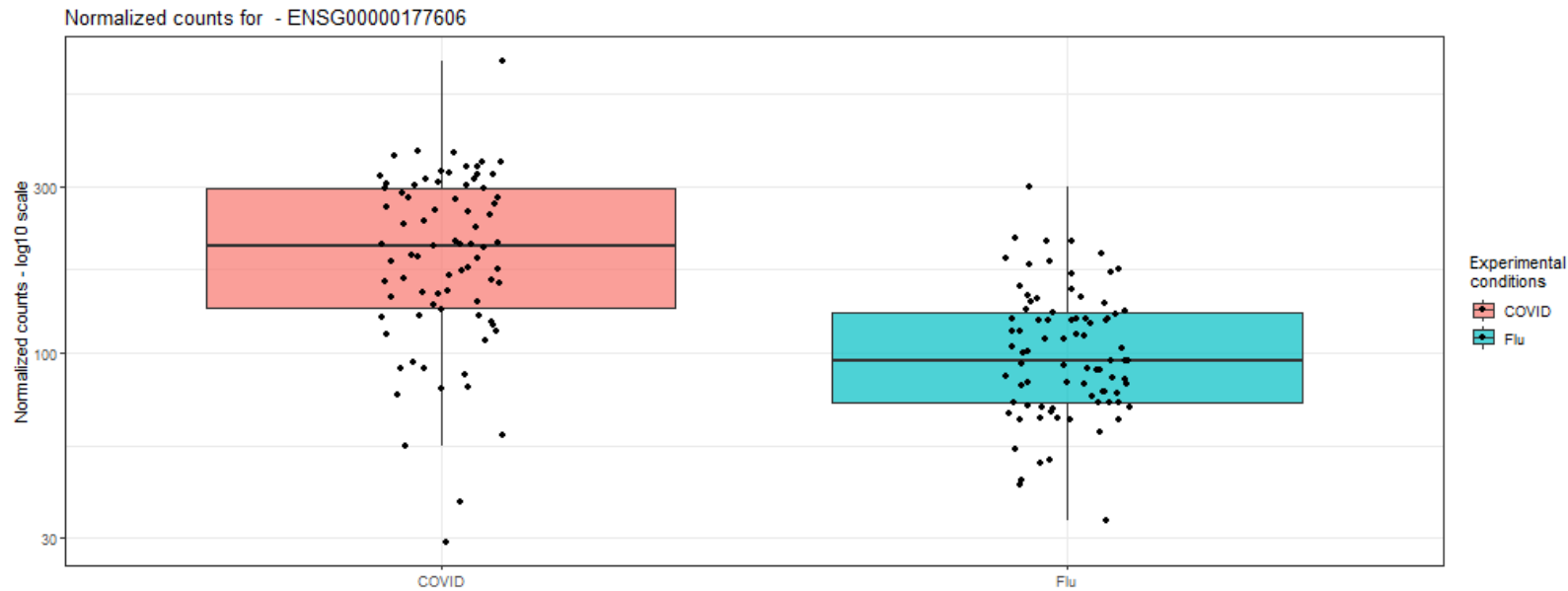
Higher in Covid19			Higher in Influenza		
Gene	logFC	P-value	Gene	logFC	P-value
<b>IGHV1-24</b>	3.035589645	5.32E-16	<b>PDZK1IP1</b>	-1.14418544	1.04E-06
<b>IGHG1</b>	2.991176278	9.40E-13	<b>CD163</b>	-1.086176454	1.63E-07
<b>IGLV3-25</b>	2.685154012	2.73E-13	<b>TTC26</b>	-0.992097771	1.03E-07
<b>IGLV3-19</b>	2.609259377	1.01E-16	<b>SLC5A9</b>	-0.970985937	1.77E-07
<b>IGLV3-10</b>	2.545445307	1.54E-11	<b>OR9A2</b>	-0.940140998	1.78E-07
<b>IGHV4-34</b>	2.388716259	2.44E-16	<b>RPS28P1</b>	-0.930112334	7.31E-08
<b>IGHG3</b>	2.384071424	3.25E-14	<b>GPER1</b>	-0.875533008	1.22E-06
<b>IGLV3-1</b>	2.291400869	9.92E-13	<b>CPM</b>	-0.800626927	1.69E-07
<b>IGLV6-57</b>	2.222716359	1.68E-11	<b>TAGLN</b>	-0.648318141	3.77E-07
<b>IGHV5-51</b>	2.08596257	5.57E-13	<b>NUDT16</b>	-0.618920455	1.44E-07
<b>IGLV3-27</b>	2.07332268	1.05E-10	<b>DLEU7</b>	-0.616446217	5.67E-08
<b>IGHV1-46</b>	2.035594432	3.81E-12	<b>GIMAP8</b>	-0.565445981	3.81E-07
<b>IGHV3-33</b>	2.033336754	2.98E-12	<b>SMIM25</b>	-0.523051552	1.06E-06
<b>IGLV4-69</b>	2.020974723	1.29E-11	<b>SIRPB2</b>	-0.450439926	1.21E-06
<b>IGHA1</b>	1.995244843	1.09E-11	<b>C1ORF162</b>	-0.447344044	4.17E-08
<b>JCHAIN</b>	1.966384828	1.94E-11	<b>MINDY3</b>	0.336067471	5.03E-07
<b>IGHV3-30</b>	1.943436086	3.02E-11	<b>ARF4</b>	0.340013552	9.97E-08
<b>IGKV1-27</b>	1.884704333	2.81E-11	<b>ARL15</b>	0.352548509	5.17E-07
<b>IGLV3-21</b>	1.86772077	1.05E-09	<b>PRUNE1</b>	0.361338574	7.86E-07
<b>IGHV2-5</b>	1.860441631	2.54E-10	<b>LINC01003</b>	0.367876437	7.87E-07

**Table 5-3 Differentially expressed genes in Covid and Influenza, ordered by P-value.**

Covid			Flu		
Gene	logFC	P-value	Gene	logFC	P-value
<b>IGLV3-19</b>	2.609259	1.01E-16	<b>C10RF162</b>	-0.44734	4.17E-08
<b>IGHV4-34</b>	2.388716	2.44E-16	<b>DLEU7</b>	-0.61645	5.67E-08
<b>IGHV1-24</b>	3.03559	5.32E-16	<b>RPS28P1</b>	-0.93011	7.31E-08
<b>IGHG3</b>	2.384071	3.25E-14	<b>TTC26</b>	-0.9921	1.03E-07
<b>JUN</b>	1.074533	4.07E-14	<b>NUDT16</b>	-0.61892	1.44E-07
<b>LAPTM4B</b>	1.013699	9.22E-14	<b>CD163</b>	-1.08618	1.63E-07
<b>LETM2</b>	0.93946	1.56E-13	<b>CPM</b>	-0.80063	1.69E-07
<b>HIST1H2BD</b>	0.798397	2.72E-13	<b>SLC5A9</b>	-0.97099	1.77E-07
<b>IGLV3-25</b>	2.685154	2.73E-13	<b>OR9A2</b>	-0.94014	1.78E-07
<b>RGS16</b>	1.135616	4.55E-13	<b>TAGLN</b>	-0.64832	3.77E-07
<b>IGHV5-51</b>	2.085963	5.57E-13	<b>GIMAP8</b>	-0.56545	3.81E-07
<b>CDC6</b>	0.946548	7.87E-13	<b>PDZK1IP1</b>	-1.14419	1.04E-06
<b>IGHG1</b>	2.991176	9.40E-13	<b>SMIM25</b>	-0.52305	1.08E-06
<b>HIST1H4H</b>	0.726816	9.66E-13	<b>SIRPB2</b>	-0.45044	1.21E-06
<b>IGLV3-1</b>	2.291401	9.92E-13	<b>GPB1</b>	-0.87553	1.22E-06
<b>HIST1H2BG</b>	1.04129	1.21E-12	<b>ZNF366</b>	-0.60365	1.42E-06
<b>CYP27A1</b>	1.436561	1.64E-12	<b>OR52N1</b>	-1.21331	1.46E-06
<b>IGHV3-33</b>	2.033337	2.98E-12	<b>RNASE6</b>	-0.55935	1.64E-06
<b>SLC44A1</b>	0.545948	3.11E-12	<b>MEF2A</b>	-0.43806	1.7E-06
<b>ISYNA1</b>	0.723301	3.13E-12	<b>PFKFB2</b>	-1.13615	1.88E-06



The top gene which was not directly a part of the humoral adaptive immune response was JUN. JUN (c-JUN) is extensively documented to be important in viral replication and mediating host inflammatory responses (389, 390), and has been implicated in cytokine storm (391). C-JUN N terminal kinase (JNK) and C-JUN are implicated in formation of inflammasomes by inducing transcription factor NF- $\kappa$ B (392). The gene which was the most differentially expressed while having the highest expression in Influenza was PDZK1IP1, a gene encoding a cargo protein which carries membrane proteins from the endoplasmic reticulum. When overexpressed inflammation is triggered (393). When expression of c-JUN is compared between the cohorts, there is again, a large amount of overlap (Figure 5-8) despite this being a gene with a high degree of differential expression.



**Figure 5-8 - Expression of c-JUN in Covid19 and Influenza cohorts**

Figure shows box-and-whisker plots produced by pcaExplorer looking specifically at the differences in expression with Covid and Influenza cohorts. The majority of Covid cohort have between 150-300 counts on this scale, with influenza patients having between 70 and 150 reads mapped to the JUN gene.

Automatic Pathway analysis was produced using the Pandaomics software. The analysis showed that 5 of the top 6 highest ranked pathways by iPanda score, in Covid19 cohort were involved in Synthesis of phosphatidylinositol-phosphates or 'PIPs' (Table 5-4). Pandaomics was also able to produce a visualisation of the most upregulated node within that pathway (Figure 5-9).

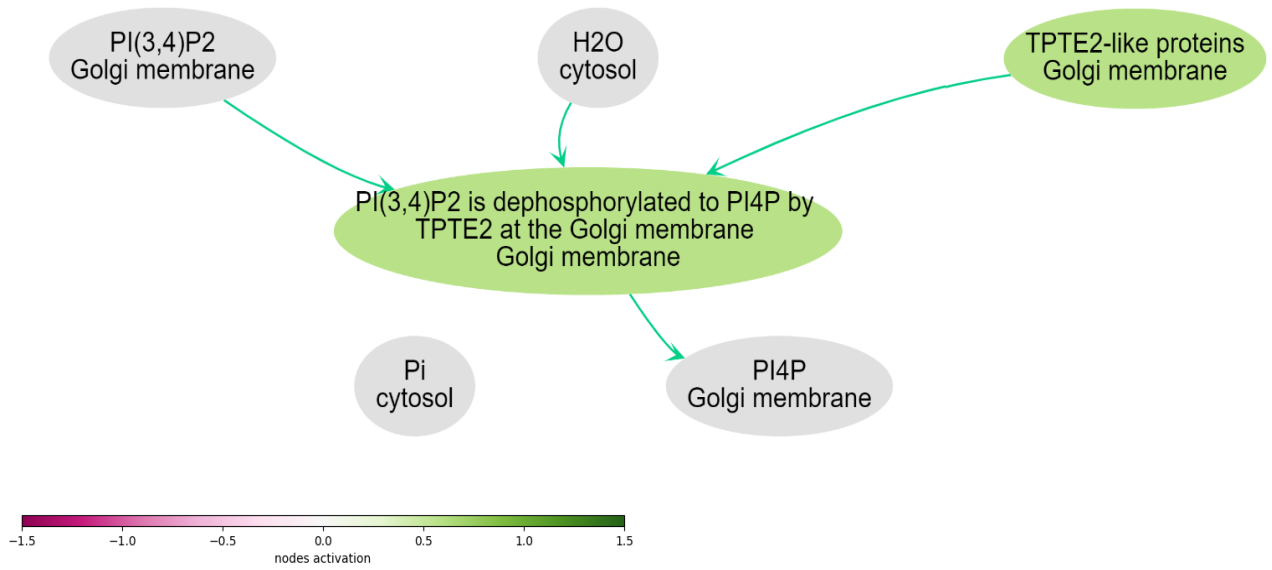
The pathways most highly expressed in Influenza cohort when compared with the Covid19 cohort was the disease-associated 'CD163 mediating an anti-inflammatory response'. CD163 has been previously associated with negative outcomes and hospitalization in Influenza patients (394).

The nodes organization and upregulation intensity were shown (Figure 5-10). IL-6 and IL-10 trigger CD163 re-localisation from the nucleolus to the plasma membrane. At this point CD163 can be cleaved and become soluble during macrophage activation. High levels of soluble CD163 (sCD163) are associated with inflammatory disease (395). CD163 expression was also plotted in box-and-whisker plots for the two cohorts to visualize the expression differences (Figure 5-11

**Table 5-4 Table of most highly expressed pathways in Covid19 cohort compared with Influenza pathways.**

Pathway	Main cellular process	iPanda activation	p-value (FDR corrected)	Database	Genes number
Synthesis of PIPs at the Golgi membrane	Metabolism	1.289	3.78E-06	Reactome	18
Synthesis of PIPs at the late endosome membrane	Metabolism	1.206	5.78E-06	Reactome	11
PI Metabolism	Metabolism	1.200	5.01E-06	Reactome	84
Synthesis of PIPs at the plasma membrane	Metabolism	1.163	8.37E-06	Reactome	53
Mtb iron assimilation by chelation	Disease	1.120	2.21E-06	Reactome	1
Synthesis of PIPs at the early endosome membrane	Metabolism	1.114	7.36E-06	Reactome	16
FCGR activation	Immune System	1.084	1.32E-06	Reactome	92
Classical antibody-mediated complement activation	Immune System	1.034	1.13E-06	Reactome	86
CDK-mediated phosphorylation and removal of Cdc6	Cell Cycle, DNA Replication	0.947	7.87E-13	Reactome	73
Role of phospholipids in phagocytosis	Immune System	0.941	5.69E-06	Reactome	105
G2-M DNA replication checkpoint	Cell Cycle	0.895	1.97E-07	Reactome	5
Events associated with phagocytolytic activity of PMN cells	Immune System	0.820	1.60E-06	Reactome	2
Fcgamma receptor (FCGR) dependent phagocytosis	Immune System	0.794	2.13E-05	Reactome	166
Scavenging of heme from plasma	Vesicle-mediated transport	0.758	3.03E-05	Reactome	90
Latent infection - Other	Disease	0.672	3.10E-04	Reactome	4

responses of Mtb to phagocytosis					
CD22 mediated BCR regulation	Immune System	0.670	9.44E-05	Reactome	68
Creation of C4 and C2 activators	Immune System	0.659	1.22E-04	Reactome	94
Phosphorylation of Emi1	Cell Cycle	0.617	1.25E-05	Reactome	6
Synthesis of PIPs at the ER membrane	Metabolism	0.606	3.76E-05	Reactome	5
Mitotic Metaphase-Anaphase Transition	Cell Cycle	0.578	2.13E-06	Reactome	2
HDACs deacetylate histones	Chromatin organization	0.559	1.68E-06	Reactome	60
p53-Independent DNA Damage Response	Cell Cycle	0.555	1.00E-06	Reactome	52
p53-Independent G1-S DNA damage checkpoint	Cell Cycle	0.555	1.00E-06	Reactome	52
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A	Cell Cycle	0.555	1.00E-06	Reactome	52
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	Immune System	0.545	2.95E-04	Reactome	93



**Figure 5-9 The most upregulated molecular expression node in the Covid19 cohort**

Figure shows the individual genes which are upregulated, their location, and amount of upregulation along with the direction of influence of expression. Green arrows indicate a downstream process regulated by the previous gene. Colour of the gene circle indicates fold change of expression in the gene.

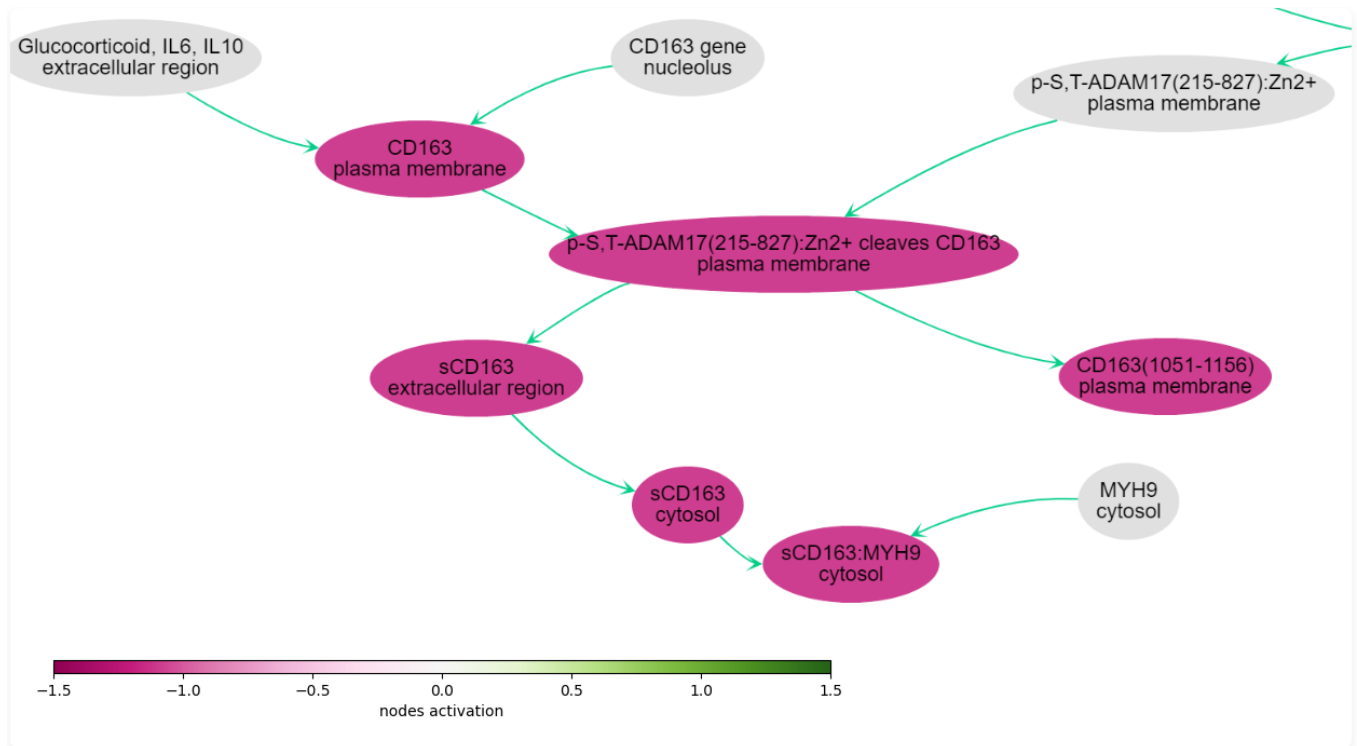
**Table 5-5 -Pathways with most upregulation in Influenza cohort when compared with Covid19 cohort.**

Pathway	Main cellular process	iPanda activation	p-value (FDR corrected)	Database	Genes number
CD163 mediating an anti-inflammatory response	Disease	-0.866	2.85E-06	Reactome	9
Interleukin-18 signaling	Immune System	-0.104	2.68E-03	Reactome	8
Interleukin-33 signaling	Immune System	-0.080	0.042	Reactome	3
Inactivation of CDC42 and RAC1	Developmental Biology	-0.057	9.71E-03	Reactome	8
Xenobiotics	Metabolism	-0.044	0.09	Reactome	24
Interleukin-36 pathway	Immune System	-0.041	0.145	Reactome	7
PAOs oxidise polyamines to amines	Metabolism	-0.029	4.35E-03	Reactome	2
Activated NTRK2 signals through FYN	Signal Transduction	-0.029	0.077	Reactome	7
LTC4-CYSLTR mediated IL4 production	Disease	-0.027	0.042	Reactome	7
Apoptotic cleavage of cell adhesion proteins	Programmed Cell Death	-0.026	0.286	Reactome	11
Interconversion of polyamines	Metabolism	-0.026	5.26E-03	Reactome	3
Collagen degradation	Extracellular matrix organization	-0.024	0.116	Reactome	64
The AIM2 inflammasome	Immune System	-0.023	0.086	Reactome	3
Negative regulation of TCF-dependent signaling by WNT ligand antagonists	Signal Transduction	-0.022	0.154	Reactome	15
ERK-MAPK targets	Immune System, Signal Transduction	-0.021	0.094	Reactome	22
The IPAF inflammasome	Immune System	-0.021	0.038	Reactome	2

Understanding transcriptomic differences in COVID-19 and Influenza: Results

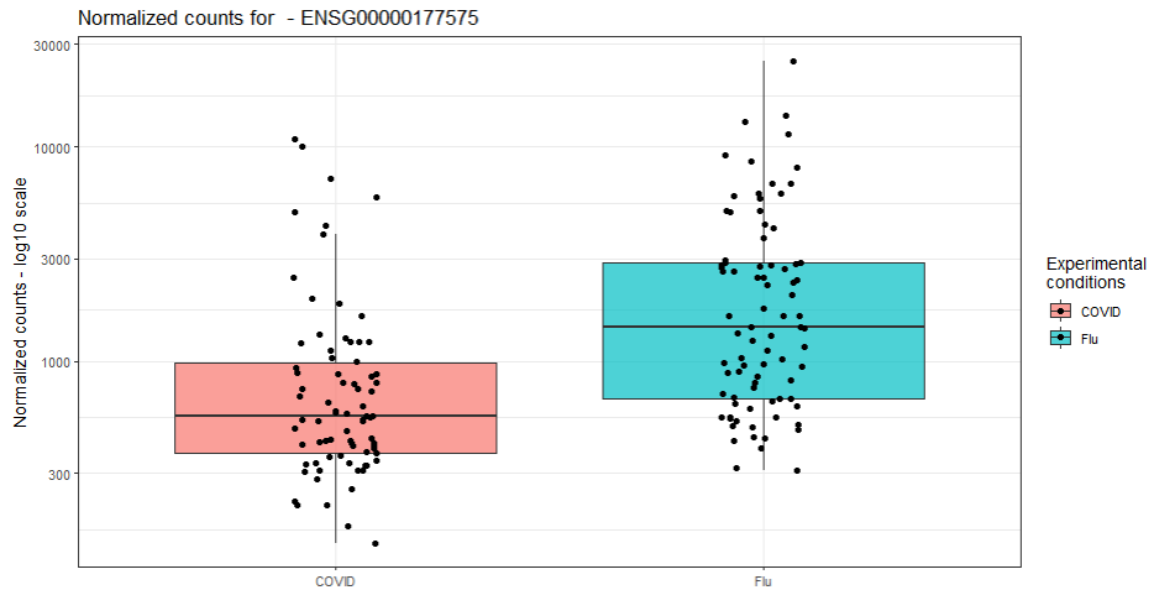
Amine Oxidase reactions	Metabolism	-0.018	0.022	Reactome	4
Relaxin receptors	Signal Transduction	-0.016	0.087	Reactome	8
cGMP effects	Hemostasis	-0.012	0.077	Reactome	16
Nuclear Events (kinase and transcription factor activation)	Signal Transduction	-0.012	0.05	Reactome	61
Scavenging by Class A Receptors	Vesicle-mediated transport	-0.011	0.026	Reactome	19
Interleukin-27 signaling	Immune System	-0.010	0.048	Reactome	11
Metallothioneins bind metals	Cellular responses to external stimuli	-8.33E-03	0.228	Reactome	11
Phase 2 - plateau phase	Muscle contraction	-8.15E-03	0.14	Reactome	25
Inhibition of nitric oxide production	Disease	-7.23E-03	9.43E-03	Reactome	3





**Figure 5-10 The most highly expressed molecular node in the Influenza cohort when compared to the Covid19 cohort**

Figure shows the individual genes which are upregulated, their location, and amount of upregulation along with the direction of influence of expression. Green arrows indicate a downstream process regulated by the previous gene. Colour of the gene circle indicates fold change of expression in the gene, colour coding is reversed here, i.e., magenta indicates expression is higher as the colour coding is comparative to the Covid19 cohort.



**Figure 5-11 CD163 expression in Covid19 and Influenza patients**

Figure shows box-and-whisker plots produced by pcaExplorer looking specifically at the differences in expression with Covid and Influenza cohorts. The majority of Covid cohort have between 400-1000 counts on this scale, with influenza patients having between 700 and 3000 reads mapped to the CD163 gene.

## 5.6 Comparing and contrasting isoform expression between COVID-19 and Influenza

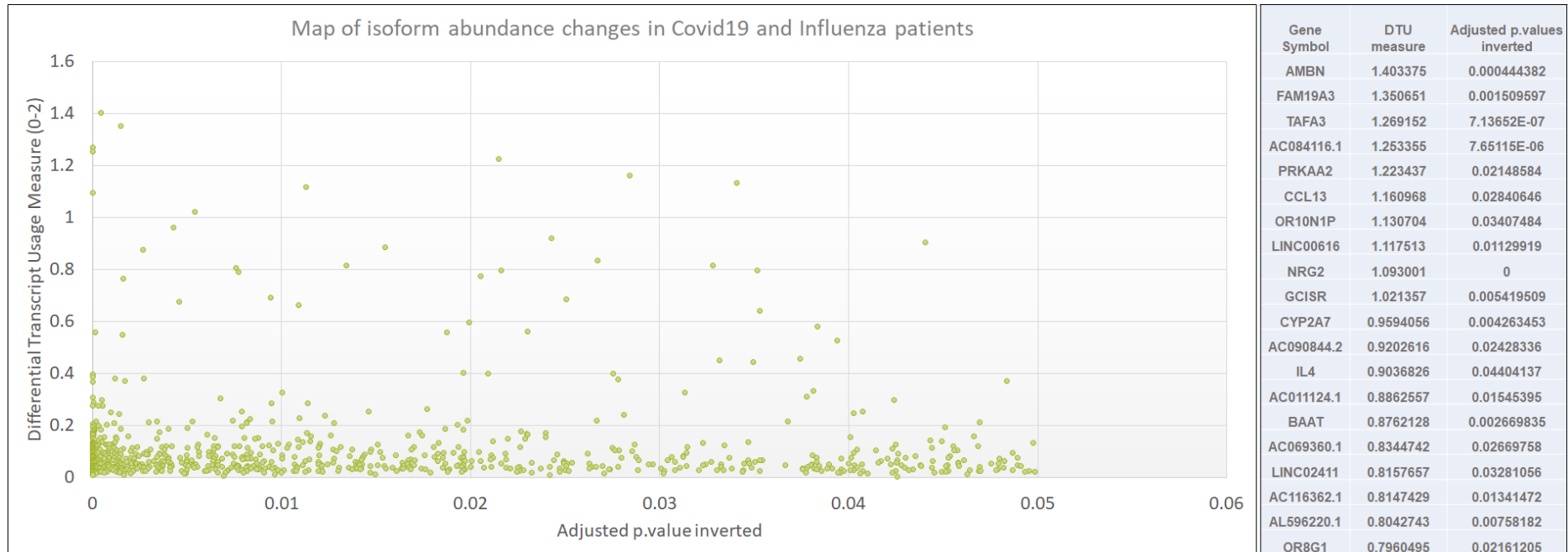
BANDITS is a software package which performs differential splicing analysis between groups. This statistical utility makes it suitable for cohort studies, unlike the tools used in the PID groups, which are aimed at outlier detection and aren't equipped with the same statistical tools. For group comparison, BANDITS outperforms alternative packages in terms of true positives vs false discovery rate, and gives transcript and gene level results, useful for follow up pathway analysis which other packages do not. BANDITS produces two sets of results, those at gene level and those at transcript level, and both are output to .txt. format. For differences in expression, p-values are produced, and adjusted using Benjamini-Hochberg correction. In addition, inverted p-values are provided which vary only when the dominant transcript remains the same in both groups despite abundance changes. This is calculated by taking the square root of p-value which results in an inflated value. This is performed to give priority ranking in results to those results in which differential splicing results in the change of a dominant transcript.

Sample 26 was unsuccessful during alignment stage for isoform abundance and excluded from further study. Initially the output file contains read counts for 207,749 possible transcripts. Many of these transcripts have none, or very few reads assigned. When filtering was performed for those transcripts which had more than or equal to 80 reads total across all samples there were 123,782 remaining. This step was performed separately from downstream analysis.

In total 1694 genes experienced statistically significant (adjusted p-value < 0.05) changes in isoform abundance. Of These 1475 experiences a change in the dominant isoform between Covid and Influenza groups, and 219 did not. When adjusting p-values for inversion based on dominant transcript changes, 843 genes had statistically significant changes in isoform abundance.

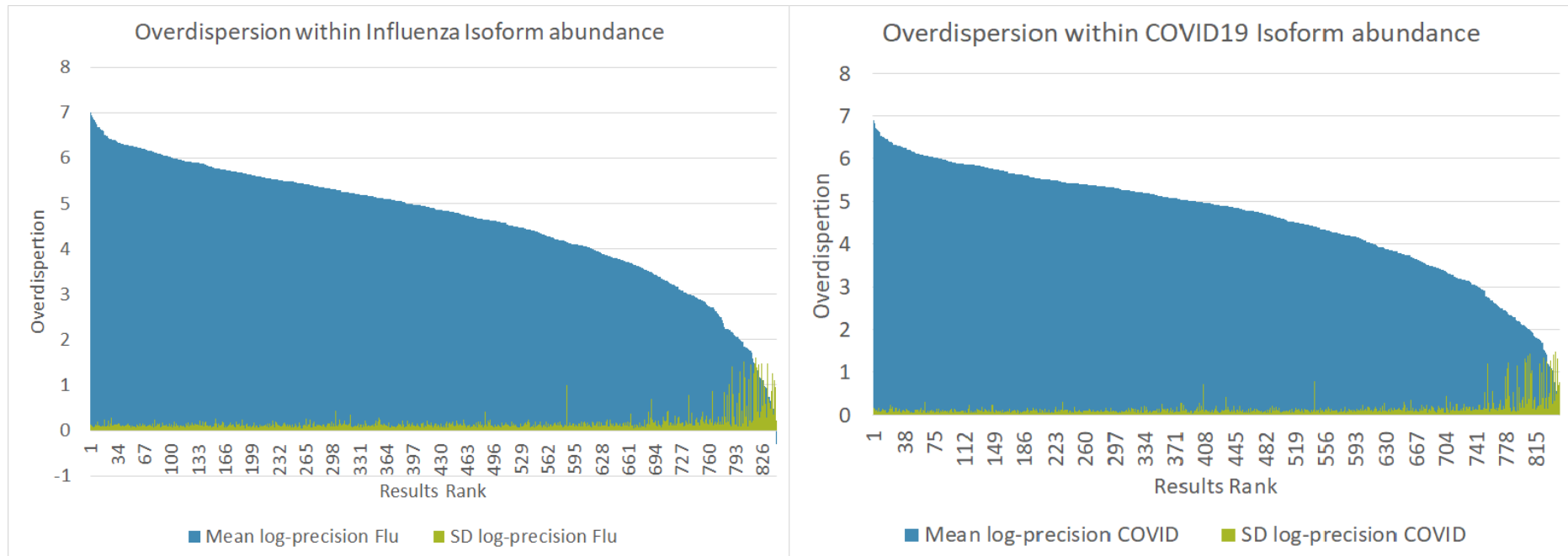
DTU change is a metric created by the BANDITS software which is similar to fold change in differential expression analysis. The DTU measure represent the sum of absolute difference between two transcripts between the groups. A value of zero represents proportions are statistically identical in each group, whereas a value of represents different transcripts are always used between the groups. Those results which had an adjusted, inverted P-value of less than 0.05 are plotted in Figure 5-12. Most transcripts which had a significant alteration in abundance had relatively low change in

DTU – under 0.2 The highest value was for the AMBN gene which had a DTU score of 1.403375. The relative abundances of isoforms in the top 20 genes were displayed in stacked bar plots (Figure 5-14, Figure 5-15, Figure 5-16, Figure 5-17). The precision of the outputs is calculated by modelling the degree of over dispersion or variation beyond what would be expected by the model. To do this the tool uses a Dirichlet-multinomial model to give a mean precision value, derived from precision values of each sample for that gene which are bases on the levels of overdispersion, and this is complemented by a standard deviation about the mean. As would be expected the high mean accuracy is met with low standard deviation, however when the overdispersion is higher, leading to less mean accuracy overall there is an increase in standard deviation about the mean (Figure 5-13).



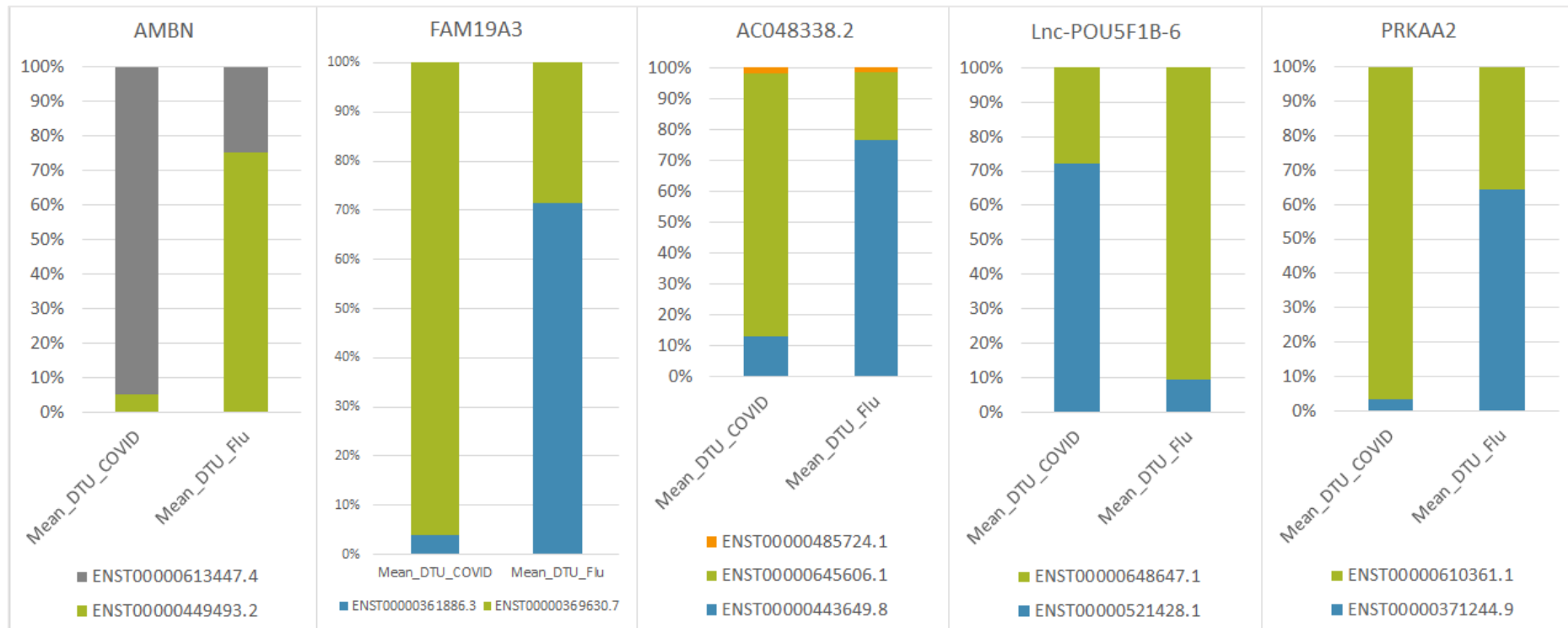
**Figure 5-12 Genes with the greatest change in transcript use between Covid19 and Influenza**

Figure displays all the genes (green circles) which had statistically significant (adjusted p-value inverted < 0.05) changes in isoform abundance. The top 20 genes are identified in the table to the right of the diagram. AMBN has the highest DTU value and so is the gene which has the most robust changes of transcript use when comparing the transcriptomes of the two cohorts.



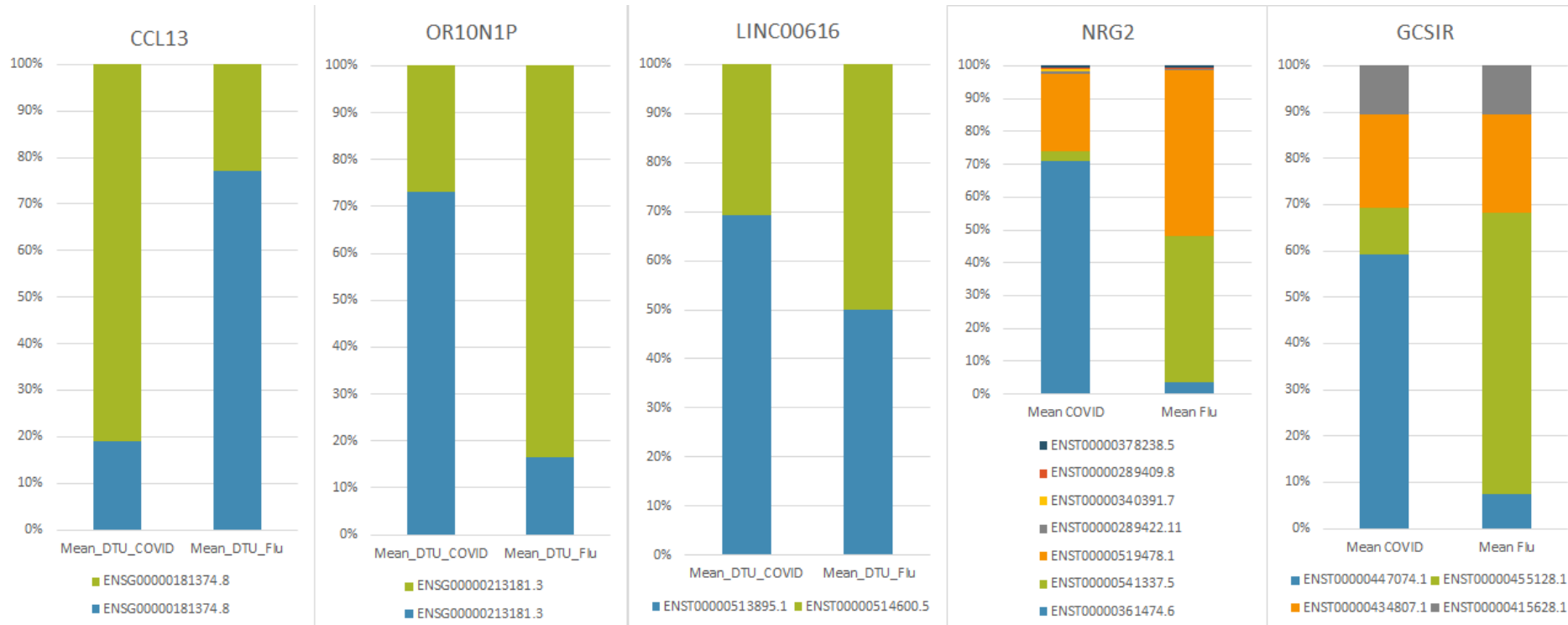
**Figure 5-13 Precision of the model in determining differential transcript use.**

Figure shows the precision of the model in estimating differential transcript use. The genes with the greatest level, and therefore more robustly discernible instances of differential transcript use, predictably produced more accurate results. Less robust changes appeared further down the ranking of genes as overdispersion caused reduced accuracy and increased standard deviation for both infection types.



**Figure 5-14 Genes with top DTU 1-5**

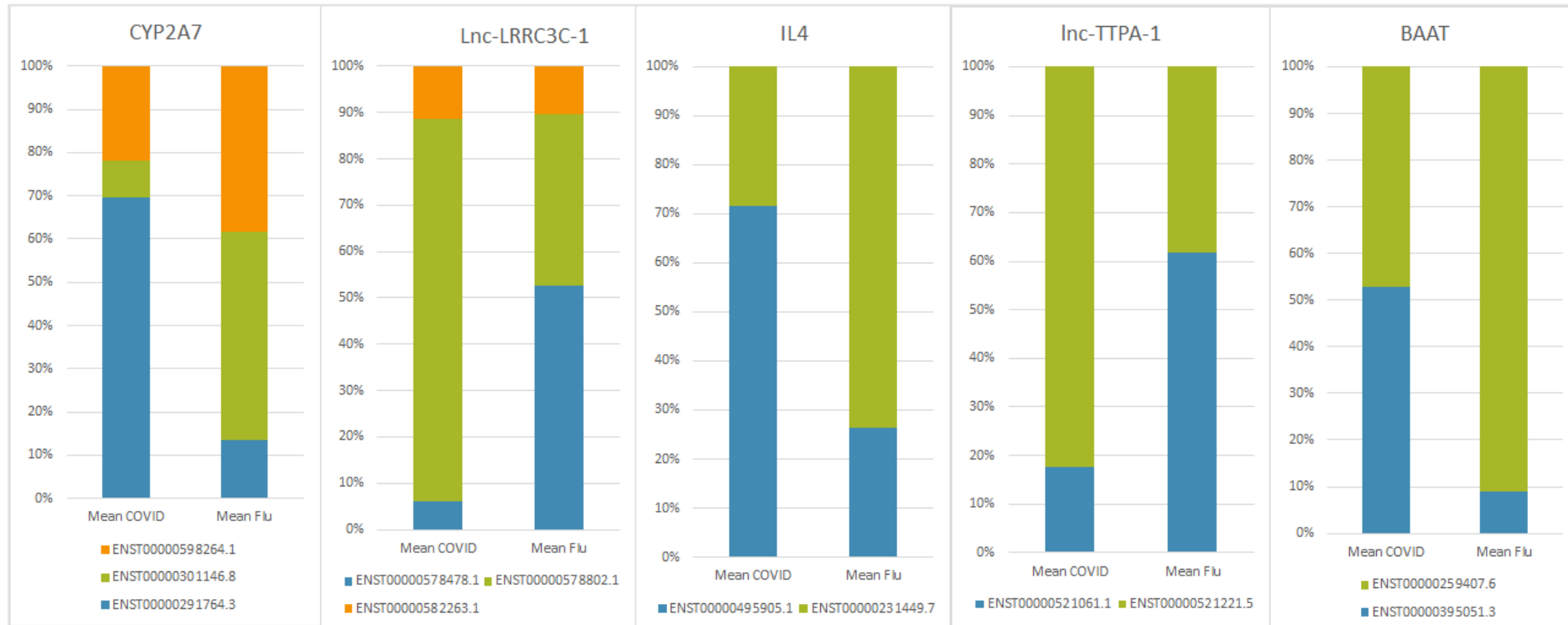
Figure shows mean differential transcript use depicted as abundance for genes ranked 1-5, being AMBN, FAM19A3, AC048338.2, Lnc-POU5F1B-6, PRKAA2. The Primary isoform is altered in all examples. Most of the genes have only 2 isoforms which have appeared at detectable levels after filtering was performed, this is with the exception of AC048338.2, which had three isoforms, although the third isoform had relatively constant expression through both examples, each colour represents a different isoform for which the identity is present in the key.



**Figure 5-15 Genes with top DTU 6-10**

Figure shows mean differential transcript use depicted as abundance for genes ranked 6-10, being CCL13, OR10N1P, LINC00616, NRG2, GCSIR. The Primary isoform is altered in all examples. CCL13, OR10N1P, LINC00616 have only 2 isoforms which have appeared at detectable levels after filtering was performed. NRG2, GCSIR had MULTIPLE isoforms, with levels varying significantly, each colour represents a different isoform for which the identity is present in the key.





**Figure 5-16 Genes with top DTU 11-15**

Figure shows mean differential transcript use depicted as abundance for genes ranked 6-10, being CYP2A7, Lnc-LRRC3C-1, IL-4, Inc-TTPA-1, and BAAT. The Primary isoform is altered in all examples. IL-4, Inc-TTPA-1, and BAAT have only 2 isoforms which have appeared at detectable levels after filtering was performed. CYP2A7, Lnc-LRRC3C-1 have 3 isoforms, each colour represents a different isoform for which the identity is present in the key.

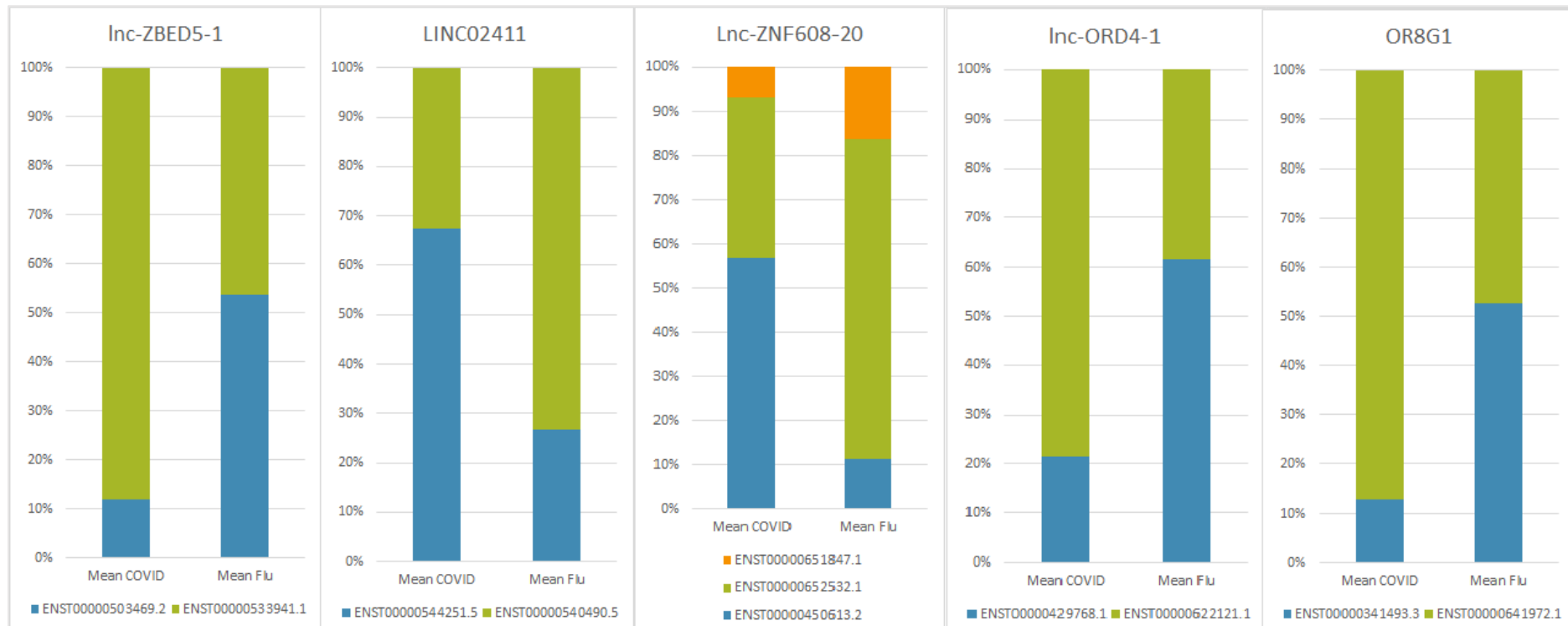


Figure 5-17 Genes with top DTU 16-20

Figure shows mean differential transcript use depicted as abundance for genes ranked 6-10, being CYP2A7, Lnc-LRRC3C-1, IL-4, Inc-TTPA-1, and BAAT. The Primary isoform is altered in all examples. IL-4, Inc-TTPA-1, and BAAT have only 2 isoforms which have appeared at detectable levels after filtering was performed. CYP2A7, Lnc-LRRC3C-1 have 3 isoforms, each colour represents a different isoform for which the identity is present in the key.

## 5.7 Analysis of differential gene expression and differential transcript use in Covid19 and Influenza

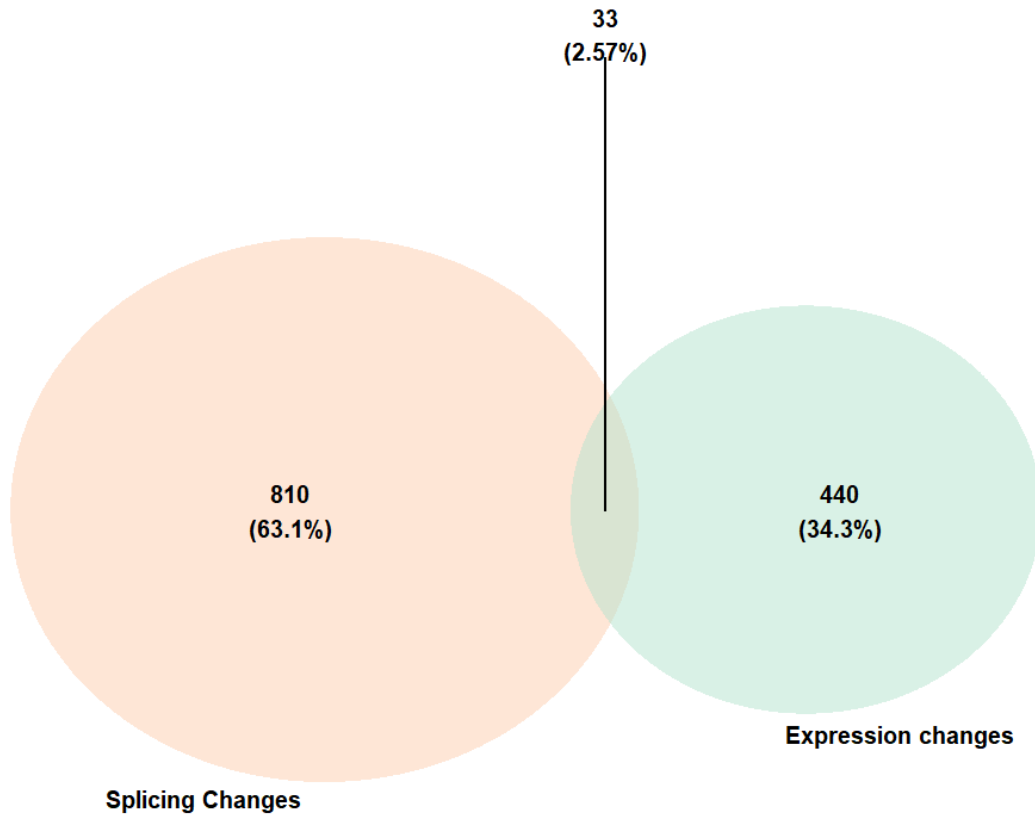
The list of genes which underwent statistically significant differential transcript use as determined by BANDITS were compared with the genes which underwent statistically significant differential gene expression as determined by Pandaomics. Genes were loaded into ToppGene software, as the free online tool requires no setup or associated code, and outputs gene lists for functional enrichment. and analysis was conducted to ascertain the pathways and biological processes which were enriched. The lists of genes, pathways, and processes for both were saved to a .txt file and Venn diagrams were produced to visualise the comparison.

In total 843 genes identified as undergoing differential transcript use from BANDITS program were loaded into the ToppGene software, of which a majority of 820 were recognised within the ToppGene database and were included in subsequent analysis. 436 of 472 of the differentially expressed gene list produced by the Pandaomics software were recognised for ToppGene. Venn diagrams were produced to compare the lists of genes, pathways and processes. The ToppGene software used Bonferroni correction with threshold of 0.05 to calculate enrichment in biological processes and pathways analysis.

Around 63%, or 2/3 of all the genes had differential transcript use only and around 34% experienced changes only in gene expression (Figure 5-18). 33 genes, or a tiny 2.57% of the genes experienced both differential transcript use and differences in gene expression. Conversely, the pathway analysis showed that a greater number of pathways were affected by the gene expression changes, whereas the large number of changes as a result of alternative splicing were concentrated in fewer pathways (Figure 5-19). 47 pathways were enriched for the differential gene expression and 14 for the differential transcript use. Therefore, around 3/4 of all genes affected, were affected by splicing changes and these were concentrated in only 1/4 of the pathways affected. Some of the top pathways affected were those of the innate immune system. 126 biological processes were enriched in the list of differential gene expression between the two infections, and 93 biological processes were enriched in the list of genes affected by alternative splicing. Only 10, or 4.37% were affected by both (Figure 5-20). However, this still strongly supports the notion that alternative splicing and differential gene expression affect separate cellular processes. To visually represent the

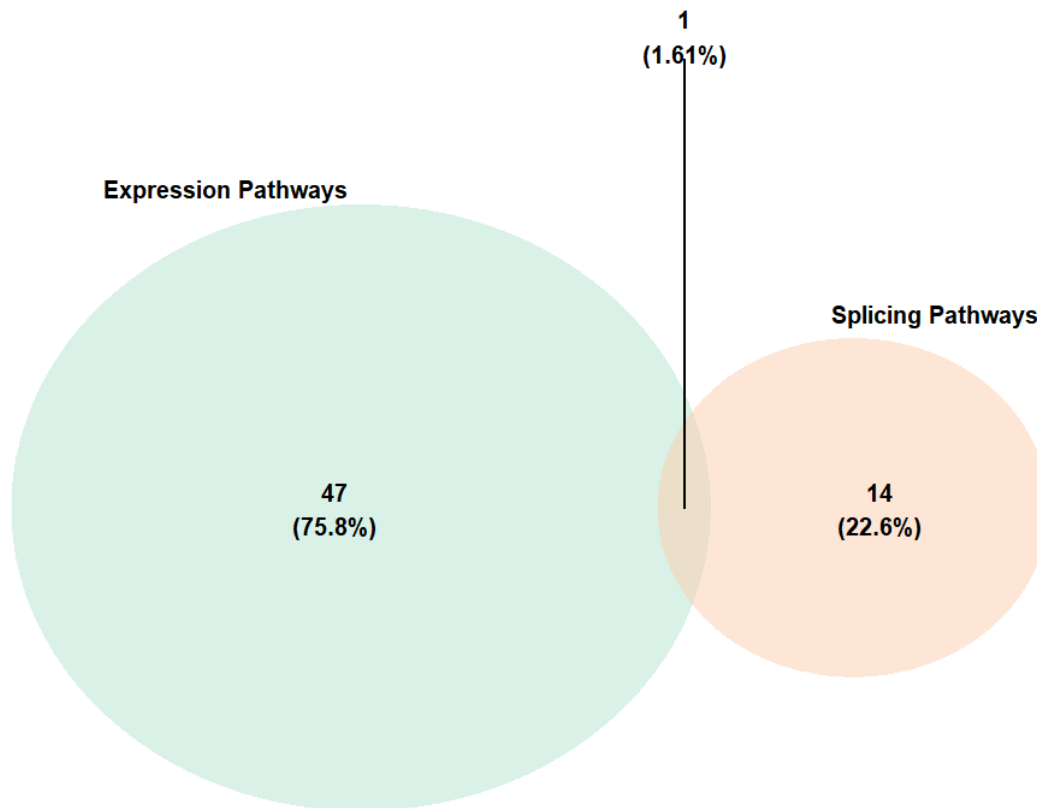
biological processes which were enriched in the two processes, DGE and DTU, the gene ontology or 'GO' terms were copied into REVIGO software and Tree Maps were produced for both DTU and DEG which show and group the biological processes affected. Data shown in the image reflects strong differences in humoral immunity, B-cell involvement and complement activation (orange sections). There are also strong signals apparent in processes involved in cell cycle control (purple), implicated in mitosis and clonal expansion, this is also reflected in DNA replication signals, which are also enriched (bright orange). The red and blue section of the Tree Map show strong enrichment of pathway involved with endocytosis, exocytosis, organelle fission and vesicle budding.

The biological processes associated with changes in alternative splicing and as such, differential transcript use, actually show less alignment with those processes canonically considered to be part of the immune process. The large blue section is heavily implicated in cell regulatory processes, including cellular metabolism, subcellular localisation, molecular metabolism and molecular catabolism and anabolism. The aqua sections highlight strong enrichment biological processes which are involved with the physical structure of the cells, cytoskeletal organisations, protein assembly, actin formation, vesicle and chromatin organisation. Light blue sections show enrichment for processes involved with cellular secretions and exports, and the green sections catabolism and autophagy. Some canonical immune processes are shown to be enriched, red and purple sections, but these appear to be a minority.



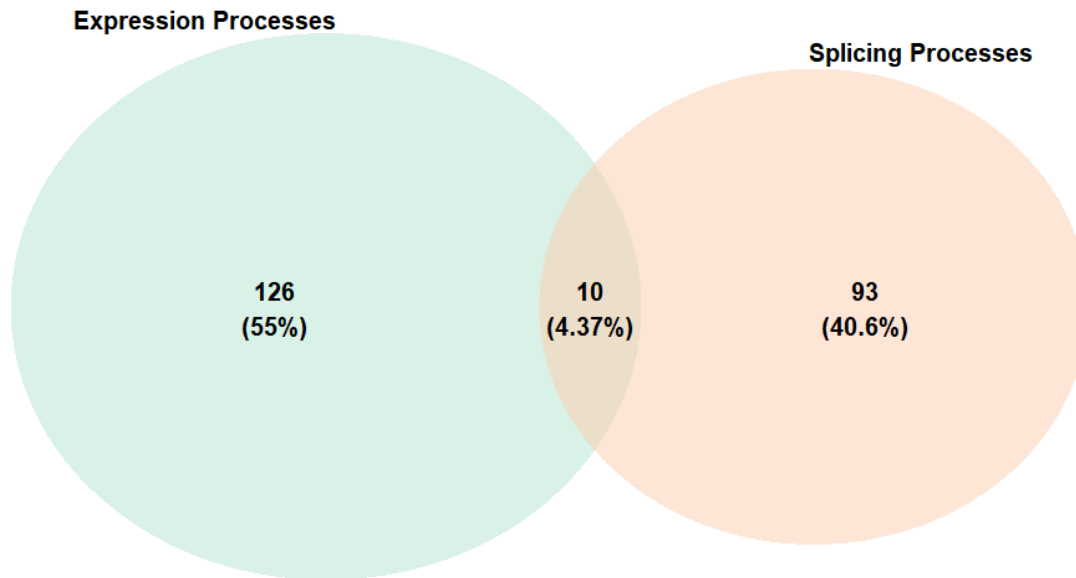
**Figure 5-18 Genes affected by changes in expression and splicing.**

Venn Diagram showing the number of genes which have significant differences in transcript use (peach) which number 810 after Bonferroni correction, and those which have significant differences in gene expression (mint) 440 after Bonferroni correction. Very few genes (33 or 2.57%) have both differences in transcript use as a result of splicing, and gene expression changes.



**Figure 5-19 Pathways enriched for differential expression and differential transcript use.**

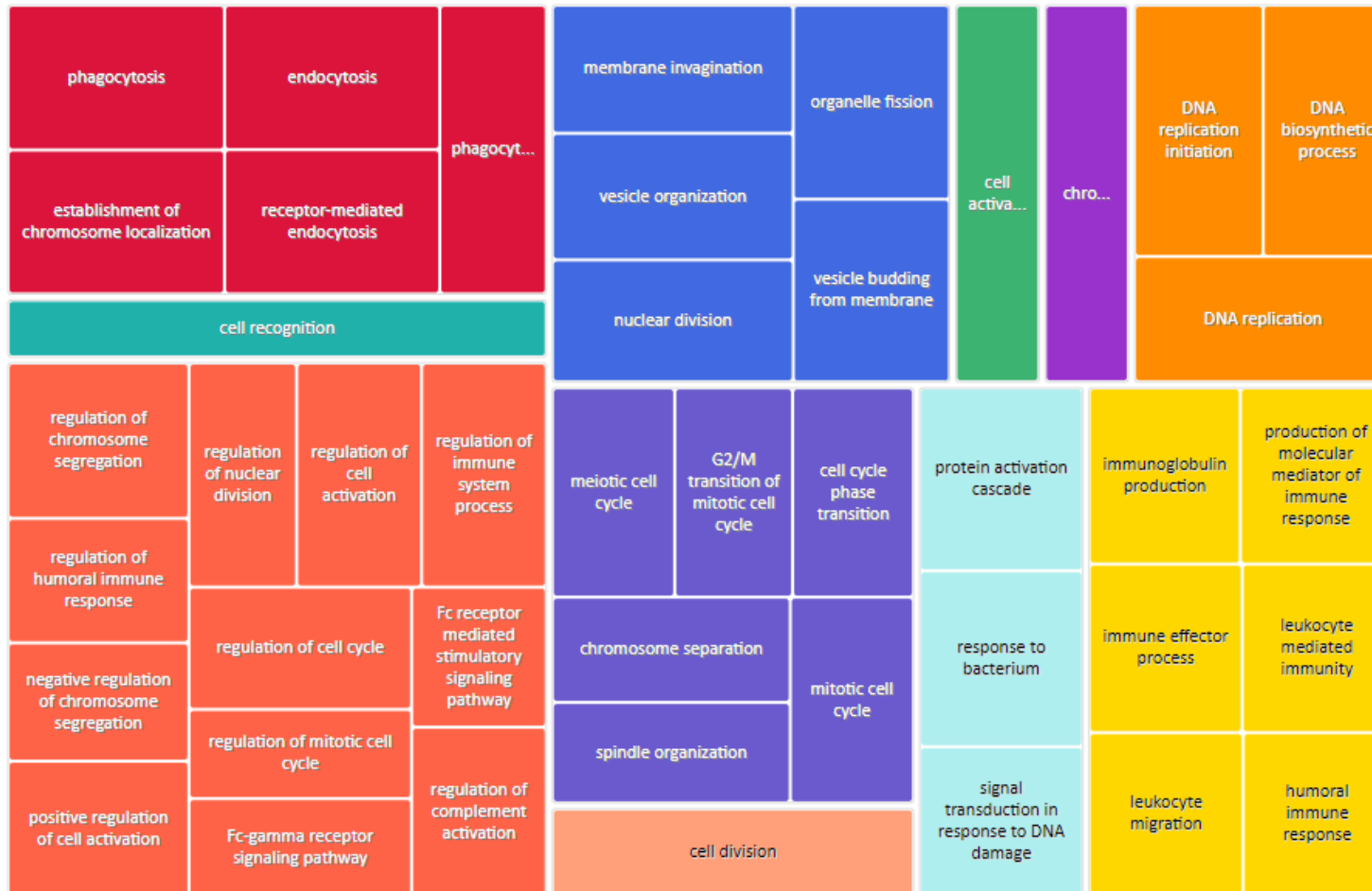
Venn Diagram showing the number of pathways which are enriched for differences in transcript use (peach) which number 14 (Bonferroni correction threshold = 0.5 %) after, and those which have significant differences in gene expression (mint) 47 (Bonferroni correction threshold = 0.5 %). Only 1 pathway (cell cycle) or 1.61% is found in both datasets.



**Figure 5-20 Biological processes enriched for differential gene expression and differential transcript use.**

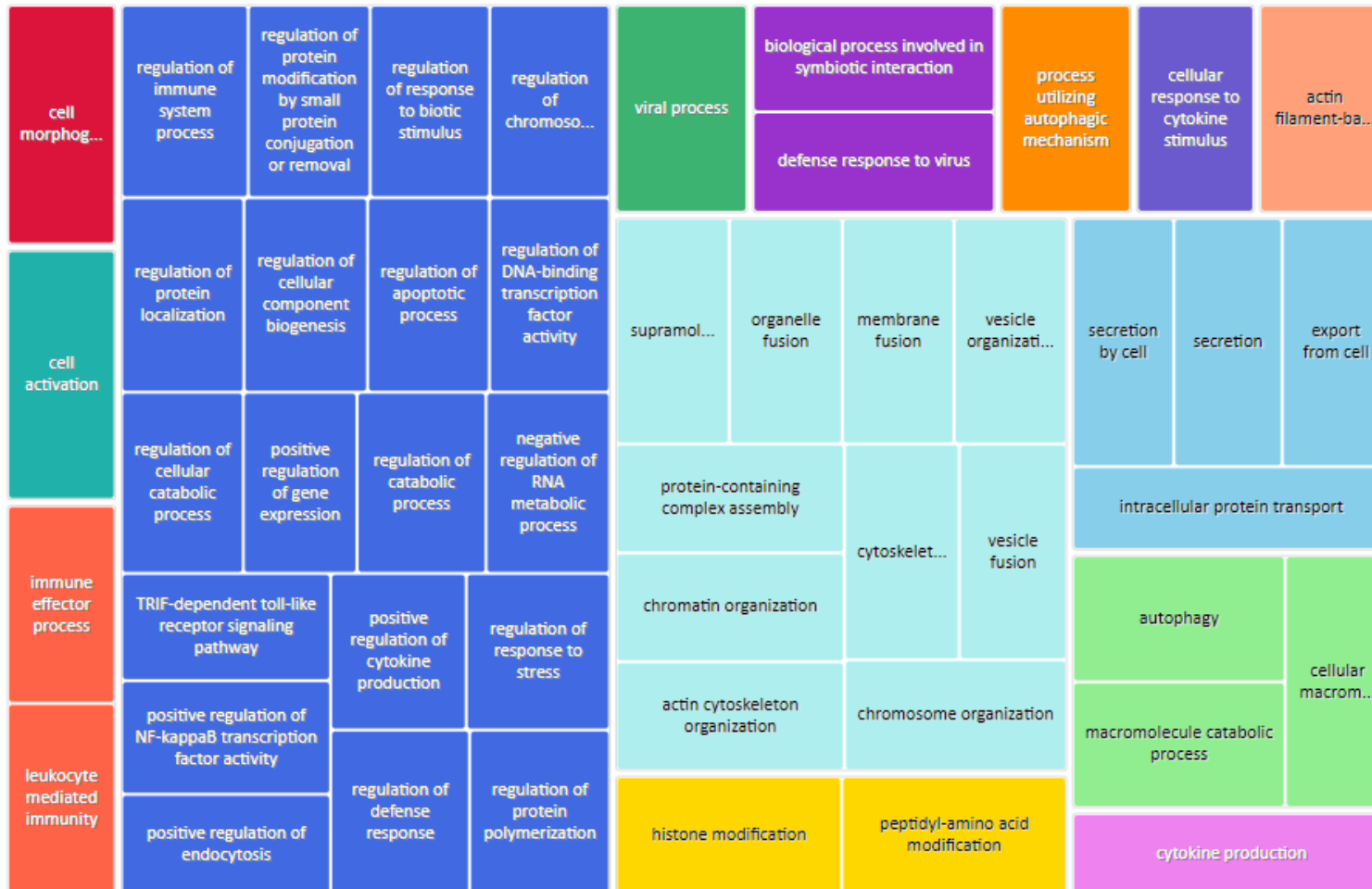
Venn Diagram showing the number of biological processes which are enriched for differences in transcript use (peach) which number 93 (Bonferroni correction threshold = 0.5 %), and those which have significant differences in gene expression (mint) 126 (Bonferroni correction threshold = 0.5 %). Only 10 pathways (or 4.37%) are found in both datasets.





**Figure 5-21 Biological processes which experience enrichment from DEG between Infectious disease.**

Tree Map showing biological processes enriched in lists of DEG between infections and how they cluster, each colour represents groups of processes which cluster under general terms.



**Figure 5-22 Biological processes which experience enrichment from DTU between Infectious disease.**

Tree Map showing biological processes enriched in lists of DEG between infections and how they cluster, each colour represents groups of processes which cluster under general terms.

## 5.8 Discussion

### 5.8.1 Findings

The work has shown the major differences in the transcriptome of infected patients are a result of differential splicing, gene expression is affected to a lesser degree. This work also demonstrates that the pathways affected by the two processes are also predominantly discrete.

The aims of this chapter were to build a transcriptomic profile of patients who were infected with influenza and SARS-CoV-2 and begin to characterise the differences between these infections. It also aimed to demonstrate the importance and utility of alternative splicing in the immune system response. In order to do this, changes in isoform abundance or 'differential transcript use' were quantified to capture the often-unseen changes in the transcriptome.

The cohorts were statistically very similar in age and sex, with the SARS-CoV-2 cohort being slightly older and marginally less dispersed. Exploratory data analysis using PCA showed that the groups were divergent but had overlapping dispersions and confidence intervals around the principal components by which the most differed. This indicates that whilst the immune responses to viral infections are distinct, as the literature suggests, personal genetics and environmental factors likely play a role in determining the response and as such PCA analysis can completely deconvolute which cohort the data comes from.

In addition to this, the transcriptome is affected by several factors including the direct effects of the pathogen on cellular RNA production, but also the innate and subsequent adaptive immune response and finally the pathological progress of resulting disease. As cohort recruitment was opportunistic, the total time between first infection and immune response cannot be accurately determined. Symptom presentation time is not uniform. It is not possible to know at which stage of the immune response the patients were at when the sample was taken, and so the transcriptomes will reflect different degrees of influence from the innate and the adaptive immune response. Moreover, the hospitalisation of these patients means there are likely both seen and unseen factors contributing to their clinical presentation including co-morbidities.

Notwithstanding these design limitations, results in differential gene expression which were both statistically robust and biologically sensible were obtained. Consistent with other literature, the

immunoglobulin genes appeared at much higher expression levels in patients harbouring SARS-CoV-2 infections. This is a significant finding as it demonstrates that the B-cell response in SARS-CoV-2 is vigorous even compared with other infectious disease models. The very high expression of a number of immune related genes in the SARS-CoV-2 cohort could be a result of the relative novelty of the infection to the immune system of the cohort. Over time, many viruses tend to become less virulent to hosts due to selection pressures (396) and thus stimulate a less robust immune response.

Although slightly different methods were utilised for the differential gene expression, work published by Legebeke J et al., had striking similarity in the lists of differentially expressed genes as a result the GO terms from gene expression data (388). However, a vast amount of response to infection is mediated by splicing and resulting the isoform abundance changes, which are rarely explored as a way to understand host responses to infection, or indeed infection pathophysiology. Certainly, the software, databases and online data infrastructure is less well developed and therefore some interoperability is lost when trying to compare findings between these methods.

Also seen was a much greater expression of genes involved in synthesis of PIPs in the SARS-CoV-2 cohort. The data from the REVIGO Tree Maps for DEG is concordant with the Pandaomics analysis which shows enrichment signals involved in phosphatidylinositol-phosphate pathways. This biological process is crucial in viral propagation, some RNA viruses manipulate the phosphatidylinositol-phosphate pathways to generate distinct RNA replication organelles (397). This observation was not found in the Legebeke et al data analysis of gene expression differences (388). Despite similarities in gene expression changes themselves. This suggests differences in the TopMD and PandaOmics molecular topographical analysis.

Downregulation of this leads to apoptosis and effectively prevents viruses from surviving by destroying host cells and preventing replication and budding of viral particles (398). Since upregulation of this has not been seen in the SARS-CoV-2 literature, this difference may be the result of the influenza virus having been present in the population at more consistent rates and host responses have adapted to sequester more effectively the influenza virus processes. Perhaps casting doubt on this hypothesis is the observation that both cohorts were hospitalised and so differences in transcriptome might be more likely to reflect successful pathogen processes as opposed to successful host responses. The biological truth is likely to be some combination of these factors, unable to be ascertained or elucidated without a robust control group. The enrichment of genes associated with CD163 mediated anti-inflammatory response in the influenza cohort is perhaps puzzling as CD163 positive macrophages/monocytes have been heavily implicated in both influenza

and Sars-Cov-2 virulence (399). However it seems that circulating levels of cd163+ M1 inflammatory type monocytes have been robustly correlated with severe and deleterious outcomes in humans with influenza (400) as has been observed in non-human primates (401). The higher levels observed then may be a result of correlation whereby hospitalised influenza patients have higher levels of M1 monocytes to start with. Interleukins 18, 33 and 36 also were upregulated in influenza compared with SARS-CoV-2. IL-18 is strongly pro-inflammatory and released by monocytes and macrophages in response infection indicating innate immune responses are present and active. IL-33 is an alarmin, triggering an immune response in the event of cells death and also is heavily pro-inflammatory (402) along with IL-36 (403). The strong pro-inflammatory signals being found primarily in influenza patients and contrasted by what appears to be primarily adaptive immune responses in the SARS-CoV-2 cohort. Recent literature suggests that the latency period between symptomatic infection is longer with SARS-CoV-2 than with Influenza however, and this might account for some of the differences here (404, 405).

Within this piece of work, differential transcript use has proven to be a remarkably successful and informative approach to understanding differences in the transcriptome which result from alternative splicing. The approach was able to reliably infer which isoform a read was generated from oftentimes, without needing to defer to the use of equivalence classes because specific isoforms were unable to be determined. Despite the conservative model, a very high number of genes which had DTU between the cohorts were identified with clear changes which oftentimes resulted in a change in the dominant isoform. Given that differing isoforms often have antagonistic functions, this finding is significant and affirms the importance of quantifying changes in splicing which occur in disease states.

The majority of these genes did not experience DGE, and so may not have been linked to these infections had this analysis not been done. It could be surmised that because these two models represent similar disease states, that is URTI resulting from RNA virus, that a case/control study for these infections which uses the same metrics and methods may find differences in isoform abundance which are even more profound.

Alternative splicing of the AMBN gene is known to produce isoforms which either promote or suppress mesenchymal stem cells proliferation and osteogenesis (406). MSC's are anti-inflammatory and drive other immune cells such as monocytes and macrophages to an anti-inflammatory/immunoregulatory type 2 state (407). AMBN<sup>WT</sup> (ENST00000613447.4/ENST00000322937.10/NM\_016519.6/AMBN-201) was expressed as the

dominant transcript in SARS-CoV-2, whereas the truncated version AMBN $\Delta$ 6<sub>1-15</sub> was the primarily expressed transcript in the Influenza cohort. This suggests that MSC proliferation was suppressed or less upregulated in the Influenza cohort. FAM19A3 or Tafa3 encodes a small secreted protein, expressed primarily in the brain and testis, which functions as a chemokine or neurokinin regulating immune and nerve cells (408). FAM19A3 or Tafa3 was also alternatively spliced to produce different dominant transcripts between the cohorts. This gene is associated with M2 polarisation in microglia and monocytes (408), specifics around isoform function are yet to be elucidated, however it could be hypothesised that the dominant transcript in the Influenza cohort, is likely to also function in the same type 1, inflammatory manner as the previous AMBN examples.

AMPK or PRKAA2 is also well known to be alternatively spliced with the ENST0000610361.1 isoform having exon 2 included, and ENST00000371244.9/NM\_006252.4/AMPK\_201 isoform having this exon skipped (409). Both isoforms are present canonically, however imbalances are associated with development of Alzheimer's (410) and AMPK $\alpha$ 2 protein subunit is imperative in cone survival and function and so the development of new neural connections (411). Olfactory receptor genes also had differential transcript abundances between the cohorts. It's therefore possible that the clinical feature of anosmia, now synonymous with COVID19, may well be a result of alternative splicing in the olfactory receptors resulting from SARS-CoV-2 infection. These examples of results from the research represent just a handful of important transcriptomic evidence which supports the parallel investigation of both alternative splicing and gene expression in infectious disease. Using these parallel lenses, it was evident that there were very significant differences in the transcriptomes of the cohorts both in terms of gene expression and isoform abundance. What was extremely surprising is that the number of alternative splicing differences greatly outnumber those arising from differential gene expression, despite the extremely conservative statistical model used to infer DTU.

These two discreet mechanisms of producing transcriptome diversity, seem to regulate entirely different processes and cellular functions without much overlap. Given the extremely strong signals found, is it highly unlikely that this might be a feature of statistical bias.

Whilst the data is clear that the distinct viral infections produce very different transcriptomic profiles, further unpicking which aspects pertain to either the viral effects directly, the immune response or disease progression will be complicated and time intensive, likely needing further experimentation. Given sufficient resources and time a transcript specific investigation would be

useful to identify potential biomarkers or therapeutic targets within the cohorts. Moreover, intentional immune challenge would likely allow for a greater degree of control of variables surrounding timing of infection. A longitudinal approach to each patient, allowing the mapping of the immune response through initial inflammation, innate immune and transition to adaptive immune responses would also be deeply informative. However even within these parameters there are still inherent limitations which arise from using bulk, whole blood RNA modalities. Multi-omics approaches would likely be extremely beneficial and help understand the effects of disrupted transcription and more specifically translation.

## Chapter 6 The Effect of Ageing on Host Transcriptomic Profiles during Viral Infection

### 6.1 Introduction

The transcriptomic profiles of hosts with infections are pathogen specific (412). Genetic heterogeneity and a lifetime of environmental differences will cause variation within these pathogen specific responses (413) and therefore individual profiles within these pathogen specific responses are seen (414). It is possible that the overlap in the principal component analysis in section 5.4 Figure 5-6 of the transcriptomes of patients with COVID19 and influenza, could result from influences at the individual level. Other sources of variation could be the specific strain of the virus with which the patient was infected (415, 416), or the different health states of the hosts. There is an established correlation between poorer outcomes and advancing age in the majority of infectious diseases (279, 280). Immunosenescence, inflammageing and chronic systemic inflammation is known to have deleterious effects on the hosts' immune response efficacy (246). The hypothesis was made, that the age of the patient might cause changes in the immune response; individual genes may have less significant changes in expression in aged individuals and the distinctive nature of the response may be lost to a more 'communal' immune response. Therefore, advanced age and immunosenescence may be a major contributor to some of the lack of differentiation between patient cohorts for these two distinct infections. This chapter presents the results of an investigation into the differences in transcriptome of patients with upper respiratory tract infections which occur with advancing age. First through exploratory data analysis of cohorts at different age ranges. This is followed with the use of linear regression to identify the relationship of all transcriptomic features with age. The work then aims to identify any loss of distinctness of transcriptome between infections which occurs with age to evaluate the impact of ageing on the immune response to infection, and the relative contribution of gene expression and splicing to this process. Finally, the research takes advantage of machine learning classification models to see if gene expression or splicing performs better for discerning infection from the transcriptome in young and old patients.



### **6.1.1 Aims**

The aims of this chapter were exploring the hypothesis that decreases in the distinctness of the immune response was occurring with advancing age, and therefore contributing, in-part, to the overlap in principal component analysis of transcriptomes of cohorts with different infections.

To explore the effects of advancing age and immunosenescence on the gene expression and isoform abundance profile of patients with infectious disease and identify opportunities for therapeutic intervention.

## 6.2 Methods

We analysed the expression data following the curation of the raw data as described in Chapter 5. To first visualise the effect of ageing on the transcriptomic response to viral infection we removed participants over 65 years of age from the cohort and the principal component analysis was repeated. The 65 year cut-off was based on previous evidence of decreased immune response to pathogen in cohort over 65 (248). To explore this at a more granular level, the expression of some of the top differentially expressed genes were then viewed individually, stratified by decade of life to see the effect on these profiles over time.

An informatic pipeline was developed to produce gene counts and relative isoform abundances, which used STAR and Salmon to perform the alignments, counts and differential transcript use values. This was designed in light of the requirement for Salmon-produced transcripts for the downstream implementation of BANDITS, and the high processing speed and low computational demands of the Salmon program. The counts were extracted and combined with clinical meta data into a data table. Meta data included patient ID, age, sex, white blood cell count, neutrophil count, lymphocyte count, C-reactive protein levels, Diabetes status, immunosuppression status, smoking status, presence of cardiovascular disease, and presence of respiratory disease. Patient age was then regressed from each individual feature from the entire transcriptome in a series of single, multiple regressions, with meta-data included as cofactors. This included 305,165 features, of which 267,543 were transcript abundances, measured in values from 0 to 1, representing the proportion of total transcripts from the gene, each transcript represented. 61,587 features were gene expression features measured in TPM. Beta-coefficients and p-values were calculated for each of these features when age was regressed.

Results from the linear regressions had 'x=zero' beta values removed, and histograms were plotted of beta coefficient distribution using Microsoft Excel, in which groupings were automatically assigned. The values of groupings were rounded to 2 decimal places. Using this data from the linear regression, volcano plots were generated using the EnhancedVolcano package in R. The volcano plots represented the gene expression or relative isoform abundance changes with age in each cohort.

Beta coefficients impute the scale of effect on the response variable in machine learning regression and can be compared across experiments. P-values are not often included in the ML packages, due to the lack of utility in ML applications. Despite the beta coefficient values being generally accepted in machine learning communities without the need for p-values as a proxy for statistical significance due to the lack of hypothesis testing (417), additional steps were taken to calculate p-values from the package and the results from the gene expression and relative isoform abundance regression were filtered with pre-determined alpha, set at 0.05 and beta VALUE OF 0.1 to ensure a conservative method.

To compare and contrast how the processes of differential gene expression and differential transcript usage affected the two infections, cross-referencing of the outputs from the previous steps was conducted. Genes which underwent differential expression (Pandaomics results), differential transcript use (BANDITS, gene-level results) and those genes and isoforms which were differentially expressed with age within the two cohorts were then combined into a .txt file via the list function in R. The file with the 6 respective lists of genes were compared and contrasted using the UpSet (418) program in R to understand the overlap in genes which experienced differential gene expression and splicing and how this was related to genes which experience differential expression and splicing with age.

In order to quantitate any convergence of transcriptomes at the gene and isoform level the following approaches were taken. The list of genes which were differentially expressed between COVID19, and influenza cohorts identified in section 5.5 were compared with the list of genes and isoforms which had contrasting beta coefficient values (positive and negative) using the 'vlookup' function in Microsoft Excel. Identification of contrasting beta values was achieved by filtering the beta values from one infection for positive values, the beta values in the opposite infection for negative values, performing a vlookup in Microsoft Excel to identify genes which appeared in both lists. These were filtered for significance using Bonferroni correction. The features were only counted if the contrasting expression changes, represented by the beta values were opposed to the initial direction of differential expression or differential isoform usage between the infections. For instance, if a differentially expressed gene was originally identified as having higher expression in COVID19 patients than in influenza, then the beta coefficient values would need to be a) in opposing directions (i.e., positive and negative), and b) with COVID19 expression trending in the opposite direction to the original differential expression, (in this example negative beta in COVID19 and a positive beta in influenza). This is to ensure convergence and not greater divergence was occurring. The resulting

gene lists were compiled into text files and loaded into the TOPPGENE online software (363) for gene set enrichment analysis.

To determine the value of gene expression changes and isoform abundance changes in diagnostics and investigation, classification machine learning was performed. Using the dataset obtained from the Pandaomics software, the top 100 differentially expressed genes were extracted to a separate text file along with the ages of the patients. This data was loaded into the MATLAB program and the gene expression data were used as independent variables or features to perform a series of classification exercises using the modelling tools available in MATLAB. All available models (n= 24) were used for the exercise to avoid any bias in user selection. Five-fold cross validation was used to evaluate the utility of the features and the models in diagnosing infection. The age of the group was systematically reduced by sequentially removing people at the highest decade of life, over 80, over 70, over 60, over 50, leaving groups of people under 81, under 71, under 61, and under 51. To avoid statistical bias from shrinking groups, the process was then repeated with the cohort using the same approach reversed, systematically making the group older for 5 groups. The performance of the models training was recorded in a table to demonstrate how the training performance changed with age, and to select the best model moving forwards. These results were also plotted to a graph to show how overall performance changed with age when training ML models.

SVM was identified as the best performing model on average for predicting the infection based on 100 top differentially expressed genes across all ages. The 77 COVID patients and 83 flu patients were separated into a test and training dataset. To partition roughly 20% of the cohort as a test dataset, but still have a representative spread of data, the two infection datasets were individually ranked by age and every 6<sup>th</sup> person was transferred to the test dataset. The two cohorts were then further divided into those which were above and below the median age: Group 1 where age >60, Group 2 where age is <61. This cut-off was applied after more recent literature suggests immunosenescence can be detected at around 60+ (419, 420). The elderly cohorts and young cohorts from each infection were then merged with each other so the resulting groups were as follows.

**Table 6-1 Cohort grouping for machine learning application.**

<p><b>Training set 1</b></p> <p>Appx 80% of samples from COVID19 and influenza cohorts under the age of 61</p>	<p><b>Test set 1</b></p> <p>Appx 20% of samples from COVID19 and influenza cohorts under the age of 61</p>
<p><b>Training set 2</b></p> <p>Appx 80% of samples from COVID19 and influenza cohorts over the age of 60</p>	<p><b>Test set 2</b></p> <p>Appx 20% of samples from COVID19 and influenza cohorts under the age of 60</p>

Using MATLAB (421), support vector machine learning models were then applied to the training datasets first, using a 5-fold cross validation for model production within the training dataset only to prevent data leak. Support Vector machine learning models were used as they were demonstrated to be the most accurate in the previous testing phase. The model's performance on the training dataset and test dataset were compared. This was repeated for using the top 100 differentially expressed isoforms with the differential transcript use metrics derived earlier.

## 6.3 Results

### 6.3.1 Exploratory data analysis

The Shapiro wilk test for normality shows that the groups are normally distributed, although the influenza cohort was approaching statistical significance alpha value of 0.05 indicating the influenza data came from a group whose ages were less well distributed. The T-test test statistic is extremely low, and as the p-value from the T-test is not below 0.05, we conclude the means between the groups are not significantly different.

PCA analysis shows clear grouping of patients about principal component 1 (PC1) in both principal component plots (Figure 6-1, Figure 6-2). After investigation in the genes which most heavily weighted PC1 this was determined to be a result of sex specific differences in gene expression, as such the two vertical clusters represent male and females. There also exists an incomplete separation of the groups whereby Influenza patients are primarily above Y axis point 0, and the COVID19 patients are primarily below Y axis point 0. 95% confidence intervals have a significant

amount of overlap between the groups. Although modest, there was a more significant separation of the groups and less significant overlap in the 95% confidence intervals about PC2 once the elderly (>65 years) individuals were removed. Aligned reads for groups once elderly individuals had been removed (Figure 6-3) carried a similar profile as seen previously (Figure 5-5), with an average count after filtering for low expressed genes of around 19-20 million reads per sample.

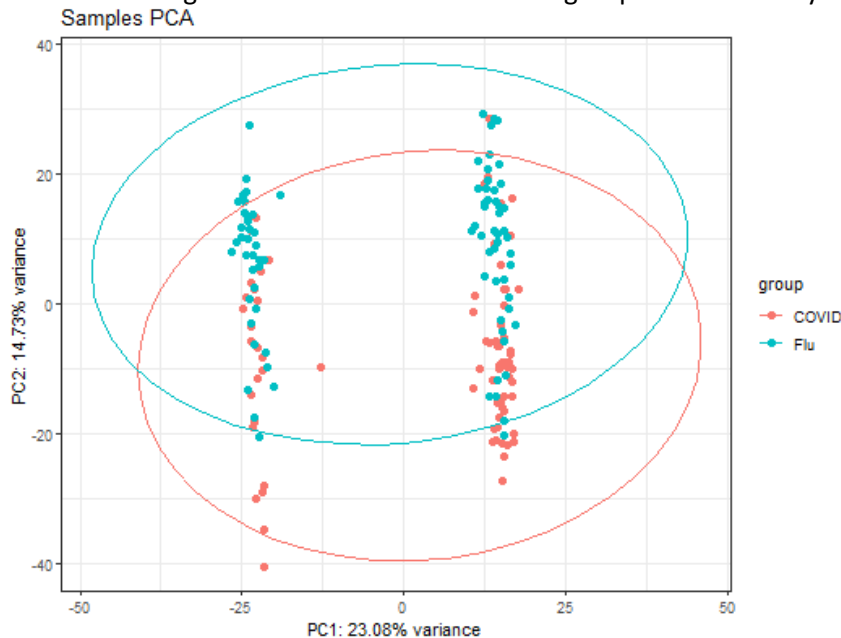
In order to increase the resolution of analysis to the gene level, we looked at some of the most significantly differentially expressed genes and stratified the expression by both decade of life and by infection type. As observed in chapter 4, one of the most significantly differentially expressed genes between the infections that was a non-immunoglobulin gene was *JUN*. Figure 6-4 shows that the expression of *JUN* is different between the cohorts in the early to middle decades. In the later decades these expression profiles appear to converge. Similar expression profiles were also seen for many of the immunoglobulin genes which were the most differentially expressed (Figure 6-5)(Appendix A.11). This convergence was not present for all the differentially expressed genes, however. CD163 was one of the most differentially expressed genes, with higher expression in influenza, and is an example of a differentially expressed gene between the infections which did not show this converging pattern of expression between the groups in any significant way (Figure 6-6).

**Table 6-2 The distribution of ages for entire Influenza and COVID19 cohorts**

<b>Influenza</b>	<b>Mean</b>	<b>57.84</b>
	Standard deviation	18.381
	Median	59
	Quartiles	42,59,73
	Shapiro-Wilk normality	W=0.97052 p =0.05327
<b>COVID19</b>	Mean	61.38095
	Standard deviation	17.555
	Median	61.5
	Quartiles	47.75, 61.50,73.25
	Shapiro-Wilk normality	W=0.9793

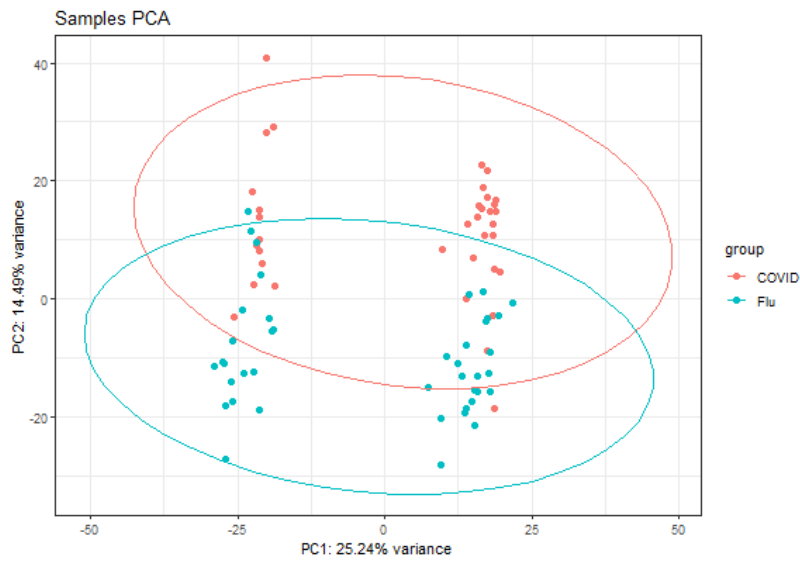
		p=0.1949
<b>T-test for normality</b>	<p><b>T = -1.2716,</b></p> <p><b>p = 0.2053</b></p> <p>Degrees of freedom = 164.45,</p> <p>95% C.I = -9.030 and 1.955</p>	
<b>Interpretation</b>	<p>T-test P &gt; 0.05 accept the null hypothesis; there is no evidence the mean age between groups are significantly different.</p> <p>Shapiro-Wilk results indicate both groups are normally distributed.</p>	

Statistical testing for normal distribution within groups and normality between groups.



**Figure 6-1 Principal component analysis of Covid19 and Influenza cohort**

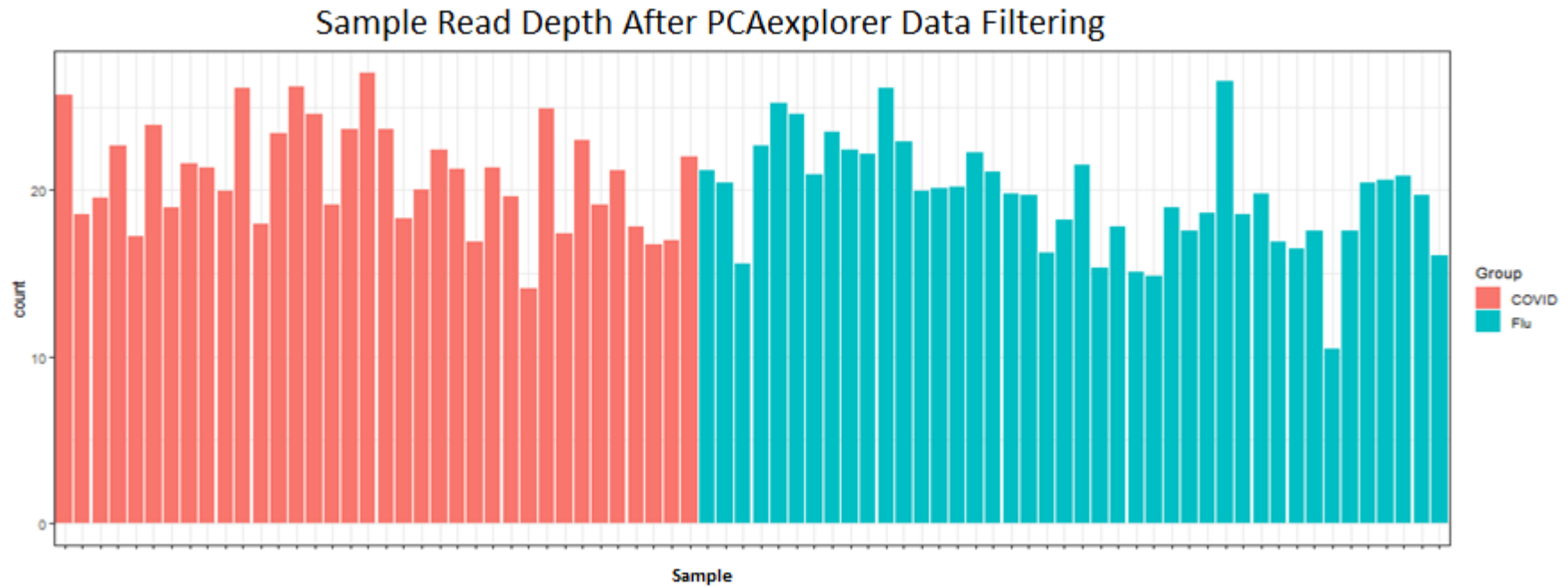
Principal component plot derived using COVID19 and Influenza cohort. Red points are COVID19 patients, blue points are Influenza patients. Circles indicate 95% confidence intervals. X axis represents principal component 1, Y axis represents principal component 2.



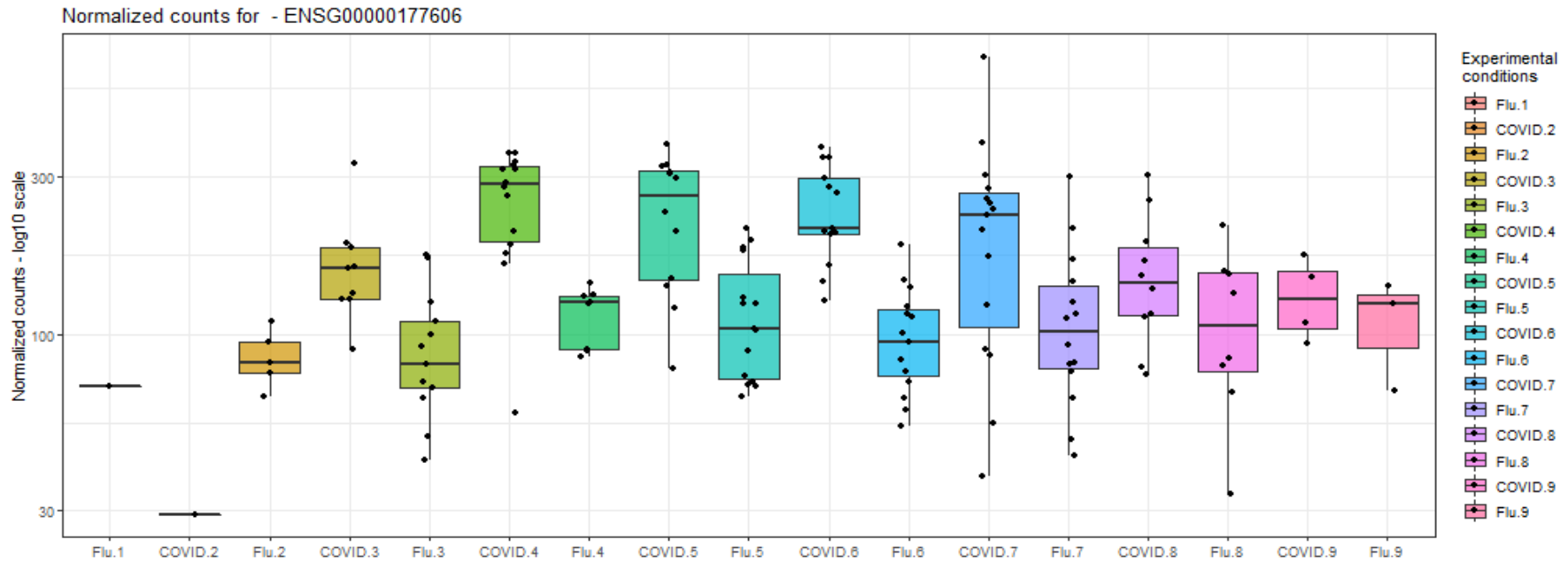
**Figure 6-2 Principal component analysis of Covid19 and Influenza cohort**

Principal component plot derived using COVID19 and Influenza cohort after removing those members of the cohort who were above the age of 65, of which there were a combined number of 62 patients. Red points are COVID19 patients, blue points are COVID19 patients. Circles indicate 95% confidence intervals. X axis represents principal component 1, y axis represents principal component 2.



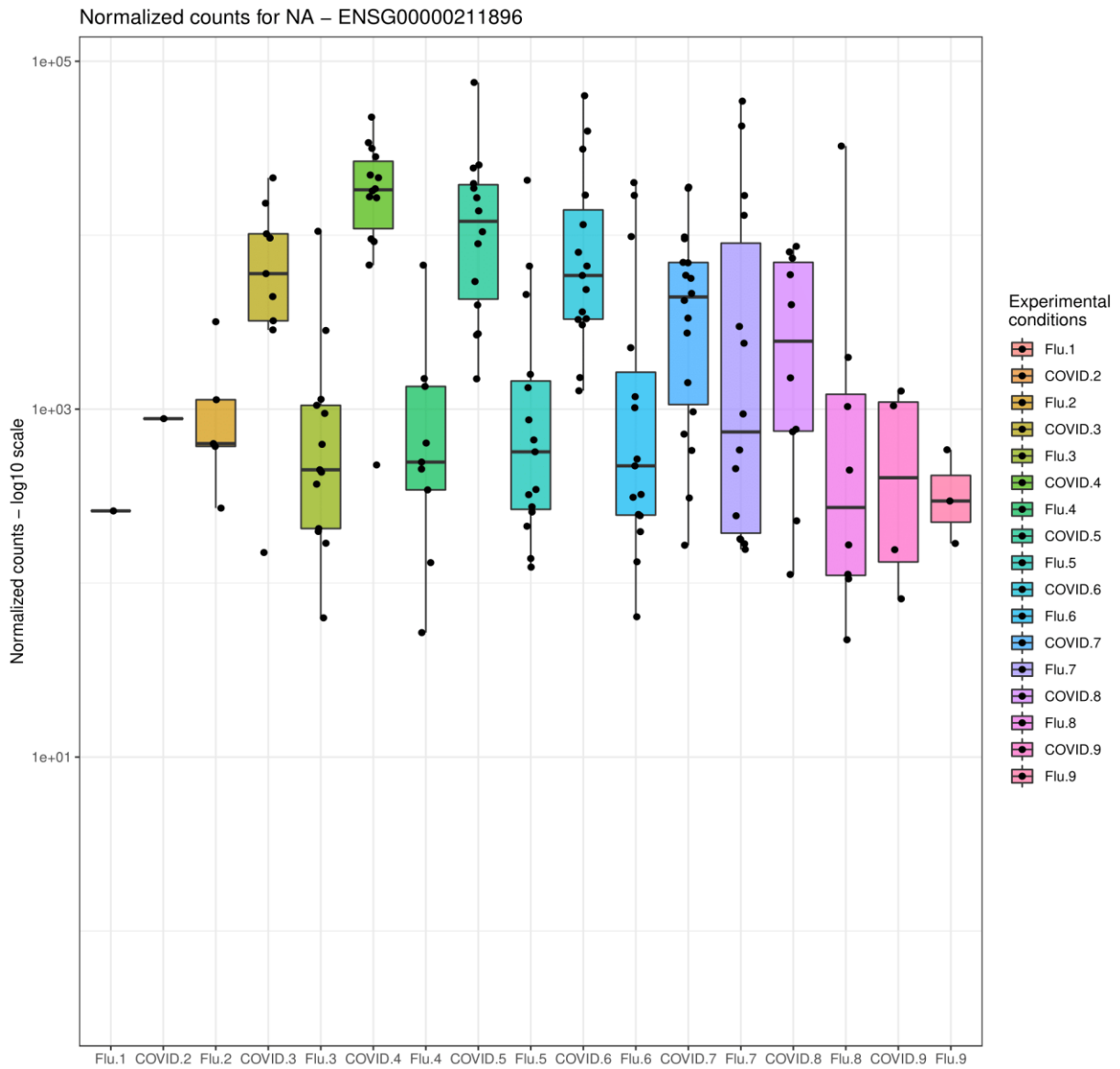


**Figure 6-3 Exploratory data analysis: total aligned reads per samples for Covid19 and Influenza samples only patients under 65 years of age**  
 The matrix displays total aligned reads per sample, after the data was cleaned using the pcaExplorer tool as per the parameters in methods section, the aligned reads appear to be approximately consistent across both data sets.



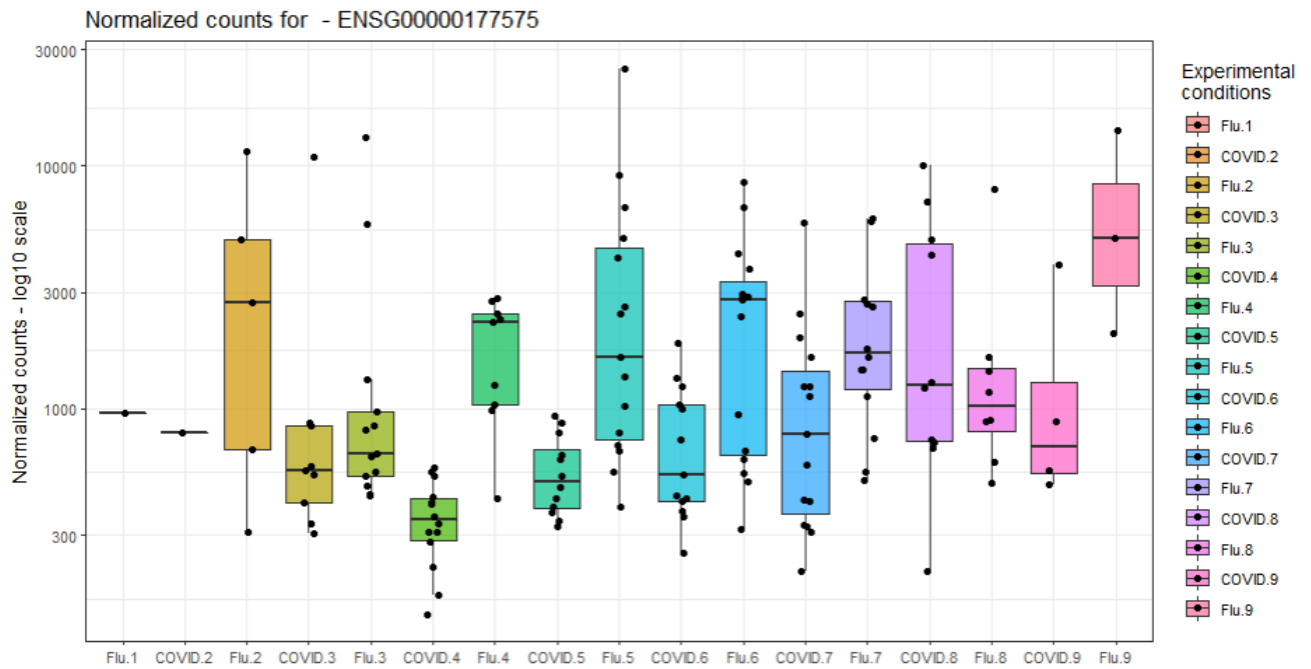
**Figure 6-4 JUN expression in the COVID19 and Influenza cohorts stratified by age.**

Figure shows box and whisker plot of the expression difference in normalised counts between the cohorts for each decade of life. Each single digit number represents a decade i.e., Flu.2 represent Influenza patients of ages from 20 to 29. log base2 Normalised read counts represented on the Y axis, and the cohort groups, stratified by infection and decade of life are represented on the X-axis.



**Figure 6-5 Differential gene expression for IGHG1 in COVID and Influenza cohorts stratified by decade of life.**

Figure shows box and whisker plot of the expression difference in normalised counts between the cohorts for each decade of life. Logbase2 transformed read counts represented on the Y axis, and the cohort groups, stratified by infection and decade of life are represented on the X-axis.



**Figure 6-6 CD163 expression in Covid19 and Influenza patients stratified by decade of life.**

Figure 5-6 shows box and whisker plot of the expression difference in normalised counts between the cohorts for each decade of life for CD163. Of all the most differentially expressed genes which had higher expression in COVID19, this was the most significant. Normalised read counts represented on the Y axis, and the cohort groups, stratified by infection and decade of life are represented on the X-axis.

### 6.3.2 Results from multiple regression performed using Python.

In total there were 2609 genes which has association ( $\text{Beta} > 0.1 / \text{Beta} < -0.1 + p < 0.05$ ) with advancing age in the COVID19 cohort, and 2524 genes which has association ( $\text{Beta} > 0.1 / \text{Beta} < -0.1 + p < 0.05$ ) with advancing age in the Influenza cohort.

Tables were produced so show this data, results for most differentially expressed genes with age. Beta values are ranked by scalar quantity and are vector agnostic (Table 6-3,

Table 6-4, Table 6-5, Table 6-6).

Within the COVID19 cohort, some extremely strong negative associations were seen with *NT5E*, *WLS* and *PSMB5*. *NT5E* otherwise known as *CD73* is an ectonucleotidase, already linked to COVID19 pathogenesis (422). *WLS* or the *WNT* Ligand Secretion Mediator is responsible for all *WNT* secretion and also linked to the orchestration of immune response through its function on dendritic cells (423). *PSMB5* is an essential subunit of the 20S proteasome used for substrate degradation and is one of the subunits not found to be downregulated in COVID19 patients compared with healthy controls in other literature (424) However with advancing age, the expression of this proteasome decreased indicating a potential mechanism for loss of function. A number of other genes were upregulated in COVID19 with advancing age including *CNTNAP3*, a neurexin associated with neural-glial cell interaction, and was also recently discovered elsewhere as a key differentially expressed gene when comparing viral infections, although the nature of its involvement remains unclear (425). *CCDC170* was the most upregulated gene with age in the COVID19 cohort, and its relationship with the infection is somewhat enigmatic for now. The gene can increase oestrogen receptor  $\alpha$  expression (286), and oestrogen production has been linked to positive outcomes in COVID19. Otherwise, no obvious link is evident (426).

Within the Influenza cohort *CD248* was found as the most downregulated gene with age, as described elsewhere in a healthy ageing cohort (310). It is also associated with naïve T-cell number proliferation and as such might be an effective marker for immunosenescence in future (427).

The *TBCK* gene experiences the greatest change in isoform abundance associated with age in influenza patients. Known to have isoforms with differing function (428), *TBCK* encodes a protein kinase involved in proteostasis and lysosomal activity (429). It is also shown to be closely linked with

the MTOR pathway (430), generally accepted to be directly related to ageing, and indeed TOR has been identified as a potential target for immunosenescence therapeutic intervention (431, 432).

CDK6-AS1 was the gene with the most differentially expressed isoform in COVID19 with age, and is an antisense regulatory RNA for CDK6 and has increased expression in ageing COVID19 patients (433). CDK6 is required for early thymocyte development (434), crucial for adaptive immune response (435) and known to be affected in immunosenescence (436). Moreover the CDK4/6 axis effects P21, a known inducer of cellular and immunosenescence (437). CDK6-AS1 has also been identified in work looking into respiratory virus' as a risk factor for COVID19 via its effect on CCL3 (438).

**Table 6-3 Top 20 genes with expression associated with age in COVID19 patients.**

COVID FEATURE		BETA	P-VALUE
ENSG00000120262.10	CCDC170	0.676949	7.9E-06
ENSG00000250158.1	Unknown	0.666418	6.5E-05
ENSG00000106714.18	CNTNAP3	0.665596	1.43E-05
ENSG00000226394.2	NDUFA4 Pseudogene	0.649011	0.000323
ENSG00000135318.12	NT5E	-0.64413	1.17E-05
ENSG00000248936.1	Lnc-RELL1-1	0.64054	7.65E-05
ENSG00000224335.2	NBPF4 Pseudogene	0.639547	0.000357
ENSG00000235040.1	MTCO3P1	0.637415	0.000268
ENSG00000078053.17	AMPH	0.630936	0.000146
ENSG00000183631.5	PRR32	0.616167	0.000256
ENSG00000231136.1	IBRDC3 Pseudogene	0.602903	0.000487
ENSG00000100804.19	PSMB5	-0.59943	0.00029
ENSG00000243658.1	MTND5P16	0.59871	0.000217
ENSG00000224986.2	PPP1R8P1	0.595046	0.000424
ENSG00000102174.10	PHEX	0.593466	0.000441
ENSG00000213390.11	ARHGAP19	0.591521	0.000191
ENSG00000279550.1	TEC	0.589743	0.000852
ENSG00000214190.2	RNF152P1	-0.58646	0.000513
ENSG00000206724.1	RNU6-756P	0.584907	0.000561
ENSG00000110079.19	MS4A4A	0.584122	0.000586

**Table 6-4 Top 20 genes with expression associated with age in influenza patients.**

INFLUENZA	GENE NAME	BETA	P-VALUE
ENSG00000174807.4	CD248	-0.53585	2.50E-07
ENSG00000232460.4	BMPR1AP2 Pseudogene	-0.50943	0.000201
ENSG00000228909.2	LINC01803	0.497777	0.000323
ENSG00000251616.1	Lnc-FSTL4-1	0.481547	0.00026
ENSG00000228814.4	HNRNPA1 Pseudogene	0.464068	0.000607
ENSG00000255078.1	OR4A6P	0.463631	0.000384
ENSG00000183239.5	RPL29 Pseudogene	-0.46297	7.61E-05
ENSG00000258469.1	CHMP4BP1	0.460285	0.000778
ENSG00000207965.1	MIR629	0.444408	0.00057
ENSG00000154143.3	PANX3	0.444394	0.001597
ENSG00000248293.1	SIN3A Pseudogene	-0.44267	0.017139
ENSG00000207946.3	MIR516B1	0.442186	0.000482
ENSG00000261395.3	Hsp10 Member 1 Pseudogene 5	0.437477	0.000382
ENSG00000276639.1	MIR7154	0.435973	0.001117
ENSG00000253678.3	PRSS52P	0.435764	0.002248
ENSG00000255606.1	LINC02952	0.434828	0.001766
ENSG00000234604.2	MTATP6 Pseudogene	-0.43178	0.000337
ENSG00000237446.1	RHEBP3	-0.43169	0.001709
ENSG00000236124.1	MGN2P23	0.431521	0.001336

ENSG00000227945.1 | PTPRK Antisense RNA 1 -0.42969 8.32E-05

**Table 6-5 Top 20 Isoform abundance changes associated with age in influenza.**

FEATURE	GENE		FLU_BETA	FLU_P-VALUE
ENST00000503832.1	ENSG00000145348.17	TBCK	-0.580749575	0.000669
ENST00000682417.1	ENSG00000172469.17	MANEA	-0.559995822	0.001606
ENST00000477751.1	ENSG00000151655.19	ITIH2	0.553075751	5.56E-06
ENST00000462464.1	ENSG00000174953.14	DHX36	-0.550575569	1.8E-05
ENST00000529543.5	ENSG00000154144.13	TBRG1	0.548267286	8.54E-06
ENST00000425828.1	ENSG00000205090.9	TMEM240	0.527267718	1.51E-05
ENST00000545746.5	ENSG00000047621.12	C12orf4	-0.518079068	0.017747
ENST00000435762.2	ENSG00000168461.13	RAB31	-0.516935081	0.007526
ENST00000483448.1	ENSG00000008282.9	SYPL1	0.511282562	5.24E-05
ENST00000507416.1	ENSG00000188725.8	SMIM15	0.504294663	0.000514
ENST00000339020.8	ENSG00000188725.8	SMIM15	-0.504294663	0.000514
ENST00000648560.1	ENSG00000119979.18	DENND10	0.503883873	1.13E-05
ENST00000649932.1	ENSG00000023839.12	ABCC2	-0.49768411	0.000235
ENST00000589946.1	ENSG00000198046.13	ZNF667	-0.495295154	0.006003
ENST00000265827.8	ENSG00000048405.11	ZNF800	-0.494110142	7.02E-05
ENST00000231061.9	ENSG00000113140.11	SPARC	-0.494002057	0.000568
ENST00000557881.1	ENSG00000225151.10	GOLGA2P7	-0.491334652	0.002774
ENST00000664206.1	ENSG00000287888.1	Unknown	0.489489795	0.000124
ENST00000489476.5	ENSG00000144034.16	TPRKB	0.485958118	0.00065

**Table 6-6 Top Isoform abundance changes associated with age in COVID19.**

FEATURE	GENE	GENE NAME	COV_BETA	COV_P-VALUE
ENST00000653602.1	ENSG00000286742.1	CDK6-AS1	0.733381932	1.5E-05
ENST00000394777.8	ENSG00000165181.17	SHOC1	0.7330571	6.9E-05
ENST00000456262.5	ENSG00000170264.13	FAM161A	0.729869138	3.8E-05
ENST00000585229.1	ENSG00000101596.17	SMCHD1	0.708467549	3.9E-05
ENST00000456665.6	ENSG00000177192.14	PUS1	0.697090972	3.65E-05
ENST00000529884.1	ENSG00000255506.1	SNODB794	0.696609179	0.000102
ENST00000344624.8	ENSG00000113360.17	DROSHA	0.696075354	6.03E-05
ENST00000373626.4	ENSG00000198034.11	RPS4X	-0.693706514	5.19E-05
ENST00000611306.1	ENSG00000148655.15	LRMDA	-0.691598478	0.000112
ENST00000509903.5	ENSG00000153113.24	CAST	0.683844526	8.09E-05
ENST00000483146.1	ENSG00000114744.9	COMMD2	-0.683471865	3.55E-05
ENST00000567879.5	ENSG00000187741.15	FANCA	0.681986388	0.000268
ENST00000484194.1	ENSG00000204592.9	HLA-E	0.678849207	0.000184
ENST00000543697.5	ENSG00000127314.18	RAP1B	0.678182079	0.000176
ENST00000423231.1	ENSG00000232037.3	RPL21P29	0.673554533	0.000254
ENST00000397588.8	ENSG00000153046.18	CDYL	-0.671209616	0.000183
ENST00000518260.1	ENSG00000253628.2	Lnc-ERGIC1-1	0.670695185	6.86E-05
ENST00000483749.1	ENSG00000157036.13	EXOG	-0.670471833	0.000197
ENST00000578490.2	ENSG00000265485.8	LINC01915	0.669557621	0.000111

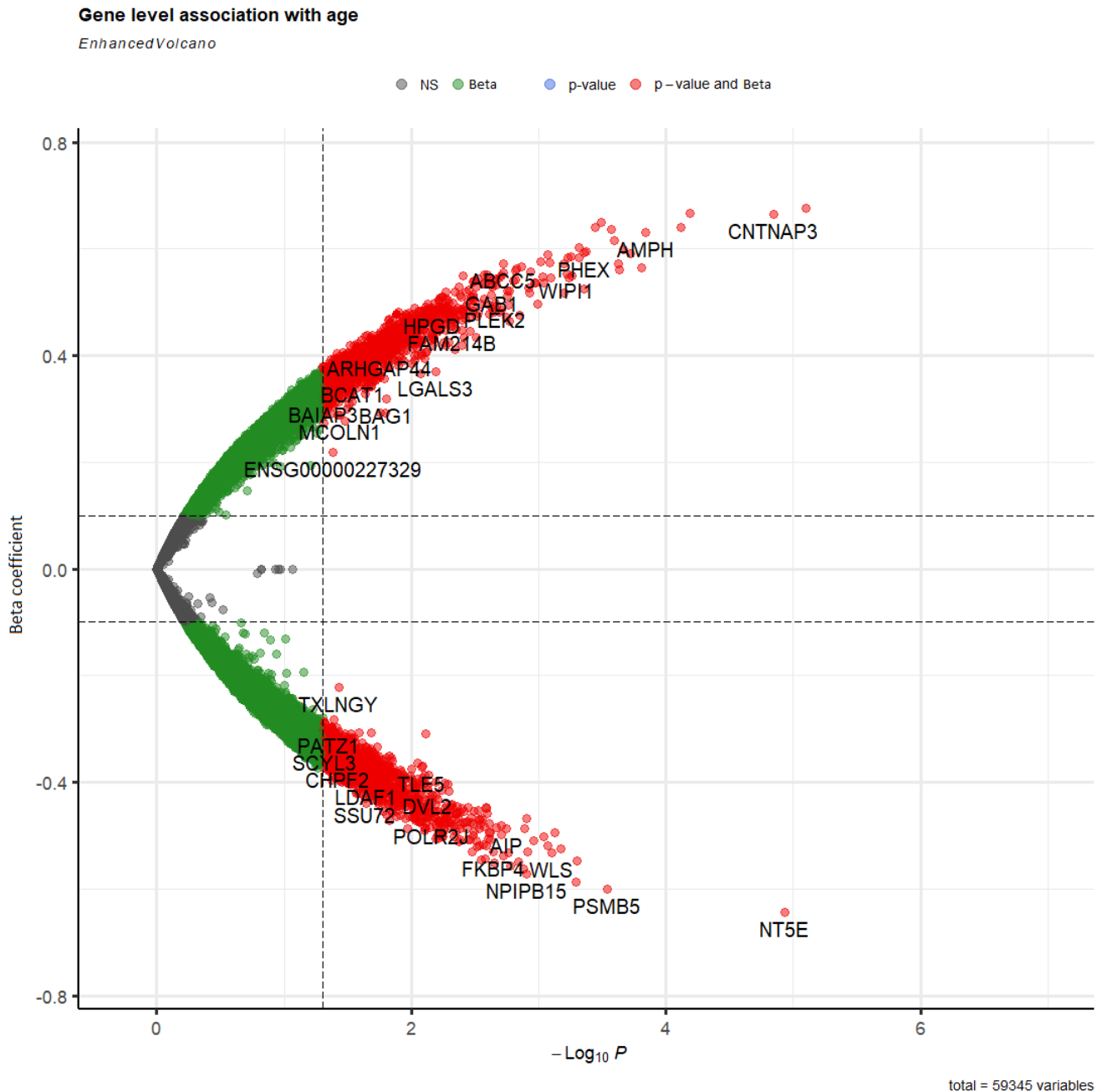


ENST00000689492.1 | ENSG00000163913.14 IFT122 0.66682713 0.00035

### 6.3.3 Volcano plots of transcriptome features association with age

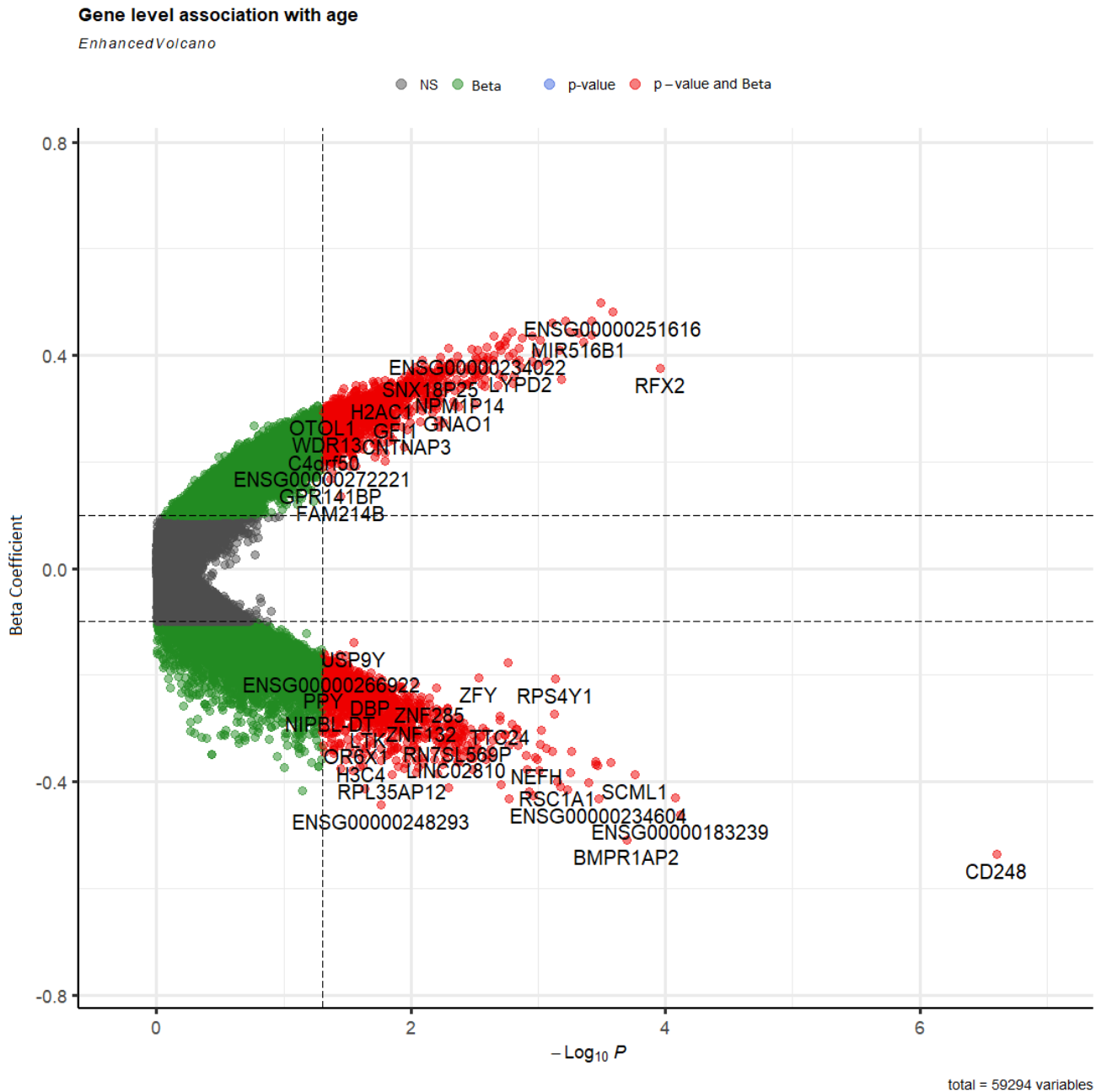
Volcano plots were produced to visualise distribution of transcriptomic features association with age (Figure 6-7, Figure 6-8, Figure 6-9, Figure 6-10). Volcano plots of all gene expression features association with age show a large number of genes to be significantly associated with age. Volcano plots for isoform abundance follow a similar distribution pattern as with gene expression in so far as the COVID19 associations appear to follow a stricter pattern of association than with gene expression. This effect is seen across all genes and isoforms, and so is unlikely to be a direct result of any biological phenomenon and instead an unknown technical artifact.

Splicing factors or trans acting RNA-binding proteins which regulate the process of alternative splicing (439). To understand the cause of changes in isoform abundance observed with age, investigation was performed into the expression levels of splicing factors with advancing age. For both cohorts, the decrease in expression of splicing factors associated with age was clear and profound (Figure 6-11, Figure 6-12). There was some difference around which were the most downregulated, but for both cohorts, the factors which regulate alternative splicing were almost ubiquitously downregulated with only the degree of severity and significance differing between splicing factors. Some of the most downregulated were *HNRNPF* for COVID19 patients and *LSM8* for Influenza patients. *HNRNPF* has been robustly associated with advancing age and associated age related inflammatory conditions previously (440, 441) and interacts with *FOXP3* to modulate alternative splicing, T-reg cell function and immunosenescence (442).



**Figure 6-7 Volcano plot for ageing gene expression in COVID19 patients**

The transverse volcano plot shows the gene expression changes in COVID19 patients which are associated with advancing age. Grey points represent non-significant changes. Green points represent a change in beta value of at least 0.1 or -0.1, red points represent genes which have a change in beta of at least 0.1/-0.1 but also statistically significant ( $p < 0.05$ ).

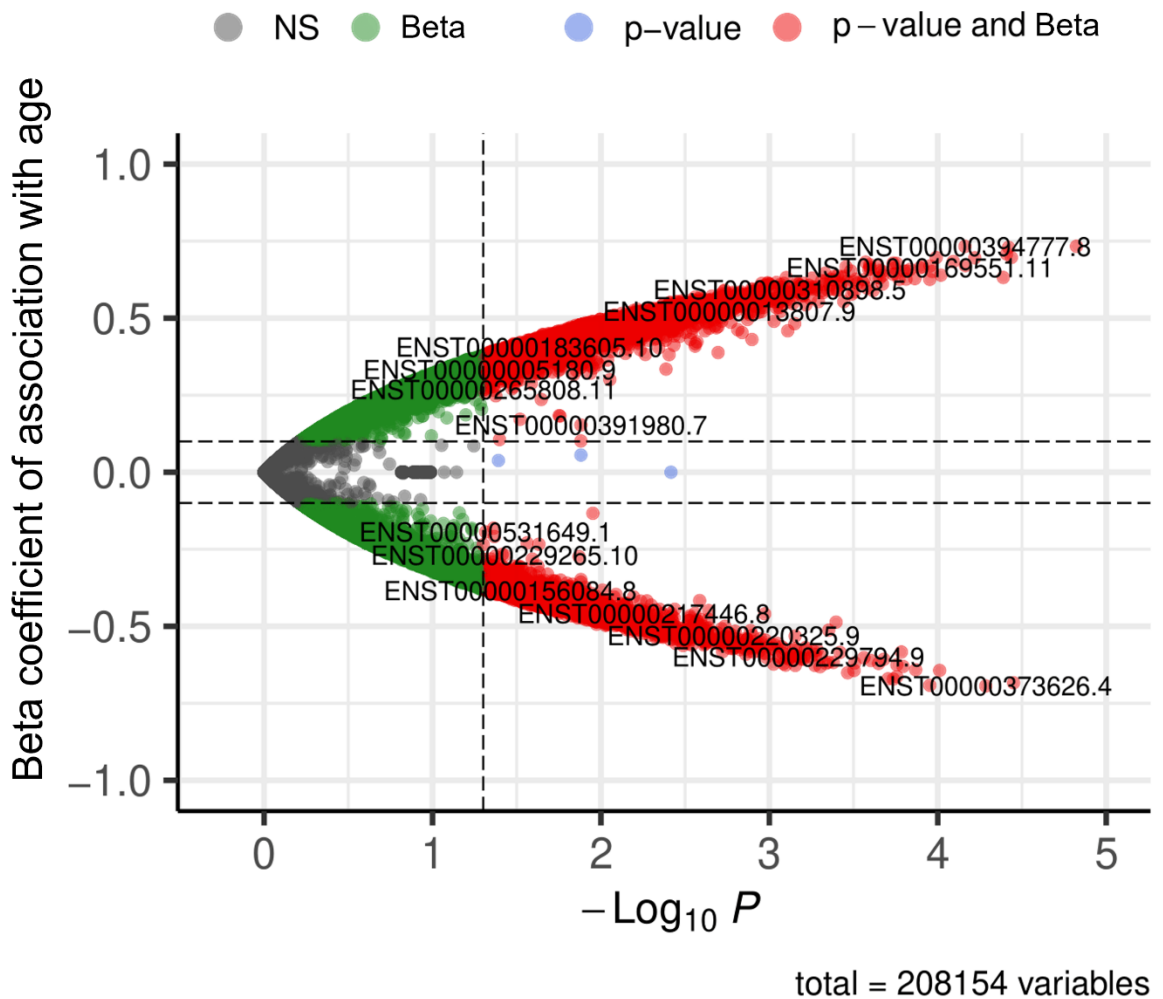


**Figure 6-8 Volcano plot for ageing gene expression in Influenza patients**

The transverse volcano plot shows the gene expression changes in Influenza patients which are associated with advancing age. Grey points represent non-significant changes. Green points represent a change in beta value of at least 0.1 or -0.1, red points represent genes which have a change in beta of at least 0.1/-0.1 but also statistically significant ( $p < 0.05$ ).

## Covid Isoform Association with Age

*EnhancedVolcano*

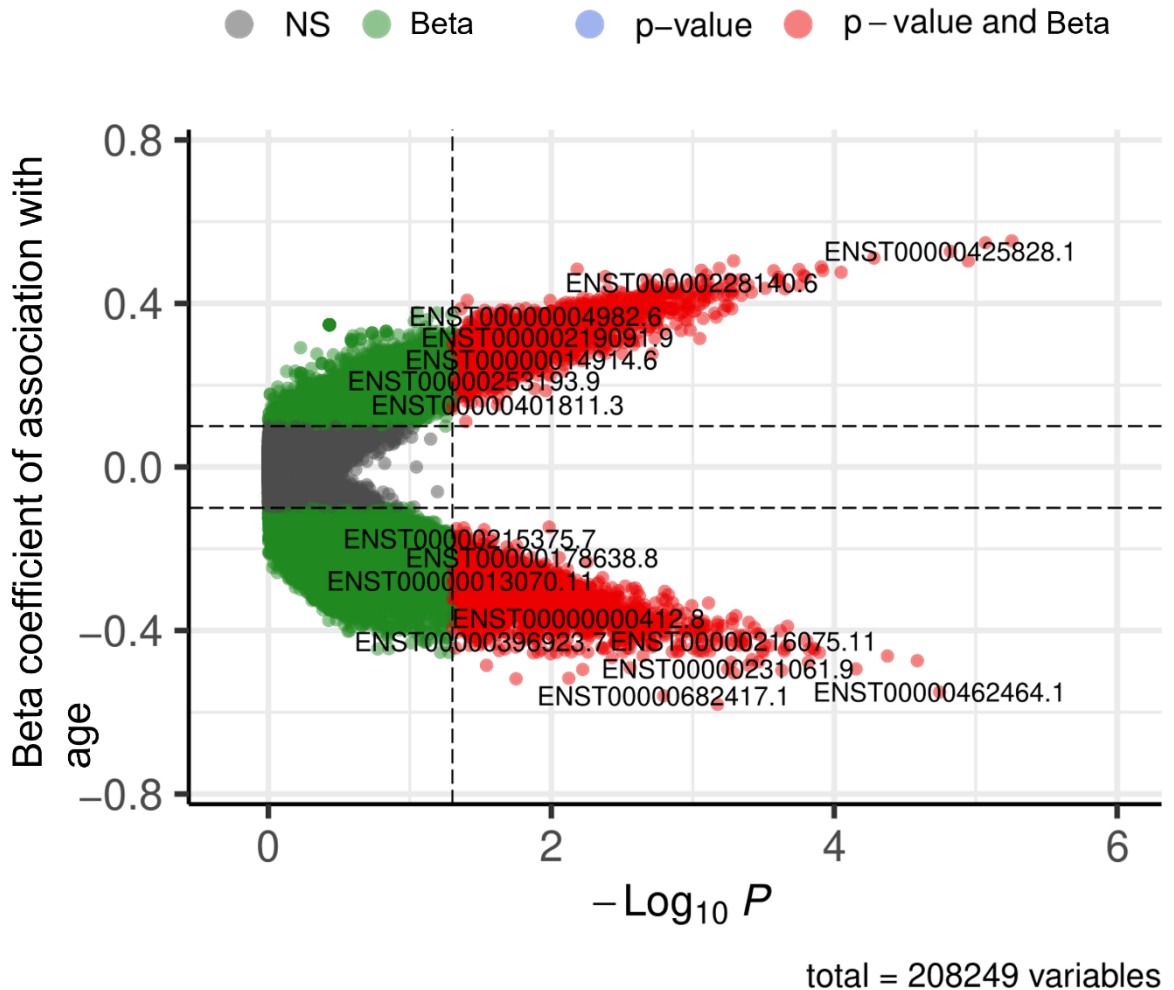


**Figure 6-9 COVID19 isoform association with age volcano plot**

This figure shows the association between changes in relative isoform abundance and advancing age in patients with COVID19. Grey dots represent non-significant changes below 0.10, the beta coefficient threshold value. All changes larger than this are shown in green. The p-value threshold value was set at 0.05, and all associations larger than this are shown in blue. If associations are above p-value and beta thresholds, they are shown in red. X axis is  $\log_{10}$  p-value. Y axis is beta coefficient.

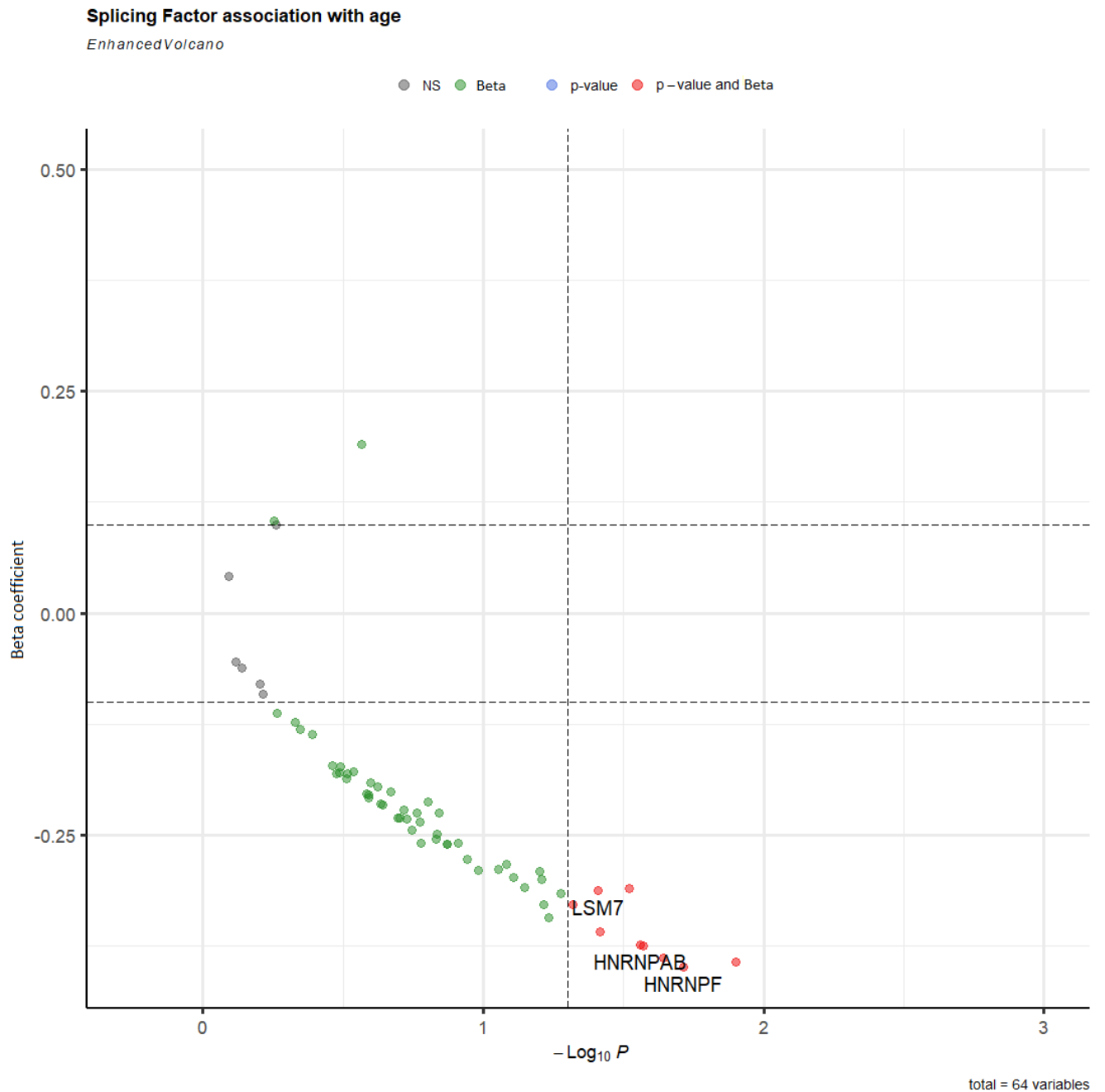
## Influenza Isoform Association with Age

*EnhancedVolcano*



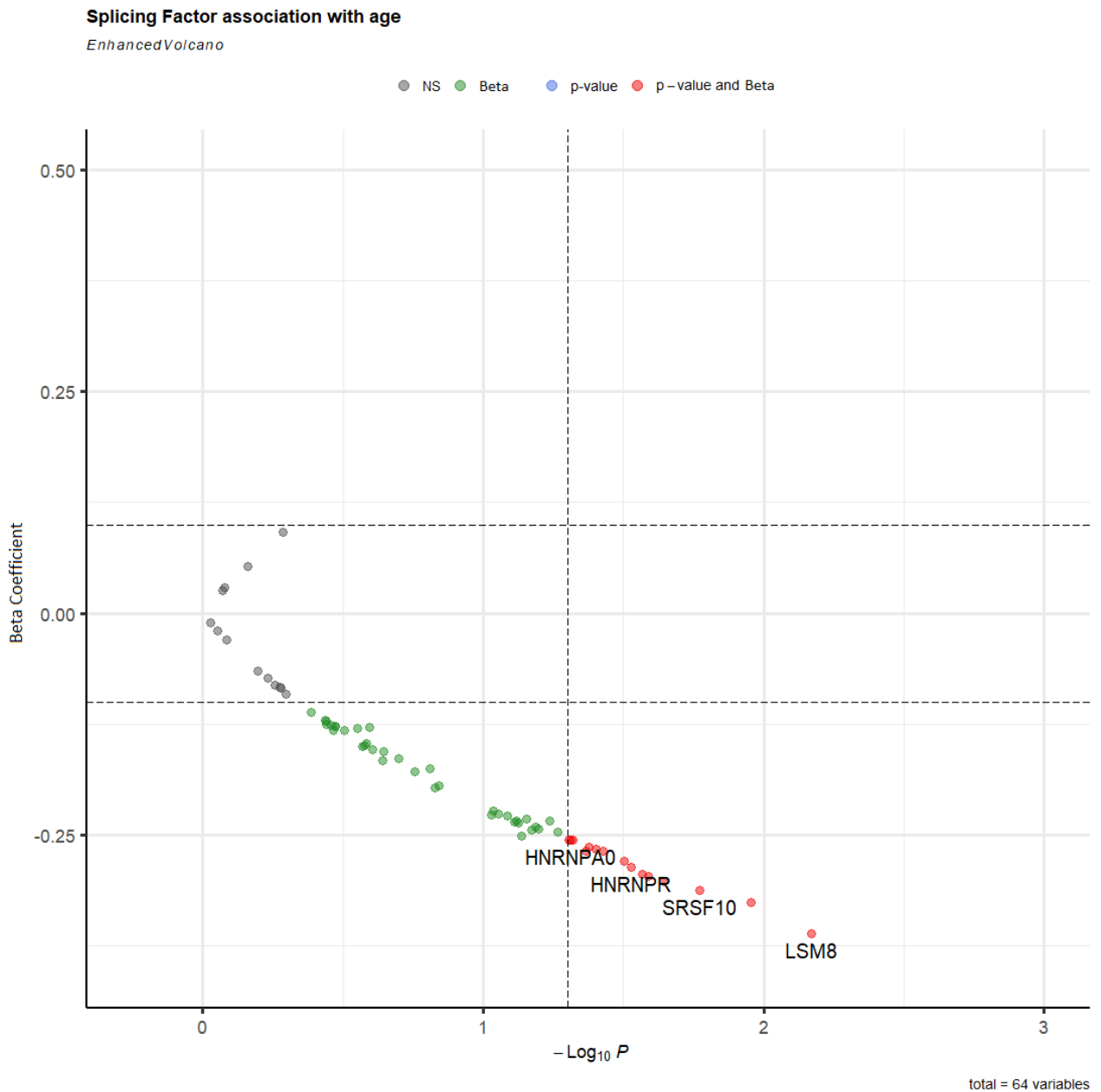
**Figure 6-10 Influenza isoform association with age volcano plot**

This figure shows the association between changes in relative isoform abundance and advancing age in patients with Influenza. Grey dots represent non-significant changes below 0.10, the beta coefficient threshold value. All changes larger than this are shown in green. The p-value threshold value was set at 0.05, and all associations larger than this are shown in blue. If associations are above p-value and beta thresholds, they are shown in red. X axis is  $\log_{10}$  p-value. Y axis is beta coefficient.



**Figure 6-11 Volcano plot for ageing splicing factor expression in Covid19 patients.**

The transverse volcano plot shows the splicing factor gene expression changes in Covid19 patients which are associated with advancing age. Grey points represent non-significant changes. Green points represent a change in beta value of at least 0.1 or -0.1, red points represent genes which have a change in beta of at least 0.1/-0.1 but also statistically significant ( $p < 0.05$ ).



**Figure 6-12 Volcano plot for ageing splicing factor expression in Influenza.**

The transverse volcano plot shows the gene expression changes in Influenza patients which are associated with advancing age. Grey points represent non-significant changes. Green points represent a change in beta value of at least 0.1 or -0.1, red points represent genes which have a change in beta of at least 0.1/-0.1 but also statistically significant ( $p < 0.05$ ).

#### 6.3.4 Beta-coefficient distribution with age

For datasets of over 5000 samples, visual interpretation of distribution is recommended. Many of the tools for distribution quantitation including the Shapiro-Wilk test in R, cannot be applied data with over 5000 samples (443).

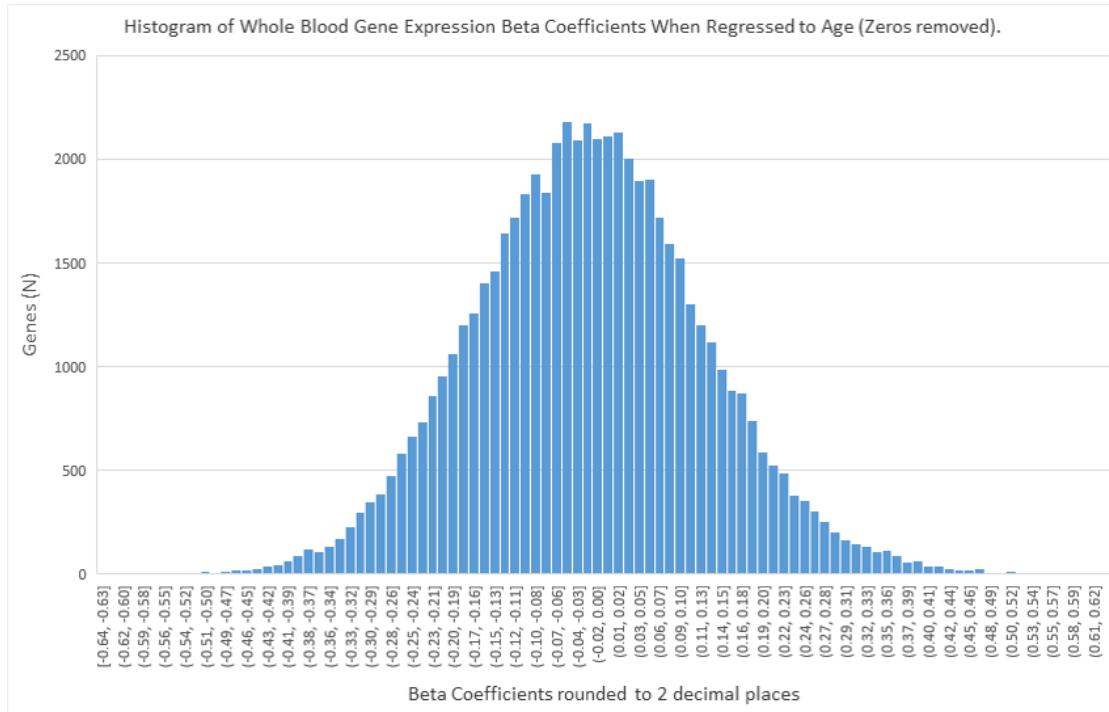
For the purposes of initial exploratory data analysis, the beta coefficients of all features were first divided into groups based on value and plotted on a bar chart to understand the overall association of gene and isoform transcripts with age in these cohorts. For COVID19, the associations of gene expression and relative isoform abundance with age appear to follow a normal distribution with little weighting in either direction. This data shows a peak at around 2100 genes which have a beta-coefficient value between -0.03 and -0.04. For relative isoform abundance, the peak was around 4600 isoforms which had a beta coefficient between 0.03 and 0.04. There was a greater degree of variation about the normal distribution for isoform abundance than there was for gene expression, despite a greater number of features.

For influenza the distribution of betas for genes specifically was less representative of a normal distribution; there appeared to be skew in the data whereby slightly more genes were negatively associated with age than were positive. The number of genes which had negative beta coefficient values decreased slightly more linearly rather than exponentially as such there appeared to be a small majority of genes which had a negative beta coefficient value and therefore had a decreased expression with age. Influenza genes' betas peaked at 0.05-0.06 with around 2100 genes in this category Isoforms in influenza followed a more normal distribution, but similar to the transcripts in COVID19 there was more deviation from the normal distribution curve, with staggered faces rather than a smooth curve. Peak around 4600 genes for the group -0.1 – 0.0.

The normal distribution of features did not provide significant new insight and suggests that while there may be some specific features and processes or pathways associated with ageing and immunosenescence to a greater degree, largely the molecular drivers of the process are related to general wholistic processes leading to dysfunction and information loss, as opposed to a specific ageing program as some have hypothesised. This highlights the importance or targeting the macro-processes such and transcriptional regulation and splicing as whole and suggests that the more reductionist strategies and approaches aimed at specific pathways will continue to have limited success in the field of immunosenescence and ageing as a whole. This information comes with the

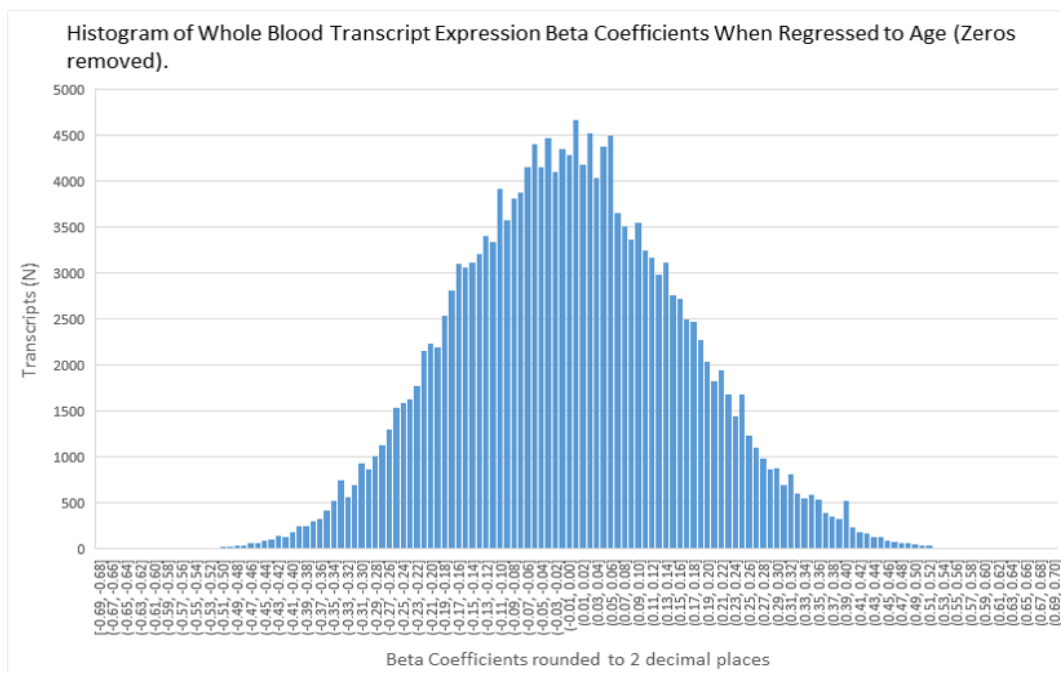


caueat, that whole blood is being for this investigation, and immune only tissues might have a more nuanced profile.



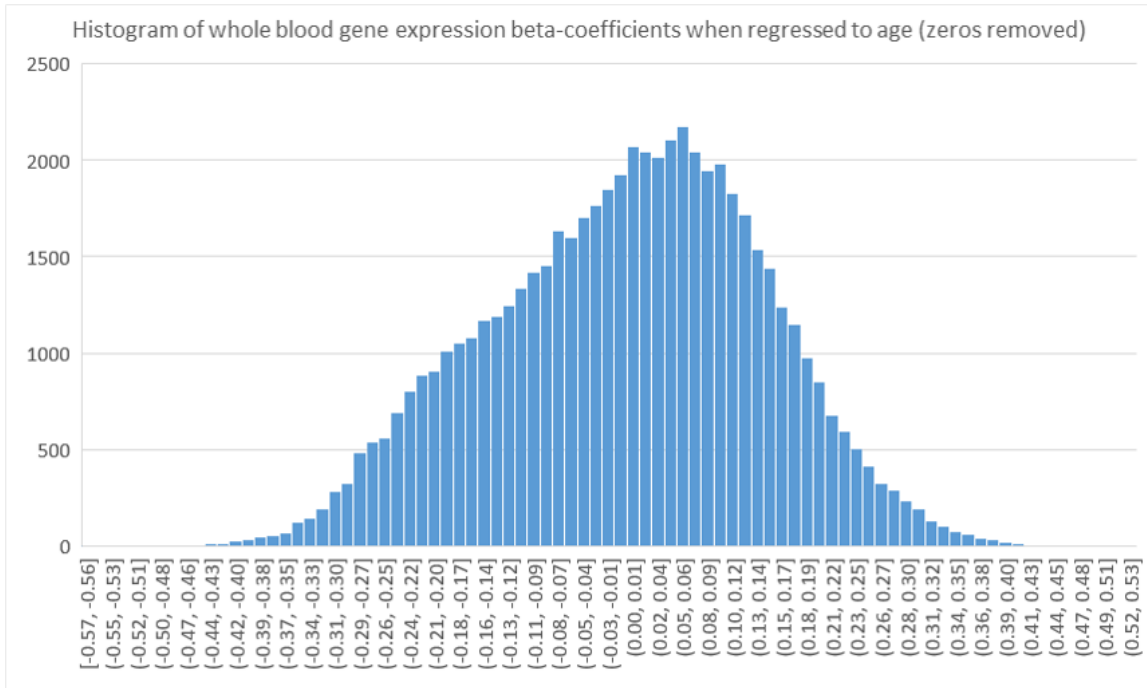
**Figure 6-13 Histogram of Beta-coefficients for genes in Covid19**

Figure shows histogram representing the number of genes which reside in each beta-coefficient interval grouping when regressed against age in Covid19 patients.



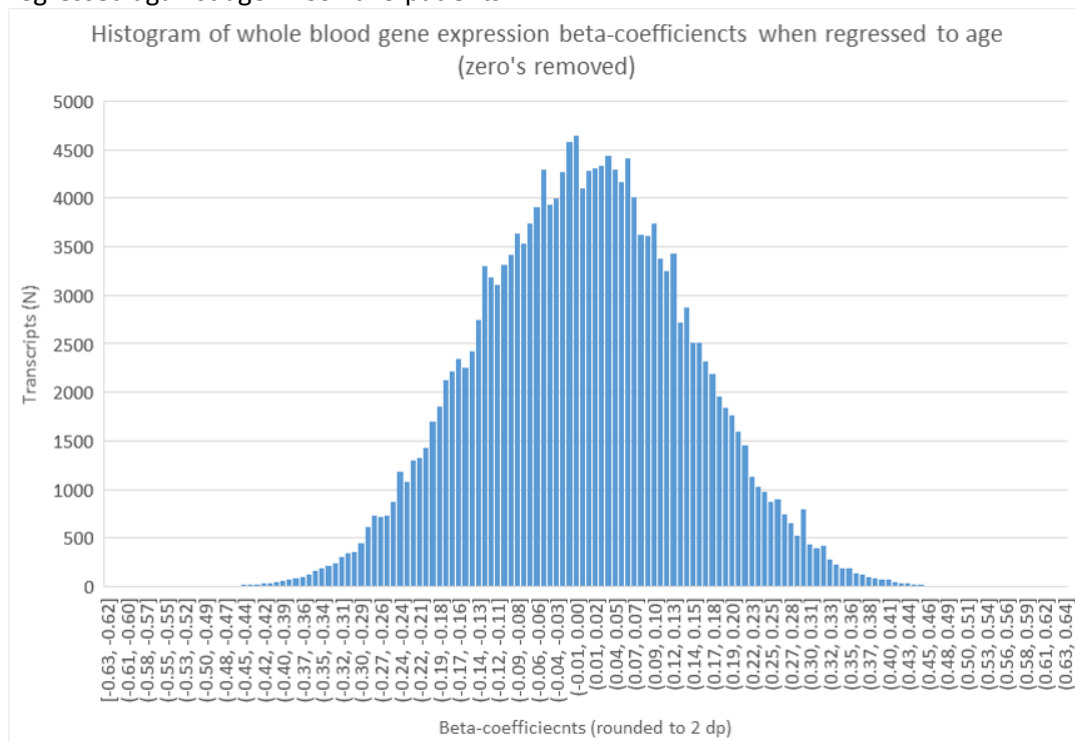
**Figure 6-14 Histogram of Beta-coefficients for genes in Covid19**

Figure shows the number of isoforms for which the relative abundance beta coefficient value corresponding to each group when regressed against age in Covid19 patients.



**Figure 6-15 Histogram of Beta-coefficients for genes in Influenza**

Figure shows the number of genes for each beta-coefficient value grouping when regressed against age in Covid19 patients.



**Figure 6-16 Histogram of Beta-coefficients for transcripts in Influenza**

Figure shows the number of isoforms for which the relative abundance beta coefficient value corresponding to each group when regressed against age in Influenza patients.

### 6.3.5 Quantifying the groups of genes associated with ageing in the cohorts.

The R package 'UpSet plot' was used to visualise where the lists of genes from the different conditions had shared values. UpSet plot acts as a scalable alternative to Venn diagrams and as such can visually demonstrate the overlap between the DEG/DTU between infections, but also compare the age associated DEG/DTU with those which make the infections responses distinct, showing the proportion of genes and isoforms which make up the distinct responses but are then affected by age. This plot also allows us to compare the proportions of information in the transcriptome which are related to gene expression and alternative splicing and the effect ageing has on these processes.

Around one third as many genes are associated with ageing as transcripts, speaking to the importance of splicing regulation and age-related disease (COVID19 AA-DTU = 7879, Influenza AA-DTU = 7377 compared with COVID19 DEG = 2609, Influenza DEG = 2524).

There were a high number of genes unique to these categories (COVID19 DTU = 4516, Influenza DTU = 4051) than gene expression (COVID19 =1662, Influenza =1720). Around 25% (2094) of the genes which underwent differential transcript use were, shared between the cohorts demonstrating that splicing underpins specific immune response processes, but that these processes are also extremely nuanced and infection specific.

In patients with COVID19, 338 genes underwent alternative splicing and differential transcript use with age. In Influenza patients 294 genes were both expressed and spliced differently with age,

There were 212 genes which exhibited both differential transcripts use between the infections, but also were affected by age and saw those same differentially expressed transcripts change with age.

Only one differentially expressed gene between the cohorts was also seen to undergo differential expression with advancing age in both cohorts. This was SATB2, a gene which produces a nuclear matrix protein and an important regulator of epigenetic chromatin remodelling

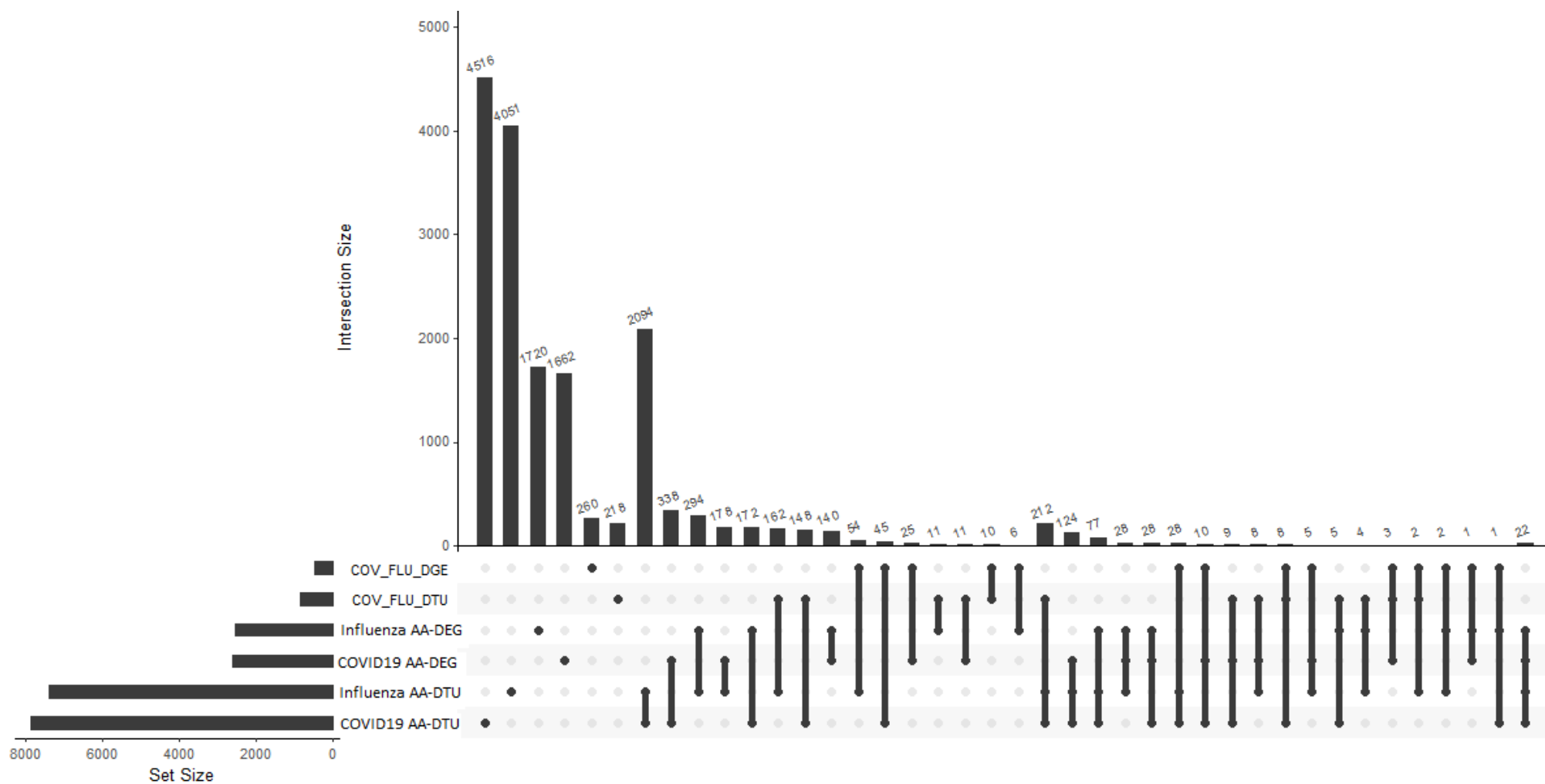


Figure 6-17 Upset plot - Matrix Based Comparison of Gene Expression Panel Data Sets

UPSET PLOT LEGEND

COV-FLU\_DGE = DEG between Covid and Influenza, COV\_FLU\_DTU= Genes with differential transcript use between infections. AA DEG = Age associated differentially expressed genes. AA-DTU Age associated Differential transcript usage. Left bars indicate size of the set. Top bars indicate size of subset/set overlap, bottom right point matrix indicates which sets (points joined by bars) are included in the analysis.

### 6.3.6 Analysis of transcriptomic convergence results

To gain further insight into the age-related changes in the transcriptomes, a conservative analysis was performed to allow identification of converging expression with age in genes and transcripts which were differentially expressed between cohorts and gene set enrichment analysis was performed.

PANDOMICS calculated 25778 genes which were differentially expressed had higher expression in COVID19, and 10995 had higher expression in influenza. These genes were cross referenced with the 9658 (COVID19) and 12037 (influenza) genes which showed beta values indicating a change in the opposite direction to which they started so were converging (e.g., genes originally up in Covid19 would have a negative beta, and vice versa). Only those genes with an absolute difference in betas greater than 0.25 were selected for. This is to represent the 0.1 beta association previous used, in both directions plus a further 25% change in betas for stringency. This yielded a list of 1456 genes for COVID19 and 937 genes for influenza. This process is represented in Table 6-7.

These gene sets were entered into the TOPPGENE online tool (363) with no further statistical filtering for gene set enrichment analysis. Over 200 processes were enriched for. The top 20 biological processes which were enriched for are shown in Table 6-9 and Table 6-10.

For those gene which were initially expressed higher in COVID19 but then converged with age, there was a strong enrichment of processes related to mitotic processes, cell cycle control and host immune adaptive response. Among the biological processes which were initially expressed high in influenza but converged with age, were regulatory genes involved in gene silencing, and microRNA genes, Toll like receptor 9, bioenergetic processes, including regulation of glucose transmembrane transport, negative regulation of cAMP-dependent protein kinase activity, and fatty acid derivative metabolic process. This process was repeated for the isoforms with differential transcript usage however the BANDITS software automatically adjusts the p-values with Bonferroni correction method, and so this analysis was completed with enhanced stringency. There were 2045 isoforms differentially expressed after correction, of these 1021, or almost exactly 50% had a higher relative abundance in COVID19 and 1024 had a higher relative abundance in influenza.

After comparing these with the isoforms which showed contrasting expression profiles with age in the two infections, 182 isoforms were present which started up in COVID19 patients and converged

with age and 176 isoforms were present which started with higher expression in influenza and converged with age. The respective genes from which the isoforms originated were entered into the TOPPGENE online tool with no further statistical filtering, for gene set enrichment analysis. Over 200 processes were enriched for in both COVID19, and influenza sets. The top 20 biological processes which were enriched for in converging splicing patterns for each cohort are shown in tables Table 6-11 and Table 6-12. For COVID19 the biological processes which were enriched in the gene set were related to phagocytosis, antibody dependant cytotoxicity, hypersensitivity reactions inflammation and protein stability, representing the macro-processes of innate and adaptive immune responses.

Interestingly, the biological processes which were initially higher in influenza but converged with age also represented hypersensitivity reactions, inflammatory responses, antibody mediated cytotoxicity, hemopoiesis, and leukocyte differentiation.

**Table 6-7 Cross referencing and comparison of gene list to establish evidence of convergence.**

Genes which were differentially expressed and initially up in Covid = <b>25788</b>	Genes which were differentially expressed and initially up in Influenza = <b>10995</b>
Genes with change in beta > 0.25 = <b>1936</b>	Genes with change in beta > 0.25 = <b>2412</b>
Genes which had negative beta values for COVID19 and positive beta values for influenza = <b>9658</b>	Genes which had positive beta values for COVID19 and negative beta values for influenza = <b>12037</b>
Number of genes which were initially significantly higher in COVID19 then displayed expression convergence = <b>1456</b>	Number of genes which were initially significantly higher in influenza then displayed expression convergence = <b>937</b>

**Table 6-8 Cross referencing and comparison of isoform list to establish evidence of convergence.**

Isoforms with significant differential isoform abundance, initially up in Covid = <b>1021</b>	Isoforms with significant differential isoform abundance, initially up in Influenza = <b>1024</b>
Isoforms which had negative beta values for COVID19 and positive beta values for influenza = <b>49582</b>	Isoforms which had positive beta values for COVID19 and negative beta values for influenza = <b>48252</b>
Number of isoforms which were initially significantly higher in COVID19 then displayed expression convergence = <b>182</b>	Number of isoforms which were initially significantly higher in influenza then displayed expression convergence = <b>176</b>



**Table 6-9 GO analysis for list of 'COVID19 genes' which showed converging expression.**

	<b>ID</b>	<b>Name</b>	<b>Source</b>	<b>p-value</b>
<b>1</b>	GO:0002250	adaptive immune response		8.43E-10
<b>2</b>	GO:1903047	mitotic cell cycle process		6.72E-08
<b>3</b>	GO:0000278	mitotic cell cycle		9.67E-08
<b>4</b>	GO:0002377	immunoglobulin production		9.99E-07
<b>5</b>	GO:0140014	mitotic nuclear division		2.21E-06
<b>6</b>	GO:0002440	production of molecular mediator of immune response		2.50E-06
<b>7</b>	GO:0010948	negative regulation of cell cycle process		3.04E-06
<b>8</b>	GO:0002449	lymphocyte mediated immunity		3.19E-06
<b>9</b>	GO:0007346	regulation of mitotic cell cycle		8.30E-06
<b>10</b>	GO:0044772	mitotic cell cycle phase transition		8.42E-06
<b>11</b>	GO:0002684	positive regulation of immune system process		8.56E-06
<b>12</b>	GO:0000070	mitotic sister chromatid segregation		8.59E-06
<b>13</b>	GO:0007093	mitotic cell cycle checkpoint signalling		8.59E-06
<b>14</b>	GO:0010564	regulation of cell cycle process		8.98E-06
<b>15</b>	GO:0000075	cell cycle checkpoint signalling		9.33E-06
<b>16</b>	GO:0022402	cell cycle process		1.13E-05
<b>17</b>	GO:1901990	regulation of mitotic cell cycle phase transition		1.14E-05
<b>18</b>	GO:1901991	negative regulation of mitotic cell cycle phase transition		1.19E-05
<b>19</b>	GO:0006260	DNA replication		1.60E-05
<b>20</b>	GO:0000819	sister chromatid segregation		1.75E-05

**Table 6-10 GO analysis for list of 'Influenza genes' which showed converging expression.**

<b>ID</b>	<b>Name</b>	<b>Source</b>	<b>p-value</b>
<b>1</b>	GO:0046324	regulation of glucose import	1.18E-04
<b>2</b>	GO:0010827	regulation of glucose transmembrane transport	1.22E-04
<b>3</b>	GO:0002320	lymphoid progenitor cell differentiation	1.25E-04
<b>4</b>	GO:1904659	glucose transmembrane transport	2.53E-04
<b>5</b>	GO:0048771	tissue remodelling	2.65E-04
<b>6</b>	GO:0008645	hexose transmembrane transport	2.97E-04
<b>7</b>	GO:0021816	extension of a leading process involved in cell motility in cerebral cortex radial glia guided migration	3.01E-04
<b>8</b>	GO:0015749	monosaccharide transmembrane transport	3.66E-04
<b>9</b>	GO:0097581	lamellipodium organization	3.86E-04
<b>10</b>	GO:0071345	cellular response to cytokine stimulus	4.24E-04
<b>11</b>	GO:0046323	glucose import	4.69E-04
<b>12</b>	GO:0034097	response to cytokine	5.73E-04
<b>13</b>	GO:0002682	regulation of immune system process	6.78E-04
<b>14</b>	GO:0034219	carbohydrate transmembrane transport	8.27E-04
<b>15</b>	GO:0002764	immune response-regulating signalling pathway	8.50E-04
<b>16</b>	GO:2001222	regulation of neuron migration	8.81E-04
<b>17</b>	GO:0033133	positive regulation of glucokinase activity	9.94E-04
<b>18</b>	GO:0070782	phosphatidylserine exposure on apoptotic cell surface	9.94E-04
<b>19</b>	GO:0045588	positive regulation of gamma-delta T cell differentiation	9.94E-04
<b>20</b>	GO:0007135	meiosis II	1.11E-03

**Table 6-11 GO analysis: isoforms which were initially higher COVID19.**

<b>ID</b>	<b>Name</b>	<b>p-value</b>
<b>1</b>	GO:0006909 phagocytosis	1.37E-06
<b>2</b>	GO:0001812 positive regulation of type I hypersensitivity	3.06E-06
<b>3</b>	GO:0001810 regulation of type I hypersensitivity	4.58E-06
<b>4</b>	GO:0016068 type I hypersensitivity	4.58E-06
<b>5</b>	GO:0007005 mitochondrion organization	2.77E-05
<b>6</b>	GO:0001798 positive regulation of type IIa hypersensitivity	2.99E-05
<b>7</b>	GO:0002894 positive regulation of type II hypersensitivity	2.99E-05
<b>8</b>	GO:0001796 regulation of type IIa hypersensitivity	3.63E-05
<b>9</b>	GO:0001788 antibody-dependent cellular cytotoxicity	3.63E-05
<b>10</b>	GO:0002892 regulation of type II hypersensitivity	3.63E-05
<b>11</b>	GO:0001794 type IIa hypersensitivity	5.14E-05
<b>12</b>	GO:0002445 type II hypersensitivity	5.14E-05
<b>13</b>	GO:0050766 positive regulation of phagocytosis	5.40E-05
<b>14</b>	GO:0031648 protein destabilization	6.05E-05
<b>15</b>	GO:0002885 positive regulation of hypersensitivity	7.01E-05
<b>16</b>	GO:0031647 regulation of protein stability	7.94E-05
<b>17</b>	GO:0002866 positive regulation of acute inflammatory response to antigenic stimulus	9.28E-05
<b>18</b>	GO:0097278 complement-dependent cytotoxicity	1.06E-04
<b>19</b>	GO:0030100 regulation of endocytosis	1.09E-04
<b>20</b>	GO:0002883 regulation of hypersensitivity	1.20E-04

**Table 6-12 GO analysis: isoforms which were initially higher Influenza.**

	<b>ID</b>	<b>Name</b>	<b>p-value</b>
<b>1</b>	GO:0001812	positive regulation of type I hypersensitivity	1.71E-06
<b>2</b>	GO:0001810	regulation of type I hypersensitivity	2.56E-06
<b>3</b>	GO:0016068	type I hypersensitivity	2.56E-06
<b>4</b>	GO:0001798	positive regulation of type IIa hypersensitivity	1.68E-05
<b>5</b>	GO:0002894	positive regulation of type II hypersensitivity	1.68E-05
<b>6</b>	GO:0001796	regulation of type IIa hypersensitivity	2.03E-05
<b>7</b>	GO:0001788	antibody-dependent cellular cytotoxicity	2.03E-05
<b>8</b>	GO:0002892	regulation of type II hypersensitivity	2.03E-05
<b>9</b>	GO:0001794	type IIa hypersensitivity	2.88E-05
<b>10</b>	GO:0002445	type II hypersensitivity	2.88E-05
<b>11</b>	GO:0002885	positive regulation of hypersensitivity	3.94E-05
<b>12</b>	GO:0050729	positive regulation of inflammatory response	4.58E-05
<b>13</b>	GO:0002521	leukocyte differentiation	4.90E-05
<b>14</b>	GO:0002866	positive regulation of acute inflammatory response to antigenic stimulus	5.22E-05
<b>15</b>	GO:0097278	complement-dependent cytotoxicity	5.96E-05
<b>16</b>	GO:0030097	hemopoiesis	6.14E-05
<b>17</b>	GO:0002861	regulation of inflammatory response to antigenic stimulus	6.19E-05
<b>18</b>	GO:0002883	regulation of hypersensitivity	6.75E-05
<b>19</b>	GO:0048534	hematopoietic or lymphoid organ development	9.42E-05
<b>20</b>	GO:0002863	positive regulation of inflammatory response to antigenic stimulus	1.06E-04

### **6.3.7 Classification of infection based on gene expression.**

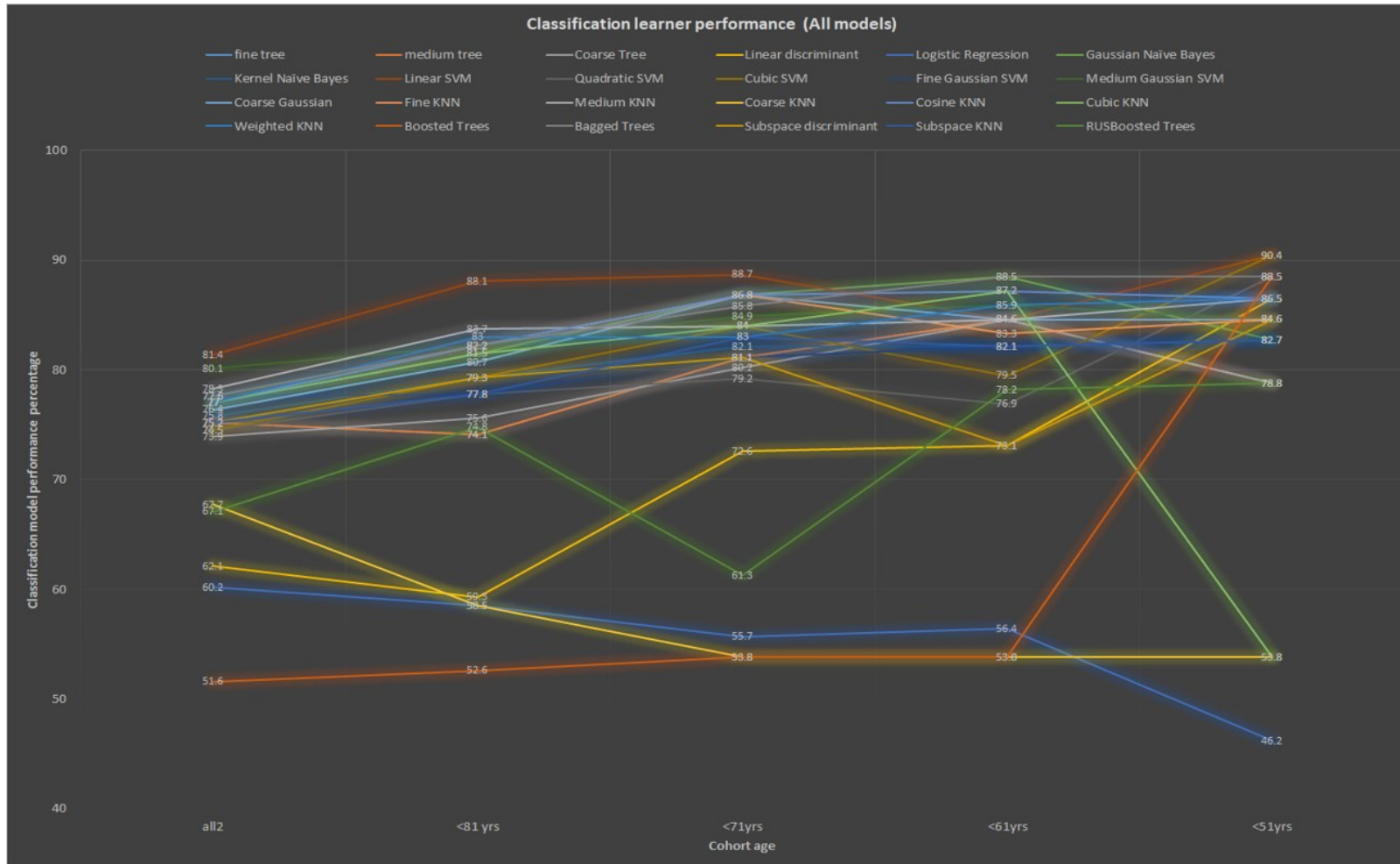
The classification using the top 100 differentially expressed genes as features in MATLAB showed that the combined and mean accuracies of all available models increased sequentially (Table 6-13), when the group was made younger by removing the eldest, despite the group becoming smaller (and so being more likely to be affected by stochasticity and having less data to train models on). When this process was reversed, the opposite was seen, and decreased accuracy was observed. The totals, the mean values and peak performance all followed the same trends. Linear support vector machines appeared to be the best performing classification model, and so were used in downstream analysis, no hyperparameter tuning was used in this classification experiment.

When the performance (measured by overall accuracy %) of these models was plotted on graphs, the classification performance percentage clearly showed a positive association with advancing age (Figure 6-18) which tended to start between 70-80% for most models and climb to between 80 and 90% when only patients under 51 were considered. Likewise, a negative association was also visible when the group became older (Figure 6-19), with most performance percentages starting at between 70-80% and falling to between 60 and 70% as the group became only populated by people of 60 years of age.

**Table 6-13 -Classification Machine Learning Model Performance**

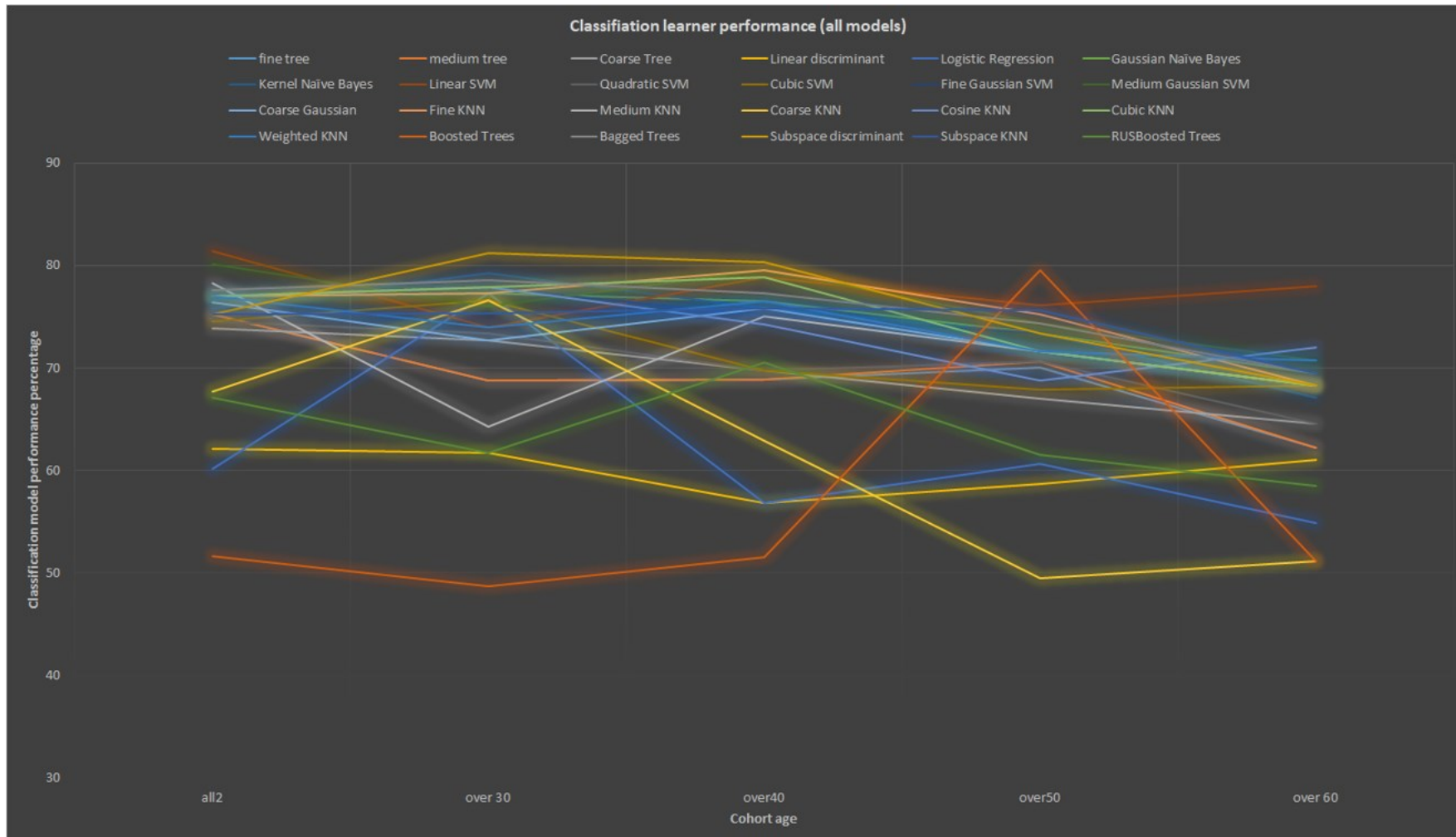
Model subtype	all	<81	<71	<61	<51	all	>30	>40	>50	>60
fine tree	75.2	74.1	81.1	84.6	78.8	75.2	68.8	68.9	70	62.2
medium tree	75.2	74.1	81.1	84.6	78.8	75.2	68.8	68.9	70.6	62.2
Coarse Tree	73.9	75.6	80.2	84.6	78.8	73.9	72.7	69.7	67	64.6
Linear discriminant	62.1	59.3	72.6	73.1	86.5	62.1	61.7	56.8	58.7	61
Logistic Regression	60.2	58.5	55.7	56.4	46.2	60.2	77.3	56.8	60.6	54.9
Gaussian Naïve Bayes	77	81.5	86.8	88.5	82.7	77	77.3	76.5	73.4	69.5
Kernel Naïve Bayes	75.8	79.3	82.1	82.1	82.7	75.8	79.2	75.8	73.4	67.1
Linear SVM	81.4	88.1	88.7	84.6	90.4	81.4	74	78.8	76.1	78
Quadratic SVM	74.5	77.8	79.2	76.9	88.5	74.5	73.4	69.7	70.6	64.6
Cubic SVM	74.5	79.3	84	79.5	90.4	74.5	76.6	69.7	67.9	68.3
Fine Gaussian SVM	77.6	77.8	81.1	82.1	82.7	77.6	77.9	75	73.4	68.3
Medium Gaussian SVM	80.1	82.2	84.9	85.9	84.6	80.1	76	79.5	74.3	70.7
Coarse Gaussian	76.4	80.7	86.8	84.6	84.6	76.4	72.7	75.8	71.6	68.3
Fine KNN	77	82.2	86.8	83.3	84.6	77	77.3	79.5	75.2	68.3
Medium KNN	78.3	83.7	84	84.6	86.5	78.3	64.3	75	71.6	68.3
Coarse KNN	67.7	58.5	53.8	53.8	53.8	67.7	76.6	62.9	49.5	51.2
Cosine KNN	77	82.2	86.8	87.2	86.5	77	77.9	74.2	68.8	72
Cubic KNN	77	81.5	84	87.2	53.8	77	77.9	78.8	71.6	68.3
Weighted KNN	77	83	83	85.9	86.5	77	74	76.5	71.6	70.7
Boosted Trees	51.6	52.6	53.8	53.8	88.5	51.6	48.7	51.5	79.5	51.2
Bagged Trees	77.6	82.2	85.8	88.5	88.5	77.6	78.6	77.3	74.3	69.5
Subspace discriminant	75.2	79.3	81.1	73.1	84.6	75.2	81.2	80.3	73.4	68.3
Subspace KNN	75.2	77.8	83	82.1	82.7	75.2	75.3	75.8	7.6	69.5
RUSBoosted Trees	67.1	74.8	61.3	78.2	78.8	67.1	61.7	70.5	61.5	58.5
<b>TOTALS</b>	<b>1764.6</b>	<b>1826.1</b>	<b>1887.7</b>	<b>1905.2</b>	<b>1930.5</b>	<b>1764.6</b>	<b>1749.9</b>	<b>1724.2</b>	<b>1612.2</b>	<b>1575.5</b>
Mean	73.5	76.1	78.7	79.4	80.4	73.5	72.9	71.8	67.2	65.6
Mode	77	82.2	81.1	84.6	78.8	77	77.3	69.7	73.4	68.3
Median	75.5	79.3	82.55	83.95	84.6	75.5	75.65	75	71.6	68.3
min	51.6	52.6	53.8	53.8	46.2	51.6	48.7	51.5	7.6	51.2
max	81.4	88.1	88.7	88.5	90.4	81.4	81.2	80.3	79.5	78
range	29.8	35.5	34.9	34.7	44.2	29.8	32.5	28.8	71.9	26.8

Performance of the classifier models. Conditional formatting is applied to vertical columns, with green indicating highest performance and red indicating lowest performance. For totals, means and max conditional formatting is applied horizontally.



**Figure 6-18 Machine Learning Classification Performance Using All Models: Decreasing Age**

Figure shows performance trends for applied models. Cohort age decreases from left to right, performance of model is along Y axis.



**Figure 6-19 Machine Learning Classification Performance Using All Models: Increasing Age**

Figure shows performance trends for applied models. Cohort age increases from left to right, performance of model is along Y axis.



**6.3.8 Application of classification learner linear support vector machine to split cohort.**

The results showed that models using isoform abundance as features always performed better than their gene expression trained counterparts. When considering the differential transcript use, the accuracy of the training models was higher in the young, as expected. However, this pattern was not observed in the results from the test cohort. When considering gene expression, the training models performed similarly but the test models were marginally (~7%) worse in the younger cohort. The performance plots can be found in appendix A.12.

**Table 6-14 Performance of classification learning machine learning models on transcriptomic data.**

Gene expression		Differential transcript use	
Cohort subset	Overall classification accuracy	Cohort subset	Overall classification accuracy
Old Training	82.6%	Old Training	85.5%
Old Test	83.3%	Old Test	91.7%
Young Training	82.6%	Young Training	91%
Young Test	75%	Young Test	91.7%

This table represents the performance of the models applied to transcriptomic data from the Influenza and COVID19 cohorts. Models were trained and tested on the top 100 features from either differential gene expression or differential transcript usage.

## 6.4 Discussion

### 6.4.1 Significance

With advancing age, chronic systemic inflammation becomes ubiquitous, which blunts the detectable response of the immune system to challenges due to the chronic low-level stimulation and also masks inflammation caused by infectious disease (444, 445). The transcriptomic profiles between COVID19 and Influenza patients have a high degree of overlap in 95% confidence interval ellipses when assessed through a PCA plot. In an attempt to explain some of this overlap, those over the age of 65 years were removed, and a reduction which can be visually observed in the overlap of 95% confidence intervals occurs. This suggests that the transcriptomic profiles of people with these two viral infections are more distinct in the young. These transcriptomic profiles of the groups are not, however, able to be completely separated in the PCA plot, despite the exclusion of the elderly in the existing cohort. This could be partially a result of the cohort still having some 'older' individuals for whom the process of immunosenescence has already begun. As such, the limited effects seen in the PCA may be in part still be reflecting the impact of age on the immune response. Age continues to be a factor, but the transcriptomes are also likely to be influenced by individual factors, technical and statistical artefacts, and limitations.

There has been some observation of gene expression profiles from different tissues converging with age and specific cellular genes are down regulated and generic genes are upregulated. In addition, immunosenescence delivers a blunted immune response with advancing age. We hypothesised that some of the convergence in gene expression with age may be a result of the loss of specific immune gene expression, leading to a loss of specificity in the response to the two infections. We further hypothesised that as alternative splicing and thus isoform abundance was demonstrated to be a large portion of the specific transcriptomic response to infection, there may be a loss in distinct profiles of isoform abundance with age in the immune genes as dysregulation sets in.

Individual gene level investigation showed that some of the most differentially expressed genes between infection cohorts had an expression profile which appeared to converge with advancing age (JUN/IGHG1), however this appear to be the result of changes only in the gene expression in COVID19. To investigate this further, genes and isoforms which had differential expression overall were cross referenced with those which had contrasting expression changes with age as determined by beta coefficient values from linear regression. Our methods were extremely conservative and

used robust P-value adjustment methods for differential expression and stringent beta value cut-offs for age association. Despite this we found many genes converged with age.

It is interesting that despite this apparent convergence with age, these genes remained some of the most differentially expressed. The statistical corollary of which is that there is some likelihood that other genes may have been significantly differentially expressed in the younger population, but lost significance with the addition of the older individuals and so have not been detected. This is supported by the finding that only a single differentially expressed gene of 472 (Bonferroni adjusted  $p < 0.05$ ) between the infections was shown to have changes in expression related to age, despite visible associations with age in many differentially expressed genes. Those genes which experienced the most robust age-related changes would likely have lost some of their statistical significance if differentially expressed in the non-aged group. It would be of interest to compare genes differentially expressed between the cohorts in the various decades of life to see how profound the loss of differentially expressed genes is in infectious disease with age.

Volcano plots of the genes undergoing differential gene expression showed strong association of *NT5E* with advancing age in COVID19. *NT5E* (AKA *CD73*) is generally accepted to be immunosuppressive (446). It is expressed on the surface of CD8+ T lymphocytes, whose numbers decline with age. *NT5E* dephosphorylates AMP into adenosine and organic phosphate. Adenosine acts through G-protein coupled receptor mediated signalling from the cell surface to regulate intracellular cyclic AMP levels, which mediates immunosuppression (447). Therefore, its decreased presence in an ageing cohort, especially one which is harbouring an infection associated with cytokine storm, is not surprising. This could represent the loss of the immunosuppressive mechanisms which prevent chronic inflammation, or it might be a result of the inappropriate hyper-inflammatory response seen in hospitalised COVID19 patients, or some combination of the two. Literature demonstrates that the *NT5E* levels are associated with mild and severe COVID19, and that blood from these patients is actually able to deplete *NT5E* levels in health controls (447). Therefore, the negative association of *NT5E* with advancing age in this cohort might be a feature of immunosenescence/inflammageing. Moreover, in combatting cytokine storm, mesenchymal stem cells and extracellular vesicles which harbour *NT5E* have been investigated and so there is a good chance that this molecule might represent an interesting therapeutic target (448, 449).

The strong negative association of *WLS* with advancing age in the cohort is also linked to these processes and *NT5E* also appears to be tightly associated with the *WLS/WNT* axis responsible for maintenance of telomeres and implicated in immunosenescence (450, 451). Overexpressed gene

*CNTNAP3* has been implicated in some viral infections, but interesting associations are also found in conditions which bridge the immune/neurological axis. The gene which mediates neuron to glial cell communications is overexpressed in PTSD with immune involvement (452), pre-eclampsia (453), autism spectrum disorder (454), major depressive disorder (455), Crohn's disease (456). This gene and its associated pathway or process may therefore be part of the way in which COVID19 elicits its effect on the neurological system.

The decrease in splicing factor expression with advancing age echoes work performed in other tissues which shows that splicing factors directly correlate to age (457), and that long lived species such as the naked mole rat have very high splicing factor expression throughout life (458). As splicing contributes to the diversity of eukaryotes by increasing number of possible of RNA species, a loss of splicing through the decline in these factors will lead to a convergence of RNA isoform transcriptome, much as we have observed the same phenomenon in the expression of key genes which mediate specific function in the immune system. These findings suggest that a pivotal event in the process of ageing is the loss of biological information through convergence of the transcriptome.

The strict adherence to normal distribution of beta-coefficients appears to support the hypothesis that through ageing, transcriptomes undergo a loss of information. One would expect that as the stochastic nature of the accumulated genetic and epigenetic lesions which lead to changes in expression with age would produce a pattern of distribution which resembles a normal distribution. Encouragingly, the isoform abundance histograms also follow this normal distribution when correlated to age. This might indicate that the loss of information and subsequent convergence of isoform level information in the transcriptomes follows this pattern. For example, if all the molecular mechanisms of regulation of gene expression, both up and down, were to be slowly lost, one would expect to see this normal distribution pattern. Whereas if specific transcriptional programs were activated during ageing and others were deactivated, the transcriptional association values would be less likely to be so representative of a normal distribution.

The UpSet matrix shows more genes experiencing age associated DTU compared to the number of genes experiencing age associated DEG in both infections. This suggests that dysregulation of the alternative splicing process may have more impact on age related decline in immune function, than does gene expression. Consequently, alternative splicing may hold more promise for developing therapeutics which mitigate the effects of ageing in the immune system. Interestingly, the majority of these changes seem to be also infection specific, as only around 20-25% of these are shared between infections (N= 2094).

The three classical forces acting on a transcriptome during infection are a) the direct effect of the virus on the host cell, b) the host immune response to the virus, c) the progression of disease in the host. As age progresses, and the immune response is blunted, a fourth pressure comes into effect. Deconvoluting which changes are a result of which process, and therefore identifying targets for therapeutic intervention is not an easy process, complicated by the lack of a control group in this setup. However, by regressing the features to age and comparing them between infections, it has been possible to identify sets of genes which are affected by age and may represent useful important novel targets in immunosenescence. An important next step to resolve targets and pathways of the highest priority would be to compare DEG and DTU in young vs older patients with infections. This approach would mitigate some of the lost signal.

Only a single gene was differentially expressed between infections but also had significant age-related changes in both infections and thus likely represents an important target in immunosenescence. This was *SATB2*; a gene which produces a nuclear matrix protein and an important regulator of epigenetic chromatin remodelling. In congruence with our findings, overexpression of this product is able to rejuvenate bone mesenchymal stem cells and regenerate the skeleton to improve bone mineral density in animal models and prevents stem cell senescence by maintaining *NANOG* expression (459).

On the splicing side, the 212 genes which experienced DTU between infections, but also experienced DTU with age, are also likely to be low hanging fruit in terms of pathways which are disease specific and also age-associated. It may therefore be possible to target these by upregulating specific splicing factors and ameliorate some of the age associated effects of the infections, especially if these genes represent immune processes.

Convergence of the transcriptome at the gene expression level has been reported for the first time in publication this year (2022). The seminal work demonstrates that different tissues in the body experience convergence towards a common transcriptome with ageing which overexpress ubiquitous genes and have reduced expression of cell specific genes. (460). This is opposed to the divergence to many distinct transcriptomes for varying cell types seen during development (460). This process is termed DiCo (divergence – convergence) by the authors who discovered it. Unlike this research, our work was able to characterise the convergence in transcriptomes of host immune response; that is the lack of ability of the immune system to launch a distinct response to specific pathogens with advancing age. We found evidence of convergence of the transcriptomes through gene expression levels, and also relative isoform abundance levels. At the gene expression level

there was a preponderance of genes involved in cell cycle control and mitosis which were initially higher in COVID19 patients. Greatly increased cellular division occurs during the acute phase of infection (461), although why this is greater in one infection than the other remains unclear. Convergence of cell cycle genes could be a result of decreased overall cellular division occurring because of increasing senescence, a blunted immune response, or indeed general convergence of ubiquitously expressed genes as outlined in the literature pertaining to DiCo. Converging gene expression also occurred in genes which showed enrichment of T-cell mediated immunity, T-cell extravasation, NK T-cell activation, B-cell mediated immunity and interestingly – olfactory learning, known to be a symptom in COVID19 (462). The COVID19 induced hypoxia, has been suggested to cause changes in the glycolytic processes of the cell, disrupting the metabolic controls. The upregulation of these genes in COVID19 is likely a response to this, and our results show that over time, these responses are lost, which likely contributes to the increased morbidity and mortality from COVID19 with advancing age.

The greatest enrichment at the relative isoform abundance level, was seen in genes which contribute to hypersensitivity reactions. Specifically, type I which are mediated by IgE, and type II which are mediated by IgG and IgM. These genes were present in lists for converging isoform expression from both infections suggesting that the elements of the pathways comprised the immune response for each infection, but with time, this was lost and the delicate balance of isoforms which were needed for the immune response deregulated in old age. Overall, the results demonstrate that age has a profound effect on the immune systems regulation and specificity, and host responses in general to infection.

This information offers important opportunities for therapeutic development. If immunosenescence is partly a result of the loss of regulation of key processes, such as transcription and splicing, re-regulation becomes a therapeutic target. If regulation is lost due to declining expression of key genes involved in the process, such as splicing factors, then upregulation of these genes through therapeutic intervention may help protect the immune system from immunosenescence. Indeed, some master regulators of splicing have already been identified as potential therapeutic targets in other literature (463). In addition to endogenous age related convergence of splicing which is a consequence of loss of splicing factors with age, the viruses themselves may also contribute to the convergence of the transcriptome as it is well established that they directly modulate splicing (322). It has also been shown that the NSP16 binds to the domains of U1 and U2 small nucleolar RNA element of the ribonucleoproteins which catalyse splicing leading to global suppression of splicing

(323). The convergence seen then is a result of ageing of the cell, but also the process of infection itself. As the immune system ages, infections will be more successful at compromising host cell machinery, they which will in turn cause further convergence of the transcriptome, resulting in synergistic effects and a loss of information in the RNA of the cell, leading to increased morbidity and mortality.

The machine learning classification models training also supported the idea that specificity of immune response was lost with advancing age. Significant improvements with decreasing age of cohort, and loss in performance with increasing age, were seen regardless of reduction in sample size. This was both for peak performance of any model, but also averaging across all models; an analysis which limits any algorithm specific changes in performance that might be related to the sample number, or any user selections bias. This gives robust support to the hypothesis that transcriptomic profiles converge with age and highlights the importance of targeting immunosenescence in the elderly to rejuvenate immune function. However, immune responses are multifactorial and heterogeneous from person to person, and this is reflected in the inability of the classification models to determine which infection an individual is suffering from with 100% accuracy at any age during model training. Although the possibility of this being possible with a larger cohort of patients with fully mature but not yet aged immune systems is not able to be ruled out. Due to time limitations, it was not possible to repeat this training step optimisation process using the splicing related differential transcript abundance data. Therefore, the work rests on the assumption that performance of the linear support vector machines would be good enough to conduct a fair assessment on both data types.

Once the optimal model was identified by performance metrics in training (linear support vector machines), the data was re-partitioned to prevent information leak, and tests were conducted for classification ability for both gene expression and differential transcript use.

The results showed that despite picking the model type best suited to gene expression data, isoform abundance data was better suited to determine the infection status of the individuals at any age. It performed equally well on the test data at both ages, despite the trained dataset yielding better results in the younger cohort. For gene expression data the results indicated the metric resulted in less reliable classification of disease. Surprisingly this seemed to be worse in the younger cohort. Unfortunately, the datasets were still relatively low in number, and test datasets were as small as  $n=6$ . Further work on larger datasets and other infections would be needed to validate these results, but it does give encouraging evidence that isoform abundance is an extremely useful metric in

determining infection type and that splicing is a critical modality for host immune response, which is underutilised and poorly understood.

#### **6.4.2 Limitations**

Factors hampering the interrogation of the transcriptomes includes the cohort being opportunistic and the study not being a controlled infection model. Only patients presenting to the clinic could be assessed and therefore already have a suboptimal immune response. The exact point of infection also cannot be determined. The data is increasingly likely to be 'noisy' as it does not necessarily reflect the exact same point in the immune response and thus innate adaptive immune responses may be at slightly different stages.

When considering the splicing changes or DTU, the genes in which the change took place were used for analysis. This means the nature of those changes cannot be fully represented. An increase in one transcript can, sometimes but not always, result in the decrease of another transcript and it is not easy to know from this type of analysis which type of change is occurring.

A more granular approach would need to be adopted as the investigation continues to identify the transcript in question, it's function and the type of approach which would be useful.

There are three pressures on the immune system transcriptome; the pathogen; the host, and the disease progression. It requires a more complex analysis to tease out which changes should be prioritised as therapeutic opportunities and which are just downstream effect. Analysis techniques which take advantage of molecular topography, pathway analysis and validation experiments are required. This work however could produce exciting and useful results with far reaching benefits, and the approach could be deployed to numerous age-related diseases.

### **6.5 Conclusion**

Immunosenescence is at least in part mediated by transcriptomic convergence. It can be detected in transcriptomes of patients with infectious disease and is represented in gene expression and isoform abundance. This research has demonstrated that these two upper respiratory tract infections elicit host responses which differ in both gene expression and splicing mediated ways. We have also shown that these differences decrease with advancing age. It's likely that much of the loss of distinct



signal for each infection results from the a few master regulators, which could be targeted to maintain regulation. This is despite the inherent limitations of using a cohort which are immunologically sub-optimal and at different stages of immune activation indicating the features are robust and conspicuous. Whilst the pathways observed may be affected by the sample, the underlying processes which we have observed and documented will be common in individuals healthy and otherwise. The work shows that studying a stimulated immune system will allow unique insights depending on the stimulation and in order to fully understand the effects of immunosenescence, activation may be necessary. This work showed that there are many processes which change with age which are largely infection specific. While common features exist between the ageing transcriptomes of these two infections, the differences are more numerous. This is the case even between viral infections of relatively similar outcomes when considered in the wider context of infectious disease. It is therefore important to consider both shared pathways and individual responses in context of immunosenescence and its treatment to mitigate infection. This approach has helped identify a number of potential therapeutic targets, and importantly developed a bioinformatic pipeline which can be redeployed to other bulk RNAseq dataset to identify targets of aging within those tissues.

### **6.5.1 'DiCo' occurs at the level of immune response to infection also.**

This work has shown that with advancing age some key differentially expressed genes responsible for responding to infection stop being differentially expressed. Much like different tissue gene expression profiles converging, the specificity of the immune responses also seems to do so adding another dimension to the DiCo hypothesis. The loss in specificity of the immune response to infection may help to explain the age associated increase in morbidity and mortality in infection with age. We have re-iterated existing evidence which provides some mechanistic insights into how isoform abundance is affected by age, namely a loss in expression of splicing factors with age.

## **6.6 Statement of contributions**

This chapter is a collaborative effort between Yaron Strauch, a PhD student in the same group, and I. The gene expression data and isoform abundance data had been created during the previous

chapters, and so 'feature engineering' for machine learning was already complete. The concept of using gene expression and isoform abundance to quantify and characterise age-related changes in the transcriptome of a given pathology or disease came in part from my previous work of a similar nature (314) but also from the work of Wang et al., (70). The exploratory data analysis was my own work, the machine learning classification work was my own work. Yaron and I helped each other to understand the multi-regression work by Wang et al. Yaron created a Python script to replicate this multi-regression machine learning and create the beta-coefficients. The creation of volcano plots, histograms, UpSet plots, pathway analysis, comparisons, and matrices which compare the regression results to that obtained through the earlier chapters is my own work.

## **Chapter 7      Results: Immune ageing and infection specific ageing clocks – machine learning application**

### **7.1      Introduction**

The effect of ageing on human health is profound, and only just now being understood in a way which allows any meaningful intervention. The ageing of the immune system is a process to which much morbidity and mortality has been attributed (259-261). The consequences of immunosenescence include inflammatory disease, cancer, neurodegenerative disease and cardiovascular disease, and poor outcomes from infectious disease. The unchallenged immune system does not always provide enough functional information for understanding discrepancies in immune function, baseline signals may not accurately reflect the problems in tackling an immune challenge. In order to understand and measure the decline in tissue function with age, ageing clocks have been developed. These are a combination of ‘features’, and an associated algorithm which incorporate patient age to generate a mathematical model which shows how those features change with advancing age. This has a range of functionality including identifying the best predictive factors of advancing age in a patient’s tissue, predicting a patient’s age based on their features and comparing this to chronological age to see if they are age faster and slower than average in this tissue. These can be compared with chronological age to find out if someone is ageing in a manner which is unhealthy or likely to cause age related disease faster than would be expected. They can also be used to measure the effectiveness of treatments for this age-related disease or be used to determine clinical action on a personalised basis (330). Above all they aid in the understanding of age-related disease and help to develop therapeutic approaches. We hypothesised that an ageing clock for a challenged immune system would give a new insight into immunosenescence. The novel outputs from the previous chapters suggested that gene expression based ageing clocks are likely sub-optimal and those which incorporated splicing data would prove to be more accurate and give a wider range of targets. Some earlier work examined the possibility of predicting age using just splice site usage or isoform abundance (70). Our previous work demonstrated that the pathways affected by ageing tend to be either occurring through gene expression changes OR splicing changes, but rarely is a gene affected by both (See section 5.7) As such, we hypothesised that a combination of gene expression and isoform abundance would provide the optimal transcriptomic ageing clock.

### 7.1.1 Aims

- To use the transcriptomic data obtained from the RNAseq of whole blood in these cohorts to build two ‘disease-specific, challenged-system immune ageing clocks’ to map how transcriptome changes with advancing age; one for COVID19 patients and one for Influenza patients. These will be the first of their kind, being both disease-specific and looking at a challenged immune system, but also combining gene expression and splicing based metrics which could be used for a variety of research and industrial applications.
- To compare these and find the contribution of splicing based and gene expression-based metrics to derive some insight into the importance of these features in immunosenescence.

## 7.2 Methods

Cohorts used were identical to the previous chapter. The final bioinformatic program was produced entirely using Python. fq files of patients with either COVID19 or influenza and these were trimmed using trimmomatic (342). Using the STAR aligner (343), these reads were then aligned to the human reference genome version GRCh38, using the GENCODE V39 annotation. STAR generates raw gene counts, which are converted to TPM values in python using the formula found in 2.4.2.

For the transcript abundance, the same informatic pipeline was used as previously whereby Salmon was utilised using the selective alignment method found in the previous section (2.4.1.1.6). Briefly, first the Salmon program builds a reference file comprising the human genome (version GRCh38) and all known transcripts are known as the ‘Gentrome’. Reads are then aligned to this reference file, first to the transcripts and then to the genome if no match is found. This provides fast and accurate alignment, which is isoform aware, but also able to align reads which span non-annotated junctions. The relative isoform abundances were then calculated by comparing the total number of reads aligned to each gene, and all the isoforms it has produced.

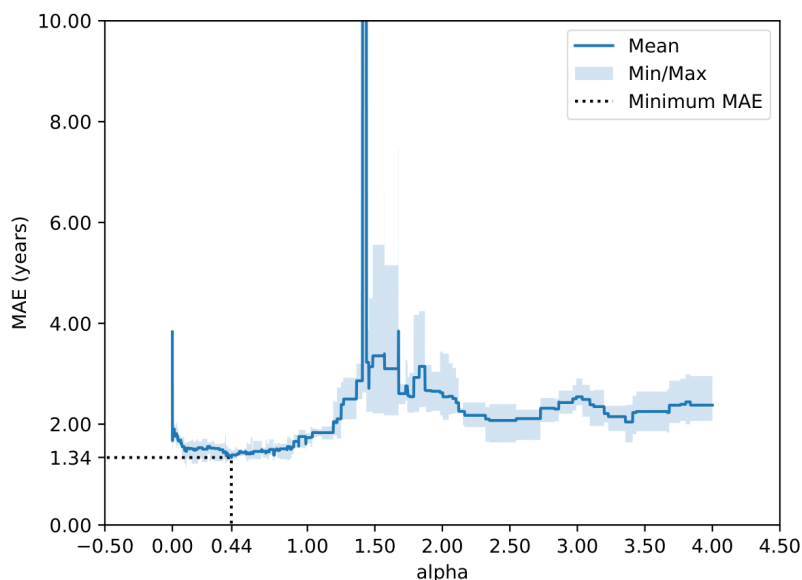
In each case these were combined into a single dataset which also included the phenotype to be predicted (patient age). Feature reduction was then implemented, first all features which only have a single value across all samples were filtered out. After this, Spearman’s correlation was applied to the complete set of features (305,165 features, of which 267,543 were transcript abundances and 61,587 were gene expression values), to reduce the set to a number which lasso regression could comfortably manage. This was set as 200,000 features based on preliminary data generated by our

group (and the experience of Yaron Strauch). The data was first standardised by subtracting the mean and dividing by the standard deviation. The resulting matrix of feature quantifications was used to regress age using a series of alpha values representing the hyperparameters for tuning model performance and using 4-fold cross validation to assess overfitting. The mean absolute expression, R-values, p-values and beta coefficients were calculated and output. The alpha value with the lowest mean absolute expression was extracted and this model was fitted to the whole dataset, as the final model. The mean absolute error is an indicator of the accuracy of the model, which is agnostic of the direction of error (i.e. +/- 5). This metric is most used when comparing the accuracy of ageing clocks (332). R squared values represent the amount of variation in the dependant variable, which can be explained by the independent variable. In this case age will not change as a direct result of changes in the transcriptome, so rather this is a measure of how well the cohorts' age can be predicted by the variation differential gene expression and differential transcript use.

## 7.3 Results Iteration 1

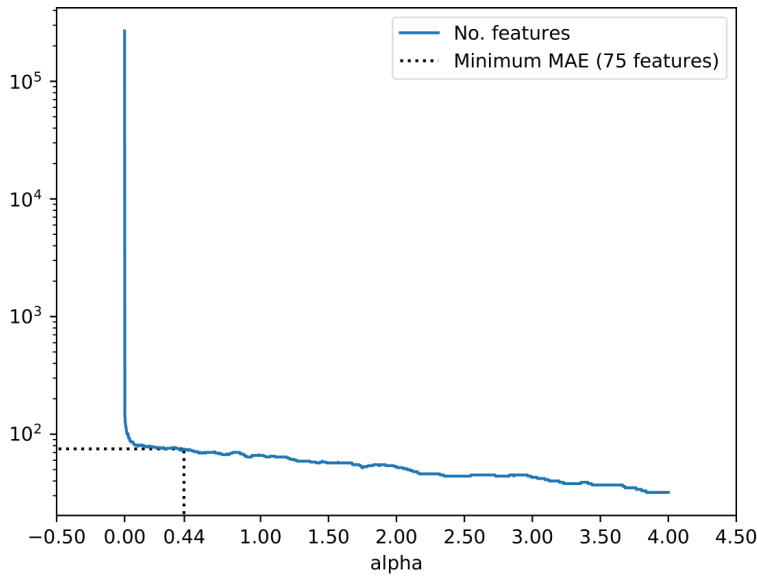
### 7.3.1 Results: First iteration

The first iteration of the ageing clocks showed performance in prediction of age superior to any currently available ageing clocks (332). For the COVID19 cohort the MAE of 1.34 years was achieved with an alpha value of 0.44 (Figure 7-1), producing 75 features (Figure 7-2) and  $R^2$  of 98% (Figure 7-3). The Influenza ageing clock was equally high performing, reaching a MAE of 1.59 years (Figure 7-4), using an alpha value of 0.55 (Figure 7-5) producing 82 features and an  $R^2$  value of 98% (Figure 7-6). These two clocks shared one (1) feature; ENST00000308873.10 an isoform of the RUNX3 gene (mRNA-RUNX family transcription factor 3, transcript variant 2, from RefSeq NM\_004350) (464). The percentages of features which were isoforms and genes were 93-7% COVID19, 91-9% Influenza and 86.6 -13.4% for combined (Figure 7-11). The peak MAE for the combined cohort was 1.32 years (Figure 7-7), with a slightly less more conservative alpha value of 0.25 giving 157 features (Figure 7-8) and an  $R^2$  value of 98% (Figure 7-9). Only a small number of features were shared with the infection specific clocks. This may mean that more of these features are linked to ageing in general and may not be disease or ageing specific (Figure 7-11).

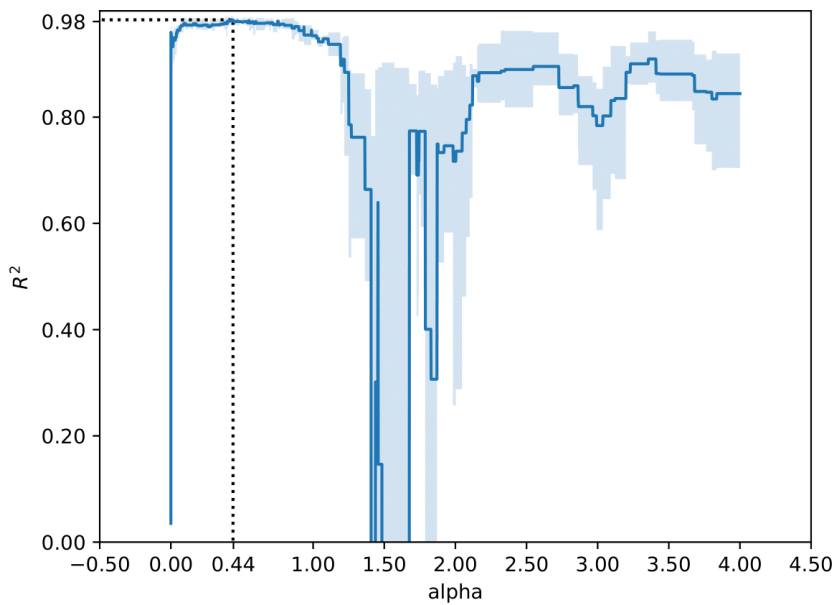


**Figure 7-1 MAE for a range of alpha values in the COVID19 cohort**

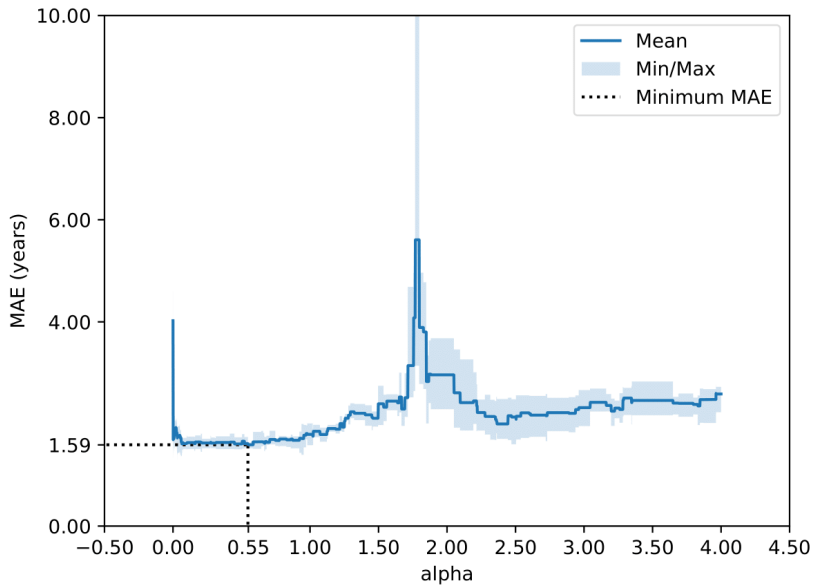
MAE represents the mean average error, this is calculated based on error for each predicted age and actual chronological age for each individual, across all four data partitions within cross validation for the cohort. Peak performance was at alpha = 0.44 with and MAE of 1.34.



**Figure 7-2 Number of features for a range of alphas in the COVID19 cohort**  
Plot of the number of features for a range of alpha values within the cohort. The peak performance was observed with 75 features at an alpha of 0.44.

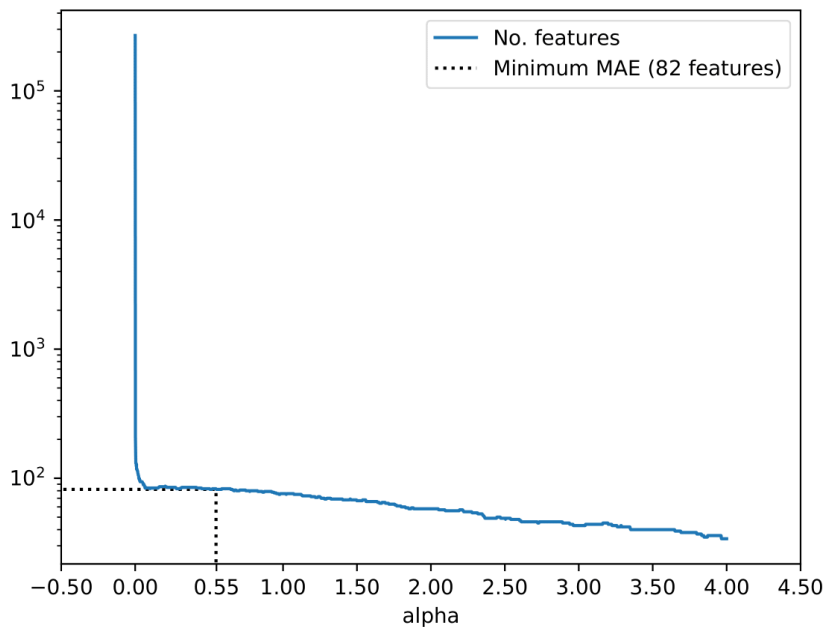


**Figure 7-3 R-squared values for a range of alphas in the COVID19 cohort**  
Plot represents that the total amount of variability in age which can be predicted by the model across the range of hyperparameter values. Most of the variability (around 98%) was predicted by the model.



**Figure 7-4 MAE for a range of alphas in the Influenza cohort**

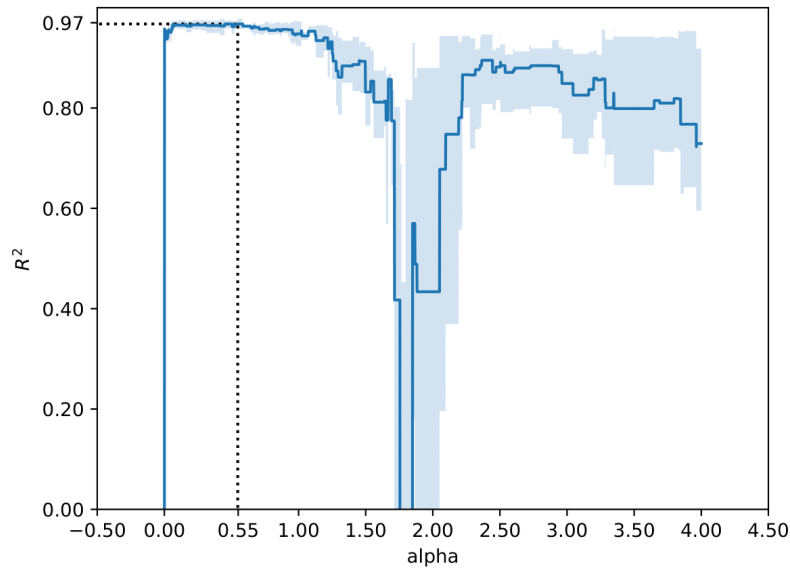
MAE represents the mean average error, this is calculated based on error for each predicted age and actual chronological age for each individual, across all four data partitions within cross validation for the cohort. Peak performance was at alpha = 0.55 with an MAE of 1.59.



**Figure 7-5 Number of features for a range of alphas in the Influenza cohort**

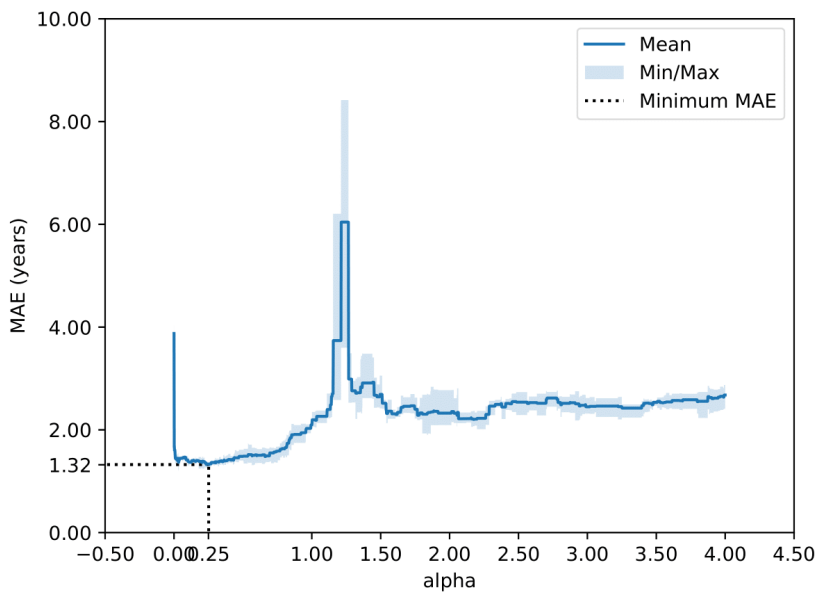
Plot of the number of features for a range of alpha values within the cohort. The peak performance was observed with 82 features at an alpha of 0.55.





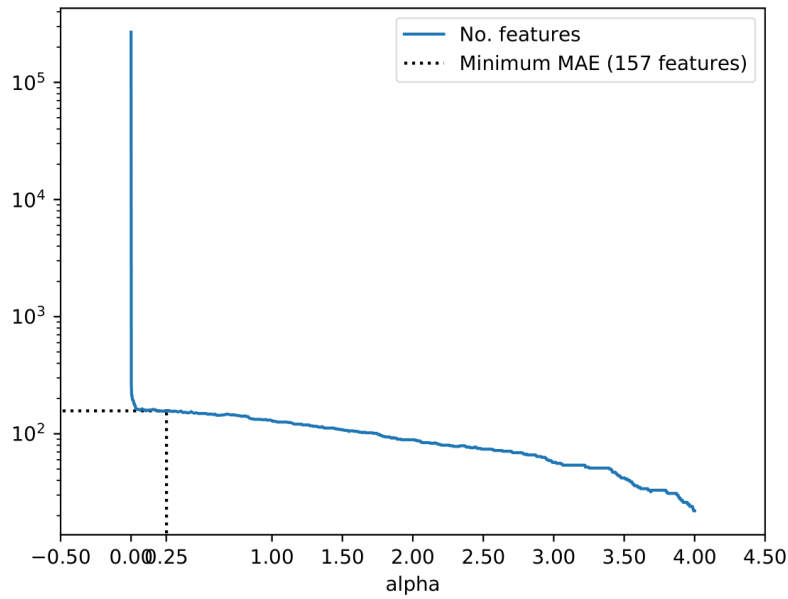
**Figure 7-6 R-square values for a range of alphas in the Influenza cohort**

Plot represents that the total amount of variability in age which can be predicted by the model across the range of hyperparameter values. Most of the variability (around 97%) was predicted by the model.



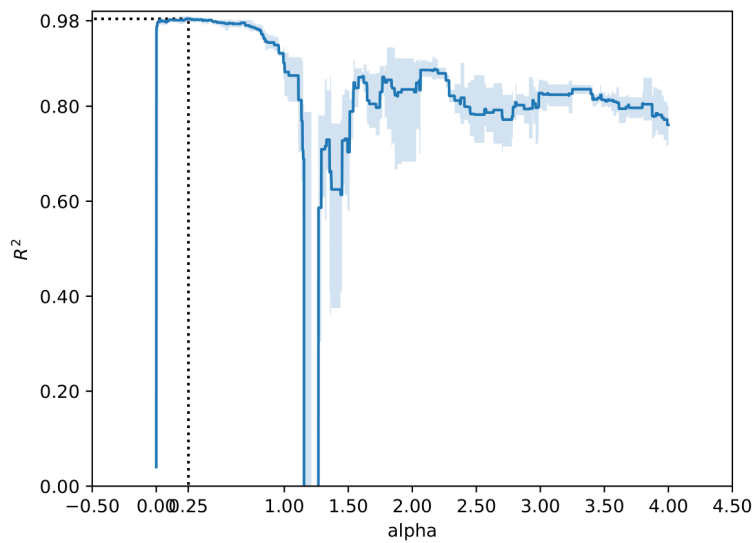
**Figure 7-7 MAE for a range of alphas in the combined cohort**

MAE represents the mean average error, this is calculated based on error for each predicted age and actual chronological age for each individual, across all four data partitions within cross validation for the cohort. Peak performance was at alpha = 0.25 with an MAE of 1.32.



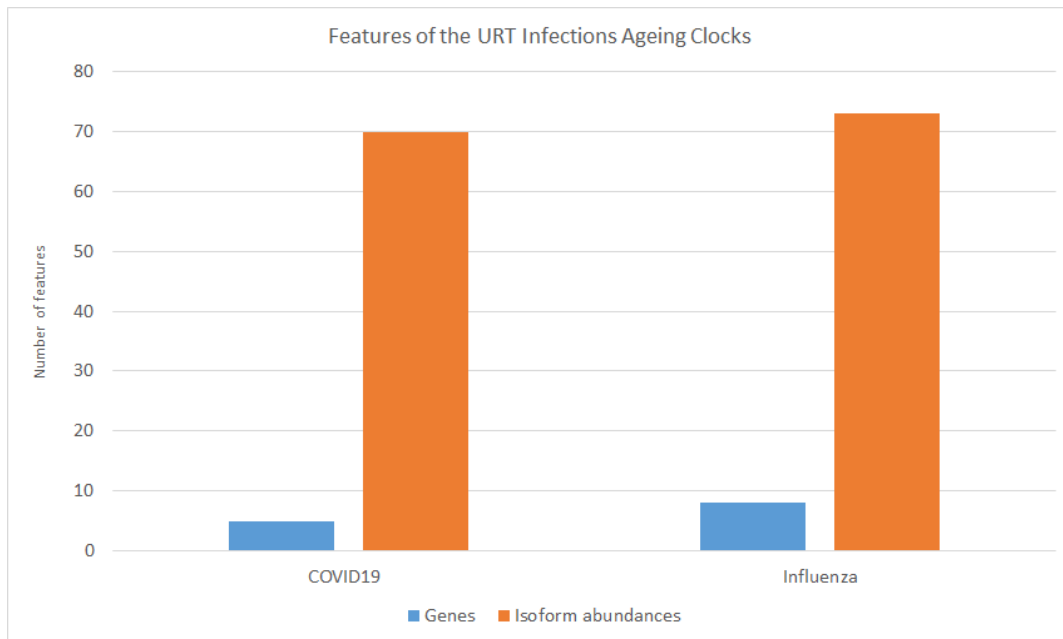
**Figure 7-8 Number of features for a range of alphas in the combined cohort**

Plot of the number of features for a range of alpha values within the cohort. The peak performance was observed with 157 features at an alpha of 0.25.



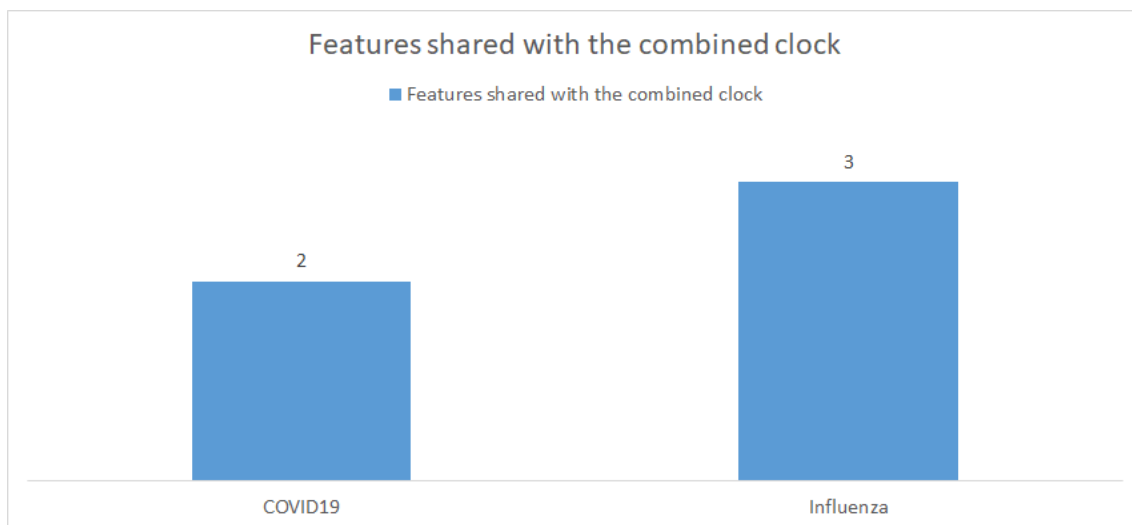
**Figure 7-9 R-squared valued for a range of alphas in the combined cohort**

Plot represents that the total amount of variability in age which can be predicted by the model across the range of hyperparameter values. Most of the variability (around 98%) was predicted by the model.



**Figure 7-10 Feature type for ageing clocks**

Features for each clock which were either genes (in blue) or isoform abundances (in orange). The COVID19 = 75 features, 5 genes, 70 Isoform abundances (93.3% splicing). Influenza = 81 features, 8 genes, 73 Isoform abundances (90.1% Splicing).



**Figure 7-11 Features shared between the infection specific and combined clock.**

Combined clock has 158 features, 16 genes and 140 isoform abundances (88.6%). A total of 4 features were shared with individual infection specific clocks when RUNX3 isoform duplication is considered. The combined clock of immunosenescence – shared 2 isoform abundances with COVID19 and 3 with Influenza.

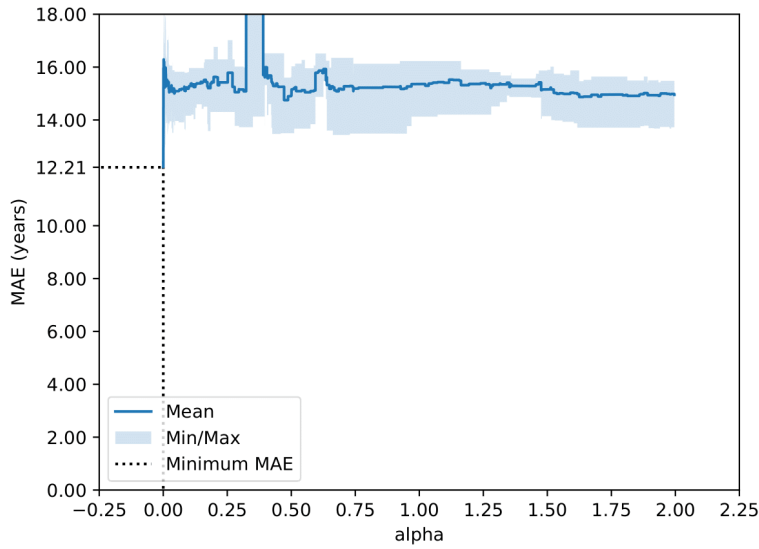
### 7.3.2 Results: Second Iteration

The early accuracy seen in the ageing clocks' ability to predict the age of a patient based on a small set of features from their transcriptome was lost after changes to the model were made. In this iteration, the model was trained only on a subset of the data and not the entirety before testing. Accuracy of prediction as measure by mean absolute error in years, dropped from around 1.3-1.7 years across all models, to 11.67 years for COVID19 (Figure 7-15), 12.21 years for Influenza (Figure 7-12), and 11.53 years for both (Figure 7-18). In addition, the peak performances obtained were now with very large numbers of features (Figure 7-16, Figure 7-13, Figure 7-19).

The peak performing clock for the influenza cohort had 1133 features (cf 82 in iteration 1), and a mean  $R^2$  value 29.9% (Figure 7-14) indicating the majority of the variance in age was not able to be predicted based on the available features. The best performing betas from the influenza clock regression were from 1129 transcript abundances and 4 gene expression values.

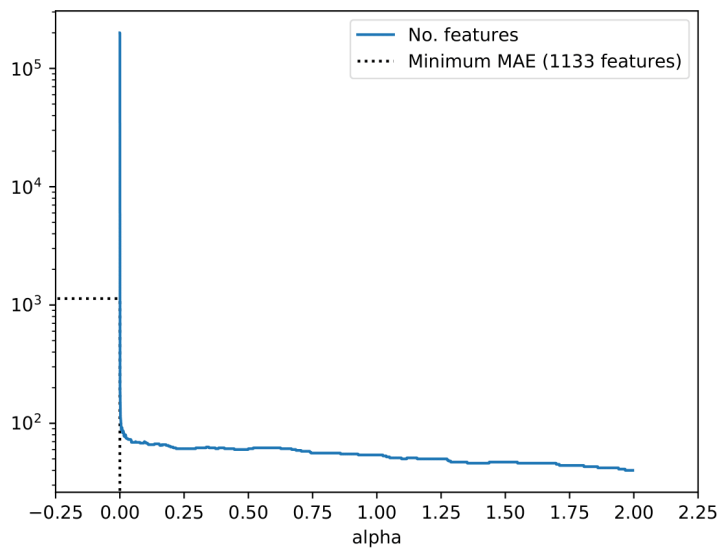
The COVID19 ageing clock and the dual clock both used 100% of the 200,000 (Figure 7-16, Figure 7-19) features in order to obtain peak predictive accuracy of the clock. The peak performing COVID19 clock had an alpha value of 0 (Figure 7-15), – meaning no hyperparameter tuning was beneficial and 100% of the 200,000 features (cf 157 features in iteration 1) were used in order to obtain peak predictive accuracy of the clock. It had an R value of 38.2% (Figure 7-17) which was slightly better than the Influenza iteration, however there was still a majority of variance in age which could not be predicted by the transcriptome of hospitalised patients.

Finally, the combined clock had a peak performance at alpha value of 0 also – meaning no hyperparameter tuning was beneficial and 100% of the 200,000 features were used in order to obtain peak predictive accuracy of the clock (Figure 7-19). It had a mean  $R^2$  value of 40.3% (Figure 7-20) suggesting the clock was able to explain more variance than the other two models, but the majority of the variance in age could not be explained by the transcriptomics ageing clock. The data supporting figures for the second iteration of ageing clocks is seen in Tables 23-25 which show the optimum models for the 3 cohorts.



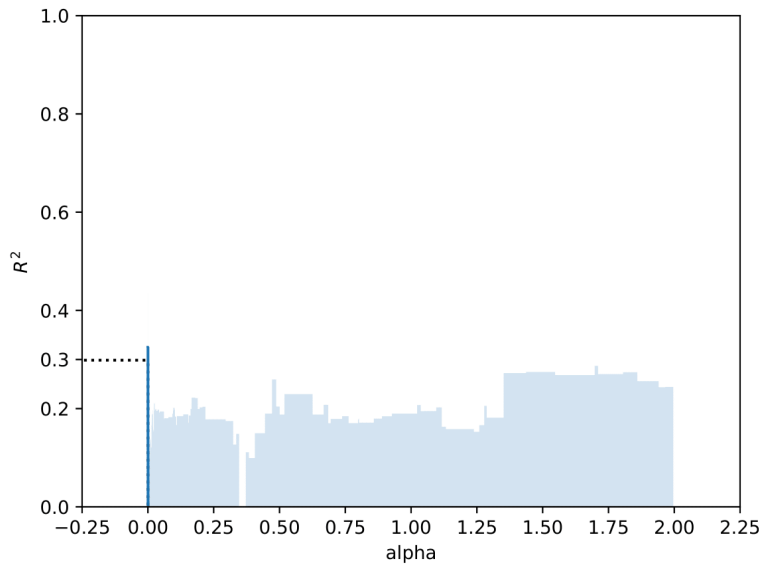
**Figure 7-12 MAE scores for a range of alpha values in Influenza**

MAE represents the mean average error, this is calculated based on error for each predicted age and actual chronological age for each individual, across all four data partitions within cross validation for the influenza cohort. Peak performance was at  $\alpha = 0.00133$  with and MAE of 12.21.



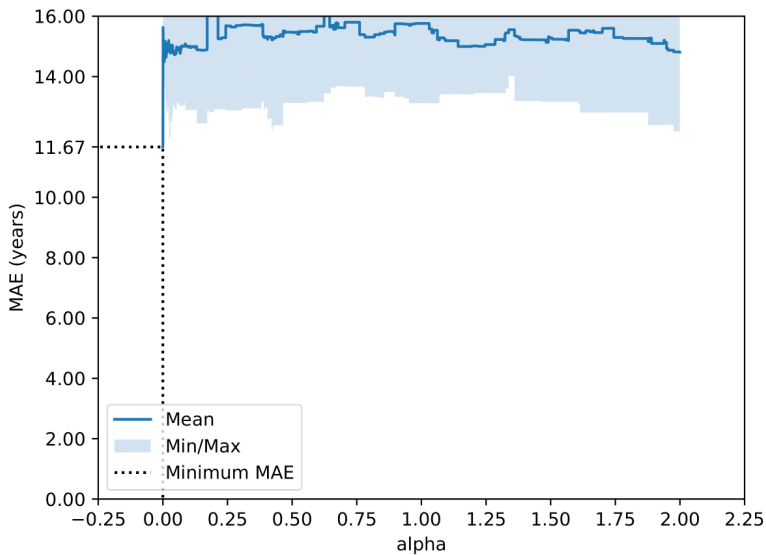
**Figure 7-13 Number of features for a range of alpha values in Influenza**

Plot of the number of features for a range of alpha values within the Influenza cohort. The peak performance was observed with 1133 features at an alpha of 0.00133.



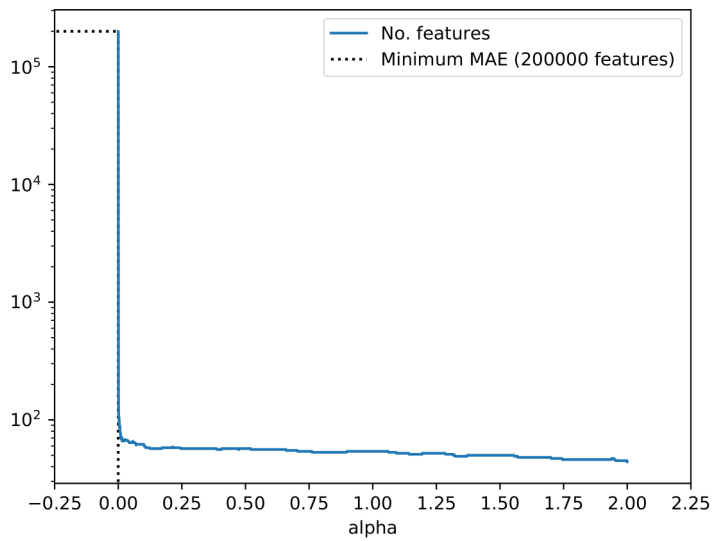
**Figure 7-14 R-squared values for range of alpha values in Influenza**

Plot represents that the total amount of variability in age which can be predicted by the model (29.9%) across the range of hyperparameter values. Most of the variability (around 70%) was not able to be predicted by the model.

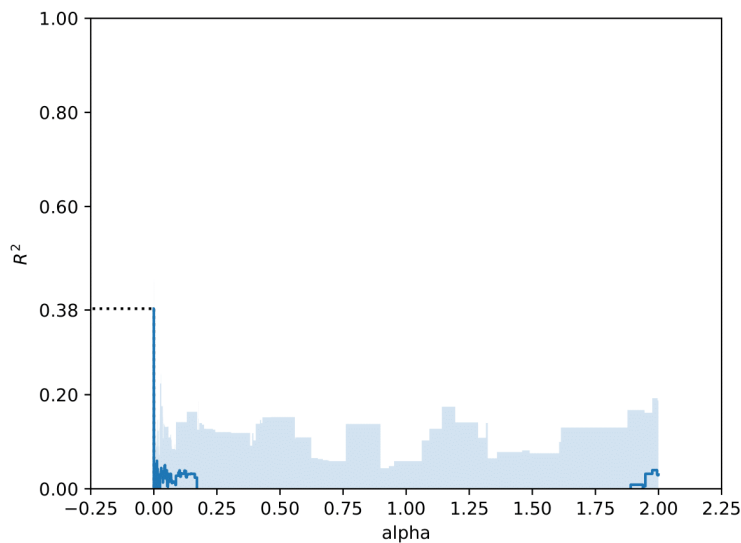


**Figure 7-15 MAE scores across a range of alpha values for COVID19**

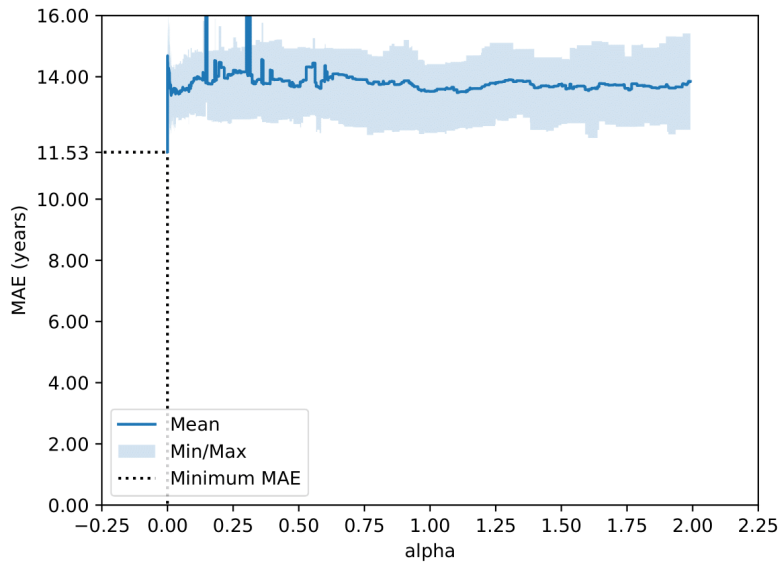
MAE represents the mean average error, this is calculated based on error for each predicted age and actual chronological age for each individual, across all four data partitions within cross validation for the COVID19 cohort. Peak performance was at alpha = 0 with and MAE of 11.67.



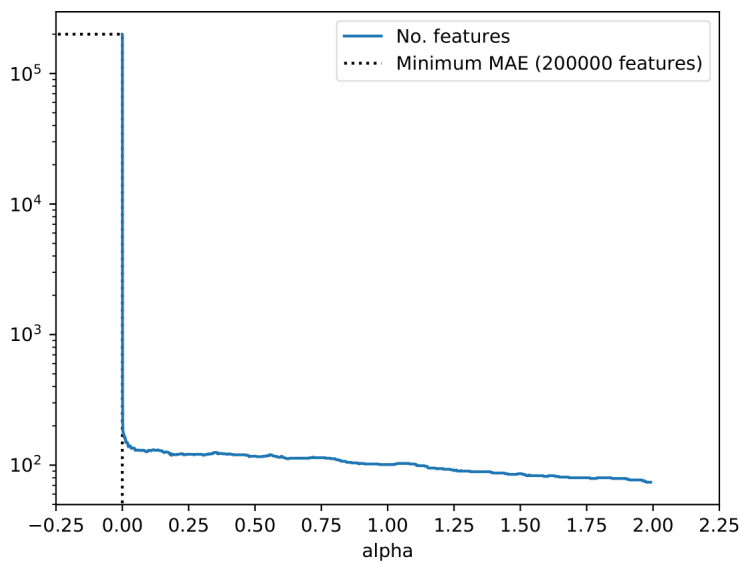
**Figure 7-16 Number of features across a range of alpha values for COVID19**  
Plot of the number of features for a range of alpha values within the COVID19 cohort. The peak performance was observed with 200,000 features at an alpha of 0.0.



**Figure 7-17 R-squared scores across a range of alpha values for COVID19**  
Plot represents that the total amount of variability in age which can be predicted by the model across the range of hyperparameter values. Most of the variability (around 62%) was not able to be predicted by the model.

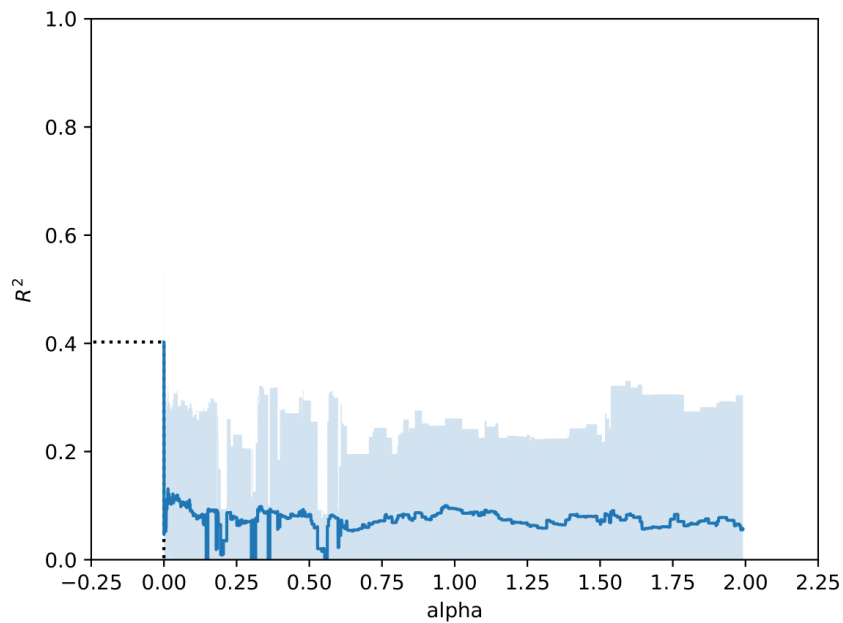


**Figure 7-18 MAE scores across a range of alpha values for combined infections**  
 MAE represents the mean average error, this is calculated based on error for each predicted age and actual chronological age for each individual, across all four data partitions within cross validation for the combined cohort. Peak performance was at alpha = 0 with and MAE of 11.53.



**Figure 7-19 Number of features across a range of alpha values for combined infections**  
 Plot of the number of features for a range of alpha values within the COVID19 cohort. The peak performance was observed with 200,000 features at an alpha of 0.0.





**Figure 7-20 R-squared values across a range of alpha values for combined infections**

Plot represents that the total amount of variability in age which can be predicted by the model across the range of hyperparameter values. Most of the variability (around 60%) was not able to be predicted by the model.

**Table 7-1 Top 10 models for ageing clock in Influenza**

Alpha	n_features	mae_mean	mae_min	mae_max	mae_std	rsquared_mean	rsquared_min	rsquared_max	rsquared_std
0.000133	1133	12.2088256	10.877702	14.7006948	1.48017839	0.298584369	0.214986111	0.43421095	0.083244749
0	200000	12.4033159	10.980112	13.9463526	1.27718588	0.324907046	0.213343192	0.415471754	0.072562027
0.471898	60	14.7502691	13.497359	15.6008059	0.83621751	-0.049441823	-0.391781158	0.259293965	0.279314833
1.624375	46	14.8702514	13.708255	15.4632184	0.69991041	-0.04210791	-0.38658232	0.268218518	0.27020365
1.641576	46	14.883233	13.708255	15.5151448	0.71117894	-0.046517208	-0.404219512	0.268218518	0.275873375
1.697713	45	14.8881331	13.708255	15.6423977	0.77374619	-0.048254477	-0.404219512	0.287038906	0.276627502
0.487366	60	14.8953251	13.497359	15.6008059	0.82906431	-0.063206561	-0.391781158	0.204235012	0.264737314
0.498167	60	14.9017975	13.497359	15.6266954	0.83462899	-0.066405739	-0.404577869	0.204235012	0.268735719
1.812254	43	14.9116811	13.708255	15.4929733	0.71599711	-0.038765779	-0.365620084	0.273916834	0.262719374
1.933729	42	14.9138751	13.67641	15.4206802	0.71767698	-0.049476552	-0.364194848	0.255770223	0.259253511

Top 10 ranked models when considering the mean performance (MAE) and power ( $R^2$ ) of the ageing clock for the influenza cohort in determining the age of the test subset.

**Table 7-2 Top 10 models for ageing clock in COVID-19**

alpha	n_features	mae_mean	mae_min	mae_max	mae_std	rsquared_mean	rsquared_min	rsquared_max	rsquared_std
0	200000	11.6676597	10.159177	13.5569883	1.46850689	0.382863339	0.325014013	0.437259281	0.041573537
<b>0.000133</b>	<b>1193</b>	<b>12.1102777</b>	<b>10.454885</b>	<b>13.3460715</b>	<b>1.091312</b>	<b>0.284376446</b>	<b>0.145798821</b>	<b>0.44499227</b>	<b>0.111079754</b>
0.0044	90	14.478736	13.077998	15.9622397	1.1820894	0.012853769	-0.078399651	0.216937324	0.118731272
0.005334	86	14.4834942	13.692714	15.9531459	0.90464554	0.02685906	-0.082685611	0.202758621	0.109932749
0.0052	90	14.5053799	13.65123	15.9622397	0.89618527	0.032064377	-0.062013313	0.216937324	0.113544848
0.005467	90	14.5058915	13.692714	15.9622397	0.91298235	0.02773507	-0.082685611	0.216937324	0.116294139
0.004667	90	14.5152301	13.251324	15.9622397	1.13604248	0.010791159	-0.080390143	0.216937324	0.119812879
0.005067	90	14.519617	13.638773	15.9622397	0.88807116	0.031944601	-0.099179585	0.216937324	0.116736771
0.004534	90	14.5285099	13.345597	15.9622397	1.12108206	0.010784688	-0.107589707	0.216937324	0.123317423
0.006	86	14.6164535	13.757725	15.9531459	0.82721785	0.019803962	-0.075588123	0.202758621	0.113603689

Top 10 ranked models when considering the mean performance (MAE) and power ( $R^2$ ) of the ageing clock for the COVID19 cohort in determining the age of the test subset.

**Table 7-3 Top 10 models for ageing clock in combined cohort.**

alpha	n_features	mae_mean	mae_min	mae_max	mae_std	rsquared_mean	rsquared_min	rsquared_max	rsquared_std
0	200000	11.5336049	10.260289	12.2323261	0.75441238	0.402519114	0.346200614	0.472430029	0.05219312
0.000133	1351	11.9099036	8.8225854	13.1571951	1.78805031	0.365150285	0.281222081	0.545974679	0.105873615
0.014001	150	13.3865583	12.153879	14.2974122	0.93268278	0.131369512	-0.004789411	0.319834305	0.119786418
0.014268	149	13.4164727	12.153879	14.4358126	0.9707133	0.127999833	-0.004789411	0.320911312	0.120225461
0.014134	150	13.4211584	12.153879	14.4358126	0.96773962	0.127730581	-0.004789411	0.319834305	0.119793554
0.042536	135	13.4288925	12.576552	14.4000225	0.6936905	0.118367525	-0.011936385	0.307905802	0.126995641
0.030669	139	13.4323984	12.329998	14.3824056	0.78393118	0.121444068	-0.013108513	0.301807678	0.117459866
0.030802	139	13.4379763	12.329998	14.4047172	0.79072089	0.121809033	-0.011648656	0.301807678	0.117042753
0.044803	134	13.4489484	12.576552	14.4359749	0.72118828	0.117677311	-0.019496988	0.313665955	0.131031378
0.04267	135	13.4492208	12.576552	14.4000225	0.70310288	0.118127424	-0.011936385	0.307905802	0.12691959

Top 10 ranked models when considering the mean performance (MAE) and power ( $R^2$ ) of the ageing clock for the combined cohort in determining the age of the test subset.

## 7.4 Discussion

We aimed to create infection specific ageing clocks that had research and industrial utility by profiling the ageing whole blood transcriptomes of people suffering with an upper respiratory tract infection; either COVID19 or Influenza.

Whilst our first iterations appeared extremely accurate, when attempting to perform optimisation the realisation was made that the lasso-based feature selection step took place using the entire dataset, before a 4-fold cross validation step was completed. This likely led to what is described as ‘information leakage’ (332), where information about the test dataset leaks into the training dataset and biases the parameters to have increased accuracy, a type of model overfitting. There are, nonetheless, useful elements of the results. There exists a preponderance of isoform abundance features in all the models, compared with gene expression changes. Notwithstanding the accuracy overestimation, this is still important information and demonstrates that the process of alternative splicing is intimately linked to the immunosenescence phenomena in an infection specific manner.

When considering differences between the results from the combined clock versus the disease specific clock, there are more features which are related to gene expression. This difference is small, but the slightly higher percentage of gene expression-based features (13.4%) could suggest that gene expression changes represent a slightly larger portion of the age-related changes in immunosenescence. This however would need to be validated by comparing the findings to a cohort of non-infected, age matched controls, to unpick those features directly linked to infection response.

It is encouraging to find that very few of the features are shared between the ageing clocks. If all features were shared, it would not be possible to discern if what was observed was simple ageing of the organism or was linked to the immune response to infection. This lack of shared features likely means that whilst the accuracy is overstated, the models are indeed comprised of disease specific features, related to the host immune response and are not simply picking up on transcriptomic features which are involved in generic ageing process. A disease-specific ageing clock can help identify key pathways which lead to increased susceptibility or deleterious outcomes and as such help find disease specific treatments.

The lack of shared features between the clocks is with the exception only of differential transcript use of the RUNX3 gene. This strongly suggests that this age-related change is likely to be an

important part of the immune response to viral infections and this factor may be a therapeutic target for immunosenescence. Interestingly, RUNX3 has been previously highlighted as a critical component of the shift from lymphoid to myeloid lineages of cells produced from stem cells in the ageing process, something known to be a high critical factor in the immunosenescence process (465, 466). In support of this, RUNX3 loss is associated with increased oxidative DNA damage, TGF- $\beta$  signalling and cellular senescence (467).

The much lower accuracy of the second iteration of ageing clocks is more likely to represent the true biological difference in transcriptomes which results from the complex interplay between the heterogeneous ageing process, immunosenescence and host response to infection. These models must be considered in the context of the opportunistic data acquisition process, the heterogeneity of the immune response acting in concert with heterogeneous ageing process, genetic and environmental differences, and not least the varying times before attendance to the clinic and sample capture which likely means different stages of immune activation. This last variable is likely adding a high degree of noise to the immune transcriptomic profile and reduced accuracy significantly. So even using a feature engineering process which is more accurately able to capture the biological age of the individuals, the likelihood is low of outperforming the other ageing clocks based on baseline whole blood gene expression. Our later models are approximately half as accurate as other in the literature which are developed on baseline signals (332). This raises an important question, if the feature engineering and machine learning process we have developed were to be deployed on a cohort at baseline, and then after immune stimulation, how accurate could the ageing clocks which the tool outputs be? Would the model have similar accuracy at both points for the cohort, or does ageing cause greater dysregulation in immune response, when the system is working harder than at baseline. It is not possible to accurately infer this between studies and instead a cross sectional cohort study in which immune systems are measured before and after stimulation would be needed.

The large error rates observed in the second iteration of the disease-specific, challenged-system immune ageing clocks likely more accurately represent the very large differences in the immune responses. The differences between the predicted and the actual chronological age also represent the scale of the opportunity for improvement of immune response via therapeutic intervention; if some people have an immune response which appears prematurely aged by a decade, there's lots of improvements to be made and our results can be further interrogated to discern what these differences are, which pathways are involved and how to modulate them. These models have given

a first pass look at some of the key features of the immune response which change with age and found that changes in isoform abundance, which is a result of dysregulated splicing, is a key factor and comprises a majority of the data which machine learning models identify as associated with increased age. Further investigation into these may yield some targets for therapeutic intervention. These models can also now be used to show the efficacy of interventions by comparing test and control sets for significant difference in response with therapeutics targeted at immunosenescence.

To further optimise these models and explore methods to improve their performance, it would be useful to repartition data so that a test set was isolated, before lasso regression and cross validation are performed within an initial data set. This would eliminate the information leak element and still prevent overfitting. Furthermore, the tool would likely demonstrate significant improvements in performance in a controlled infection or immunisation model, in which it could be applied at the same time for all samples, reducing the exposure time variable to enhance accuracy. In this example it could be used to measure the early, innate immune response or track the late adaptive immune response. Tracking immune responses to vaccination in the elderly or ageing also presents as an important space which specific challenged immune system again clocks can be of great use. Examples of blunted effects in the elderly in response to vaccinations are frequent (468) and vaccination specific effects of immunosenescence have been observed (469). Understanding these differences and testing adjuvant therapies to support vaccination or immune rejuvenation strategies are both potential applications of the tool we have developed.

This tool is tissue agnostic, and so could also be applied to a number of other tissues to develop similar models for various age-related diseases. For example, if the desire were to measure ageing in another tissue or organ (muscle, brain, lung etc.), the developed tool is able to identify the association of all transcriptomic features with age and identify those most closely linked to advancing age. The tool can then identify how important alternative splicing is to this ageing process in a specific tissue. Then an ageing clock would be developed which could serve to identify the most important pathways, or to measure the efficacy of treatment in those tissues aimed at restoring transcriptome to an earlier state.

After consultation with industry leaders in drug discovery, some early plans for enhancement of functionality of the tool have also been developed. These involve target prioritisation, and de-risking based on a number of accepted factors which can be converted to metrics. These range from molecule type (kinase, GPCR etc.) level of neglect in the literature, clinical trials involving the feature or pathway, and gene co-expression analysis. We also aim to link the genome wide isoform abundance

changes with expression changes of specific splicing factors using a combination of correlation analysis and software which maps protein binding to RNA sequences (470).

## 7.5 Conclusion

The ageing clocks we have produced suffer from either over-fitting or low accuracy. The over fitting can be outengineered with more time and/or access to other datasets. The low accuracy will likely be improved by using a controlled infection model and further model optimisation. The informatic tool itself, however, represents the foundation to a useful research / industrial pipeline for target identification, prioritisation and therapeutic validation. Our unique approach of combining the gene expression abundance and isoform abundance is derived from the understanding of contribution of splicing to the complexity of the organism, and the important interplay between gene expression and splicing and separate processes in often separate pathways. Our results show that both isoform abundance and gene expression are critical in understanding the ageing process, and that isoform abundance is considerably over-represented in age related changes in the transcriptome during immunosenescence. Whilst individual targets worth of exploration such as RUNX3 and those in the previous chapter have been identified, our results also highlight the distinctiveness of alternative splicing as a variable category when investigating disease, especially with regards to age-related disease.

## 7.6 Statement of contributions

This chapter is a collaborative effort between Yaron Strauch, a PhD student in the same group, and I. The concept of building an ageing clock for a challenged immune system which incorporated relative isoform abundance mixed with gene expression data was my own. First iterations of the bioinformatic pipeline for feature engineering was created by myself using various bioinformatic tools, Microsoft excel and manual steps. Yaron much improved this using a complete end-to-end Python script. Yaron wrote the code for this, and the machine learning steps himself with some high-level input from myself regarding what the tool should do at each step (align reads, calculate isoform abundance or TPM, combine these, perform feature reduction/selection, be able to take an individual sample and output its age based on the model). Yaron designed the regression models, incorporated an early feature selection step to optimise performance and more.

We have agreed on a contribution of 75% Myself /25% Yaron.



## Chapter 8 Discussion

The primary objectives of this research were to explore the utility of RNAseq based approaches to Immunodeficiency diagnostics and investigations. This was in the context of primary and secondary immunodeficiencies, that is, Mendelian and specifically acquired immunodeficiency as a result of ageing. Whilst many of the results of the research have been informative and helped understand immune deficiencies and the importance of the contribution of these features to immunity, many of the aims of this project were not met.

For primary immunodeficiencies, the research aimed to show the benefit of RNAseq approaches in identifying causal variants for Mendelian disease, which was to be achieved through the study of gene expression, alternative splicing and allelic imbalance, this returned mixed results.

The single patient for whom a potential cause was identified through gene expression outlier detection appeared to have complete ablation of expression. With a logical gene panel design this expression loss could have been detected with a cheaper and simpler approach, such as a microarray or multiplexed PCR platform. It is therefore suggested that the benefits of the RNAseq approach have not been realised in this instance, and no advantage was had. The work demonstrated that the batch effect is challenging to overcome if an identical protocol is not followed. It may also be true that the batch correction step has also prevented detection of the outlier genes.

Despite fruitful investigation using the technique, it has been highlighted by the literature that FPKM or TPM values are not interoperable metrics between samples, even if sequenced on the same run. As such in approaching differential expression as a metric for diagnosis, calculating the TPM's of a gene relative to other gene expression within the individual – such as housekeeping genes or a general geomean of gene expression might provide more reliable and robust metrics for interpretation.

A more optimal approach might be to measure the expression change of genes before and after an *in-vitro* immune challenge, as was part of the original planned work. The utility of control datasets would then be measuring a normal immune response as fold changes in expression from baseline within a sample, as opposed to compared to alternative samples. This would alleviate the need for batch effect correction and a blunted immune response would be observable and statistically discernible.

When considering the investigation into splicing, there were a number of events discovered which may well have been causative. However clinical phenotypes were not present for those with potentially causative splicing events and in some other cases, genomes also were not present. With incomplete complimentary data, it was not possible to complete an evaluation of the ability of splicing analysis using the Mendelian RNAseq tool to support or inform diagnosis in this instance. However, alternative splicing data is more interoperable, and the fidelity or homogeneity of control datasets are less critical. The methods were able to identify aberrant splicing patterns in genes which were in close proximity to variants which would likely cause changes in the gene function such as with patient SRB0013 (Section 4.4.4), in genes which may produce the expected phenotype. As a result of this work, some patients have been recalled for genomic sequencing, which will allow validation of the results via alternate approach.

RNAseq certainly has potential to aid in clinical diagnostics of T-cell specific primary immunodeficiency, and whilst that potential has begun to be illuminated through this work, it has not been completely realised. Challenges with using gene expression as a metric come from sequencing batch effects, and interoperable metrics and quantitative methods. These are likely to be mitigated in part by selecting for PBMC's at earlier stages or conducting T-cell activation assay and comparing baseline to activated signals in dynamic RNAseq. Alternative splicing changes do not suffer from the same interoperability issues and gene expression, and the methods were arguably more successful in detecting potential causative events, however this awaits further validation. The filtering process is however a complicated one, and different types of events are missed. An automated series of filters which are specific to certain types of events would be useful. This approach is applicable then but remains overcomplicated for clinical application. There is also possibility that certain types of splicing event, which occur in response to t-cell activation, may also not be detectable until an immune challenge is present. Therefore, it is likely that some genetic variants will remain undiagnosed unless splicing analysis is also used in conjunction with dynamic RNAseq.

The immunosenescence related secondary immunodeficiency results required complex interpretation. The major transcriptomic differences between infectious disease, for both gene and isoform level were identified and explored to some degree, illuminating some of the biological processes affected. This produced new insight into how alternative splicing and gene expression are used as cellular mechanisms to respond to infection, and better established the importance of alternative splicing in host response to infection.

The majority of studies, bioinformatic tools and therapeutic interventions are designed around expression changes, and isoform abundance is often overlooked. Perhaps most interestingly, this work has shown that the vast majority of splicing changes in response to infection are not in the same genes which have expression changes, even in similar upper respiratory tract infections. The processes themselves are used to regulate distinct molecular programs, and it is likely that both will yield interesting therapeutic targets.

Multiple orthogonal investigations conducted during this project showed that these transcriptomic differences vary significantly with age: 95% confidence interval ellipse overlap on principal component plots was reduced when the cohort became younger, suggesting the older transcriptomes are more similar, or lose some distinct features. Some of the most differentially expressed genes between infections appear to converge when viewed individually. Indeed, many of the differentially expressed genes, and differential transcript abundances also have coefficients of expression which suggest convergence with advancing age, even with the most stringent filtering applied. The classification machine learning models gave the best performance when being trained on younger cohorts (although the testing results were not congruent with this, this was probably a result of the very small test data). This work perhaps supports the recent finding there is a loss of transcriptomic identity with age in bodily tissues, contrasting the increase in tissue diversity which happens in development. However, we believe this is the first time that a convergence in transcriptome has been observed and can be distinguished specifically as convergence in host immune response. This aspect of immunosenescence has not been documented before. These processes present new opportunity for therapeutic intervention in immunosenescence driven immunodeficiency.

The lack of overlap between infections for differentially expressed genes with age and differential transcript use with age, supports the notion that the distinct responses to different infections themselves are also subject to a loss of fidelity with advancing age. As the majority of this loss occurs in differential transcript abundance, it could be argued that re-regulating splicing process may represent a key therapeutic target for combatting immunosenescence, over targeting gene expression. Whilst it is hard to deconvolute key cellular pathways or processes from bulk RNAseq data, the additional information that splicing factors themselves decrease with age as we have demonstrated, harmonises with this evidence to suggest that upregulation of splicing factors may lead to re-regulation of splicing in aged immune systems, rejuvenating the responses by allowing the transcriptional programs to once again produce distinct, infection specific responses. The research was also able to identify some more specific novel targets for immune senescence. As an example, SATB2 was identified as a potential target for immunosenescence modulation and interestingly had already been identified as a potential regeneration triggering molecule for the skeleton.

The production of ageing clocks also presented unforeseen challenges in managing data partitioning. This needed to be completed in a manner which allowed a set of features to be identified and validated. In each of the two approaches taken we faced trade-offs. These were between accuracy and robust feature selection. A next iteration would benefit from having a subset of data removed in the first instance, before lasso is conducted and cross-validation on the majority of the data. Once this optimisation step has been completed, comparing this tool to other transcriptomic ageing clocks is required. For this to be reliably done, the tool would require re-deployment on the same dataset as an existing model. This would provide strong evidence that the combined features provide superior insight that gene expression alone. Nonetheless, the existing results strongly suggest that the process of alternative splicing is a predominant feature in the ageing immune system and mediate its blunted response to viral infection. The tool itself can now be redeployed to large datasets to extract important information about tracking ageing in tissues with novel ageing clocks, whilst also mapping the association of every feature individually to advancing age.

### **8.1.1 Limitations**

Many of the limitations have been discussed earlier. This work does not benefit from a multi-omic lens. Whilst RNA modalities are the most mature and arguably the most sensitive, phenotypes observed in the clinic are a result of proteome. Whilst the transcriptome is the greatest contributing

factor, the correlation is not always accurate, and many other factors contribute to the translation and protein abundance. The mendelian work on PID was limited primarily by not have genomes present. When these are acquired, they may well demonstrate no variants were present to precipitate these effects. Having only two healthy controls also presents as a potential source of error. A larger control group could demonstrate large effects across all PID patients. This lack of healthy controls was also a pervasive problem for the infectious disease cohorts. Comparing only one with the other meant it was not possible to discern if a gene was truly upregulated in one infection or down in another, and as such the evaluation could only give relative conclusions. We attempted to overcome this with linear regression comparisons and observing a loss in distinct signals but comparing both infections with control data would have been optimal and allowed complete separation if host response data with systemic inflammation data and general ageing transcriptome features. Finally, the ageing clocks would have benefitted from having other sets or larger sets of data for redeployment and validation, which could have demonstrated the accuracy of the clock. The cohort was at the lower threshold for machine learning approaches, and whilst the tool developed is robust and agile, the conclusions from the dataset may be subject to errors resulting from a small sample size.



## Appendix A

### A.1 IUIS PID gene list.

*IL2RG, JAK3, IL7R, PTPRC, CD3D, CD3E, CD247, CORO1A, LAT, RAG1, RAG2, DCLRE1C, PRKDC, NHEJ1, LIG4, AK2, ADA, DOCK2, CD40LG (TNFSF5), CD40 (TNFRSF5), ICOS, CD3G, CD8A, ZAP70, TAP1, TAP2, TAPBP, B2M, CIITA, RFXANK, RFX5, RFXAP, DOCK8, RHOH, STK4, TRAC, LCK, MALT1, CARD11, BCL10, BCL11B, IL21, IL21R, TNFRSF4, IKBKB, MAP3K14, RELB, MSN, TFRC, WAS, WIPF1, ARPC1B, ATM, NBS1, BLM (RECQL3), DNMT3B, ZBTB24, CDCA7, HELLS, PMS2, RNF168, MCM4, POLE, POLE2, LIG1, NSMCE3, ERCC6L2, GINS1, TBX1, CHD7, SEMA3E, FOXN1, Del10p13-p14, RMRP, SMARCAL1, MYSM1, RNU4ATAC, EXTL3, STAT3, SPINK5, PGM3, CARD11, DKC1, NHP2, NOP10, RTEL1, TERC, TERT, TINF2, TPP1, DCLRE1B/ SNM1/APOLLO:, PARN, WRAP53, STN1, CTC1, SAMD9, SAMD9L, TCN2, SLC46A1, MTHFD1, NEMO (IKBKG), IKBA (NFKBIA), ORAI1, STIM1, PNP, TTC7A, SP110, EPG5, HOIL1 (RBCK1), HOIP1 (RNF31), CCBE1, FAT4, STAT5B, KMT2D (MLL2), KDM6A, BTK, IGHM, IGLL1, CD79A, CD79B, BLNK, PIK3R1, TCF3, Unknown, PIK3CD GOF, PIK3R1, PTEN, CD19, CD81, MS4A1, CR2, TNFRSF13B (TACI), TNFRSF13C (BAFF-R), TNFSF12, MOGS (GCS1), TRNT1, TTC37, NFKB1, NFKB2, IKZF1, IRF2BP2, ATP6AP1, AICDA, UNG, INO80, MSH6, Mutation or chromosomal deletion at 14q32, IGKC, CARD11, PRF1, UNC13D, STX11, STXBP2, FAAP24, LYST, RAB27A, AP3B1, AP3D1, FOXP3, IL2RA, CTLA4, LRBA, STAT3, BACH2, AIRE, ITCH, ZAP70, TPP2, JAK1, PEPD, DNASE1L3, TNFRSF6, FASLG, CASP10, CASP8, FADD, IL10, IL10RA, IL10RB, NFAT5, SH2D1A, XIAP, CD27, CTPS1, RASGRP1, CD70 (TNFSF7), RLTPR, ITK, MAGT1, PRKCD, ELANE, GFI1, HAX1, G6PC3, VPS45, G6PT1, WAS, LAMTOR2, TAZ, VPS13B,*

*USB1, JAGN1, CLPB, CSF3R, SMARCD2, HYOU1, ITGB2, SLC35C1, FERMT3, RAC2, ACTB, FPR1, CTSC, CEBPE, SBDS, WDR1, CFTR, DNAJC21, SRP54, MKL1, CYBB, CYBA, NCF1, NCF2, NCF4, G6PD, GATA2: loss of stem cells, CSF2RB, CSF2RA, IL12RB1, IL12B, IFNGR1, IFNGR2, STAT1, CYBB, IRF8, IRF8, TYK2, ISG15, RORC, JAK1, TMC6, TMC8, CXCR4, STAT1, STAT2, IRF7, IFNAR2, FCGR3A, IFIH1, TLR3, UNC93B1, TRAF3, TICAM1, TBK1, IRF3, CARD9, IL17RA, IL17RC, IL17F, STAT1, TRAF3IP2, IRAK4, MYD88, IRAK1, TIRAP, RPSA, HMOX, APOL1, NBAS, RANBP2, CLCN7, SNX10, OSTM1, PLEKHM1, TCIRG1, TNFRSF11A, TNFSF11, NCSTN, PSEN, PSENEN, TREX1, RNASEH2B, RNASEH2C, RNASEH2A, SAMHD1, ADAR1, IFIH1 (GOF), ACP5, TMEM173, POLA1, USP18, PSMB8\*, DNASE2, MEFV, MVK, NLRP3 (also called NALP3 CIAS1 or PYPAF1), NLRP3, NLRP12, NLRP3, NLRC4, PLCG2, NLRP1, TNFRSF1A, PSTPIP1 (also called C2BP1), NOD2 (also called CARD15), ADAM17, LPIN2, IL1RN, IL36RN, SLC29A3, CARD14, SH3BP2, COPA, OTULIN, TNFAIP3, CECR1, AP1S3, C1QA, C1QB, C1QC, C1R, C1S, C4A+C4B, C2, C3, C3, C5, C6, C7, C8A, C8G, C8B:, C9, MASP2, FCN3, SERPING1, CFB, CFB, CFD, CFP:, CFI:, CFH, CFHR1-5, THBD, CD46, CD59, CD55*



## A.2 Whole blood RNAseq data processing syntax

```
#!/bin/bash
#PBS -N WholeBloodProcessing
#PBS -l walltime=15:00:00
#PBS -l nodes=1:ppn=16
#PBS -l mem=40000m
#PBS -e stderr-$PBS_JOBID.$PBS_ARRAYID.log
#PBS -o stdout-$PBS_JOBID.$PBS_ARRAYID.log
#PBS -t 2
cd $PBS_O_WORKDIR/$PBS_ARRAYID
module load biobuilds/2017.11
## Names of fastq files and location in filestore
fq1="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"$PBS_ARRAYID"_1.fq.gz"
fq2="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"$PBS_ARRAYID"_2.fq.gz"
O1="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"$PBS_ARRAYID"_1.Ptrim.fq.gz"
O2="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"$PBS_ARRAYID"_1.Utrim.fq.gz"
O3="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"$PBS_ARRAYID"_2.Ptrim.fq.gz"
O4="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"$PBS_ARRAYID"_2.Utrim.fq.gz"
## Trimmomatic v0.3.6 via biobuilds/2017.11
trimmomatic PE $fq1 $fq2 $O1 $O2 $O3 $O4
ILLUMINACLIP:"/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/Novoadap.fa":2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
## Make output directory structure to put results into
outdir="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID/"
fastqc_out_T="/scratch/jl5e18/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/fastqc_trim/"
STAR_out="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/STAR/"
mkdir $outdir

## run fastqc (v0.11.3)
mkdir $fastqc_out_T
module load fastqc/0.11.3
fastqc $O1 $O3 --threads 16 --outdir $fastqc_out_T

## STAR alignment (STAR v2.6.1c)
mkdir $STAR_out
cd $STAR_out
/scratch/*username*/RNA_SEQ/Tools/Star/STAR-2.6.1c/bin/Linux_x86_64_static/STAR --genomeDir
/scratch/*username*/RNA_SEQ/Tools/genomedirgencodev30/ --readFilesCommand zcat --readFilesIn $O1 $O3
--runThreadN 16 --twopassMode Basic \
--twopass1readsN -1 --outSAMmapqUnique 60 --outFilterType BySJout --outFilterMultimapNmax 20 --
alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 \
--outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax
1000000 \
```

```
--quantMode TranscriptomeSAM GeneCounts --outReadsUnmapped Fastx --outSAMtype BAM Unsorted
```

```
##Samtools (v1.3.2) Sort and Index
```

```
module load samtools/1.3.2
```

```
bamfile="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/STAR/Aligned.out.bam"
```

```
sorted="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/STAR/"$PBS_ARRAYID"_sorted.bam"
```

```
samtools sort -@ 4 $bamfile > $sorted
```

```
samtools index $sorted
```

```
## Unload everything that's currently loaded and re-add as necessary
```

```
module purge
```

```
## Picard (v2.8.3) - AddOrReplaceReadGroups + MarkDuplicates
```

```
module load jdk/1.8.0
```

```
module load picard/2.8.3
```

```
module load samtools/1.3.2
```

```
RG="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/STAR/"$PBS_ARRAYID"_RG.bam"
```

```
dups="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/STAR/"$PBS_ARRAYID"_MD.bam"
```

```
java -jar /local/software/picard-tools/2.8.3/jarlib/picard.jar AddOrReplaceReadGroups I=$sorted O=$RG SO=coordinate RGID="$PBS_ARRAYID" RGLB="$PBS_ARRAYID" RGPL=illumina RGPU=machine RGSB="$PBS_ARRAYID" TMP_DIR=/scratch/jle18
```

```
samtools index $RG
```

```
java -jar /local/software/picard-tools/2.8.3/jarlib/picard.jar MarkDuplicates I=$RG O=$dups CREATE_INDEX=true VALIDATION_STRINGENCY=SILENT M=output.metrics TMP_DIR=/scratch/*username*
```

```
samtools index $MD
```

```
## RSEM (v1.3.1)
```

```
trans_bam="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/STAR/Aligned.toTranscriptome.out.bam"
```

```
RSEMout="/scratch/*username*/RNA_SEQ/PIDproject/Novogene/WholeBlood/raw_data/raw_data/"$PBS_ARRAYID"/RSEM/"
```

```
mkdir $RSEMout
```

```
/scratch/*username*/RNA_SEQ/Tools/RSEM-1.3.2/rsem-calculate-expression --paired-end --alignments --num-threads 16 \
```

```
$trans_bam /scratch/*username*/RNA_SEQ/Tools/RSEM-1.3.2/human_gencode $RSEMout
```

```
Fi
```

### A.3 Novogene RNAseq QC methods

#### Novogene next generation RNAseq QC

The sequencing quality is calculated as the sequencing takes place in the Illumina next generation sequencing platform. These are denoted as follows. “e” is representative of the sequencing error rate.  $Q_{\text{phred}}$  is indicative of the base quality value.  $Q_{\text{phred}} = -10\text{LOG}_{10}(e)$ . These distribution of quality scores are then plotted against base position for each sample in the report to the user (Figure 8-1) for the user to identify any overarching problems with the sequence quality. It is normal so see a slight gradual reduction in quality; the product of the increasing likelihood of error due to stochastic error rates increasing as reagents are consumed within the platform. Phred score cut-offs are commonly deemed as acceptable at around 30, giving a 99.9% confidence score.

The distribution of error rate can also be observed to increase as the fragment sequencing progresses for the same reasons. The error rate percentage and position data are also visualised on a graph (Figure 8-2) by Novogene in the reports. There is an initial high error rate in the first six bases, as the random hex primers bind incompletely to the RNA template during cDNA synthesis.

Distribution of bases graphs are included (Figure 8-3) as GC rich and GC-poor fragments are under-represented in RNAseq, affecting the ability to perform DGE and outlier detection (471).

A final data filtering step is then conducted to remove those reads which: contain adapters, have more than 10% of bases with undetermined identity, or Q score less than 5 across 50% of the total base number. The pre-filtering statistics are presented in the Novogene report (Figure 8-4).

Phred score	error base	right base	Q-score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

The distribution of quality score is shown in Fig.1:

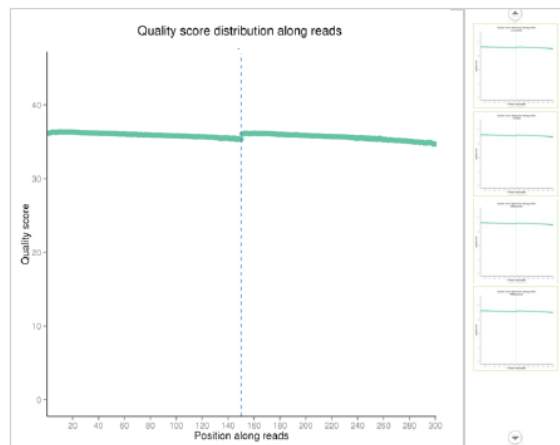


Figure 8-1 - Novogene Q-score distribution

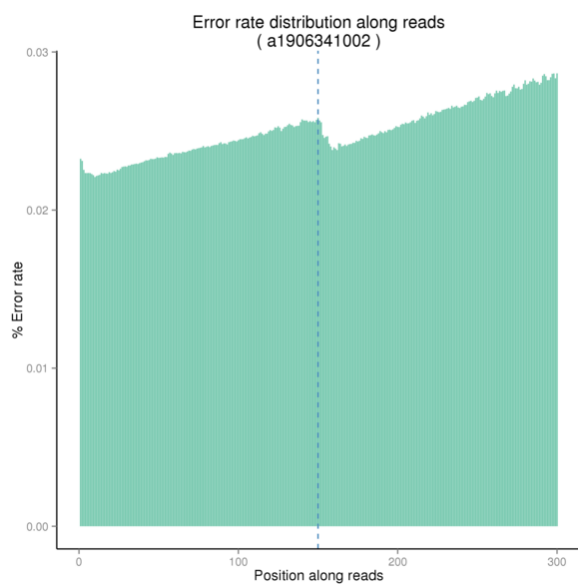


Figure 8-2 Error rate distribution: Novogene

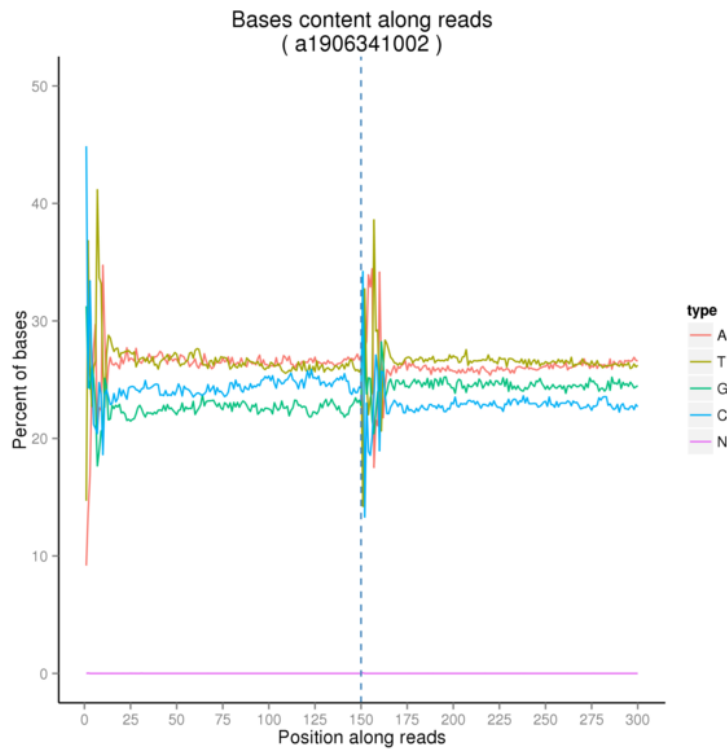


Figure 8-3 GC content distribution: Novogene

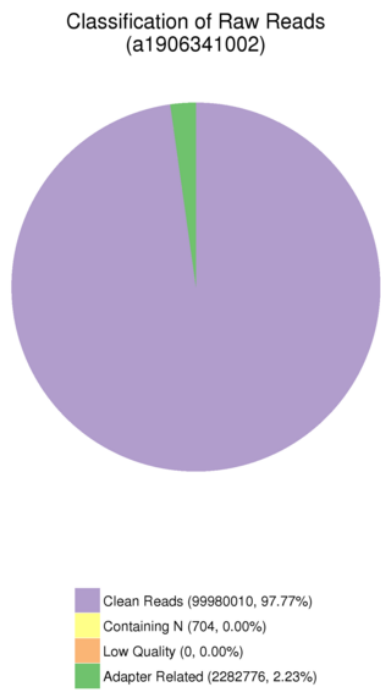
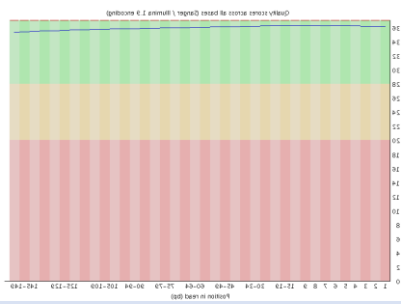
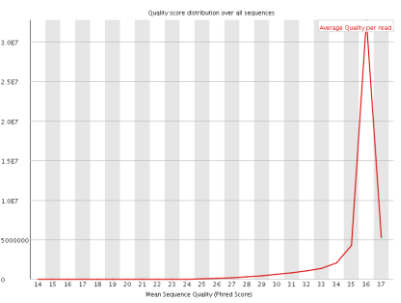
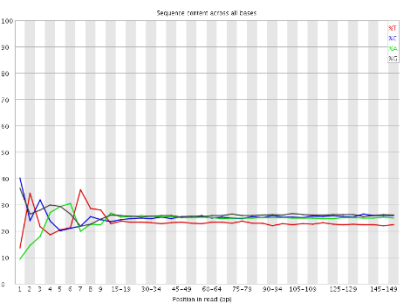
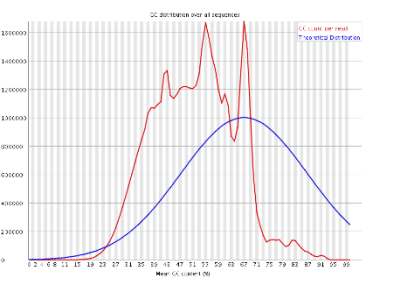


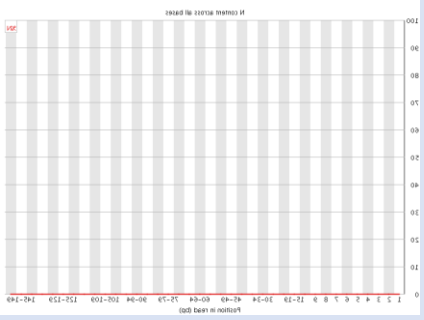
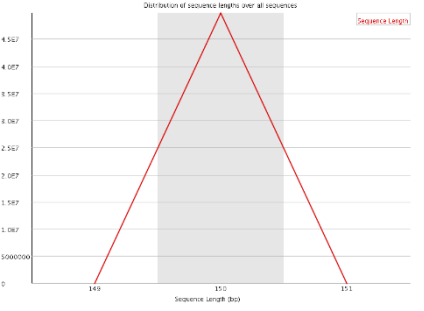
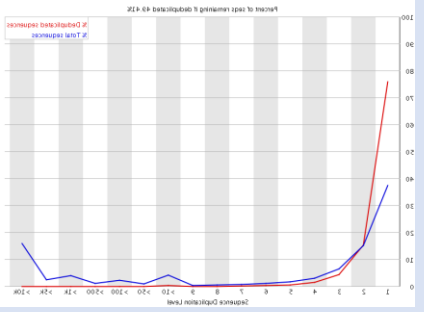
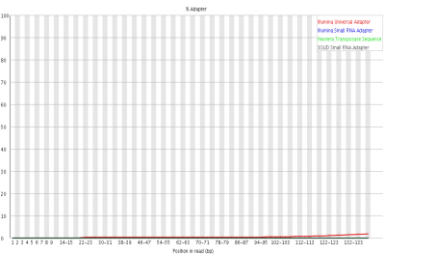
Figure 8-4 Raw read classification QC: Novogene



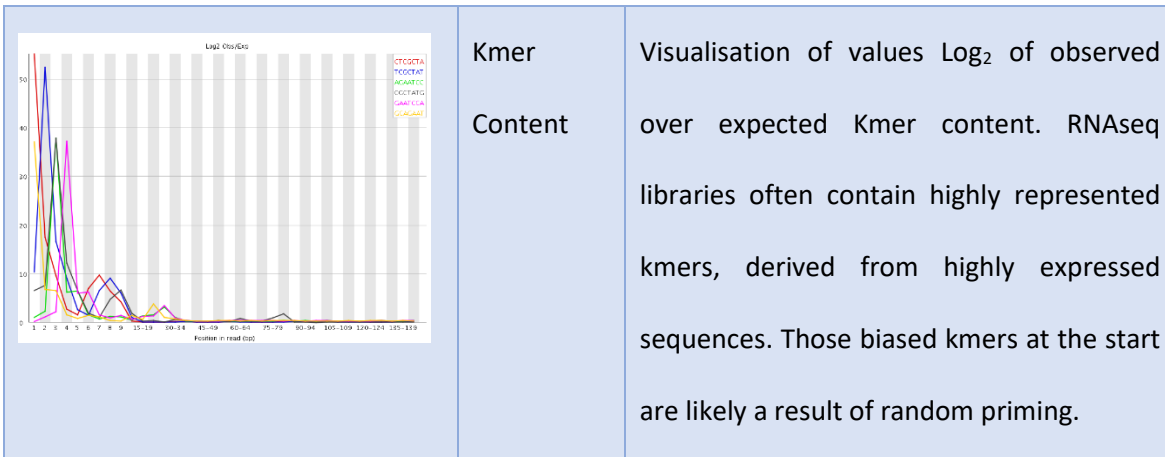
A brief explanation of the fastQC outputs with examples can be found in Table 8-1

**Table 8-1 FastQC output explanation**

Novogene Visualisation example	Name	Description
	<p>Per base sequence quality</p>	<p>A plot of total number of reads vs average quality score over full length. Distribution should have little variance and be high scoring, as seen in this example.</p>
	<p>Per sequence quality score</p>	<p>The distribution of quality scores over all sequences. Ideally, one tight peak at the upper range of quality is desirable.</p>
	<p>Per base sequence content</p>	<p>The sequence content in terms of base identity across all bases. As little deviation from the expected is ideal. FastQC will often assign a warning flag or fail to RNAseq data here as expression is not consistent across the genome.</p>
	<p>Per sequence GC content</p>	<p>A specific look at the GC content distribution over all sequences. Spikes, as seen on the right-hand side of the red curve at this location are often indicative of</p>

	<p>Per base N content</p>	<p>adapter contamination. The distribution of N across all bases. N being an unidentified base.</p>
	<p>Sequence Length distribution</p>	<p>Distribution of sequence lengths over all sequences.</p>
	<p>Sequence Duplication levels</p>	<p>Percent of sequences which are remaining if all sequences are de-duplicated. Not as helpful with RNAseq data as very abundant transcripts can be overrepresented.</p>
	<p>Adapter Content</p>	<p>Percentage of sequences which were found to be an adapter.</p>







#### A.4 – Genomics England PID gene list

*ACP5, ADA, ADA2, ADAR, AICDA, AIRE, AK2, AP3B1, ARPC1B, ATM, B2M, BLM, BLNK, BTK, C1QA, C1QB, C1QC, C1R, C1S, C2, C4A, C4B, C5, C6, C7, C8A, C8B, C9, CARD9, CARMIL2, CASP10, CASP8, CCBE1, CD19, CD27, CD3D, CD3E, CD3G, CD40, CD40LG, CD46, CD55, CD59, CD70, CD79A, CD79B, CDCA7, CFD, CFH, CFI, CFP, CHD7, CIITA, CLPB, COPA, CORO1A, CSF2RA, CSF3R, CTLA4, CTPS1, CXCR4, CYBA, DCLRE1B, DCLRE1C, DNMT3B, DOCK2, DOCK8, ELANE, EPG5, EXTL3, FADD, FAS, FASLG, FERMT3, FOXP1, FOXP3, G6PC3, G6PD, GATA2, GFI1, GINS1, HAX1, HELLS, HTRA2, ICOS, IFNGR1, IFNGR2, IGHM, IGLL1, IKBKB, IKBKG, IKZF1, IL10, IL10RA, IL10RB, IL12B, IL12RB1, IL17RA, IL17RC, IL1RN, IL21R, IL2RA, IL2RG, IL36RN, IL7R, INO80, IRAK4, IRF8, ISG15, ITCH, ITGB2, ITK, JAGN1, JAK3, LAMTOR2, LCK, LIG4, LPIN2, LRBA, LYST, MAGT1, MALT1, MAP3K14, MCM4, MEFV, MOGS, MSN, MTHFD1, MVK, MYD88, MYSM1, NBN, NCF1, NCF2, NFKB1, NFKB2, NFKBIA, NHEJ1, NLRC4, NLRP12, NLRP3, NOD2, ORAI1, OTULIN, PARN, PGM3, PIK3CD, PIK3R1, PLCG2, PNP, PRF1, PRKCD, PRKDC, PSMB8, PSTPIP1, PTPRC, RAB27A, RAG1, RAG2, RFX5, RFXANK, RFXAP, RMRP, RNASEH2A, RNASEH2B, RNASEH2C, RNF168, RORC, RPSA, RTEL1, SAMHD1, SBDS, SERPING1, SGPL1, SH2D1A, SLC29A3, SLC35C1, SLC37A4, SLC46A1, SMARCAL1, SP110, SPINK5, STAT1, STAT2, STAT3, STAT5B, STIM1, STK4, STX11, STXBP2, TAP1, TAP2, TAZ, TBK1, TCF3, TCN2, TICAM1, TLR3, TMC6, TMC8, TNFAIP3, TNFRSF1A, TPP2, TREX1, TRNT1, TTC37, TTC7A, TYK2, UNC13D, UNC93B1, UNG, USB1, VPS13B, VPS45, WAS, XIAP, ZAP70, ZBTB24, ACD, ATP6AP1, BACH2, C3, CARD11, CEBPE, CSF2RB, CTSC, CYBB, DKC1, DNAJC21, DNASE2, ERCC6L2, F12, FAT4, GATA1, IFIH1, LAT, MYO5B, NCF4, NHP2, NSMCE3, PEPD, POLA1, RASGRP1, RBCK1, RIPK1, SKIV2L, SPPL2A, TMEM173, TRAC, WIPF1, ADAM17, AP1S3, BCL10, CARD14, CD247, CD81, CD8A, CFB, CFHR1, CFHR3, CFHR4, CFHR5, CFTR, CR2, DNASE1L3, FCGR3A, FPR1, IGKC, IL17F, IL21, IRF3, IRF7, KRAS, MBL2, NCSTN, NOP10, NRAS, PMS2, POLE, PSENEN, RAC2, RHOH, SAMD9, TAPBP, TBX1, TERC, TERT, TIN2, TNFRSF13C, ACTB, AP3D1, APOL1, BCL11B, BLOC1S6, C8G, CD4, CFHR2, CLCN7, CNBP, COLEC11, CTC1, ELF4, EPCAM, ERCC2,*

*ERCC3, FAAP24, FBF1, FCGR1A, FCGR2A, FCGR2B, FCGR3B, FCGRT, FCN3, FPR2, FPR3, GAD1, GTF2H5, GUCY2C, HMOX1, HPS1, HPS4, HPS6, HYOU1, ICOSLG, IFNAR2, IGHG2, IL17A, IL18, IL22, IL23A, IRAK1, IRF2BP2, ITGAM, JAK1, KDM6A, KMT2A, KMT2D, LIG1, LRRC8A, MASP1, MASP2, MKL1, MPI, MPO, MRE11, MS4A1, MSH6, NBAS, NFAT5, NFKBID, NLRP1, OSTM1, PLEKHM1, POLE2, PSEN1, PSMA3, PSMB4, PSMB9, PTEN, RANBP2, RECQL4, RELB, RET, RNF31, RNU4ATAC, SAMD9L, SART3, SEMA3E, SH3BP2, SMARCD2, SNX10, SRP54, STAT5A, STN1, TCIRG1, TFRC, THBD, TIRAP, TNFRSF11A, TNFRSF13B, TNFRSF4, TNFSF11, TNFSF12, TRAF3, TRAF3IP2, UNC119, USP18, WDR1, WRAP53, ISCA-37433-Loss, ISCA-37446-Loss*

## A.5 HTG T-Cell gene list.

**Table 8-2 HTG T-Cell gene list.**

Gene name	Gene stable ID	Gene type	Gene description
<i>PTPRC</i>	ENSG00000081237	protein_coding	protein tyrosine phosphatase receptor type C [Source:HGNC Symbol;Acc:HGNC:9666]
<i>CD1A</i>	ENSG00000158477	protein_coding	CD1a molecule [Source:HGNC Symbol;Acc:HGNC:1634]
<i>PDCD1LG2</i>	ENSG00000197646	protein_coding	programmed cell death 1 ligand 2 [Source:HGNC Symbol;Acc:HGNC:18731]
<i>CD274</i>	ENSG00000120217	protein_coding	CD274 molecule [Source:HGNC Symbol;Acc:HGNC:17635]
<i>NFATC1</i>	ENSG00000131196	protein_coding	nuclear factor of activated T-cells 1 [Source:HGNC Symbol;Acc:HGNC:7775]
<i>GATA3</i>	ENSG00000107485	protein_coding	GATA binding protein 3 [Source:HGNC Symbol;Acc:HGNC:4172]
<i>NFATC3</i>	ENSG00000072736	protein_coding	nuclear factor of activated T-cells 3 [Source:HGNC Symbol;Acc:HGNC:7777]
<i>BTLA</i>	ENSG00000186265	protein_coding	B and T lymphocyte associated [Source:HGNC Symbol;Acc:HGNC:21087]
<i>CD28</i>	ENSG00000178562	protein_coding	CD28 molecule [Source:HGNC Symbol;Acc:HGNC:1653]
<i>CTLA4</i>	ENSG00000163599	protein_coding	cytotoxic T-lymphocyte associated protein 4 [Source:HGNC Symbol;Acc:HGNC:2505]
<i>CD40</i>	ENSG00000101017	protein_coding	CD40 molecule [Source:HGNC Symbol;Acc:HGNC:11919]

<b><i>GNLY</i></b>	ENSG00000115523	protein_coding	granulysin [Source:HGNC Symbol;Acc:HGNC:4414]
<b><i>PDCD1</i></b>	ENSG00000188389	protein_coding	programmed cell death 1 [Source:HGNC Symbol;Acc:HGNC:8760]
<b><i>ICOS</i></b>	ENSG00000163600	protein_coding	inducible T-cell costimulator [Source:HGNC Symbol;Acc:HGNC:5351]
<b><i>ICOSLG</i></b>	ENSG00000160223	protein_coding	inducible T-cell costimulator ligand [Source:HGNC Symbol;Acc:HGNC:17087]
<b><i>FOXP3</i></b>	ENSG00000049768	protein_coding	forkhead box P3 [Source:HGNC Symbol;Acc:HGNC:6106]
<b><i>CD99</i></b>	ENSG00000002586	protein_coding	CD99 molecule (Xg blood group) [Source:HGNC Symbol;Acc:HGNC:7082]
<b><i>CD80</i></b>	ENSG00000121594	protein_coding	CD80 molecule [Source:HGNC Symbol;Acc:HGNC:1700]
<b><i>LAG3</i></b>	ENSG00000089692	protein_coding	lymphocyte activating 3 [Source:HGNC Symbol;Acc:HGNC:6476]
<b><i>SLAMF6</i></b>	ENSG00000162739	protein_coding	SLAM family member 6 [Source:HGNC Symbol;Acc:HGNC:21392]
<b><i>SLAMF1</i></b>	ENSG00000117090	protein_coding	signaling lymphocytic activation molecule family member 1 [Source:HGNC Symbol;Acc:HGNC:10903]
<b><i>CD4</i></b>	ENSG00000010610	protein_coding	CD4 molecule [Source:HGNC Symbol;Acc:HGNC:1678]
<b><i>SLAMF7</i></b>	ENSG00000026751	protein_coding	SLAM family member 7 [Source:HGNC Symbol;Acc:HGNC:21394]
<b><i>CD276</i></b>	ENSG00000103855	protein_coding	CD276 molecule [Source:HGNC Symbol;Acc:HGNC:19137]
<b><i>CD40LG</i></b>	ENSG00000102245	protein_coding	CD40 ligand [Source:HGNC Symbol;Acc:HGNC:11935]
<b><i>DPP4</i></b>	ENSG00000197635	protein_coding	dipeptidyl peptidase 4 [Source:HGNC

			Symbol;Acc:HGNC:3009]
<b>TCF7</b>	ENSG00000081059	protein_coding	transcription factor 7 [Source:HGNC Symbol;Acc:HGNC:11639]
<b>CD70</b>	ENSG00000125726	protein_coding	CD70 molecule [Source:HGNC Symbol;Acc:HGNC:11937]
<b>NFATC4</b>	ENSG00000100968	protein_coding	nuclear factor of activated T-cells 4 [Source:HGNC Symbol;Acc:HGNC:7778]
<b>LAT</b>	ENSG00000213658	protein_coding	linker for activation of T-cells [Source:HGNC Symbol;Acc:HGNC:18874]
<b>RAG1</b>	ENSG00000166349	protein_coding	recombination activating 1 [Source:HGNC Symbol;Acc:HGNC:9831]
<b>CD3D</b>	ENSG00000167286	protein_coding	CD3d molecule [Source:HGNC Symbol;Acc:HGNC:1673]
<b>SPN</b>	ENSG00000197471	protein_coding	sialoporphin [Source:HGNC Symbol;Acc:HGNC:11249]
<b>TCL1B</b>	ENSG00000213231	protein_coding	T-cell leukemia/lymphoma 1B [Source:HGNC Symbol;Acc:HGNC:11649]
<b>CD8A</b>	ENSG00000153563	protein_coding	CD8a molecule [Source:HGNC Symbol;Acc:HGNC:1706]
<b>CD86</b>	ENSG00000114013	protein_coding	CD86 molecule [Source:HGNC Symbol;Acc:HGNC:1705]
<b>CD27</b>	ENSG00000139193	protein_coding	CD27 molecule [Source:HGNC Symbol;Acc:HGNC:11922]
<b>CD2</b>	ENSG00000116824	protein_coding	CD2 molecule [Source:HGNC Symbol;Acc:HGNC:1639]
<b>CD69</b>	ENSG00000110848	protein_coding	CD69 molecule [Source:HGNC Symbol;Acc:HGNC:1694]
<b>SUSD3</b>	ENSG00000157303	protein_coding	sushi domain containing 3 [Source:HGNC Symbol;Acc:HGNC:28391]
<b>EOMES</b>	ENSG00000163508	protein_coding	eomesodermin [Source:HGNC Symbol;Acc:HGNC:3372]

## A.6 Combined IUIS, GeCIP, T-cell panel from HTG EdgeSeq panel

BTLA, CD1A, CD2, CD27, CD274, CD276, CD28, CD3D, CD4, CD40, CD40LG, CD69, CD70, CD80, CD86, CD8A, CD99, CTLA4, DPP4, EOMES, FOXP3, GATA3, GNLY, ICOS, ICOSLG, LAG3, LAT, NFATC1, NFATC3, NFATC4, PDCD1, PDCD1LG2, PTPRC, RAG1, SLAMF1, SLAMF6, SLAMF7, SPN, SUSD3, TCF7, TCL1B, CD3E, CD3Z, CORO1A, IL2RG, IL7R, JAK3, ADA, AK2, DCLRE1C, LIG4, NHEJ1, PRKDC, RAC2, RAG2, B2M, BCL10, CARD11, CD3G, TNFRSF5, TNFSF5, CIITA, DOCK2, DOCK8, FCHO1, IKBKB, IKZF1, IL21, IL21R, ITK, LCK, MALT1, MAP3K14, MSN, POLD1, POLD2, REL, RELA, RELB, RFX5, RFXANK, RFXAP, RHOH, STK4, TAP1, TAP2, TAPBP, TFRC, TNFRSF4, TRAC, ZAP70, ARPC1B, WAS, WIPF1, ATM, BLM, RECQL3, CDCA7, DNMT3B, GINS1, HELLS, LIG1, MCM4, NBS1, NSMCE3, PMS2, POLE1, POLE2, RNF168, ZBTB24, CHD7, FOXN1, SEMA3E, TBX1, EXTL3, MYSM1, RMRP, RNU4ATAC, SMARCAL1, ERBB21P, IL6R, IL6ST, PGM3, SPINK5, STAT3, TGFBR1, TGFBR2, ZNF341, MTHFD1, SLC46A1, TCN2, IKBKG, NFKBIA, ORAI1, STIM1, BCL11B, CCBE1, EPG5, FAT4, KDM6A, KMT2A, KMT2D, MLL2, NFE2L2, PNP, RBCK1, RNF31, SKIV2L, SP110, STAT5B, TTC37, TTC7A, BLNK, BTK, CD79A, CD79B, IGHM, IGLL1, PIK3CD, PIK3R1, SLC39A7, TCF3, TOP2B, ARHGEF1, ATP6AP1, CD19, CD20, CD21, CD81, IRF2BP2, MOGS, GCS1, NFKB1, NFKB2, PIK3CD, PTEN, SEC61A1, SH3KBP1, TNFRSF13B, TNFRSF13C, TNFSF12, TRNT1, AICDA, INO80, MSH6, UNG, IGKC, FAAP24, PRF1, SLC7A7, STX11, STXBP2, UNC13D, AP3B1, AP3D1, LYST, RAB27A, BACH2, DEF6, FERMT1, IL2RA, IL2RB, LRBA, AIRE, ITCH, JAK1, PEPD, TPP2, IL10, IL10RA, IL10RB, NFAT5, RIPK1, TGFB1, CASP10, CASP8, FADD, TNFRSF6, TNFSF6, CARMIL2, CTPS1, MAGT1, PRKCD, RASGRP1, SH2D1A, TNFRSF9, XIAP, CEBPE, CLPB, CSF3R, DNAJC21, EFL1, ELANE, G6PC3, G6PT1, GFI1, HAX1, HYOU1, JAGN1, LAMTOR2, SBDS, SMARCD2, SRP54, TAZ, USB1, VPS13B, VPS45, ACTB, CFTR, CTSC, FERMT3, FPR1, ITGB2, MKL1, SLC35C1, WDR1, CYBA, CYBB, NCF1, NCF2, NCF4, CYBC1, G6PD, GATA2, CSF2RA, CSF2RB, IFNGR1, IFNGR2, IL12B, IL12RB1, IL12RB2, IL23R, IRF8, ISG15, RORC, SPPL2A, STAT1, TYK2, CIB1, CXCR4, TMC6, TMC8, FCGR3A, IFIH1, IFNAR1, IFNAR2, IRF7, IRF9, POLR3A, POLR3C, POLR3F, STAT2, DBR1, IRF3, TBK1, TICAM1, TLR3, TRAF3,



UNC93B1, CARD9, IL17F, IL17RA, IL17RC, TRAF3IP2, IRAK1, IRAK4, MYD88, TIRAP, APOL1, CLCN7, HMOX, NBAS, NCSTN, OSTM1, PLEKHM1, PSEN, PSENE1, RANBP2, RPSA, SNX10, TCIRG1, TNFRSF11A, TNFSF11, IL18BP, IRF4, ACP5, ADA2, ADAR1, DNASE1L3, DNASE2, OAS1, RNASEH2A, RNASEH2B, RNASEH2C, SAMHD1, TMEM173, TREX1, USP18, POLA1, MEFV, MVK, NLRC4, NLRP1, NLRP12, NLRP3, PLCG2, ADAM17, ALPI, AP1S3, CARD14, COPA, HAVCR2, IL1RN, IL36RN, LPIN2, NOD2, OTULIN, PSMB8, PSMG2, PSTPIP1, SH3BP2, SLC29A3, TNFAIP3, TNFRSF1A, TRIM22, C1QA, C1QB, C1QC, C1R, C1S, C2, C3, C4A, C4B, C5, C6, C7, C8A, C8B, C8G, C9, CD46, CD55, CD59, CFB, CFD, CFH, CFHR1, CFHR2, CFHR3, CFHR4, CFHR5, CFI, CFP, FCN3, MASP2, SERPING1, THBD, ACD, BRCA1, BRCA2, BRIP1, CTC1, DKC1, ERCC4, ERCC6L2, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCI, FANCL, FANCM, MAD2L2, NOLA2, NOLA3, PALB2, PARN, RAD51, RAD51C, RFWF3, RTEL1, SAMD9, SAMD9L, SLX4, SRP72, STN1, TERC, TERT, TINF2, TP53, UBE2T, WRAP53, XRCC2, XRCC3, ADAR, DCLRE1B, FAS, FASLG, HTRA2, IL12B, NBN, SGPL1, SLC37A4, F12, GATA1, MYO5B, NHP2, CD247, CR2, KRAS, MBL2, NOP10, NRAS, POLE, BLOC1S6, CNBP, COLEC11, ELF4, EPCAM, ERCC2, ERCC3, FBF1, FCGR1A, FCGR2A, FCGR2B, FCGR3B, FCGRT, FPR2, FPR3, GAD1, GTF2H5, GUCY2C, HMOX1, HPS1, HPS4, HPS6, IGHG2, IL17A, IL18, IL22, IL23A, ITGAM, LRRC8A, MASP1, MPI, MPO, MRE11, MS4A1, NFKBID, PSEN1, PSMA3, PSMB4, PSMB9, RECQL4, RET, SART3, STAT5A, UNC119,

## A.7 OTRIDER syntax

```

Library (OUTRIDER)
# small testing data set
odsSmall <- makeExampleOutriderDataSet(dataset="Kremer")
# full data set from Kremer et al.
baseUrl <- paste0("https://static-content.springer.com/esm/",
"art%3A10.1038%2Fncmms15824/MediaObjects/")
count_URL <- paste0(baseUrl, "41467_2017_BFncmms15824_MOESM390_ESM.txt")
anno_URL <- paste0(baseUrl, "41467_2017_BFncmms15824_MOESM397_ESM.txt")
ctsTable <- read.table(count_URL, sep="\t")
annoTable <- read.table(anno_URL, sep="\t", header=TRUE)
annoTable$sampleID <- annoTable$RNA_ID

# create OutriderDataSet object
ods <- OutriderDataSet(countData=ctsTable, colData=annoTable)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(org.Hs.eg.db)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
map <- select(org.Hs.eg.db, keys=keys(txdb, keytype = "GENEID"),
keytype="ENTREZID", columns=c("SYMBOL"))
try({
library(RMariaDB)
library(AnnotationDbi)
con <- dbConnect(MariaDB(), host='genome-mysql.cse.ucsc.edu',
dbname="hg19", user='genome')
map <- dbGetQuery(con, 'select kgId AS TXNAME, geneSymbol from kgXref')
txdbUrl <- paste0("https://cmm.in.tum.de/public/",
"paper/mitoMultiOmics/ucsc.knownGenes.db")
download.file(txdbUrl, "ucsc.knownGenes.db")
txdb <- loadDb("ucsc.knownGenes.db")
})

# calculate FPKM values and label not expressed genes
ods <- filterExpression(ods, txdb, mapping=map,
filterGenes=FALSE, savefpkm=TRUE)

# display the FPKM distribution of counts.
plotFPKM(ods)

# do the actual subsetting based on the filtering labels
ods <- ods[mcols(ods)$passedFilter,]
ods <- estimateSizeFactors(ods)
ods <- controlForConfounders(ods, q=21, iterations=3)

```

```
ods <- fit(ods)
hist(theta(ods))

# compute P-values (nominal and adjusted)
ods <- computeP-values(ods, alternative="two.sided", method="BY")
# compute the Z-scores
ods <- computeZscores(ods)

res <- results(ods)
head(res)
```

## A.8 Salmon Script

```
#!/bin/bash
#PBS -N Salmon index creation
#PBS -l walltime=15:00:00
#PBS -l nodes=1:ppn=16
#PBS -l mem=40000m
#PBS -t 1,2,5,6,7,8,9,10,11,12

module load conda/4.4.0
source activate salmon

cd /scratch/jl5e18/RNA_SEQ/build38/Build38

fastq_files="/scratch/jl5e18/RNA_SEQ/build38/Build38/FASTQ"
idx="/scratch/jl5e18/RNA_SEQ/build38/Build38/salmon_index/"

## Names of fq files and location in filestore
fq1="/scratch/jl5e18/RNA_SEQ/build38/Build38/FASTQ/"$PBS_ARRAYID"_1.fq.gz"
fq2="/scratch/jl5e18/RNA_SEQ/build38/Build38/FASTQ/"$PBS_ARRAYID"_2.fq.gz"
OUT="/scratch/jl5e18/RNA_SEQ/build38/Build38/Salmon_Out/"$PBS_ARRAYID""

##salmon index -t "/scratch/jl5e18/RNA_SEQ/build38/Build38/gentrome.fa" -d
"/scratch/jl5e18/RNA_SEQ/build38/Build38/decoys.txt" -p 12 -i salmon_index --gencode

salmon quant -i $idx -l A -1 $fq1 -2 $fq2 \
-p 4 -o $OUT --seqBias --gcBias --dumpEq

fi
```



```

write.table(counts, file = "counts.txt", sep = "\t",
            row.names = TRUE)

transcripts_to_keep = filter_transcripts(gene_to_transcript = id_table, transcript_counts = counts,
min_transcript_proportion = 0.01, min_transcript_counts = 10, min_gene_counts = 20)
head(transcripts_to_keep)
equiv_classes_files = file.path(data_dir, "STAR-salmon2", sample_names, "aux_info",
"eq_classes.txt")
file.exists(equiv_classes_files)
equiv_classes_files
samples_design$sample_idhead

input_data = create_data(salmon_or_kallisto = "salmon",
                        gene_to_transcript = id_table,
                        salmon_path_to_eq_classes = equiv_classes_files,
                        eff_len = eff_len,
                        n_cores = 120,
                        transcripts_to_keep = transcripts_to_keep)
##input filter
input_data = filter_genes(input_data, min_counts_per_gene = 20)

set.seed(61217)
precision = prior_precision(gene_to_transcript = id_table,
                            transcript_counts = counts,
                            n_cores = 120,
                            transcripts_to_keep = transcripts_to_keep)
precision$prior

png(filename="precision.png")
plot_precision(precision)
dev.off()
set.seed(61217)
results = test_DTU(BANDITS_data = input_data, precision = precision$prior, samples_design =
samples_design, group_col_name = "group", R = 10^4, burn_in = 2*10^3, gene_to_transcript =
gene_tr_id)
results
write.table(results, file = "covid_vs_flu_DTU.txt", sep = "\t", row.names = TRUE, col.names = TRUE)

write.table(top_genes(results), file = "covid_vs_flu_DTU.txt", sep = "\t", row.names = TRUE,
col.names = TRUE)

write.table(convergence(results), file = "covid_vs_flu_DTU.txt", sep = "\t", row.names = TRUE,
col.names = TRUE)

write.table(top_transcripts(results, sort_by = "transcript"), file = "covid_vs_flu_DTU.txt", sep = "\t",
row.names = TRUE, col.names = TRUE)

```

```
head(top_genes(results))
head(top_genes(results, sort_by = "DTU_measure"))
head(top_transcripts(results, sort_by = "transcript"))
head(convergence(results))
top_gene = top_genes(results, n = 1)
gene(results, top_gene$Gene_id)

top_transcript = top_transcripts(results, n = 1)
transcript(results, top_transcript$Transcript_id)

png(filename="proportions.png")
plot_proportions(results, top_gene$Gene_id, CI = TRUE, CI_level = 0.95)
dev.off()
```

## A.10 OUTRIDER results, in full

geneID		sampleID	P adjust	Z Score	l2fc	Norm counts	Aberrant sample	Aberrant Gene
ENSG00000012124.17	CD22	SRB0017	0.009097	-6.12	-2.26	124.93	6	1
ENSG00000080293.9	SCTR	SOT140	0.039617	3.89	2.56	50.01	1	1
ENSG00000084693.16	AGBL5	SOT102	0.021283	5.02	0.65	1057.59	3	1
ENSG00000092978.11	GPATCH2	SOT38	0.04585	-5.12	-0.64	347.7	2	1
ENSG00000099864.18	PALM	SOT104	0.002437	2.41	1.69	54.72	3	1
ENSG00000112238.11	PRDM13	SOT104	0.015301	3.21	2.39	57.78	3	1
ENSG00000114374.13	USP9Y	SOT33	0.006873	2.5	1.58	73.95	1	1
ENSG00000115234.11	SNX17	SOT10	0.033616	-5.46	-0.57	1327.15	1	1
ENSG00000117868.16	ESYT2	SOT19	0.012232	5.09	0.78	6405.29	2	1
ENSG00000123130.17	ACOT9	SRB0012	3.66E-06	-7.09	-4.09	62.34	1	1
ENSG00000132275.11	RRP8	SRB0006	0.006004	-5.93	-1.95	174.26	3	1
ENSG00000132704.16	FCRL2	SRB0017	0.043928	-5.79	-2.38	103.59	6	1
ENSG00000133195.11	SLC39A11	SOT18	0.028063	4.81	0.65	723.54	3	1
ENSG00000134313.15	KIDINS220	SOT18	0.001479	-5.87	-0.5	3419.64	3	1
ENSG00000140157.14	NIPA2	SOT117	0.000754	-6.13	-0.93	417.19	4	1
ENSG00000142611.17	PRDM16	SOT104	0.002437	3.3	2.15	109.54	3	1
ENSG00000156738.17	MS4A1	SRB0017	0.013149	-5.92	-2.76	252.66	6	1
ENSG00000160208.13	RRP1B	SRB0006	0.007118	-5.77	-1.47	240.76	3	1
ENSG00000164403.14	SHROOM1	SOT102	0.021283	-6.15	-3.93	4.54	3	1
ENSG00000170113.16	NIPA1	SOT117	0.032775	-5.31	-1.01	222.06	4	1
ENSG00000196092.13	PAX5	SRB0017	0.000307	-6.51	-3	52.29	6	1
ENSG00000198818.10	SFT2D1	SOT102	0.000794	-6.19	-1.18	412.4	3	1
ENSG00000211592.8	IGKC	SRB0017	0.049605	-6.31	-5.98	0	6	1
ENSG00000211899.10	IGHM	SRB0017	0.000519	-6.53	-3.67	227.34	6	1
ENSG00000230847.4	OCLNP1	SOT130	0.009216	1.21	1.4	32.06	1	14
ENSG00000230847.4	OCLNP1	SOT152	0.002884	-2.36	-4.69	0	1	14
ENSG00000230847.4	OCLNP1	SOT17	2.99E-08	-0.99	-2.35	2.15	1	14
ENSG00000230847.4	OCLNP1	SOT18	0.000266	0.93	0.92	21.95	3	14
ENSG00000230847.4	OCLNP1	SOT19	4.04E-12	-1	-2.38	2.13	2	14
ENSG00000230847.4	OCLNP1	SOT20	0.000441	0.89	0.85	20.91	1	14
ENSG00000230847.4	OCLNP1	SOT38	0.029999	0.85	0.78	19.98	2	14
ENSG00000230847.4	OCLNP1	SOT45	6.49E-08	-0.68	-1.82	3.17	1	14
ENSG00000230847.4	OCLNP1	SOT49	0.000205	0.81	0.72	18.98	1	14
ENSG00000230847.4	OCLNP1	SOT58	7.37E-15	-2.56	-5.04	0.18	1	14
ENSG00000230847.4	OCLNP1	SOT69	0.002781	1.06	1.14	26.07	1	14
ENSG00000230847.4	OCLNP1	SRB0006	6.80E-20	-3.27	-6.25	0	3	14
ENSG00000230847.4	OCLNP1	SRB0011	9.25E-14	1.44	1.79	41.56	1	14
ENSG00000230847.4	OCLNP1	SRB0013	0.00021	1.3	1.56	35.99	1	14
ENSG00000273749.5	CYFIP1	SOT117	0.019793	-5.56	-1	433.15	4	1

ENSG00000275835.5	TUBGCP5	SOT117	0.020515	-5.39	-0.97	217.67	4	1
-------------------	---------	--------	----------	-------	-------	--------	---	---

### A.11 Differentially Expressed Gene graphs from Covid19 and Influenza cohorts stratified by age from pcaExplorer

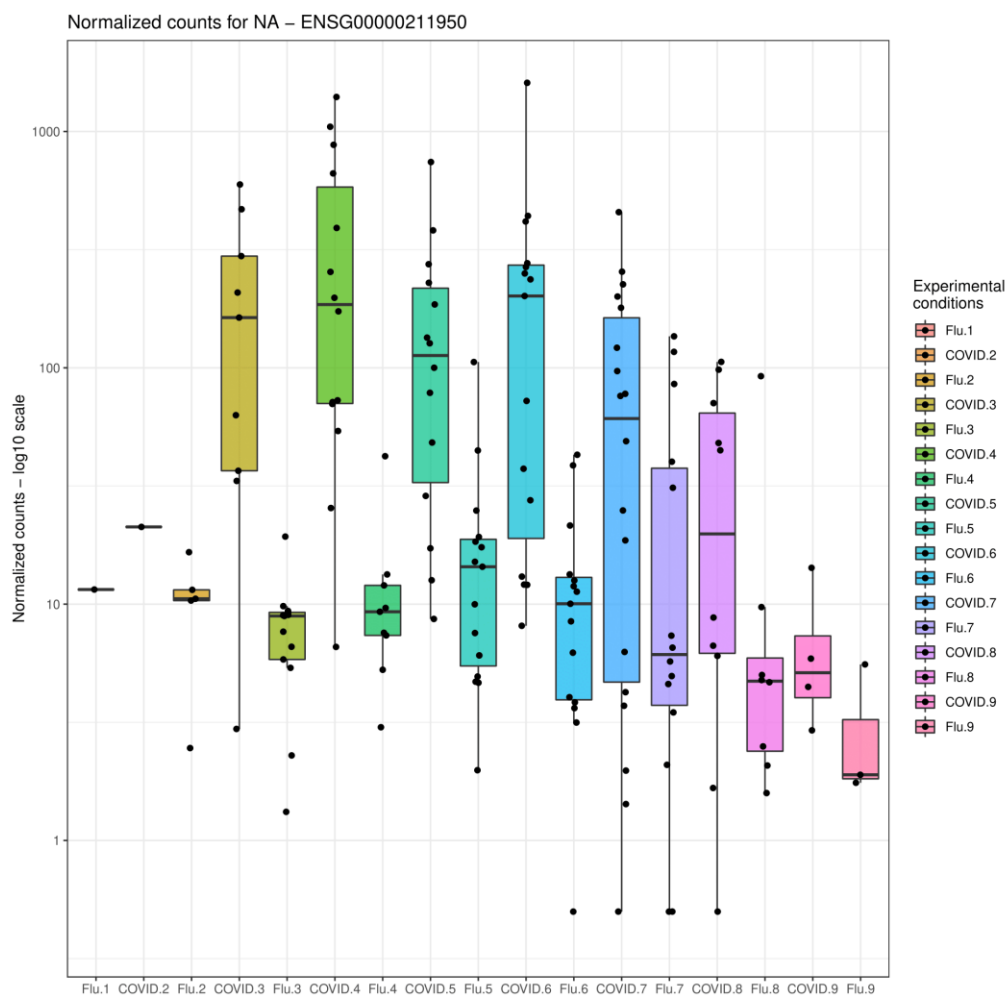
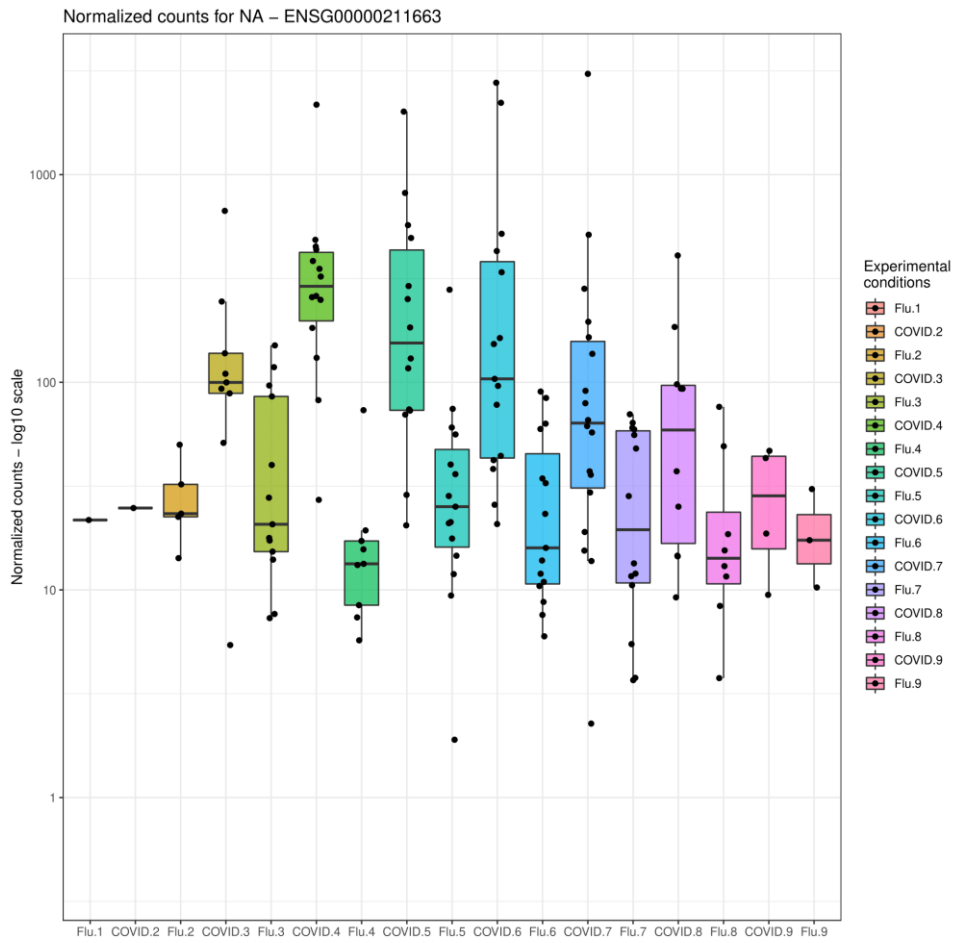
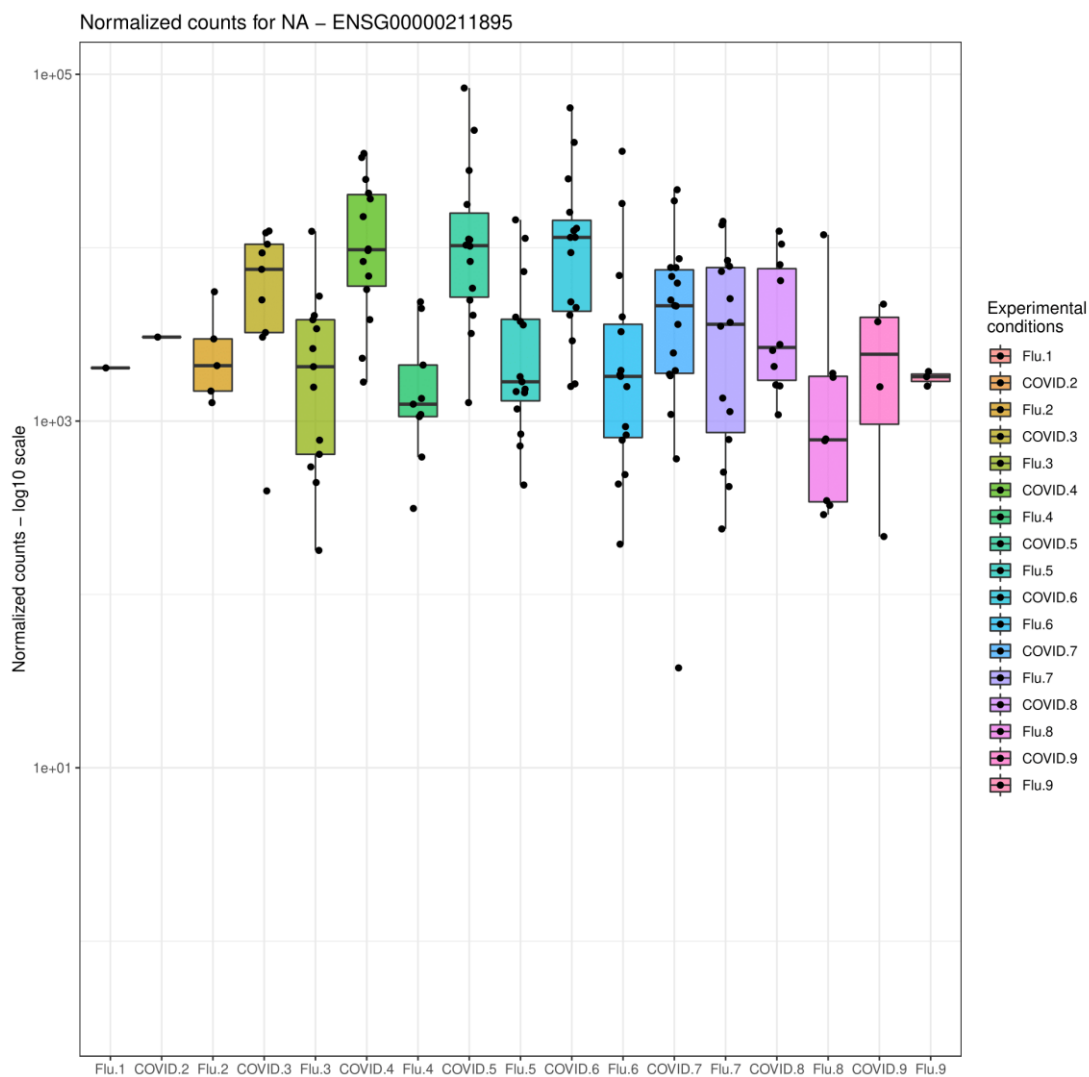


Figure 8-5 Differential gene expression for IGHG 1-24 in Covid and Influenza cohorts stratified by decade of life



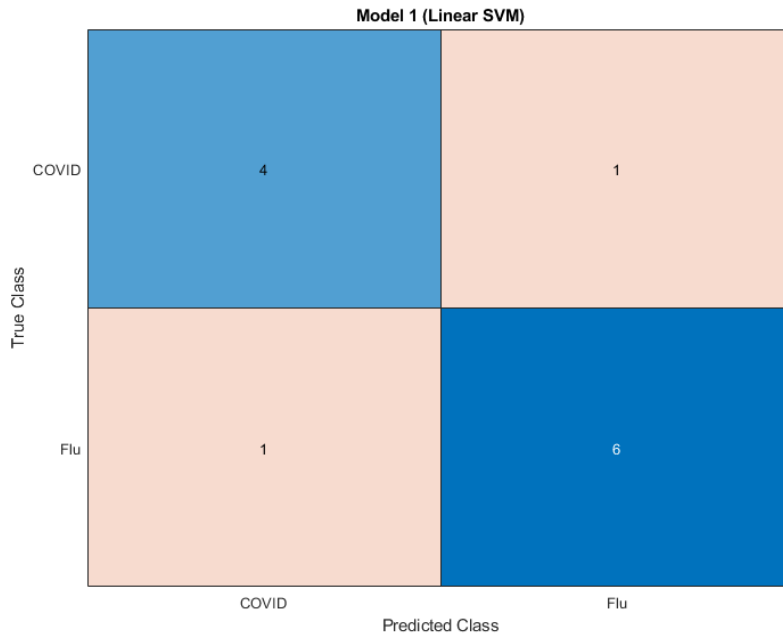


**Figure 8-6 Differential gene expression for IGLV 3-19 in Covid and Influenza cohorts stratified by decade of life.**

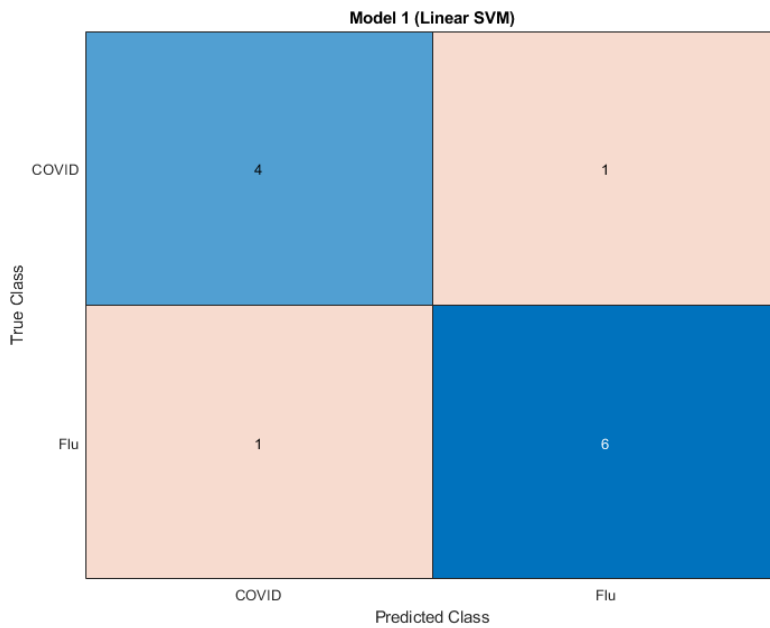


**Figure 8-7 Differential gene expression for IGHA1 in Covid and Influenza cohorts stratified by decade of life**

## A.12 Machine Learning Classification Performance Plots



**Figure 8-8 Classification matrix for old cohort, using gene expression.**



**Figure 8-9 Classification matrix for young cohort, using gene expression.**

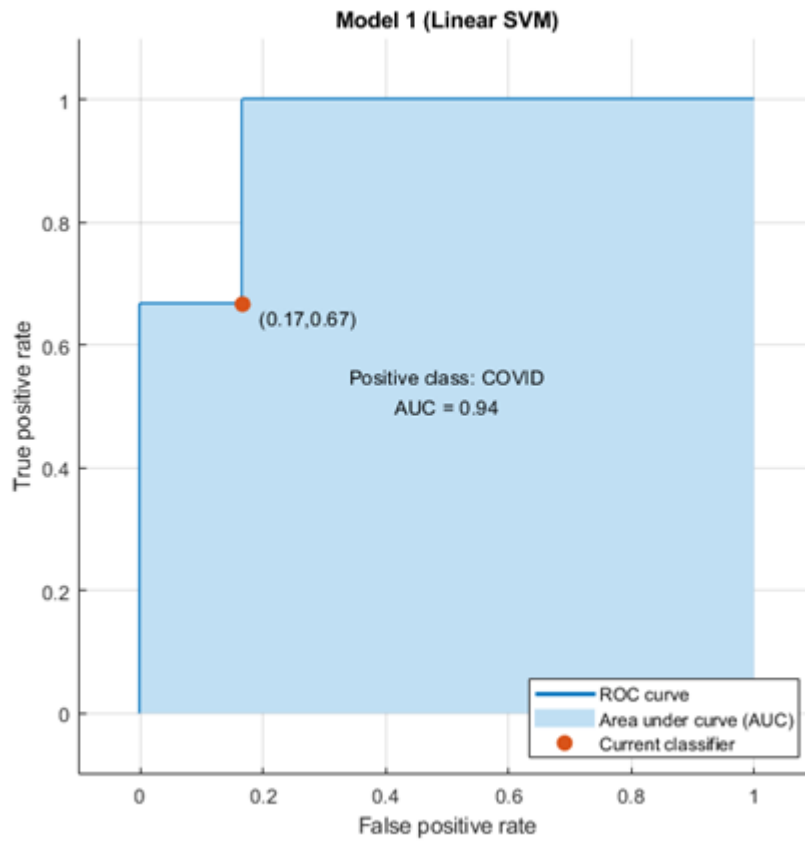


Figure 8-10 ROC plot - old test cohort, using gene expression to predict COVID19

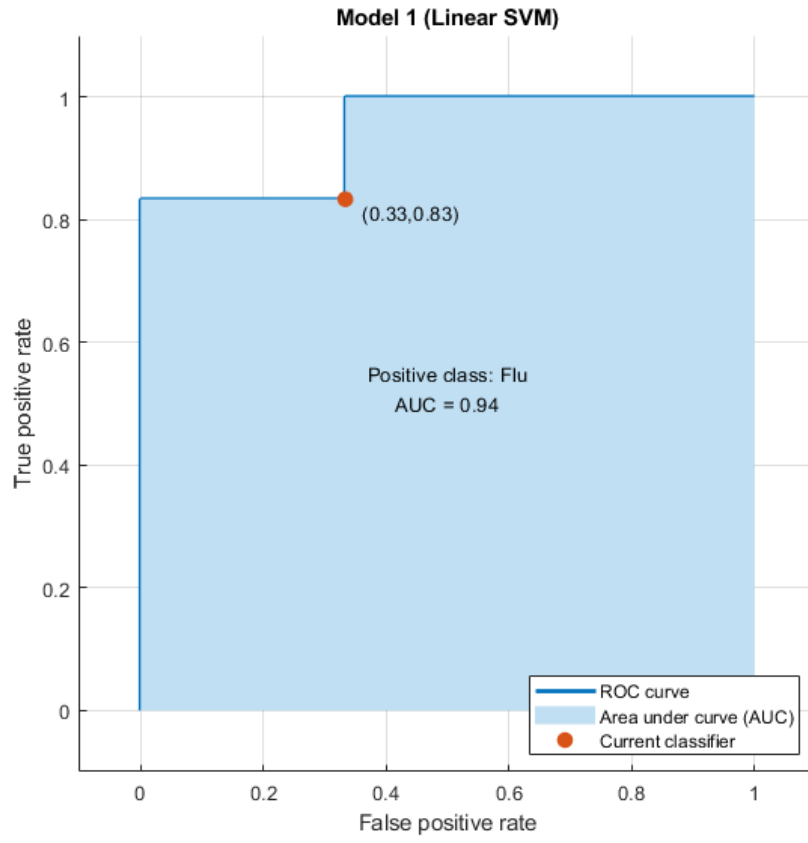


Figure 8-11 ROC plot - old test cohort, using gene expression to predict Influenza

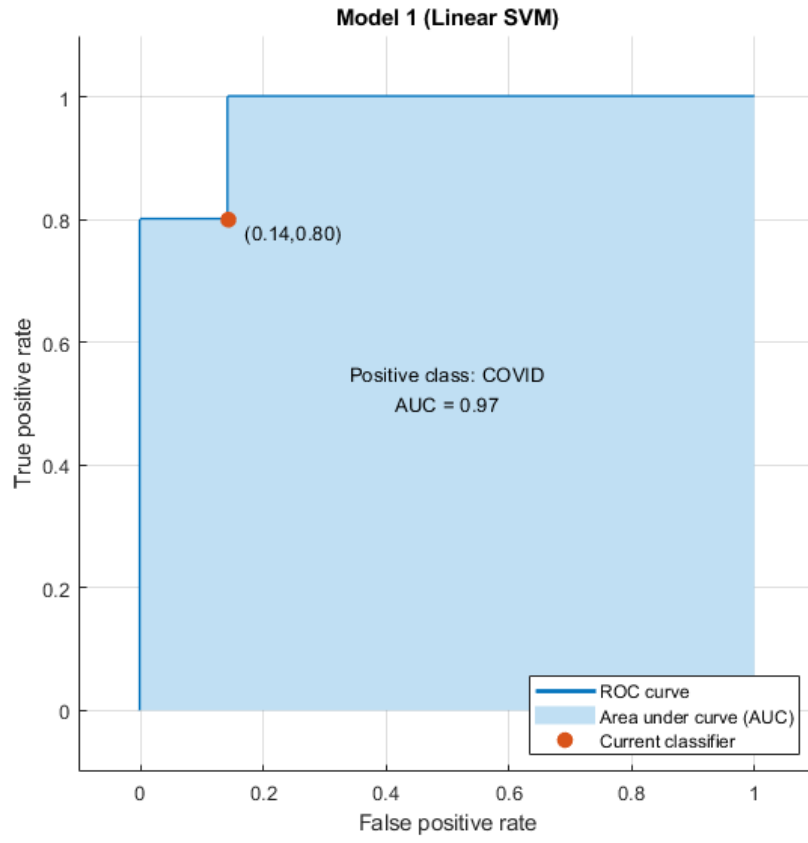


Figure 8-12 ROC plot - young test cohort, using gene expression to predict COVID19

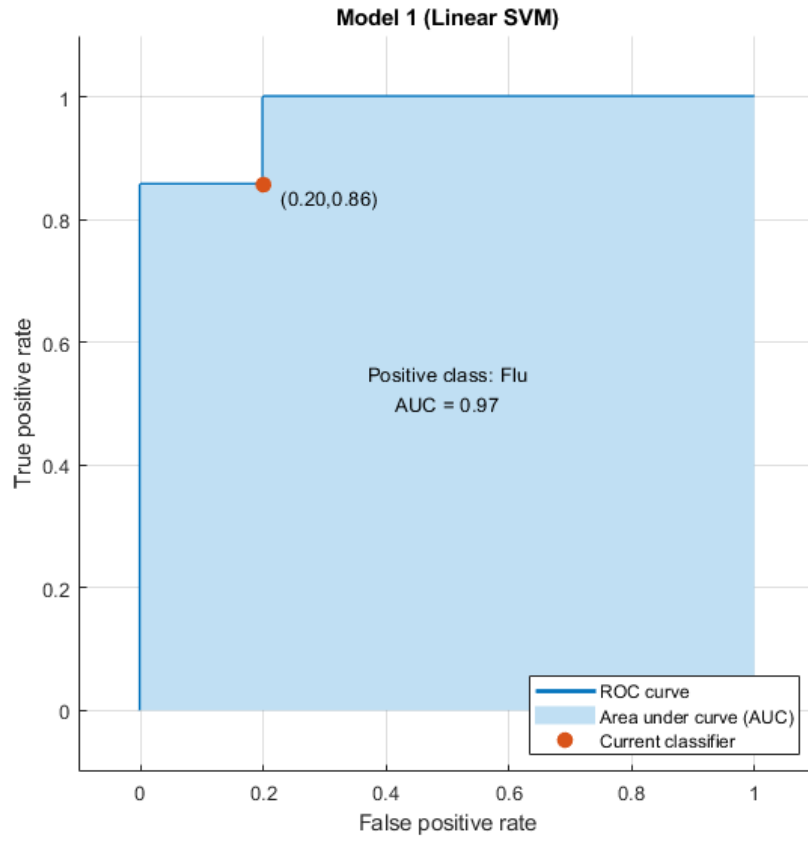
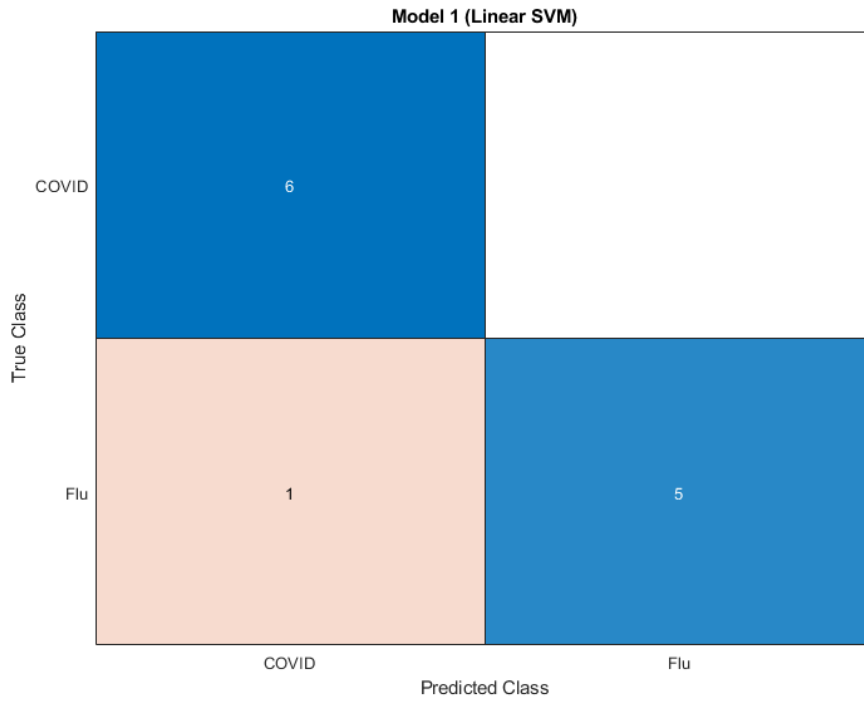
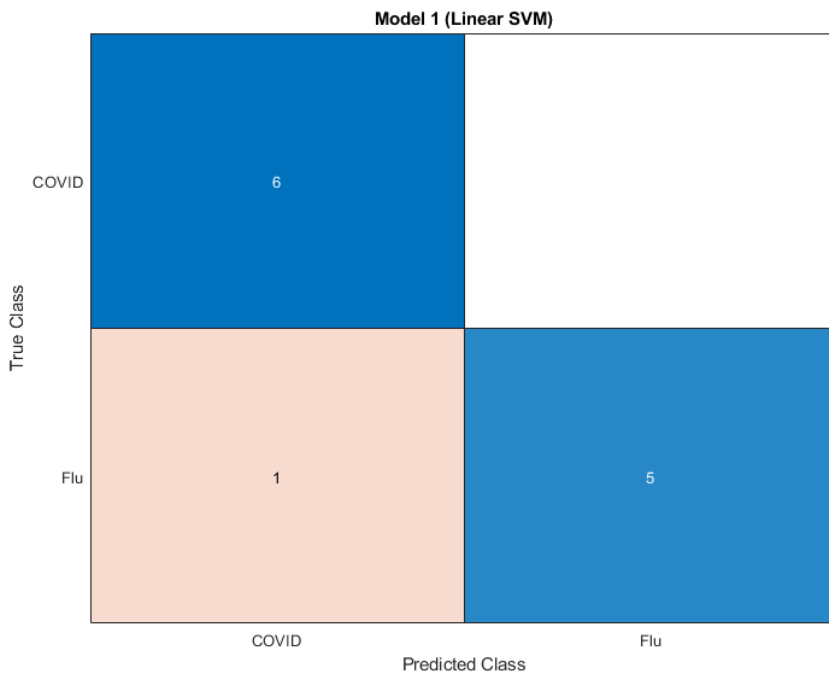


Figure 8-13 ROC plot - young test cohort, using gene expression to predict Influenza

]



**Figure 8-14** Classification matrix for old cohort, using isoform abundance.



**Figure 8-15** Classification matrix for young cohort, using isoform abundance.



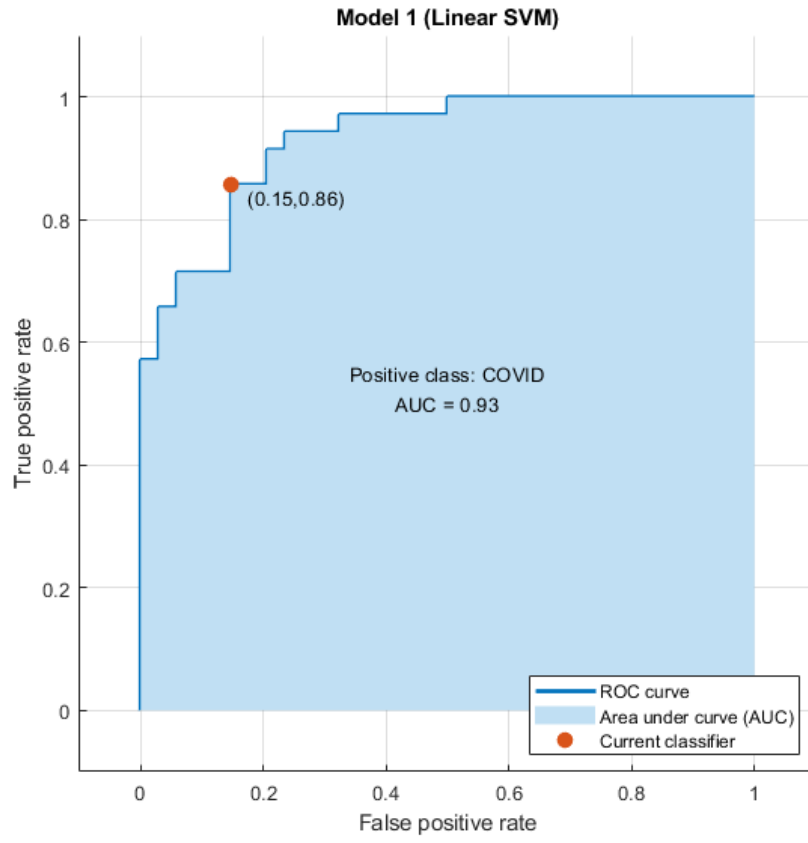


Figure 8-16 ROC plot - old test cohort, using Isoform expression to predict COVID19

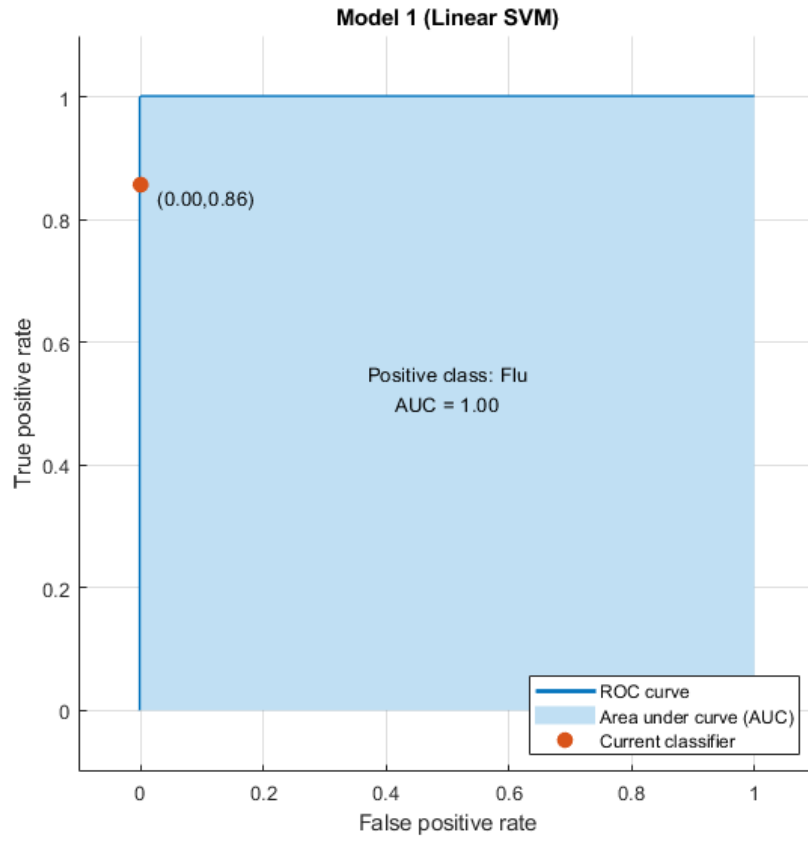
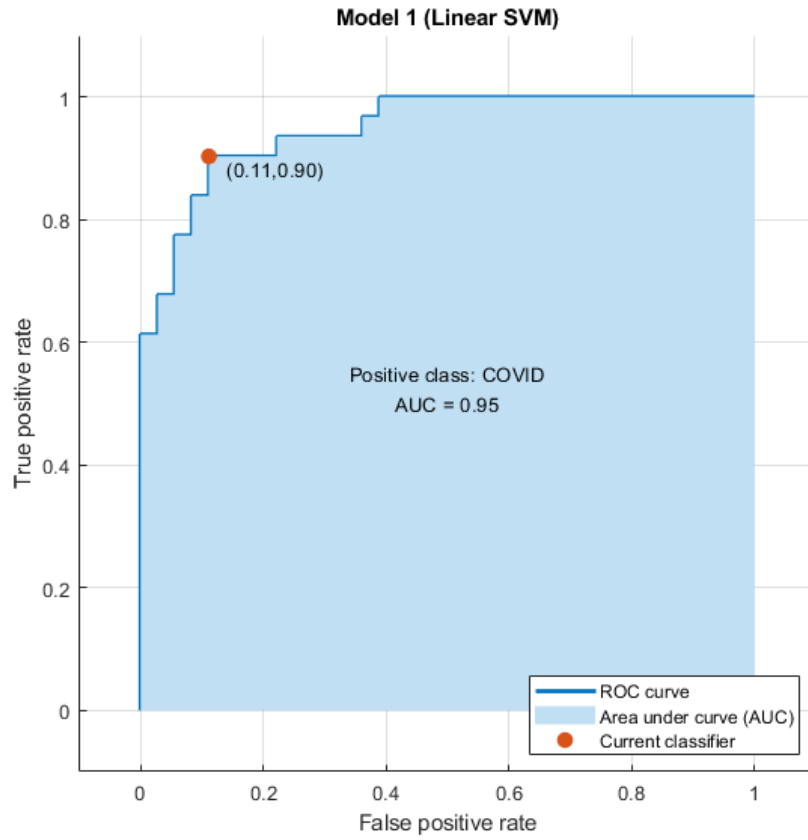


Figure 8-17 ROC plot - old test cohort, using Isoform expression to predict Influenza



**Figure 8-18 ROC plot - young test cohort, using Isoform expression to predict COVID19**

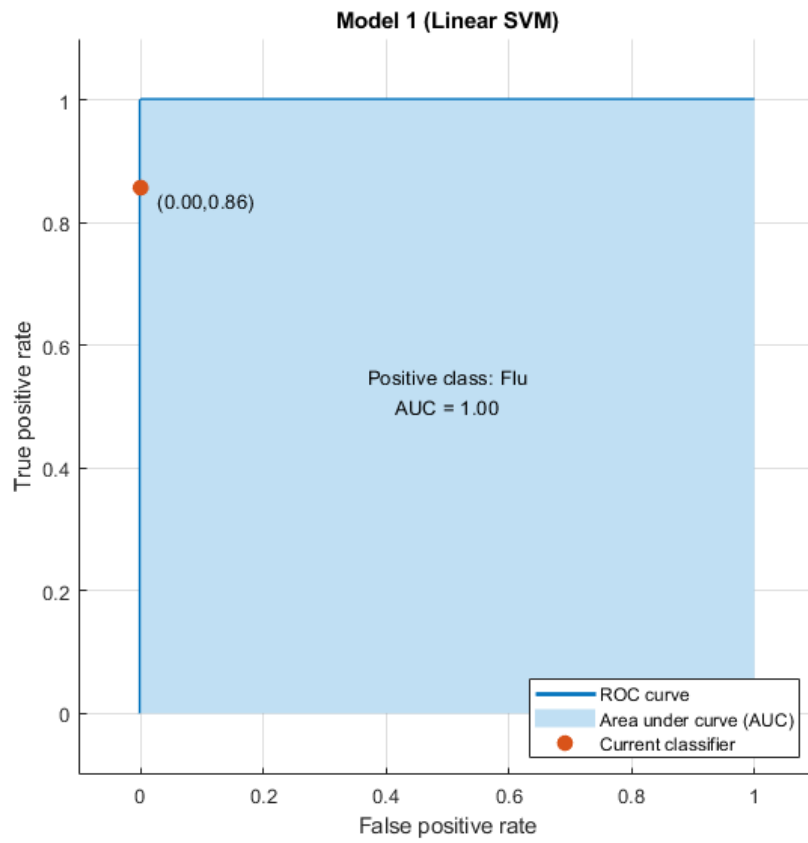


Figure 8-19 ROC plot - young test cohort, using Isoform expression to predict Influenza

### Processes which showed converging gene expression, originally higher in influenza

ID	Name	p-value
1	GO:0046324 regulation of glucose import	1.18E-04
2	GO:0010827 regulation of glucose transmembrane transport	1.22E-04
3	GO:0002320 lymphoid progenitor cell differentiation	1.25E-04
4	GO:1904659 glucose transmembrane transport	2.53E-04
5	GO:0048771 tissue remodeling	2.65E-04
6	GO:0008645 hexose transmembrane transport	2.97E-04
	extension of a leading process involved in cell motility in	
7	GO:0021816 cerebral cortex radial glia guided migration	3.01E-04
8	GO:0015749 monosaccharide transmembrane transport	3.66E-04
9	GO:0097581 lamellipodium organization	3.86E-04
10	GO:0071345 cellular response to cytokine stimulus	4.24E-04
11	GO:0046323 glucose import	4.69E-04
12	GO:0034097 response to cytokine	5.73E-04
13	GO:0002682 regulation of immune system process	6.78E-04
14	GO:0034219 carbohydrate transmembrane transport	8.27E-04
15	GO:0002764 immune response-regulating signaling pathway	8.50E-04
16	GO:2001222 regulation of neuron migration	8.81E-04
17	GO:0033133 positive regulation of glucokinase activity	9.94E-04
18	GO:0070782 phosphatidylserine exposure on apoptotic cell surface	9.94E-04
19	GO:0045588 positive regulation of gamma-delta T cell differentiation	9.94E-04
20	GO:0007135 meiosis II	1.11E-03
21	GO:0061983 meiosis II cell cycle process	1.11E-03
22	GO:0050776 regulation of immune response	1.21E-03
23	GO:0002328 pro-B cell differentiation	1.28E-03
24	GO:1903301 positive regulation of hexokinase activity	1.28E-03
25	GO:0046645 positive regulation of gamma-delta T cell activation	1.60E-03
26	GO:0061754 negative regulation of circulating fibrinogen levels	1.74E-03
27	GO:0045586 regulation of gamma-delta T cell differentiation	1.98E-03
28	GO:0002521 leukocyte differentiation	2.04E-03
	antigen processing and presentation of exogenous	
29	GO:0019886 peptide antigen via MHC class II	2.13E-03
30	GO:0042551 neuron maturation	2.27E-03
31	GO:0008643 carbohydrate transport	2.28E-03
32	GO:0033131 regulation of glucokinase activity	2.40E-03
33	GO:1903131 mononuclear cell differentiation	2.61E-03
34	GO:0002221 pattern recognition receptor signaling pathway	2.65E-03
35	GO:1990830 cellular response to leukemia inhibitory factor	2.81E-03
36	GO:1905605 positive regulation of blood-brain barrier permeability	2.86E-03
37	GO:1904347 regulation of small intestine smooth muscle contraction	2.86E-03
38	GO:1990770 small intestine smooth muscle contraction	2.86E-03
	modulation of microtubule cytoskeleton involved in	
39	GO:0021815 cerebral cortex radial glia guided migration	2.86E-03
40	GO:0044537 regulation of circulating fibrinogen levels	2.86E-03
41	GO:0046643 regulation of gamma-delta T cell activation	2.88E-03
42	GO:1903299 regulation of hexokinase activity	2.88E-03

43	GO:1990823	response to leukemia inhibitory factor regulation of pattern recognition receptor signaling	2.90E-03
44	GO:0062207	pathway antigen processing and presentation of peptide antigen	3.13E-03
45	GO:0002495	via MHC class II adaptive immune response based on somatic recombination of immune receptors built from	3.30E-03
46	GO:0002460	immunoglobulin superfamily domains	3.37E-03
47	GO:0071514	genomic imprinting	3.42E-03
48	GO:0035195	miRNA-mediated gene silencing	3.84E-03
49	GO:0030098	lymphocyte differentiation	3.95E-03
50	GO:0042415	norepinephrine metabolic process	4.00E-03
51	GO:0031269	pseudopodium assembly antigen processing and presentation of peptide or	4.00E-03
52	GO:0002504	polysaccharide antigen via MHC class II	4.03E-03
53	GO:0032765	positive regulation of mast cell cytokine production	4.24E-03
54	GO:0051490	negative regulation of filopodium assembly	4.24E-03
55	GO:0046951	ketone body biosynthetic process	4.24E-03
56	GO:0034121	regulation of toll-like receptor signaling pathway	4.39E-03
57	GO:0034505	tooth mineralization	4.43E-03
58	GO:0042492	gamma-delta T cell differentiation	4.65E-03
59	GO:0035194	post-transcriptional gene silencing by RNA	4.74E-03
60	GO:0055075	potassium ion homeostasis	4.86E-03
61	GO:0016441	post-transcriptional gene silencing	5.16E-03
62	GO:0035234	ectopic germ cell programmed cell death	5.36E-03
63	GO:0031268	pseudopodium organization	5.36E-03
64	GO:0034162	toll-like receptor 9 signaling pathway	5.36E-03
65	GO:0007186	G protein-coupled receptor signaling pathway	5.60E-03
66	GO:0002238	response to molecule of fungal origin	5.87E-03
67	GO:1902463	protein localization to cell leading edge	5.87E-03
68	GO:2000405	negative regulation of T cell migration	5.87E-03
69	GO:0070167	regulation of biomineral tissue development	6.25E-03
70	GO:0002684	positive regulation of immune system process	6.32E-03
71	GO:1902622	regulation of neutrophil migration	6.74E-03
72	GO:0110149	regulation of biomineralization	6.79E-03
73	GO:0001764	neuron migration	7.19E-03
74	GO:0048871	multicellular organismal homeostasis	7.33E-03
75	GO:0034763	negative regulation of transmembrane transport	7.33E-03
76	GO:0097009	energy homeostasis	7.71E-03
77	GO:1905603	regulation of blood-brain barrier permeability cell motility involved in cerebral cortex radial glia guided	7.74E-03
78	GO:0021814	migration	7.74E-03
79	GO:0010562	positive regulation of phosphorus metabolic process	7.86E-03
80	GO:0045937	positive regulation of phosphate metabolic process	7.86E-03
81	GO:0045321	leukocyte activation	7.92E-03
82	GO:0030183	B cell differentiation	8.06E-03
83	GO:0090022	regulation of neutrophil chemotaxis	8.63E-03
84	GO:1902624	positive regulation of neutrophil migration	8.63E-03
85	GO:0009617	response to bacterium	9.10E-03

86	GO:0097529	myeloid leukocyte migration	9.26E-03
87	GO:0008038	neuron recognition	9.53E-03
88	GO:0002683	negative regulation of immune system process	9.72E-03
89	GO:0017121	plasma membrane phospholipid scrambling	9.83E-03
90	GO:0046325	negative regulation of glucose import regulation of gastro-intestinal system smooth muscle contraction	9.83E-03
91	GO:1904304	heart rudiment formation	9.84E-03
92	GO:0003315	regulation of mast cell cytokine production	9.84E-03
93	GO:0032763	intestine smooth muscle contraction	9.84E-03
94	GO:0014827	positive regulation of killing of cells of another organism	9.84E-03
95	GO:0051712	pronephric duct morphogenesis	9.84E-03
96	GO:0039023	response to lead ion	9.98E-03
97	GO:0010288	antigen processing and presentation of exogenous peptide antigen	9.98E-03
98	GO:0002478	negative regulation of toll-like receptor signaling pathway	9.98E-03
99	GO:0034122	regulation of transmembrane transport	1.05E-02
100	GO:0034762		

#### Processes which showed converging gene expression, originally higher in COVID19

	ID	Name	Source	p-value
1	GO:0002250	adaptive immune response		3.22E-10
2	GO:1903047	mitotic cell cycle process		1.01E-08
3	GO:0000278	mitotic cell cycle		1.40E-08
4	GO:0002449	lymphocyte mediated immunity		5.58E-07
5	GO:0140014	mitotic nuclear division		5.80E-07
6	GO:0000070	mitotic sister chromatid segregation		7.26E-07
7	GO:0007093	mitotic cell cycle checkpoint signaling		9.60E-07
8	GO:0000819	sister chromatid segregation		1.36E-06
9	GO:0022402	cell cycle process		1.39E-06
10	GO:0002377	immunoglobulin production		1.42E-06
11	GO:0010948	negative regulation of cell cycle process		2.13E-06
12	GO:0006275	regulation of DNA replication		2.43E-06
13	GO:0006260	DNA replication		2.58E-06
14	GO:0010564	regulation of cell cycle process		2.58E-06
15	GO:0044772	mitotic cell cycle phase transition		2.64E-06
16	GO:0000075	cell cycle checkpoint signaling		2.64E-06
17	GO:1901990	regulation of mitotic cell cycle phase transition		3.01E-06
18	GO:1901991	negative regulation of mitotic cell cycle phase transition		3.38E-06
19	GO:0007346	regulation of mitotic cell cycle		3.97E-06
20	GO:0051983	regulation of chromosome segregation		4.58E-06
21	GO:0007059	chromosome segregation		5.29E-06
22	GO:0000280	nuclear division		1.03E-05
23	GO:0044770	cell cycle phase transition		1.16E-05
24	GO:0051276	chromosome organization		1.22E-05
25	GO:0098749	cerebellar neuron development		1.24E-05

26	GO:1905784	regulation of anaphase-promoting complex-dependent catabolic process	1.24E-05
27	GO:0051726	regulation of cell cycle	1.25E-05
28	GO:0051301	cell division	1.54E-05
29	GO:0006259	DNA metabolic process	1.76E-05
30	GO:1905818	regulation of chromosome separation	2.20E-05
31	GO:0098813	nuclear chromosome segregation	2.90E-05
32	GO:0002460	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	3.44E-05
33	GO:0002443	leukocyte mediated immunity	3.55E-05
34	GO:1901987	regulation of cell cycle phase transition	3.84E-05
35	GO:0033045	regulation of sister chromatid segregation	4.12E-05
36	GO:0042770	signal transduction in response to DNA damage	5.71E-05
37	GO:0044784	metaphase/anaphase transition of cell cycle	6.25E-05
38	GO:0010965	regulation of mitotic sister chromatid separation	6.25E-05
39	GO:1904668	positive regulation of ubiquitin protein ligase activity	6.42E-05
40	GO:0045786	negative regulation of cell cycle	6.45E-05
41	GO:0035685	helper T cell diapedesis	6.71E-05
42	GO:1901988	negative regulation of cell cycle phase transition	8.47E-05
43	GO:0051306	mitotic sister chromatid separation	8.60E-05
44	GO:0002440	production of molecular mediator of immune response	9.14E-05
45	GO:0002684	positive regulation of immune system process	9.99E-05
46	GO:0048285	organelle fission	1.01E-04
47	GO:0051052	regulation of DNA metabolic process	1.19E-04
48	GO:0006974	cellular response to DNA damage stimulus	1.21E-04
49	GO:0042267	natural killer cell mediated cytotoxicity	1.29E-04
50	GO:0033043	regulation of organelle organization	1.30E-04
51	GO:0006270	DNA replication initiation	1.32E-04
52	GO:0044774	mitotic DNA integrity checkpoint signaling	1.46E-04
53	GO:0010972	negative regulation of G2/M transition of mitotic cell cycle	1.48E-04
54	GO:0044839	cell cycle G2/M phase transition	1.55E-04
55	GO:1902099	regulation of metaphase/anaphase transition of cell cycle	1.62E-04
56	GO:0002228	natural killer cell mediated immunity	1.72E-04
57	GO:0007091	metaphase/anaphase transition of mitotic cell cycle	1.79E-04
58	GO:0010389	regulation of G2/M transition of mitotic cell cycle	1.89E-04
59	GO:0090329	regulation of DNA-templated DNA replication	1.90E-04
60	GO:1902750	negative regulation of cell cycle G2/M phase transition	1.90E-04
61	GO:0030174	regulation of DNA-templated DNA replication initiation	2.20E-04
62	GO:0072683	T cell extravasation	2.20E-04
63	GO:0002287	alpha-beta T cell activation involved in immune response	2.22E-04
64	GO:0031570	DNA integrity checkpoint signaling	2.40E-04
65	GO:0051304	chromosome separation	2.40E-04
66	GO:0000086	G2/M transition of mitotic cell cycle	2.45E-04
67	GO:0035684	helper T cell extravasation	2.50E-04
68	GO:0002488	antigen processing and presentation of endogenous peptide antigen via MHC class Ib via ER pathway	2.50E-04
69	GO:0002489	antigen processing and presentation of endogenous peptide antigen via MHC class Ib via ER pathway, TAP-dependent	2.50E-04



70	GO:0050778	positive regulation of immune response	2.69E-04
71	GO:0045930	negative regulation of mitotic cell cycle	2.74E-04
72	GO:0031343	positive regulation of cell killing	2.75E-04
73	GO:0072540	T-helper 17 cell lineage commitment	3.10E-04
74	GO:0002711	positive regulation of T cell mediated immunity	3.19E-04
75	GO:0006261	DNA-templated DNA replication	3.23E-04
76	GO:1902749	regulation of cell cycle G2/M phase transition	3.27E-04
77	GO:0002708	positive regulation of lymphocyte mediated immunity	3.27E-04
78	GO:0002456	T cell mediated immunity	3.29E-04
79	GO:0031577	spindle checkpoint signaling	3.58E-04
80	GO:0044773	mitotic DNA damage checkpoint signaling	3.73E-04
81	GO:0019724	B cell mediated immunity	4.06E-04
82	GO:0044818	mitotic G2/M transition checkpoint	4.10E-04
83	GO:0046649	lymphocyte activation	4.20E-04
84	GO:0030071	regulation of mitotic metaphase/anaphase transition	4.53E-04
85	GO:1905819	negative regulation of chromosome separation	4.69E-04
86	GO:0033044	regulation of chromosome organization	5.29E-04
87	GO:0050776	regulation of immune response	5.33E-04
88	GO:0051985	negative regulation of chromosome segregation	5.34E-04
89	GO:0045740	positive regulation of DNA replication	5.34E-04
90	GO:0050871	positive regulation of B cell activation	5.35E-04
91	GO:0016064	immunoglobulin mediated immune response	5.50E-04
92	GO:0048635	negative regulation of muscle organ development	5.68E-04
93	GO:0051132	NK T cell activation	5.68E-04
94	GO:0051251	positive regulation of lymphocyte activation	5.78E-04
95	GO:0009298	GDP-mannose biosynthetic process	6.05E-04
96	GO:0008355	olfactory learning	6.05E-04
97	GO:0002481	antigen processing and presentation of exogenous protein antigen via MHC class Ib, TAP-dependent	6.05E-04
98	GO:0070309	lens fiber cell morphogenesis	6.76E-04
99	GO:0072539	T-helper 17 cell differentiation	6.92E-04
100	GO:0050650	chondroitin sulfate proteoglycan biosynthetic process	7.22E-04

### Processes which showed converging isoform abundance, originally higher in influenza

ID	Name	p-value	
1	GO:0001812	positive regulation of type I hypersensitivity	1.71E-06
2	GO:0001810	regulation of type I hypersensitivity	2.56E-06
3	GO:0016068	type I hypersensitivity	2.56E-06
4	GO:0001798	positive regulation of type IIa hypersensitivity	1.68E-05
5	GO:0002894	positive regulation of type II hypersensitivity	1.68E-05
6	GO:0001796	regulation of type IIa hypersensitivity	2.03E-05
7	GO:0001788	antibody-dependent cellular cytotoxicity	2.03E-05
8	GO:0002892	regulation of type II hypersensitivity	2.03E-05
9	GO:0001794	type IIa hypersensitivity	2.88E-05
10	GO:0002445	type II hypersensitivity	2.88E-05
11	GO:0002885	positive regulation of hypersensitivity	3.94E-05
12	GO:0050729	positive regulation of inflammatory response	4.58E-05

13	GO:0002521	leukocyte differentiation	4.90E-05
		positive regulation of acute inflammatory response to	
14	GO:0002866	antigenic stimulus	5.22E-05
15	GO:0097278	complement-dependent cytotoxicity	5.96E-05
16	GO:0030097	hemopoiesis	6.14E-05
17	GO:0002861	regulation of inflammatory response to antigenic stimulus	6.19E-05
18	GO:0002883	regulation of hypersensitivity	6.75E-05
19	GO:0048534	hematopoietic or lymphoid organ development	9.42E-05
		positive regulation of inflammatory response to antigenic	
20	GO:0002863	stimulus	1.06E-04
21	GO:0031016	pancreas development	1.10E-04
22	GO:0002524	hypersensitivity	1.18E-04
		regulation of acute inflammatory response to antigenic	
23	GO:0002864	stimulus	1.43E-04
24	GO:1903131	mononuclear cell differentiation	1.45E-04
25	GO:0042904	9-cis-retinoic acid biosynthetic process	1.49E-04
26	GO:0042905	9-cis-retinoic acid metabolic process	1.49E-04
27	GO:0002520	immune system development	1.55E-04
28	GO:0002888	positive regulation of myeloid leukocyte mediated immunity	2.41E-04
29	GO:0030098	lymphocyte differentiation	3.83E-04
30	GO:0002437	inflammatory response to antigenic stimulus	4.82E-04
31	GO:0002682	regulation of immune system process	4.93E-04
32	GO:0048703	embryonic viscerocranium morphogenesis	4.98E-04
33	GO:0007389	pattern specification process	5.41E-04
34	GO:1903432	regulation of TORC1 signaling	5.44E-04
35	GO:0032008	positive regulation of TOR signaling	5.44E-04
36	GO:0032103	positive regulation of response to external stimulus	5.60E-04
37	GO:0001819	positive regulation of cytokine production	5.72E-04
38	GO:0045022	early endosome to late endosome transport	5.76E-04
39	GO:0002438	acute inflammatory response to antigenic stimulus	5.76E-04
40	GO:0002675	positive regulation of acute inflammatory response	6.45E-04
41	GO:0032006	regulation of TOR signaling	6.61E-04
		vesicle-mediated transport between endosomal	
42	GO:0098927	compartments	6.82E-04
43	GO:0031017	exocrine pancreas development	7.19E-04
44	GO:0038202	TORC1 signaling	9.74E-04
45	GO:0045807	positive regulation of endocytosis	1.04E-03
46	GO:0001935	endothelial cell proliferation	1.04E-03
47	GO:0046822	regulation of nucleocytoplasmic transport	1.07E-03
48	GO:0002714	positive regulation of B cell mediated immunity	1.07E-03
		positive regulation of immunoglobulin mediated immune	
49	GO:0002891	response	1.07E-03
50	GO:0080134	regulation of response to stress	1.08E-03
51	GO:0031929	TOR signaling	1.28E-03
52	GO:0009952	anterior/posterior pattern specification	1.35E-03
53	GO:0050727	regulation of inflammatory response	1.39E-03
54	GO:0031349	positive regulation of defense response	1.44E-03
55	GO:0048732	gland development	1.45E-03
56	GO:0090090	negative regulation of canonical Wnt signaling pathway	1.74E-03

57	GO:0009617	response to bacterium	1.76E-03
58	GO:0002138	retinoic acid biosynthetic process	1.83E-03
59	GO:0043009	chordate embryonic development	1.93E-03
60	GO:0046824	positive regulation of nucleocytoplasmic transport	1.98E-03
61	GO:0048010	vascular endothelial growth factor receptor signaling pathway	2.05E-03
62	GO:0002673	regulation of acute inflammatory response	2.05E-03
63	GO:0010171	body morphogenesis	2.05E-03
64	GO:0002377	immunoglobulin production	2.11E-03
65	GO:0050871	positive regulation of B cell activation	2.17E-03
66	GO:0045598	regulation of fat cell differentiation	2.17E-03
67	GO:0045600	positive regulation of fat cell differentiation	2.21E-03
68	GO:0045087	innate immune response	2.21E-03
69	GO:0016102	diterpenoid biosynthetic process	2.22E-03
70	GO:0009792	embryo development ending in birth or egg hatching	2.24E-03
71	GO:0008333	endosome to lysosome transport	2.37E-03
72	GO:1904263	positive regulation of TORC1 signaling	2.42E-03
73	GO:0002440	production of molecular mediator of immune response	2.69E-03
74	GO:0002712	regulation of B cell mediated immunity	2.70E-03
75	GO:0002889	regulation of immunoglobulin mediated immune response	2.70E-03
76	GO:1901679	nucleotide transmembrane transport	2.86E-03
77	GO:0042574	retinal metabolic process	2.86E-03
78	GO:0090316	positive regulation of intracellular protein transport	2.94E-03
79	GO:0048705	skeletal system morphogenesis	2.97E-03
80	GO:0048546	digestive tract morphogenesis	2.98E-03
81	GO:0055123	digestive system development	3.12E-03
82	GO:0048706	embryonic skeletal system development	3.17E-03
83	GO:0051622	negative regulation of norepinephrine uptake	3.20E-03
84	GO:0035543	positive regulation of SNARE complex assembly	3.20E-03
85	GO:1904609	cellular response to monosodium L-glutamate	3.20E-03
86	GO:1904608	response to monosodium L-glutamate	3.20E-03
87	GO:0060691	epithelial cell maturation involved in salivary gland development	3.20E-03
88	GO:1903284	positive regulation of glutathione peroxidase activity	3.20E-03
89	GO:0002014	vasoconstriction of artery involved in ischemic response to lowering of systemic arterial blood pressure	3.20E-03
90	GO:0015779	glucuronoside transport	3.20E-03
91	GO:0060096	serotonin secretion, neurotransmission	3.20E-03
92	GO:0051585	negative regulation of dopamine uptake involved in synaptic transmission	3.20E-03
93	GO:0002509	central tolerance induction to self-antigen	3.20E-03
94	GO:0051945	negative regulation of catecholamine uptake involved in synaptic transmission	3.20E-03
95	GO:0002886	regulation of myeloid leukocyte mediated immunity	3.27E-03
96	GO:0009894	regulation of catabolic process	3.30E-03
97	GO:0019731	antibacterial humoral response	3.37E-03
98	GO:0016114	terpenoid biosynthetic process	3.58E-03
99	GO:0006837	serotonin transport	3.58E-03
10			
0	GO:0001502	cartilage condensation	3.58E-03

## Processes which showed converging isoform abundance, originally higher in COVID19

ID	Name	p-value	
1	GO:0006909	phagocytosis	1.37E-06
2	GO:0001812	positive regulation of type I hypersensitivity	3.06E-06
3	GO:0001810	regulation of type I hypersensitivity	4.58E-06
4	GO:0016068	type I hypersensitivity	4.58E-06
5	GO:0007005	mitochondrion organization	2.77E-05
6	GO:0001798	positive regulation of type IIa hypersensitivity	2.99E-05
7	GO:0002894	positive regulation of type II hypersensitivity	2.99E-05
8	GO:0001796	regulation of type IIa hypersensitivity	3.63E-05
9	GO:0001788	antibody-dependent cellular cytotoxicity	3.63E-05
10	GO:0002892	regulation of type II hypersensitivity	3.63E-05
11	GO:0001794	type IIa hypersensitivity	5.14E-05
12	GO:0002445	type II hypersensitivity	5.14E-05
13	GO:0050766	positive regulation of phagocytosis	5.40E-05
14	GO:0031648	protein destabilization	6.05E-05
15	GO:0002885	positive regulation of hypersensitivity	7.01E-05
16	GO:0031647	regulation of protein stability	7.94E-05
17	GO:0002866	positive regulation of acute inflammatory response to antigenic stimulus	9.28E-05
18	GO:0097278	complement-dependent cytotoxicity	1.06E-04
19	GO:0030100	regulation of endocytosis	1.09E-04
20	GO:0002883	regulation of hypersensitivity	1.20E-04
21	GO:0002863	positive regulation of inflammatory response to antigenic stimulus	1.88E-04
22	GO:0019882	antigen processing and presentation	1.90E-04
23	GO:0002524	hypersensitivity	2.09E-04
24	GO:0045807	positive regulation of endocytosis	2.10E-04
25	GO:0010639	negative regulation of organelle organization	2.53E-04
26	GO:0001919	regulation of receptor recycling	2.54E-04
27	GO:0002864	regulation of acute inflammatory response to antigenic stimulus	2.54E-04
28	GO:0006911	phagocytosis, engulfment	2.63E-04
29	GO:0050764	regulation of phagocytosis	2.72E-04
30	GO:0045321	leukocyte activation	3.48E-04
31	GO:1903421	regulation of synaptic vesicle recycling	3.61E-04
32	GO:0099024	plasma membrane invagination	3.78E-04
33	GO:0002888	positive regulation of myeloid leukocyte mediated immunity	4.24E-04
34	GO:0001775	cell activation	4.29E-04
35	GO:0031397	negative regulation of protein ubiquitination	4.45E-04
36	GO:0061024	membrane organization	4.54E-04
37	GO:0010324	membrane invagination	4.86E-04
38	GO:0021543	pallium development	5.10E-04
39	GO:0060627	regulation of vesicle-mediated transport	5.14E-04
40	GO:1901137	carbohydrate derivative biosynthetic process	5.77E-04
41	GO:0001881	receptor recycling	7.00E-04
42	GO:0046649	lymphocyte activation	7.17E-04
43	GO:1903321	negative regulation of protein modification by small protein	7.33E-04

		conjugation or removal	
44	GO:0031349	positive regulation of defense response	7.40E-04
45	GO:0090324	negative regulation of oxidative phosphorylation	7.92E-04
46	GO:0090527	actin filament reorganization	7.92E-04
47	GO:1905232	cellular response to L-glutamate	7.92E-04
48	GO:0031400	negative regulation of protein modification process	8.40E-04
49	GO:0046785	microtubule polymerization	8.43E-04
50	GO:0009617	response to bacterium	8.93E-04
51	GO:1901135	carbohydrate derivative metabolic process	9.90E-04
52	GO:0045022	early endosome to late endosome transport	1.01E-03
53	GO:0002438	acute inflammatory response to antigenic stimulus	1.01E-03
54	GO:1904667	negative regulation of ubiquitin protein ligase activity	1.12E-03
55	GO:1901856	negative regulation of cellular respiration	1.12E-03
56	GO:0010040	response to iron (II) ion	1.12E-03
57	GO:0002675	positive regulation of acute inflammatory response	1.13E-03
58	GO:0050729	positive regulation of inflammatory response	1.14E-03
59	GO:0051345	positive regulation of hydrolase activity	1.19E-03
		vesicle-mediated transport between endosomal	
60	GO:0098927	compartments	1.19E-03
61	GO:0002694	regulation of leukocyte activation	1.21E-03
62	GO:0010942	positive regulation of cell death	1.27E-03
63	GO:0006910	phagocytosis, recognition	1.28E-03
64	GO:0010950	positive regulation of endopeptidase activity	1.47E-03
65	GO:0021766	hippocampus development	1.53E-03
66	GO:0010752	regulation of cGMP-mediated signaling	1.71E-03
67	GO:0001961	positive regulation of cytokine-mediated signaling pathway	1.78E-03
68	GO:0019220	regulation of phosphate metabolic process	1.79E-03
69	GO:0043524	negative regulation of neuron apoptotic process	1.79E-03
70	GO:0051174	regulation of phosphorus metabolic process	1.80E-03
71	GO:0043523	regulation of neuron apoptotic process	1.83E-03
72	GO:0002714	positive regulation of B cell mediated immunity	1.86E-03
		positive regulation of immunoglobulin mediated immune	
73	GO:0002891	response	1.86E-03
74	GO:0034097	response to cytokine	1.89E-03
75	GO:0016192	vesicle-mediated transport	1.89E-03
76	GO:0043065	positive regulation of apoptotic process	1.93E-03
77	GO:0001921	positive regulation of receptor recycling	1.93E-03
78	GO:0032103	positive regulation of response to external stimulus	2.00E-03
79	GO:0030162	regulation of proteolysis	2.03E-03
80	GO:0031399	regulation of protein modification process	2.05E-03
81	GO:0006956	complement activation	2.07E-03
82	GO:0050865	regulation of cell activation	2.11E-03
83	GO:0010952	positive regulation of peptidase activity	2.13E-03
84	GO:0043068	positive regulation of programmed cell death	2.19E-03
85	GO:0002861	regulation of inflammatory response to antigenic stimulus	2.22E-03
86	GO:1901216	positive regulation of neuron death	2.29E-03
87	GO:0060760	positive regulation of response to cytokine stimulus	2.42E-03
88	GO:0055086	nucleobase-containing small molecule metabolic process	2.88E-03

89	GO:0006897	endocytosis	2.96E-03
90	GO:0010035	response to inorganic substance	3.02E-03
91	GO:0051701	biological process involved in interaction with host	3.08E-03
92	GO:0016079	synaptic vesicle exocytosis	3.21E-03
93	GO:0031109	microtubule polymerization or depolymerization	3.28E-03
94	GO:0051402	neuron apoptotic process	3.29E-03
95	GO:0043112	receptor metabolic process	3.30E-03
96	GO:0060456	positive regulation of digestive system process	3.53E-03
97	GO:0002673	regulation of acute inflammatory response	3.55E-03
98	GO:0006487	protein N-linked glycosylation	3.55E-03
99	GO:0051648	vesicle localization	3.81E-03
100	GO:1902065	response to L-glutamate	3.85E-03

## Bibliography

1. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes and Development*. 2018;32(19-20):1267-84.
2. Song P, An J, Zou M-H. Immune Clearance of Senescent Cells to Combat Ageing and Chronic Diseases. *Cells*. 2020;9(3):671.
3. McNeela EA, Mills KH. Manipulating the immune system: humoral versus cell-mediated immunity. *Advanced drug delivery reviews*. 2001;51(1-3):43-54.
4. Medzhitov R, Janeway Jr C. Innate immunity. *New England Journal of Medicine*. 2000;343(5):338-44.
5. Sathaliyawala T, Kubota M, Yudanin N, Turner D, Camp P, Thome JJ, et al. Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. *Immunity*. 2013;38(1):187-97.
6. Melchers F. Checkpoints that control B cell development. *The Journal of Clinical Investigation*. 2015;125(6):2203-10.
7. Anderson G, Takahama Y. Thymic epithelial cells: working class heroes for T cell development and repertoire selection. *Trends in Immunology*. 2012;33(6):256-63.
8. Mishra AK, Mariuzza RA. Insights into the Structural Basis of Antibody Affinity Maturation from Next-Generation Sequencing. *Frontiers in Immunology*. 2018;9(117).
9. Forthal DN. Functions of Antibodies. *Microbiology spectrum*. 2014;2(4):1-17.
10. Birnbaum ME, Berry R, Hsiao Y-S, Chen Z, Shingu-Vazquez MA, Yu X, et al. Molecular architecture of the  $\alpha\beta$  T cell receptor-CD3 complex. *Proceedings of the National Academy of Sciences*. 2014;111(49):17576-81.
11. Wieczorek M, Abualrous ET, Sticht J, Álvaro-Benito M, Stolzenberg S, Noé F, et al. Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Frontiers in Immunology*. 2017;8:292-.
12. Germain R. T-cell development and the CDT CD8 lineage decision. *Nature Reviews Immunology*. 2002;2:309-22.
13. Vrisekoop N, den Braber I, de Boer AB, Ruiters AFC, Ackermans MT, van der Crabben SN, et al. Sparse production but preferential incorporation of recently produced naïve T cells in the human peripheral pool. *Proceedings of the National Academy of Sciences*. 2008;105(16):6115.
14. Obst R. The Timing of T Cell Priming and Cycling. *Frontiers in Immunology*. 2015;6(563).
15. Mulder DJ, Pooni A, Mak N, Hurlbut DJ, Basta S, Justinich CJ. Antigen presentation and MHC class II expression by human esophageal epithelial cells: role in eosinophilic esophagitis. *American Journal of Pathology*. 2011;178(2):744-53.
16. Lanzavecchia A, Sallusto F. Regulation of T cell immunity by dendritic cells. *Cell*. 2001;106(3):263-6.

17. Bromley SK, Burack WR, Johnson KG, Somersalo K, Sims TN, Sumen C, et al. The immunological synapse. *Annual Review of Immunology*. 2001;19:375-96.
18. Varga G, Nippe N, Balkow S, Peters T, Wild MK, Seeliger S, et al. LFA-1 Contributes to Signal I of T-Cell Activation and to the Production of Th1 Cytokines. *Journal of Investigative Dermatology*. 2010;130(4):1005-12.
19. Chou C, Li MO. Tissue-Resident Lymphocytes Across Innate and Adaptive Lineages. *Frontiers in Immunology*. 2018;9(2104).
20. Sagerström CG, Kerr EM, Allison JP, Davis MM. Activation and differentiation requirements of primary T cells in vitro. *Proceedings of the National Academy of Sciences of the United States of America*. 1993;90(19):8987-91.
21. Pennock ND, White JT, Cross EW, Cheney EE, Tamburini BA, Kedl RM. T cell responses: naive to memory and everything in between. *Advances in Physiology Education*. 2013;37(4):273-83.
22. Laidlaw BJ, Craft JE, Kaech SM. The multifaceted role of CD4+ T cells in CD8+ T cell memory. *Nature Reviews Immunology*. 2016;16(2):102-11.
23. Kumar BV, Connors TJ, Farber DL. Human T Cell Development, Localization, and Function throughout Life. *Immunity*. 2018;48(2):202-13.
24. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Medical Journal*. 2007;48(1):11-23.
25. Hughes CE, Benson RA, Bedaj M, Maffia P. Antigen-Presenting Cells and Antigen Presentation in Tertiary Lymphoid Organs. *Frontiers in Immunology*. 2016;7(481).
26. Kolb-Mäurer A, Bröcker EB. The role of dendritic cells during infection. *Journal der Deutschen Dermatologischen Gesellschaft*. 2003;1(6):438-42.
27. Mayer A, Zhang Y, Perelson AS, Wingreen NS. Regulation of T cell expansion by antigen presentation dynamics. *Proceedings of the National Academy of Sciences*. 2019;116(13):5914-9.
28. Victora GD, Nussenzweig MC. Germinal Centers. *Annual Review of Immunology*. 2012;30(1):429-57.
29. Sallusto F, Lenig D, Forster R, Lipp M, Lanzavecchia A. Two subsets of memory T lymphocytes with distinct homing potentials and effector functions. *Nature*. 1999;401(6754):708-12.
30. Teijaro JR, Turner D, Pham Q, Wherry EJ, Lefrancois L, Farber DL. Cutting edge: Tissue-retentive lung memory CD4 T cells mediate optimal protection to respiratory virus infection. *Journal of Immunology*. 2011;187(11):5510-4.
31. Mackay LK, Rahimpour A, Ma JZ, Collins N, Stock AT, Hafon ML, et al. The developmental pathway for CD103(+)CD8+ tissue-resident memory T cells of skin. *Nature Immunology*. 2013;14(12):1294-301.
32. Schenkel JM, Masopust D. Tissue-resident memory T cells. *Immunity*. 2014;41(6):886-97.
33. Baliu-Piqué M, Verheij MW, Drylewicz J, Ravesloot L, de Boer RJ, Koets A, et al. Short Lifespans of Memory T-cells in Bone Marrow, Blood, and Lymph Nodes Suggest That T-cell Memory Is Maintained by Continuous Self-Renewal of Recirculating Cells. *Frontiers in Immunology*. 2018;9(2054).



34. Romagnani S. T-cell subsets (Th1 versus Th2). *Annals of Allergy, Asthma & Immunology*. 2000;85(1):9-21.
35. Lee YK, Mukasa R, Hatton RD, Weaver CT. Developmental plasticity of Th17 and Treg cells. *Current Opinion in Immunology*. 2009;21(3):274-80.
36. von Essen MR, Kongsbak M, Geisler C. Mechanisms behind functional avidity maturation in T cells. *Clinical & Developmental Immunology*. 2012;2012:163453-.
37. Sprent J, Cho J-H, Boyman O, Surh CD. T cell homeostasis. *Immunology and Cell Biology*. 2008;86(4):312-9.
38. Nutt SL, Taubenheim N, Hasbold J, Corcoran LM, Hodgkin PD. The genetic network controlling plasma cell differentiation. *Seminars in Immunology*. 2011;23(5):341-9.
39. Jaigirdar SA, MacLeod MKL. Development and Function of Protective and Pathologic Memory CD4 T Cells. *Frontiers in Immunology*. 2015;6(456).
40. Palm A-KE, Henry C. Remembrance of Things Past: Long-Term B Cell Memory After Infection and Vaccination. *Frontiers in Immunology*. 2019;10:1787.
41. Ahrends T, Spanjaard A, Pilzecker B, Bąbała N, Bovens A, Xiao Y, et al. CD4+ T Cell Help Confers a Cytotoxic T Cell Effector Program Including Coinhibitory Receptor Downregulation and Increased Tissue Invasiveness. *Immunity*. 2017;47(5):848-61.e5.
42. Zhang Z, Xiong D, Wang X, Liu H, Wang T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nature Methods*. 2021;18(1):92-9.
43. Chopp LB, Gopalan V, Ciucci T, Ruchinskas A, Rae Z, Lagarde M, et al. An Integrated Epigenomic and Transcriptomic Map of Mouse and Human  $\alpha\beta$  T Cell Development. *Immunity*. 2020;53(6):1182-201.e8.
44. Chen Y, Zander R, Khatun A, Schauder DM, Cui W. Transcriptional and Epigenetic Regulation of Effector and Memory CD8 T Cell Differentiation. *Frontiers in Immunology*. 2018;9(2826).
45. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*. 2017;18(7):437-51.
46. Primary immunodeficiency diseases. Report of an IUIS Scientific Committee. International Union of Immunological Societies. *Clinical and Experimental Immunology*. 1999;118 Suppl 1(Suppl 1):1-28.
47. da Glória VG, Martins de Araújo M, Mafalda Santos A, Leal R, de Almeida SF, Carmo AM, et al. T Cell Activation Regulates CD6 Alternative Splicing by Transcription Dynamics and SRSF1. *The Journal of Immunology*. 2014;193(1):391-9.
48. Fan J, Hu J, Xue C, Zhang H, Susztak K, Reilly MP, et al. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *Plos Genetics*. 2020;16(5):e1008786-e.
49. Gutierrez-Arcelus M, Baglaenko Y, Arora J, Hannes S, Luo Y, Amariuta T, et al. Allele-specific expression changes dynamically during T cell activation in HLA and other autoimmune loci. *Nature Genetics*. 2020;52(3):247-53.

50. Szabo PA, Levitin HM, Miron M, Snyder ME, Senda T, Yuan J, et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications*. 2019;10(1):4706.
51. Schaub A, Glasmacher E. Splicing in immune cells-mechanistic insights and emerging topics. *International Immunology*. 2017;29(4):173-81.
52. Fedor MJ. Alternative Splicing Minireview Series: Combinatorial Control Facilitates Splicing Regulation of Gene Expression and Enhances Genome Diversity. *Journal of Biological Chemistry*. 2008;283(3):1209-10.
53. Ward AJ, Cooper TA. The pathobiology of splicing. *The Journal of pathology*. 2010;220(2):152-63.
54. Lei Q, Li C, Zuo Z, Huang C, Cheng H, Zhou R. Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biology and Evolution*. 2016;8(3):562-77.
55. Eshel D, Toporik A, Efrati T, Nakav S, Chen A, Douvdevani A. Characterization of natural human antagonistic soluble CD40 isoforms produced through alternative splicing. *Molecular Immunology*. 2008;46(2):250-7.
56. Verma B, Akinyi MV, Norppa AJ, Frilander MJ. Minor spliceosome and disease. *Seminars in Cell and Developmental Biology*. 2018;79:103-12.
57. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*. 2008;40(12):1413-5.
58. Orvain C, Matre V, Gabrielsen OS. The transcription factor c-Myb affects pre-mRNA splicing. *Biochemical and Biophysical Research Communications*. 2008;372(2):309-13.
59. Heyd F, ten Dam G, Moroy T. Auxiliary splice factor U2AF26 and transcription factor Gfi1 cooperate directly in regulating CD45 alternative splicing. *Nature Immunology*. 2006;7(8):859-67.
60. Alkhatib A, Werner M, Hug E, Herzog S, Eschbach C, Faraidun H, et al. FoxO1 induces Ikaros splicing to promote immunoglobulin gene recombination. *Journal of Experimental Medicine*. 2012;209(2):395-406.
61. Ergun A, Doran G, Costello JC, Paik HH, Collins JJ, Mathis D, et al. Differential splicing across immune system lineages. *Proceedings of the National Academy of Sciences of the United States of America*. 2013;110(35):14324-9.
62. Reynaud D, Demarco IA, Reddy KL, Schjerven H, Bertolino E, Chen Z, et al. Regulation of B cell fate commitment and immunoglobulin heavy-chain gene rearrangements by Ikaros. *Nature Immunology*. 2008;9(8):927-36.
63. Zikherman J, Weiss A. Alternative splicing of CD45: the tip of the iceberg. *Immunity*. 2008;29(6):839-41.
64. Bentley DL. Coupling mRNA processing with transcription in time and space. *Nature reviews Genetics*. 2014;15(3):163-75.
65. Bonilla FA, Oettgen HC. Adaptive immunity. *Journal of Allergy and Clinical Immunology*. 2010;125(2 Suppl 2):S33-40.

66. Oeckinghaus A, Wegener E, Welteke V, Ferch U, Arslan SC, Ruland J, et al. Malt1 ubiquitination triggers NF-kappaB signaling upon T-cell activation. *EMBO Journal*. 2007;26(22):4634-45.
67. Meininger I, Griesbach RA, Hu D, Gehring T, Seeholzer T, Bertossi A, et al. Alternative splicing of MALT1 controls signalling and activation of CD4(+) T cells. *Nature Communications*. 2016;7:11292-.
68. Matlin AJ, Clark F, Smith CW. Understanding alternative splicing: towards a cellular code. *Nature Reviews: Molecular Cell Biology*. 2005;6(5):386-98.
69. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. *Genome Biology*. 2004;5(10):R74.
70. Wang K, Wu D, Zhang H, Das A, Basu M, Malin J, et al. Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases. *Scientific Reports*. 2018;8(1):10929.
71. Lynch KW. Consequences of regulated pre-mRNA splicing in the immune system. *Nature Reviews: Immunology*. 2004;4(12):931-40.
72. Lynch KW, Weiss A. A model system for activation-induced alternative splicing of CD45 pre-mRNA in T cells implicates protein kinase C and Ras. *Molecular and Cellular Biology*. 2000;20(1):70-80.
73. Trowbridge IS, Thomas ML. CD45: an emerging role as a protein tyrosine phosphatase required for lymphocyte activation and development. *Annual Review of Immunology*. 1994;12:85-116.
74. Ip JY, Tong A, Pan Q, Topp JD, Blencowe BJ, Lynch KW. Global analysis of alternative splicing during T-cell activation. *RNA (New York, NY)*. 2007;13(4):563-72.
75. Oaks MK, Hallett KM, Penwell RT, Stauber EC, Warren SJ, Tector AJ. A native soluble form of CTLA-4. *Cellular Immunology*. 2000;201(2):144-53.
76. Magistrelli G, Jeannin P, Herbault N, Benoit De Coignac A, Gauchat JF, Bonnefoy JY, et al. A soluble form of CTLA-4 generated by alternative splicing is expressed by nonstimulated human T cells. *European Journal of Immunology*. 1999;29(11):3596-602.
77. Candotti F, Notarangelo L, Visconti R, O'Shea J. Molecular aspects of primary immunodeficiencies: lessons from cytokine and other signaling pathways. *The Journal of Clinical Investigation*. 2002;109(10):1261-9.
78. Tangye SG, Al-Herz W, Bousfiha A, Chatila T, Cunningham-Rundles C, Etzioni A, et al. Human Inborn Errors of Immunity: 2019 Update on the Classification from the International Union of Immunological Societies Expert Committee. *Journal of Clinical Immunology*. 2020;40(1):24-64.
79. McCusker C, Upton J, Warrington R. Primary immunodeficiency. *Allergy, Asthma & Clinical Immunology*. 2018;14(2):61.
80. Bousfiha A, Jeddane L, Picard C, Ailal F, Bobby Gaspar H, Al-Herz W, et al. The 2017 IUIS Phenotypic Classification for Primary Immunodeficiencies. *Journal of Clinical Immunology*. 2018;38(1):129-43.

81. Picard C, Bobby Gaspar H, Al-Herz W, Bousfiha A, Casanova J-L, Chatila T, et al. International Union of Immunological Societies: 2017 Primary Immunodeficiency Diseases Committee Report on Inborn Errors of Immunity. *Journal of Clinical Immunology*. 2018;38(1):96-128.
82. Shillitoe B, Bangs C, Guzman D, Gennery AR, Longhurst HJ, Slatter M, et al. The United Kingdom Primary Immune Deficiency (UKPID) registry 2012 to 2017. *Clinical and Experimental Immunology*. 2018;192(3):284-91.
83. Al-Herz W, Chou J, Delmonte OM, Massaad MJ, Bainter W, Castagnoli R, et al. Comprehensive Genetic Results for Primary Immunodeficiency Disorders in a Highly Consanguineous Population. *Frontiers in Immunology*. 2019;9(3146).
84. Lagresle-Peyrou C, Luce S, Ouchani F, Soheili TS, Sadek H, Chouteau M, et al. X-linked primary immunodeficiency associated with hemizygous mutations in the moesin (MSN) gene. *Journal of Allergy and Clinical Immunology*. 2016;138(6):1681-9.e8.
85. McCusker C, Warrington R. Primary immunodeficiency. *Allergy, asthma, and clinical immunology : official journal of the Canadian Society of Allergy and Clinical Immunology*. 2011;7 Suppl 1(Suppl 1):S11-S.
86. Boyle JM, Buckley RH. Population Prevalence of Diagnosed Primary Immunodeficiency Diseases in the United States. *Journal of Clinical Immunology*. 2007;27(5):497-502.
87. Yazdani R, Azizi G, Abolhassani H, Aghamohammadi A. Selective IgA Deficiency: Epidemiology, Pathogenesis, Clinical Phenotype, Diagnosis, Prognosis and Management. *Scandinavian Journal of Immunology*. 2017;85(1):3-12.
88. Lewkonia RM, Gairdner D, Doe WF. IgA deficiency in one of identical twins. *British Medical Journal*. 1976;1(6005):311-3.
89. Zhang J, van Oostrom D, Li J, Savelkoul HF. Innate mechanisms in selective IgA deficiency. *Frontiers in Immunology*. 2021;12:649112.
90. Gleeson M. Mucosal immunity and respiratory illness in elite athletes. *International Journal of Sports Medicine*. 2000;21(Sup. 1):33-43.
91. Oen K, Petty RE, Schroeder ML. Immunoglobulin A deficiency: genetic studies. *Tissue Antigens*. 1982;19(3):174-82.
92. Ferreira RC, Pan-Hammarström Q, Graham RR, Gateva V, Fontán G, Lee AT, et al. Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nature Genetics*. 2010;42(9):777-80.
93. Vořechovský I, Cullen M, Carrington M, Hammarström L, Webster ADB. Fine Mapping of <em>IGAD1</em> in IgA Deficiency and Common Variable Immunodeficiency: Identification and Characterization of Haplotypes Shared by Affected Members of 101 Multiple-Case Families. *The Journal of Immunology*. 2000;164(8):4408.
94. Yel L. Selective IgA Deficiency. *Journal of Clinical Immunology*. 2010;30(1):10-6.
95. Europe PIPDDfOCi. European Reference Paper. [worldpiweek.org](http://worldpiweek.org).
96. Edgar JD. T cell immunodeficiency. *Journal of Clinical Pathology*. 2008;61(9):988-93.

97. Eades-Perner A-M, Gathmann B, Knerr V, Guzman D, Veit D, Kindle G, et al. The European internet-based patient and research database for primary immunodeficiencies: results 2004–06. *Clinical and Experimental Immunology*. 2007;147(2):306-12.
98. Conley ME, Notarangelo LD, Etzioni A. Diagnostic criteria for primary immunodeficiencies. Representing PAGID (Pan-American Group for Immunodeficiency) and ESID (European Society for Immunodeficiencies). *Clinical Immunology*. 1999;93(3):190-7.
99. Geha RS, Notarangelo LD, Casanova JL, Chapel H, Conley ME, Fischer A, et al. Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. *Journal of Allergy and Clinical Immunology*. 2007;120(4):776-94.
100. Gaspar HB, Gilmour KC, Jones AM. Severe combined immunodeficiency—molecular pathogenesis and diagnosis. *Archives of Disease in Childhood*. 2001;84(2):169-73.
101. Waickman AT, Park JY, Park JH. The common  $\gamma$ -chain cytokine receptor: tricks-and-treats for T cells. *Cellular and Molecular Life Sciences*. 2016;73(2):253-69.
102. Cheng G, Yu A, Dee MJ, Malek TR. IL-2R Signaling Is Essential for Functional Maturation of Regulatory T Cells during Thymic Development. *The Journal of Immunology*. 2013;190(4):1567-75.
103. Roifman CM. 35 - Primary T-Cell Immunodeficiencies. In: Rich RR, Fleisher TA, Shearer WT, Schroeder HW, Frew AJ, Weyand CM, editors. *Clinical Immunology (Fifth Edition)*. London: Content Repository Only!; 2019. p. 489-508.e1.
104. Puck JM, Pepper AE, Henthorn PS, Candotti F, Isakov J, Whitwam T, et al. Mutation analysis of IL2RG in human X-linked severe combined immunodeficiency. *Blood*. 1997;89(6):1968-77.
105. Noguchi M, Yi H, Rosenblatt HM, Filipovich AH, Adelstein S, Modi WS, et al. Interleukin-2 receptor  $\gamma$  chain mutation results in X-linked severe combined immunodeficiency in humans. *Cell*. 1993;73(1):147-57.
106. Macchi P, Villa A, Giliani S, Sacco MG, Frattini A, Porta F, et al. Mutations of Jak-3 gene in patients with autosomal severe combined immune deficiency (SCID). *Nature*. 1995;377(6544):65-8.
107. Zhong L, Wang W, Ma M, Gou L, Tang X, Song H. Chronic active Epstein-Barr virus infection as the initial symptom in a Janus kinase 3 deficiency child: Case report and literature review. *Medicine*. 2017;96(42):e7989-e.
108. Aiuti A, Cattaneo F, Galimberti S, Benninghoff U, Cassani B, Callegaro L, et al. Gene Therapy for Immunodeficiency Due to Adenosine Deaminase Deficiency. *New England Journal of Medicine*. 2009;360(5):447-58.
109. Whitmore KV, Gaspar HB. Adenosine Deaminase Deficiency – More Than Just an Immunodeficiency. *Frontiers in Immunology*. 2016;7(314).
110. Tasher D, Dalal I. The genetic basis of severe combined immunodeficiency and its variants. The application of clinical genetics. 2012;5:67-80.
111. Kumrah R, Vignesh P, Patra P, Singh A, Anjani G, Saini P, et al. Genetics of severe combined immunodeficiency. *Genes & Diseases*. 2020;7(1):52-61.

112. Ijspeert H, Lankester AC, van den Berg JM, Wiegant W, van Zelm MC, Weemaes CMR, et al. Artemis splice defects cause atypical SCID and can be restored in vitro by an antisense oligonucleotide. *Genes and Immunity*. 2011;12(6):434-44.
113. Nelson DL, Terhorst C. X-linked lymphoproliferative syndrome. *Clinical and Experimental Immunology*. 2000;122(3):291-5.
114. Booth C, Gilmour KC, Veys P, Gennery AR, Slatter MA, Chapel H, et al. X-linked lymphoproliferative disease due to SAP/SH2D1A deficiency: a multicenter study on the manifestations, management and outcome of the disease. *Blood*. 2011;117(1):53-62.
115. Lyu X, Guo Z, Li Y, Fan R, Song Y. Identification of a novel nonsense mutation in SH2D1A in a patient with X-linked lymphoproliferative syndrome type 1: a case report. *BMC Medical Genetics*. 2018;19(1):60.
116. Crotty S, Kersh EN, Cannons J, Schwartzberg PL, Ahmed R. SAP is required for generating long-term humoral immunity. *Nature*. 2003;421(6920):282-7.
117. Cannons JL, Yu LJ, Jankovic D, Crotty S, Horai R, Kirby M, et al. SAP regulates T cell-mediated help for humoral immunity by a mechanism distinct from cytokine regulation. *Journal of Experimental Medicine*. 2006;203(6):1551-65.
118. Panchal N, Booth C, Cannons JL, Schwartzberg PL. X-Linked Lymphoproliferative Disease Type 1: A Clinical and Molecular Perspective. *Frontiers in Immunology*. 2018;9(666).
119. Stavnezer J, Guikema JEJ, Schrader CE. Mechanism and regulation of class switch recombination. *Annual Review of Immunology*. 2008;26:261-92.
120. Notarangelo LD, Hayward AR. X-linked immunodeficiency with hyper-IgM (XHIM). *Clinical and Experimental Immunology*. 2000;120(3):399-405.
121. Fuleihan R, Ramesh N, Geha RS. Role of CD40-CD40-ligand interaction in Ig-isotype switching. *Current Opinion in Immunology*. 1993;5(6):963-7.
122. Davies EG. Immunodeficiency in DiGeorge Syndrome and Options for Treating Cases with Complete Athymia. *Frontiers in Immunology*. 2013;4:322-.
123. Gill J, Malin M, Sutherland J, Gray D, Hollander G, Boyd R. Thymic generation and regeneration. *Immunological Reviews*. 2003;195:28-50.
124. Poliani PL, Facchetti F, Ravanini M, Gennery AR, Villa A, Roifman CM, et al. Early defects in human T-cell development severely affect distribution and maturation of thymic stromal cells: possible implications for the pathophysiology of Omenn syndrome. *Blood*. 2009;114(1):105-8.
125. Nola M, Dotlić S. Chapter 9 - The Hematopoietic and Lymphoid Systems. In: Damjanov I, editor. *Pathology Secrets (Third Edition)*. Philadelphia: Mosby; 2009. p. 161-202.
126. Yutzey KE. DiGeorge syndrome, Tbx1, and retinoic acid signaling come full circle. *Circulation Research*. 2010;106(4):630-2.
127. Taylor AMR, Byrd PJ. Molecular pathology of ataxia telangiectasia. *Journal of Clinical Pathology*. 2005;58(10):1009.
128. Bakkenist CJ, Kastan MB. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature*. 2003;421(6922):499-506.

129. Shiloh Y. ATM and related protein kinases: safeguarding genome integrity. *Nature Reviews Cancer*. 2003;3(3):155-68.
130. Gilad S, Khosravi R, Shkedy D, Uziel T, Ziv Y, Savitsky K, et al. Predominance of Null Mutations in Ataxia-Telangiectasia. *Human Molecular Genetics*. 1996;5(4):433-9.
131. Pagani F, Buratti E, Stuani C, Bendix R, Dörk T, Baralle FE. A new type of mutation causes a splicing defect in ATM. *Nature Genetics*. 2002;30(4):426-9.
132. Brechtmann F, Mertes C, Matusevičiūtė A, Yépez VA, Avsec Ž, Herzog M, et al. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *American Journal of Human Genetics*. 2018;103(6):907-17.
133. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare variation on gene expression across tissues. *bioRxiv*. 2016:074443.
134. Thormann A, Halachev M, McLaren W, Moore DJ, Svinti V, Campbell A, et al. Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nature Communications*. 2019;10(1):2373-.
135. Fodil N, Langlais D, Gros P. Primary Immunodeficiencies and Inflammatory Disease: A Growing Genetic Intersection. *Trends in Immunology*. 2016;37(2):126-40.
136. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *American Journal of Human Genetics*. 2012;90(1):7-24.
137. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*. 2012;30(11):1095-106.
138. Arand J, Wossidlo M, Lepikhov K, Peat JR, Reik W, Walter J. Selective impairment of methylation maintenance is the major cause of DNA methylation reprogramming in the early embryo. *Epigenetics Chromatin*. 2015;8(1):1.
139. Fagny M, Paulson JN, Kuijjer ML, Sonawane AR, Chen C-Y, Lopes-Ramos CM, et al. Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences*. 2017;114(37):E7841-E50.
140. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*. 2014;343(6175):1246949.
141. Casamassimi A, Federico A, Rienzo M, Esposito S, Ciccodicola A. Transcriptome Profiling in Human Diseases: New Advances and Perspectives. *International Journal of Molecular Sciences*. 2017;18(8):1652.
142. Piasecka B, Duffy D, Urrutia A, Quach H, Patin E, Posseme C, et al. Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proceedings of the National Academy of Sciences of the United States of America*. 2018;115(3):E488-e97.
143. Khokhar W, Hassan MA, Reddy ASN, Chaudhary S, Jabre I, Byrne LJ, et al. Genome-Wide Identification of Splicing Quantitative Trait Loci (sQTLs) in Diverse Ecotypes of *Arabidopsis thaliana*. *Frontiers in Plant Science*. 2019;10(1160).

144. Romo L, Ashar-Patel A, Pfister E, Aronin N. Alterations in mRNA 3' UTR Isoform Abundance Accompany Gene Expression Changes in Human Huntington's Disease Brains. *Cell Reports*. 2017;20(13):3057-70.
145. Fernández-Nogales M, Lucas JJ. Altered Levels and Isoforms of Tau and Nuclear Membrane Invaginations in Huntington's Disease. *Frontiers in Cellular Neuroscience*. 2020;13(574).
146. Thijssen-Timmer DC, Schiphorst MP, Kwakkel J, Emter R, Kralli A, Wiersinga WM, et al. PGC-1alpha regulates the isoform mRNA ratio of the alternatively spliced thyroid hormone receptor alpha transcript. *Journal of Molecular Endocrinology*. 2006;37(2):251-7.
147. Gamundi MJ, Hernan I, Muntanyola M, Maseras M, López-Romero P, Álvarez R, et al. Transcriptional expression of cis-acting and trans-acting splicing mutations cause autosomal dominant retinitis pigmentosa. *Human Mutation*. 2008;29(6):869-78.
148. Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, et al. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. *Human Mutation*. 2007;28(2):150-8.
149. Krawczak M, Reiss J, Cooper DN. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Human Genetics*. 1992;90(1-2):41-54.
150. Grodecká L, Hujová P, Kramárek M, Kršjaková T, Kováčová T, Vondrášková K, et al. Systematic analysis of splicing defects in selected primary immunodeficiencies-related genes. *Clinical Immunology*. 2017;180:33-44.
151. Ohno K, Takeda JI, Masuda A. Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley interdisciplinary reviews RNA*. 2018;9(1).
152. Wai HA, Lord J, Lyon M, Gunning A, Kelly H, Cibin P, et al. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genetics in Medicine*. 2020;22(6):1005-14.
153. Wang Y, Liu J, Huang BO, Xu Y-M, Li J, Huang L-F, et al. Mechanism of alternative splicing and its regulation. *Biomedical Reports*. 2015;3(2):152-8.
154. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*. 2002;3(4):285-98.
155. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Science Translational Medicine*. 2017;9(386):eaal5209.
156. Platt CD, Massaad MJ, Cangemi B, Schmidt B, Aldhekri H, Geha RS. Janus kinase 3 deficiency caused by a homozygous synonymous exonic mutation that creates a dominant splice site. *The Journal of allergy and clinical immunology*. 2017;140(1):268-71.e6.
157. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*. 2019;176(3):535-48.e24.
158. Jia C, Hu Y, Liu Y, Li M. Mapping Splicing Quantitative Trait Loci in RNA-Seq. *Cancer Informatics*. 2015;14(Suppl 1):45-53.



159. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*. 2017;14(4):417-9.
160. Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications*. 2017;8:14519.
161. Punwani D, Wang H, Chan AY, Cowan MJ, Mallott J, Sunderam U, et al. Combined immunodeficiency due to MALT1 mutations, treated by hematopoietic cell transplantation. *Journal of Clinical Immunology*. 2015;35(2):135-46.
162. Scacheri CA, Scacheri PC. Mutations in the noncoding genome. *Current Opinion in Pediatrics*. 2015;27(6):659-64.
163. Turunen JJ, Niemela EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome. *Wiley interdisciplinary reviews RNA*. 2013;4(1):61-76.
164. Heremans J, Garcia-Perez JE, Turro E, Schlenner SM, Casteels I, Collin R, et al. Abnormal differentiation of B cells and megakaryocytes in patients with Roifman syndrome. *Journal of Allergy and Clinical Immunology*. 2018;142(2):630-46.
165. Merico D, Roifman M, Braunschweig U, Yuen RK, Alexandrova R, Bates A, et al. Compound heterozygous mutations in the noncoding RNU4ATAC cause Roifman Syndrome by disrupting minor intron splicing. *Nature communications*. 2015;6:8718.
166. Abolhassani H, Naseri A, Rezaei N, Aghamohammadi A. Economic burden of common variable immunodeficiency: annual cost of disease AU - Sadeghi, Bamdad. *Expert Review of Clinical Immunology*. 2015;11(5):681-8.
167. Condino-Neto A, Espinosa-Rosales FJ. Changing the Lives of People With Primary Immunodeficiencies (PI) With Early Testing and Diagnosis. *Frontiers in Immunology*. 2018;9(1439).
168. UK P. Patients' experience survey of Primary Immunodeficiency Disorders services. [www.piduk.org](http://www.piduk.org); 2016 September 2016.
169. Walter JE, Farmer JR, Foldvari Z, Torgerson TR, Cooper MA. Mechanism-Based Strategies for the Management of Autoimmunity and Immune Dysregulation in Primary Immunodeficiencies. *The journal of allergy and clinical immunology In practice*. 2016;4(6):1089-100.
170. Heimall J, Keller M, Saltzman R, Bunin N, McDonald-McGinn D, Zakai E, et al. Diagnosis of 22q11.2 deletion syndrome and artemis deficiency in two children with T-B-NK+ immunodeficiency. *Journal of Clinical Immunology*. 2012;32(5):1141-4.
171. Bonilla FA, Khan DA, Ballas ZK, Chinen J, Frank MM, Hsu JT, et al. Practice parameter for the diagnosis and management of primary immunodeficiency. *Journal of Allergy and Clinical Immunology*. 2015;136(5):1186-205.e1-78.
172. Lenardo M, Lo B, Lucas CL. Genomics of Immune Diseases and New Therapies. *Annual Review of Immunology*. 2016;34:121-49.
173. Ramakrishnan KA, Pengelly RJ, Gao Y, Morgan M, Patel SV, Davies EG, et al. Precision Molecular Diagnosis Defines Specific Therapy in Combined Immunodeficiency with Megaloblastic Anemia Secondary to MTHFD1 Deficiency. *The journal of allergy and clinical immunology In practice*. 2016;4(6):1160-6.e10.

174. Fischer A. Primary T-cell immunodeficiencies. *Current Opinion in Immunology*. 1993;5(4):569-78.
175. Meyts I, Bosch B, Bolze A, Boisson B, Itan Y, Belkadi A, et al. Exome and genome sequencing for inborn errors of immunity. *The Journal of allergy and clinical immunology*. 2016;138(4):957-69.
176. Richardson AM, Moyer AM, Hasadsri L, Abraham RS. Diagnostic Tools for Inborn Errors of Human Immunity (Primary Immunodeficiencies and Immune Dysregulatory Diseases). *Current Allergy and Asthma Reports*. 2018;18(3):19.
177. Hernandez-Trujillo V, Ballow M. Diagnosing primary immunodeficiency: a practical approach for the non-immunologist AU - Lehman, Heather. *Current Medical Research and Opinion*. 2015;31(4):697-706.
178. Rudilla F, Franco-Jarava C, Martínez-Gallo M, Garcia-Prat M, Martín-Nalda A, Rivière J, et al. Expanding the Clinical and Genetic Spectra of Primary Immunodeficiency-Related Disorders With Clinical Exome Sequencing: Expected and Unexpected Findings. *Frontiers in Immunology*. 2019;10:2325.
179. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics*. 2017;100(5):695-705.
180. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Unlocking Mendelian disease using exome sequencing. *Genome Biology*. 2011;12(9):228.
181. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*. 2011;12:745.
182. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? *Journal of Medical Genetics*. 2011;48(9):580-9.
183. Kremer LS, Wortmann SB, Prokisch H. "Transcriptomics": molecular diagnosis of inborn errors of metabolism via RNA-sequencing. *Journal of Inherited Metabolic Disease*. 2018;41(3):525-32.
184. Heimall J. Now Is the Time to Use Molecular Gene Testing for the Diagnosis of Primary Immune Deficiencies. *The journal of allergy and clinical immunology In practice*. 2019.
185. Meyts I, Bosch B, Bolze A, Boisson B, Itan Y, Belkadi A, et al. Exome and genome sequencing for inborn errors of immunity. *Journal of Allergy and Clinical Immunology*. 2016;138(4):957-69.
186. Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nature Genetics*. 2015;47(7):717-26.
187. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine*. 2013;369(16):1502-11.
188. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*. 2014;312(18):1870-9.

189. Stray-Pedersen A, Sorte HS, Samarakoon P, Gambin T, Chinn IK, Coban Akdemir ZH, et al. Primary immunodeficiency diseases: Genomic approaches delineate heterogeneous Mendelian disorders. *The Journal of allergy and clinical immunology*. 2017;139(1):232-45.
190. Moens LN, Falk-Sörqvist E, Asplund AC, Bernatowska E, Smith CIE, Nilsson M. Diagnostics of primary immunodeficiency diseases: a sequencing capture approach. *PLoS One*. 2014;9(12):e114901-e.
191. Yska HAF, Elsink K, Kuijpers TW, Frederix GWJ, van Gijn ME, van Montfrans JM. Diagnostic Yield of Next Generation Sequencing in Genetically Undiagnosed Patients with Primary Immunodeficiencies: a Systematic Review. *Journal of Clinical Immunology*. 2019;39(6):577-91.
192. Philippidis A. The 100,000 Genomes club [www.genenews.com](http://www.genenews.com) 2018 [Available from: <https://www.genengnews.com/insights/the-100000-genomes-club/>]. Accessed on 25/07/2019
193. Rae W, Ward D, Mattocks C, Pengelly RJ, Eren E, Patel SV, et al. Clinical efficacy of a next-generation sequencing gene panel for primary immunodeficiency diagnostics. *Clinical Genetics*. 2018;93(3):647-55.
194. Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, et al. Cis and Trans Effects of Human Genomic Variants on Gene Expression. *Plos Genetics*. 2014;10(7):e1004461.
195. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*. 2016;17:53-.
196. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS Computational Biology*. 2017;13(5):e1005457-e.
197. Wirka RC, Pjanic M, Quertermous T. Advances in Transcriptomics: Investigating Cardiovascular Disease at Unprecedented Resolution. *Circulation Research*. 2018;122(9):1200-20.
198. Feng Y, Zhang Y, Ying C, Wang D, Du C. Nanopore-based Fourth-generation DNA Sequencing Technology. *Genomics, Proteomics & Bioinformatics*. 2015;13(1):4-16.
199. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*. 2015;13(5):278-89.
200. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harbor protocols*. 2015;2015(11):951-69.
201. Neums L, Suenaga S, Beyerlein P, Anders S, Koestler D, Mariani A, et al. VaDiR: an integrated approach to Variant Detection in RNA. *GigaScience*. 2017;7(2).
202. Zeng Y, Wang G, Yang E, Ji G, Brinkmeyer-Langford CL, Cai JJ. Aberrant Gene Expression in Humans. *Plos Genetics*. 2015;11(1):e1004942.
203. Zhao J, Akinsanmi I, Arafat D, Cradick TJ, Lee CM, Banskota S, et al. A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *American Journal of Human Genetics*. 2016;98(2):299-309.
204. Takeda N, O'Dea EL, Doedens A, Kim J-w, Weidemann A, Stockmann C, et al. Differential activation and antagonistic function of HIF- $\alpha$  isoforms in macrophages are essential for NO homeostasis. *Genes and Development*. 2010;24(5):491-501.

205. Chen S, Townsend K, Goldberg TE, Davies P, Conejero-Goldberg C. MAPT isoforms: differential transcriptional profiles related to 3R and 4R splice variants. *Journal of Alzheimer's Disease*. 2010;22(4):1313-29.
206. Kim HK, Pham MHC, Ko KS, Rhee BD, Han J. Alternative splicing isoforms in health and disease. *Pflügers Archiv - European Journal of Physiology*. 2018;470(7):995-1016.
207. DiStefano JK. The Emerging Role of Long Noncoding RNAs in Human Disease. *Methods in Molecular Biology*. 2018;1706:91-110.
208. Kramer NJ, Wang W-L, Reyes EY, Kumar B, Chen C-C, Ramakrishna C, et al. Altered lymphopoiesis and immunodeficiency in *miR-142* null mice. *Blood*. 2015;125(24):3720.
209. Chitnis N, Clark PM, Kamoun M, Stolle C, Brad Johnson F, Monos DS. An Expanded Role for HLA Genes: HLA-B Encodes a microRNA that Regulates IgA and Other Immune Response Transcripts. *Frontiers in Immunology*. 2017;8(583).
210. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *American Journal of Human Genetics*. 2019;104(3):466-83.
211. Han H, Jiang X. Disease Biomarker Query from RNA-Seq Data. *Cancer Informatics*. 2014;13(Suppl 1):81-94.
212. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine*. 2018;50(8):96.
213. See P, Lum J, Chen J, Ginhoux F. A Single-Cell Sequencing Guide for Immunologists. *Frontiers in Immunology*. 2018;9:2425-.
214. Martkamchan S, Onlamoon N, Wang S, Pattanapanyasat K, Ammaranond P. The Effects of Anti-CD3/CD28 Coated Beads and IL-2 on Expanded T Cell for Immunotherapy. *Advances in Clinical Experimental Medicine*. 2016;25(5):821-8.
215. Corley SM, MacKenzie KL, Beverdam A, Roddam LF, Wilkins MR. Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols. *BMC Genomics*. 2017;18(1):399-.
216. Shin H, Shannon CP, Fishbane N, Ruan J, Zhou M, Balshaw R, et al. Variation in RNA-Seq transcriptome profiles of peripheral whole blood from healthy individuals with and without globin depletion. *PloS One*. 2014;9(3):e91041-e.
217. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports*. 2018;8(1):4781.
218. Kumar A, Kankainen M, Parsons A, Kallioniemi O, Mattila P, Heckman CA. The impact of RNA sequence library construction protocols on transcriptomic profiling of leukemia. *BMC Genomics*. 2017;18(1):629.
219. Aluri J, Gupta MR, Dalvi A, Mhatre S, Kulkarni M, Desai M, et al. Lymphopenia and Severe Combined Immunodeficiency (SCID) - Think Before You Ink. *Indian Journal of Pediatrics*. 2019;86(7):584-9.

220. Fresard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nature Medicine*. 2019;25(6):911-9.
221. Gonorazky H, Liang M, Cummings B, Lek M, Micallef J, Hawkins C, et al. RNAseq analysis for the diagnosis of muscular dystrophy. *Annals of clinical and translational neurology*. 2016;3(1):55-60.
222. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Frontiers in Endocrinology*. 2018;9(402).
223. Šponer J, Bussi G, Krepl M, Banáš P, Bottaro S, Cunha RA, et al. RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview. *Chemical Reviews*. 2018;118(8):4177-338.
224. Ryazansky S, Radion E, Mironova A, Akulenko N, Abramov Y, Morgunova V, et al. Natural variation of piRNA expression affects immunity to transposable elements. *Plos Genetics*. 2017;13(4):e1006731-e.
225. Yizhak K, Aguet F, Kim J, Hess JM, Kübler K, Grimsby J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*. 2019;364(6444):eaaw0726.
226. Blomberg KE, Smith CI, Lindvall JM. Gene expression signatures in primary immunodeficiencies: the experience from human disease and mouse models. *Current Molecular Medicine*. 2007;7(6):555-66.
227. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*. 2015;31(2):166-9.
228. Zhang Z, Pan Z, Ying Y, Xie Z, Adhikari S, Phillips J, et al. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nature Methods*. 2019;16(4):307-10.
229. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Research*. 2012;22(10):2008-17.
230. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 2010;28(5):511-5.
231. Fresard L, Smail C, Smith KS, Ferraro NM, Teran NA, Kernohan KD, et al. Identification of rare-disease genes in diverse undiagnosed cases using whole blood transcriptome sequencing and large control cohorts. *bioRxiv*. 2018.
232. Brodin P, Jojic V, Gao T, Bhattacharya S, Angel CJ, Furman D, et al. Variation in the human immune system is largely driven by non-heritable influences. *Cell*. 2015;160(1-2):37-47.
233. Duffy D, Rouilly V, Braudeau C, Corbière V, Djebali R, Ungeheuer MN, et al. Standardized whole blood stimulation improves immunomonitoring of induced immune responses in multi-center study. *Clinical Immunology*. 2017;183:325-35.
234. Lee MN, Ye C, Villani A-C, Raj T, Li W, Eisenhaure TM, et al. Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. *Science*. 2014;343(6175):1246980.
235. Rotival M, Quach H, Quintana-Murci L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nature communications*. 2019;10(1):1671.

236. Urrutia A, Duffy D, Rouilly V, Posseme C, Djebali R, Illanes G, et al. Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses. *Cell Reports*. 2016;16(10):2777-91.
237. Khan S, Kuruvilla M, Hagin D, Wakeland B, Liang C, Vishwanathan K, et al. RNA sequencing reveals the consequences of a novel insertion in dedicator of cytokinesis-8. *Journal of Allergy and Clinical Immunology*. 2016;138(1):289-92.e6.
238. Hsu AP, Johnson KD, Falcone EL, Sanalkumar R, Sanchez L, Hickstein DD, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood*. 2013;121(19):3830-7, s1-7.
239. Starokadomskyy P, Gemelli T, Rios JJ, Xing C, Wang RC, Li H, et al. DNA polymerase- $\alpha$  regulates the activation of type I interferons through cytosolic RNA:DNA synthesis. *Nature Immunology*. 2016;17(5):495-504.
240. Chintapalli VR, Wang J, Herzyk P, Davies SA, Dow JAT. Data-mining the FlyAtlas online resource to identify core functional motifs across transporting epithelia. *BMC Genomics*. 2013;14:518-.
241. Weinstein JN. Searching for pharmacogenomic markers: the synergy between omic and hypothesis-driven research. *Disease Markers*. 2001;17(2):77-88.
242. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. *Proceedings of the National Academy of Sciences*. 2019;116(13):6491.
243. Fei T, Yu T. Batch Effect Correction of RNA-seq Data through Sample Distance Matrix Adjustment. *bioRxiv*. 2019:669739.
244. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. *The Journal of molecular diagnostics : JMD*. 2003;5(2):73-81.
245. Chinen J, Shearer WT. Secondary immunodeficiencies, including HIV infection. *The Journal of allergy and clinical immunology*. 2010;125(2 Suppl 2):S195-S203.
246. Montecino-Rodriguez E, Berent-Maoz B, Dorshkind K. Causes, consequences, and reversal of immune system aging. *The Journal of Clinical Investigation*. 2013;123(3):958-65.
247. Weng N-P. Aging of the immune system: how much can the adaptive immune system adapt? *Immunity*. 2006;24(5):495-9.
248. Aiello A, Farzaneh F, Candore G, Caruso C, Davinelli S, Gambino CM, et al. Immunosenescence and Its Hallmarks: How to Oppose Aging Strategically? A Review of Potential Options for Therapeutic Intervention. *Frontiers in Immunology*. 2019;10(2247).
249. Nations U. *World Population Ageing 2019 Highlights*. New York; 2019.
250. Chung HY, Kim DH, Lee EK, Chung KW, Chung S, Lee B, et al. Redefining Chronic Inflammation in Aging and Age-Related Diseases: Proposal of the Senoinflammation Concept. *Aging and Disease*. 2019;10(2):367-82.
251. Hou Y, Dan X, Babbar M, Wei Y, Hasselbalch SG, Croteau DL, et al. Ageing as a risk factor for neurodegenerative disease. *Nature Reviews Neurology*. 2019;15(10):565-81.
252. Anan JR, Cho WC, Sørreide K. The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging and Disease*. 2017;8(5):628-42.

253. Bijkerk P, van Lier EA, van Vliet JA, Kretzschmar ME. [Effects of ageing on infectious disease]. *Nederlands Tijdschrift voor Geneeskunde*. 2010;154:A1613.
254. W.H.O. The top 10 causes of death 2020 [Available from: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>. Accessed on.18/32021
255. Fleming DM, Elliot AJ. The impact of influenza on the health and health care utilisation of elderly people. *Vaccine*. 2005;23 Suppl 1:S1-9.
256. Chen Y, Klein SL, Garibaldi BT, Li H, Wu C, Osevala NM, et al. Aging in COVID-19: Vulnerability, immunity and intervention. *Ageing research reviews*. 2021;65:101205-.
257. Collier DA, Ferreira IATM, Kotagiri P, Datir RP, Lim EY, Touizer E, et al. Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature*. 2021;596(7872):417-22.
258. Yousefzadeh MJ, Flores RR, Zhu Y, Schmiechen ZC, Brooks RW, Trussoni CE, et al. An aged immune system drives senescence and ageing of solid organs. *Nature*. 2021;594(7861):100-5.
259. Shirakawa K, Sano M. T Cell Immunosenescence in Aging, Obesity, and Cardiovascular Disease. *Cells*. 2021;10(9):2435.
260. Fulop T, Larbi A, Dupuis G, Le Page A, Frost EH, Cohen AA, et al. Immunosenescence and Inflamm-Aging As Two Sides of the Same Coin: Friends or Foes? *Frontiers in Immunology*. 2018;8(1960).
261. Rezuş E, Cardoneanu A, Burlui A, Luca A, Codreanu C, Tamba BI, et al. The Link Between Inflammaging and Degenerative Joint Diseases. *International Journal of Molecular Sciences*. 2019;20(3):614.
262. Aw D, Silva AB, Palmer DB. Immunosenescence: emerging challenges for an ageing population. *Immunology*. 2007;120(4):435-46.
263. Cox LS, Bellantuono I, Lord JM, Sapey E, Mannick JB, Partridge L, et al. Tackling immunosenescence to improve COVID-19 outcomes and vaccine response in older adults. *Lancet Healthy Longevity*. 2020;1(2):e55-e7.
264. Vijg J, Dong X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell*. 2020;182(1):12-23.
265. Yang J-H, Hayano M, Griffin PT, Amorim JA, Bonkowski MS, Apostolides JK, et al. Loss of epigenetic information as a cause of mammalian aging. *Cell*. 2023;186(2):305-26. e27.
266. Schmauck-Medina T, Molière A, Lautrup S, Zhang J, Chlopicki S, Madsen HB, et al. New hallmarks of ageing: a 2022 Copenhagen ageing meeting summary. *Ageing*. 2022;14(16):6829-39.
267. Franceschi C, Capri M, Monti D, Giunta S, Olivieri F, Sevini F, et al. Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mechanisms of Ageing and Development*. 2007;128(1):92-105.
268. Goronzy JJ, Weyand CM. Immune aging and autoimmunity. *Cellular and Molecular Life Sciences*. 2012;69(10):1615-23.
269. Watad A, Bragazzi NL, Adawi M, Amital H, Toubi E, Porat BS, et al. Autoimmunity in the Elderly: Insights from Basic Science and Clinics - A Mini-Review. *Gerontology*. 2017;63(6):515-23.

270. Furman D, Chang J, Lartigue L, Bolen CR, Haddad F, Gaudilliere B, et al. Expression of specific inflammasome gene modules stratifies older individuals into two extreme clinical and immunological states. *Nature Medicine*. 2017;23(2):174-84.
271. Shen-Orr SS, Furman D, Kidd BA, Hadad F, Lovelace P, Huang YW, et al. Defective Signaling in the JAK-STAT Pathway Tracks with Chronic Inflammation and Cardiovascular Risk in Aging Humans. *Cell Systems*. 2016;3(4):374-84.e4.
272. Costantini E, D'Angelo C, Reale M. The Role of Immunosenescence in Neurodegenerative Diseases. *Mediators of Inflammation*. 2018;2018:6039171-.
273. Pang WW, Price EA, Sahoo D, Beerman I, Maloney WJ, Rossi DJ, et al. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proceedings of the National Academy of Sciences*. 2011;108(50):20012.
274. Rossi DJ, Bryder D, Zahn JM, Ahlenius H, Sonu R, Wagers AJ, et al. Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(26):9194.
275. Ostan R, Bucci L, Capri M, Salvioli S, Scurti M, Pini E, et al. Immunosenescence and immunogenetics of human longevity. *Neuroimmunomodulation*. 2008;15(4-6):224-40.
276. Solana R, Tarazona R, Aiello AE, Akbar AN, Appay V, Beswick M, et al. CMV and Immunosenescence: from basics to clinics. *Immunity & Ageing*. 2012;9(1):23.
277. Sadighi Akha AA. Aging and the immune system: An overview. *Journal of Immunological Methods*. 2018;463:21-6.
278. Shukla AK, Johnson K, Giniger E. Common features of aging fail to occur in *Drosophila* raised without a bacterial microbiome. *iScience*. 2021;24(7):102703.
279. Effros RB. Role of T lymphocyte replicative senescence in vaccine efficacy. *Vaccine*. 2007;25(4):599-604.
280. McElhaney JE. The unmet need in the elderly: Designing new influenza vaccines for older adults. *Vaccine*. 2005;23:S10-S25.
281. Trzonkowski P, Myśliwska J, Pawelec G, Myśliwski A. From bench to bedside and back: the SENIEUR Protocol and the efficacy of influenza vaccination in the elderly. *Biogerontology*. 2009;10(1):83-94.
282. Dugan HL, Henry C, Wilson PC. Aging and influenza vaccine-induced immunity. *Cellular Immunology*. 2020;348:103998.
283. Smetana J, Chlibek R, Shaw J, Splino M, Prymula R. Influenza vaccination in the elderly. *Human Vaccines & Immunotherapeutics*. 2018;14(3):540-9.
284. Poland GA. Influenza vaccine failure: failure to protect or failure to understand? *Expert Review of Vaccines*. 2018;17(6):495-502.
285. Haq K, McElhaney JE. Immunosenescence: Influenza vaccination and the elderly. *Current Opinion in Immunology*. 2014;29:38-42.
286. Wu C, Chen X, Cai Y, Xia Ja, Zhou X, Xu S, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Internal Medicine*. 2020;180(7):934-43.



287. COVID Data Tracker [Internet]. 2021 [cited 1/9/2021]. Available from: <https://covid.cdc.gov/covid-data-tracker>. Accessed on: 1/9/2021
288. Ho FK, Petermann-Rocha F, Gray SR, Jani BD, Katikireddi SV, Niedzwiedz CL, et al. Is older age associated with COVID-19 mortality in the absence of other risk factors? General population cohort study of 470,034 participants. *PLoS One*. 2020;15(11):e0241824.
289. Fourati S, Cristescu R, Loboda A, Talla A, Filali A, Railkar R, et al. Pre-vaccination inflammation and B-cell signalling predict age-related hyporesponse to hepatitis B vaccination. *Nature Communications*. 2016;7:10369-.
290. Verschoor CP, Lelic A, Parsons R, Eveleigh C, Bramson JL, Johnstone J, et al. Serum C-Reactive Protein and Congestive Heart Failure as Significant Predictors of Herpes Zoster Vaccine Response in Elderly Nursing Home Residents. *The Journal of infectious diseases*. 2017;216(2):191-7.
291. Yang J, Sakai J, Siddiqui S, Lee RC, Ireland DDC, Verthelyi D, et al. IL-6 Impairs Vaccine Responses in Neonatal Mice. *Frontiers in Immunology*. 2018;9:3049.
292. Bekele Y, Sui Y, Berzofsky JA. IL-7 in SARS-CoV-2 Infection and as a Potential Vaccine Adjuvant. *Frontiers in Immunology*. 2021;12(3796).
293. Budamagunta V, Foster TC, Zhou D. Cellular senescence in lymphoid organs and immunosenescence. *Aging*. 2021;13(15):19920-41.
294. Solana R, Tarazona R, Gayoso I, Lesur O, Dupuis G, Fulop T. Innate immunosenescence: Effect of aging on cells and receptors of the innate immune system in humans. *Seminars in Immunology*. 2012;24(5):331-41.
295. Pinti M, Appay V, Campisi J, Frasca D, Fülöp T, Sauce D, et al. Aging of the immune system: Focus on inflammation and vaccination. *European Journal of Immunology*. 2016;46(10):2286-301.
296. Solana C, Tarazona R, Solana R. Immunosenescence of Natural Killer Cells, Inflammation, and Alzheimer's Disease. *International Journal of Alzheimer's Disease*. 2018;2018:3128758-.
297. Pawelec G. Hallmarks of human "immunosenescence": adaptation or dysregulation? *Immunity & Ageing*. 2012;9(1):15.
298. Nguyen V, Mendelsohn A, Larrick JW. Interleukin-7 and Immunosenescence. *Journal of Immunology Research*. 2017;2017:4807853.
299. Berben L, Antoranz A, Kenis C, Smeets A, Vos H, Neven P, et al. Blood Immunosenescence Signatures Reflecting Age, Frailty and Tumor Immune Infiltrate in Patients with Early Luminal Breast Cancer. *Cancers*. 2021;13(9):2185.
300. Huff WX, Kwon JH, Henriquez M, Fetcko K, Dey M. The Evolving Role of CD8(+)CD28(-) Immunosenescent T Cells in Cancer Immunology. *International Journal of Molecular Sciences*. 2019;20(11):2810.
301. Chen X, Liu Q, Xiang AP. CD8+CD28- T cells: not only age-related cells but a subset of regulatory T cells. *Cellular & Molecular Immunology*. 2018;15(8):734-6.
302. Vallejo AN. CD28 extinction in human T cells: altered functions and the program of T-cell senescence. *Immunological Reviews*. 2005;205(1):158-69.

303. Schulz AR, Mälzer JN, Domingo C, Jürchott K, Grützkau A, Babel N, et al. Low Thymic Activity and Dendritic Cell Numbers Are Associated with the Immune Response to Primary Viral Infection in Elderly Humans. *The Journal of Immunology*. 2015;195(10):4699-711.
304. Jagger A, Shimojima Y, Goronzy JJ, Weyand CM. Regulatory T cells and the immune aging process: a mini-review. *Gerontology*. 2014;60(2):130-7.
305. Dowling MR, Kan A, Heinzel S, Marchingo JM, Hodgkin PD, Hawkins ED. Regulatory T Cells Suppress Effector T Cell Proliferation by Limiting Division Destiny. *Frontiers in Immunology*. 2018;9(2461).
306. Collison LW, Workman CJ, Kuo TT, Boyd K, Wang Y, Vignali KM, et al. The inhibitory cytokine IL-35 contributes to regulatory T-cell function. *Nature*. 2007;450(7169):566-9.
307. Annacker O, Asseman C, Read S, Powrie F. Interleukin-10 in the regulation of T cell-induced colitis. *Journal of Autoimmunity*. 2003;20(4):277-9.
308. Wang L, Wang Y, Su B, Yu P, He J, Meng L, et al. Transcriptome-wide analysis and modelling of prognostic alternative splicing signatures in invasive breast cancer: a prospective clinical study. *Scientific Reports*. 2020;10(1):16504.
309. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. *Science*. 2020;369(6509):eaba3066.
310. Harries LW, Hernandez D, Henley W, Wood AR, Holly AC, Bradley-Smith RM, et al. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell*. 2011;10(5):868-78.
311. Balliu B, Durrant M, Goede Od, Abell N, Li X, Liu B, et al. Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome Biology*. 2019;20(1):230.
312. Latorre E, Birar VC, Sheerin AN, Jeynes JCC, Hooper A, Dawe HR, et al. Small molecule modulation of splicing factor expression is associated with rescue from cellular senescence. *BMC Cell Biology*. 2017;18(1):31.
313. Georgilis A, Klotz S, Hanley CJ, Herranz N, Weirich B, Morancho B, et al. PTBP1-Mediated Alternative Splicing Regulates the Inflammatory Secretome and the Pro-tumorigenic Effects of Senescent Cells. *Cancer Cell*. 2018;34(1):85-102.e9.
314. Lye JJ, Latorre E, Lee BP, Bandinelli S, Holley JE, Gutowski NJ, et al. Astrocyte senescence may drive alterations in GFAP $\alpha$ , CDKN2A p14ARF, and TAU3 transcript expression and contribute to cognitive decline. *GeroScience*. 2019;41(5):561-73.
315. Ubaida-Mohien C, Lyashkov A, Gonzalez-Freire M, Tharakan R, Shardell M, Moaddel R, et al. Discovery proteomics in aging human skeletal muscle finds change in spliceosome, immunity, proteostasis and mitochondria. *Elife*. 2019;8:e49874.
316. Lee BP, Pilling LC, Emond F, Flurkey K, Harrison DE, Yuan R, et al. Changes in the expression of splicing factor transcripts and variations in alternative splicing are associated with lifespan in mice and humans. *Aging Cell*. 2016;15(5):903-13.
317. Kwon SM, Min S, Jeoun U-w, Sim MS, Jung GH, Hong SM, et al. Global spliceosome activity regulates entry into cellular senescence. *The FASEB Journal*. 2021;35(1):e21204.

318. Martinez NM, Lynch KW. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunological Reviews*. 2013;253(1):216-36.
319. Martinez NM, Pan Q, Cole BS, Yarosh CA, Babcock GA, Heyd F, et al. Alternative splicing networks regulated by signaling in human T cells. *RNA (New York, NY)*. 2012;18(5):1029-40.
320. Ren P, Lu L, Cai S, Chen J, Lin W, Han F. Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease. *Frontiers in Immunology*. 2021;12(3284).
321. Chauhan K, Kalam H, Dutt R, Kumar D. RNA Splicing: A New Paradigm in Host–Pathogen Interactions. *Journal of Molecular Biology*. 2019;431(8):1565-75.
322. Thompson MG, Dittmar M, Mallory MJ, Bhat P, Ferretti MB, Fontoura BM, et al. Viral-induced alternative splicing of host genes promotes influenza replication. *Elife*. 2020;9.
323. Banerjee AK, Blanco MR, Bruce EA, Honson DD, Chen LM, Chow A, et al. SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. *Cell*. 2020;183(5):1325-39.e21.
324. Metcalf TU, Wilkinson PA, Cameron MJ, Ghneim K, Chiang C, Wertheimer AM, et al. Human Monocyte Subsets Are Transcriptionally and Functionally Altered in Aging in Response to Pattern Recognition Receptor Agonists. *The Journal of Immunology*. 2017;199(4):1405-17.
325. Lorusso JS, Sviderskiy OA, Labunskyy VM. Emerging Omics Approaches in Aging Research. *Antioxidants & redox signaling*. 2018;29(10):985-1002.
326. Kennedy BK, Berger SL, Brunet A, Campisi J, Cuervo AM, Epel ES, et al. Geroscience: linking aging to chronic disease. *Cell*. 2014;159(4):709-13.
327. Sierra F, Kohanski R. Geroscience and the trans-NIH Geroscience Interest Group, GSIG. *GeroScience*. 2017;39(1):1-5.
328. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell*. 2013;153(6):1194-217.
329. Liu Z, Zhu Y. Epigenetic clock: a promising mirror of ageing. *The Lancet Healthy Longevity*. 2021;2(6):e304-e5.
330. Zhavoronkov A, Mamoshina P. Deep Aging Clocks: The Emergence of AI-Based Biomarkers of Aging and Longevity. *Trends in Pharmacological Sciences*. 2019;40(8):546-9.
331. Mamoshina P, Volosnikova M, Ozerov IV, Putin E, Skibina E, Cortese F, et al. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. *Frontiers in Genetics*. 2018;9(242).
332. Meyer DH, Schumacher B. BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell*. 2021;20(3):e13320.
333. Chilunga FP, Henneman P, Elliott HR, Cronjé HT, Walia GK, Meeks KAC, et al. Epigenetic-age acceleration in the emerging burden of cardiometabolic diseases among migrant and non-migrant African populations: a population-based cross-sectional RODAM substudy. *The Lancet Healthy Longevity*. 2021;2(6):e327-e39.
334. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biology*. 2019;20(1):249.

335. Sayed N, Huang Y, Nguyen K, Krejciova-Rajaniemi Z, Grawe AP, Gao T, et al. An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging. *Nature Aging*. 2021;1(7):598-615.
336. Bartleson JM, Radenkovic D, Covarrubias AJ, Furman D, Winer DA, Verdin E. SARS-CoV-2, COVID-19 and the aging immune system. *Nature Aging*. 2021;1(9):769-82.
337. BD. Product Catalogue 2009/10. Accessed on: 2010 11.5.2020. Available from: [https://www.bd.com/documents/guides/directions-for-use/PAS\\_BC\\_Barricor-Lithium-Heparin-Plasma-Tubes\\_DF\\_EN.pdf](https://www.bd.com/documents/guides/directions-for-use/PAS_BC_Barricor-Lithium-Heparin-Plasma-Tubes_DF_EN.pdf).
338. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*. 2013;45(6):580-5.
339. Ferreira PG, Muñoz-Aguirre M, Reverter F, Sá Godinho CP, Sousa A, Amadoz A, et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nature Communications*. 2018;9(1):490.
340. Andrews S. FastQC a Quality Control Tool for High Throughput Sequence Data 2010 [cited 2023 05/04]. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed on.05/042020
341. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-8.
342. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20.
343. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*. 2013;29(1):15-21.
344. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
345. Picard. Broad Institute, GitHub repository: Broad Institute; 2019.
346. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
347. Federico Marini HB. pcaExplorer an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*. 2019(1):331.
348. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinformatics*. 2020;2(3):lqaa078.
349. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*. 2018;46(D1):D794-D801.
350. Dobin A. STAR Manual github.com: Alexander Dobin; 2020 [updated 11/8/2020. Available from: <https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf>. Accessed on.10/04/2021
351. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*. 2019;51(11):1560-5.

352. Felix Brechtmann CM, Agne Matuseviciute, Vicente Yepez, Julien Gagneur. OUTRIDER - OUTlier in RNA-Seq flNDER. [bioconductor.org](https://bioconductor.org); 2021.
353. Cummings B. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing – a walk through <https://macarthurlab.org/2015> [Available from: <https://macarthurlab.org/2017/05/31/improving-genetic-diagnosis-in-mendelian-disease-with-transcriptome-sequencing-a-walk-through/>]. Accessed on.07/01/2020
354. Cummings B. MendelianRNA-seq Github2018 [Available from: <https://github.com/berylc/MendelianRNA-seq>]. Accessed on.07/01/2020
355. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29(1):24-6.
356. Brendish NJ, Poole S, Naidu VV, Mansbridge CT, Norton NJ, Wheeler H, et al. Clinical impact of molecular point-of-care testing for suspected COVID-19 in hospital (COV-19POC): a prospective, interventional, non-randomised, controlled study. *The Lancet Respiratory medicine*. 2020;8(12):1192-200.
357. Clark TW, Beard KR, Brendish NJ, Malachira AK, Mills S, Chan C, et al. Clinical impact of a routine, molecular, point-of-care, test-and-treat strategy for influenza in adults admitted to hospital (FluPOC): a multicentre, open-label, randomised controlled trial. *Lancet Respiratory Medicine*. 2021;9(4):419-29.
358. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal*. 2011;17(1):3.
359. Ozerov IV, Lezhnina KV, Izumchenko E, Artemov AV, Medintsev S, Vanhaelen Q, et al. In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development. *Nature Communications*. 2016;7(1):13427.
360. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47-e.
361. Lab C. Selective Alignment 2019 [Available from: <https://combine-lab.github.io/alevin-tutorial/2019/selective-alignment/>]. Accessed on.20/7/2020
362. Tiberi S, Robinson MD. BANDITS: Bayesian differential splicing accounting for sample-to-sample variability and mapping uncertainty. *Genome Biology*. 2020;21(1):69.
363. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*. 2009;37(suppl\_2):W305-W11.
364. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS One*. 2011;6(7):e21800.
365. Blighe K, S Rana, and M Lewis. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. Github2018.
366. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011;12:2825-30.
367. Marini F, Binder H. pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components. *BMC Bioinformatics*. 2019;20(1):331.

368. Syndromic X-Linked Intellectual Disability Turner Type (MRXST) [Internet]. GeneCards. [cited 3.9.2021]. Available from: [https://www.malacards.org/card/syndromic\\_x\\_linked\\_intellectual\\_disability\\_turner\\_type](https://www.malacards.org/card/syndromic_x_linked_intellectual_disability_turner_type). Accessed on: 3.9.2021
369. Mental Retardation, X-Linked, Syndromic, Claes-Jensen Type (MRXSCJ) [Internet]. GeneCards. 2021 [cited 3/9/2021]. Available from: [https://www.malacards.org/card/mental\\_retardation\\_x\\_linked\\_syndromic\\_claes\\_jensen\\_type](https://www.malacards.org/card/mental_retardation_x_linked_syndromic_claes_jensen_type). Accessed on: 3/9/2021
370. Labory J, Le Bideau G, Pratella D, Yao J-E, Ait-El-Mkadem Saadi S, Bannwarth S, et al. ABEILLE: a novel method for ABerrant Expression Identification empLoying machine LEarning from RNA-sequencing data. *Bioinformatics*. 2022;38(20):4754-61.
371. van de Winkel JG, de Wit TP, Ernst LK, Capel PJ, Ceuppens JL. Molecular basis for a familial defect in phagocyte expression of IgG receptor I (CD64). *Journal of Immunology*. 1995;154(6):2896-903.
372. Bousfiha A, Jeddane L, Picard C, Al-Herz W, Ailal F, Chatila T, et al. Human Inborn Errors of Immunity: 2019 Update of the IUIS Phenotypical Classification. *Journal of Clinical Immunology*. 2020;40(1):66-81.
373. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*. 2016;54:1.30.1-1..3.
374. Brix TH, Kyvik KO, Christensen K, Hegedüs L. Evidence for a major role of heredity in Graves' disease: a population-based study of two Danish twin cohorts. *Journal of Clinical Endocrinology and Metabolism*. 2001;86(2):930-4.
375. Ohtsuka R, Abe Y, Shiratsuchi M, Suehiro Y, Karube K, Muta K, et al. [Graves' disease with splenomegaly and pancytopenia, mimicking B-cell lymphoproliferative disease]. *Rinsho Ketsueki Japanese Journal of Clinical Hematology*. 2008;49(2):104-8.
376. Vedeler C, Ulvestad E, Bjørge L, Conti G, Williams K, Mørk S, et al. The expression of CD59 in normal human nervous tissue. *Immunology*. 1994;82(4):542-7.
377. Kinoshita T. Congenital Defects in the Expression of the Glycosylphosphatidylinositol-Anchored Complement Regulatory Proteins CD59 and Decay-Accelerating Factor. *Seminars in Hematology*. 2018;55(3):136-40.
378. He Q, Bo J, Shen R, Li Y, Zhang Y, Zhang J, et al. S1P Signaling Pathways in Pathogenesis of Type 2 Diabetes. *Journal of Diabetes Research*. 2021;2021:1341750.
379. Allende ML, Dreier JL, Mandala S, Proia RL. Expression of the sphingosine 1-phosphate receptor, S1P1, on T-cells controls thymic emigration. *Journal of Biological Chemistry*. 2004;279(15):15396-401.
380. Garris CS, Blaho VA, Hla T, Han MH. Sphingosine-1-phosphate receptor 1 signalling in T cells: trafficking and beyond. *Immunology*. 2014;142(3):347-53.
381. Pereira JP, Cyster JG, Xu Y. A role for S1P and S1P1 in immature-B cell egress from mouse bone marrow. *PLoS One*. 2010;5(2):e9277.

382. Sun X, Ma SF, Wade MS, Flores C, Pino-Yanes M, Moitra J, et al. Functional variants of the sphingosine-1-phosphate receptor 1 gene associate with asthma susceptibility. *Journal of Allergy and Clinical Immunology*. 2010;126(2):241-9, 9.e1-3.
383. García-Clemente M, Enríquez-Rodríguez AI, Iscar-Urrutia M, Escobar-Mallada B, Arias-Guillén M, López-González FJ, et al. Severe asthma and bronchiectasis. *Journal of Asthma*. 2020;57(5):505-9.
384. Frugoni F, Dobbs K, Felgentreff K, Aldhekri H, Al Saud BK, Arnaout R, et al. A novel mutation in the POLE2 gene causing combined immunodeficiency. *Journal of Allergy and Clinical Immunology*. 2016;137(2):635-8.e1.
385. Lempainen J, Korhonen LS, Kantojärvi K, Heinonen S, Toivonen L, Rätty P, et al. Associations between IFI44L gene variants and rates of respiratory tract infections during early childhood. *The Journal of infectious diseases*. 2021;223(1):157-65.
386. Zhao M, Zhou Y, Zhu B, Wan M, Jiang T, Tan Q, et al. IFI44L promoter methylation as a blood biomarker for systemic lupus erythematosus. *Annals of the Rheumatic Diseases*. 2016;75(11):1998-2006.
387. Mertes C, Scheller IF, Yépez VA, Çelik MH, Liang Y, Kremer LS, et al. Detection of aberrant splicing events in RNA-seq data using FRASER. *Nature Communications*. 2021;12(1):529.
388. Legebeke J, Lord J, Penrice-Randal R, Vallejo AF, Poole S, Brendish NJ, et al. Evaluating the Immune Response in Treatment-Naive Hospitalised Patients With Influenza and COVID-19. *Frontiers in Immunology*. 2022;13:853265.
389. Zhang J, Ruan T, Sheng T, Wang J, Sun J, Wang J, et al. Role of c-Jun terminal kinase (JNK) activation in influenza A virus-induced autophagy and replication. *Virology*. 2019;526:1-12.
390. Chen J, Ye C, Wan C, Li G, Peng L, Peng Y, et al. The Roles of c-Jun N-Terminal Kinase (JNK) in Infectious Diseases. *International Journal of Molecular Sciences*. 2021;22(17):9640.
391. Xie J, Zhang S, Hu Y, Li D, Cui J, Xue J, et al. Regulatory roles of c-jun in H5N1 influenza virus replication and host inflammation. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. 2014;1842(12, Part A):2479-88.
392. Chu X, Wang C, Wu Z, Fan L, Tao C, Lin J, et al. JNK/c-Jun-driven NLRP3 inflammasome activation in microglia contributed to retinal ganglion cells degeneration induced by indirect traumatic optic neuropathy. *Experimental Eye Research*. 2021;202:108335.
393. García-Heredia JM, Carnero A. The cargo protein MAP17 (PDZK1IP1) regulates the immune microenvironment. *Oncotarget*. 2017;8(58):98580-97.
394. Hasegawa S, Matsushige T, Inoue H, Takahara M, Kajimoto M, Momonaka H, et al. Serum soluble CD163 levels in patients with influenza-associated encephalopathy. *Brain and Development*. 2013;35(7):626-9.
395. Buechler C, Eisinger K, Krautbauer S. Diagnostic and prognostic potential of the macrophage specific receptor CD163 in inflammatory diseases. *Inflammation and Allergy - Drug Targets*. 2013;12(6):391-402.
396. Delgado-Eckert E, Ojosnegros S, Beerenwinkel N. The Evolution of Virulence in RNA Viruses under a Competition–Colonization Trade-Off. *Bulletin of Mathematical Biology*. 2011;73(8):1881-908.

397. Hsu N-Y, Illynska O, Belov G, Santiana M, Chen Y-H, Takvorian PM, et al. Viral Reorganization of the Secretory Pathway Generates Distinct Organelles for RNA Replication. *Cell*. 2010;141(5):799-811.
398. Diehl N, Schaal H. Make yourself at home: viral hijacking of the PI3K/Akt signaling pathway. *Viruses*. 2013;5(12):3192-212.
399. Zhang J, Wu H, Yao X, Zhang D, Zhou Y, Fu B, et al. Pyroptotic macrophages stimulate the SARS-CoV-2-associated cytokine storm. *Cellular & Molecular Immunology*. 2021;18(5):1305-7.
400. Cole SL, Dunning J, Kok WL, Benam KH, Benlahrech A, Repapi E, et al. M1-like monocytes are a major immunological determinant of severity in previously healthy adults with life-threatening influenza. *JCI Insight*. 2017;2(7).
401. Corry J, Kettenburg G, Upadhyay AA, Wallace M, Marti MM, Wonderlich ER, et al. Infiltration of inflammatory macrophages and neutrophils and widespread pyroptosis in lung drive influenza lethality in nonhuman primates. *PLoS Pathogens*. 2022;18(3):e1010395.
402. Mehraj V, Ponte R, Routy J-P. The Dynamic Role of the IL-33/ST2 Axis in Chronic Viral-infections: Alarming and Adjuvanting the Immune Response. *EBioMedicine*. 2016;9:37-44.
403. Wang P, Gamero AM, Jensen LE. IL-36 promotes anti-viral immunity by boosting sensitivity to IFN- $\alpha/\beta$  in IRF1 dependent and independent manners. *Nature Communications*. 2019;10(1):4700.
404. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*. 2020;172(9):577-82.
405. Ghebrehewet S, MacPherson P, Ho A. Influenza. *BMJ (Clinical research ed)*. 2016;355:i6258-i.
406. Park JW, Fu S, Huang B, Xu R-H. Alternative splicing in mesenchymal stem cell differentiation. *Stem Cells*. 2020;38(10):1229-40.
407. Weiss ARR, Dahlke MH. Immunomodulation by Mesenchymal Stem Cells (MSCs): Mechanisms of Action of Living, Apoptotic, and Dead MSCs. *Frontiers in Immunology*. 2019;10.
408. Shao Y, Deng T, Zhang T, Li P, Wang Y. FAM19A3, a novel secreted protein, modulates the microglia/macrophage polarization dynamics and ameliorates cerebral ischemia. *FEBS Letters*. 2015;589(4):467-75.
409. CSC Genome Browser on Human (GRCh38/hg38) [Internet]. 2022 [cited 19/06/2022]. Available from: [http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr1%3A56627901%2D56732821&hgid=1382601287\\_Z4cv8tKAzmVNMVTXla49ALvct4Qs](http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr1%3A56627901%2D56732821&hgid=1382601287_Z4cv8tKAzmVNMVTXla49ALvct4Qs). Accessed on: 19/06/2022
410. Zhao F, Wang C, Zhu X. Isoform-specific roles of AMPK catalytic  $\alpha$  subunits in Alzheimer's disease. *The Journal of Clinical Investigation*. 2020;130(7):3403-5.
411. Xu L, Ash J. AMPK $\alpha$ 2 plays a unique role in Cone function and survival. *Investigative Ophthalmology and Visual Science*. 2016;57(12):177-.
412. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, 3rd, Lucas J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host & Microbe*. 2009;6(3):207-17.



413. Scepanovic P, Alanio C, Hammer C, Hodel F, Bergstedt J, Patin E, et al. Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. *Genome Medicine*. 2018;10(1):59.
414. Leligdowicz A, Matthay MA. Heterogeneity in sepsis: new biological evidence with clinical applications. *Critical Care*. 2019;23(1):80.
415. Mangold CA, Rathbun MM, Renner DW, Kuny CV, Szpara ML. Viral infection of human neurons triggers strain-specific differences in host neuronal and viral transcriptomes. *PLoS Pathogens*. 2021;17(3):e1009441.
416. Josset L, Zeng H, Kelly Sara M, Tumpey Terrence M, Katze Michael G, Buchmeier Michael J. Transcriptomic Characterization of the Novel Avian-Origin Influenza A (H7N9) Virus: Specific Host Response and Responses Intermediate between Avian (H5N1 and H7N7) and Human (H3N2) Viruses and Implications for Treatment Options. *mBio*. 5(1):e01102-13.
417. Chen JM. Interpreting Linear Beta Coefficients Alongside Feature Importances in Machine Learning. *Atlantic Economic Journal*. 2021;49(2):245-7.
418. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*. 2017;33(18):2938-40.
419. Andryukov BG, Besednova NN. Older adults: panoramic view on the COVID-19 vaccination. *AIMS Public Health*. 2021;8(3):388.
420. Lawton G. You're only as young as your immune system. *New Scientist*. 2020;245(3275):44-8.
421. Sobie EA. An introduction to MATLAB. *Science Signaling*. 2011;4(191):tr7.
422. Ahmadi P, Hartjen P, Kohsar M, Kummer S, Schmiedel S, Bockmann JH, et al. Defining the CD39/CD73 Axis in SARS-CoV-2 Infection: The CD73(-) Phenotype Identifies Polyfunctional Cytotoxic Lymphocytes. *Cells*. 2020;9(8).
423. Ljungberg JK, Kling JC, Tran TT, Blumenthal A. Functions of the WNT Signaling Network in Shaping Host Responses to Infection. *Frontiers in Immunology*. 2019;10:2521.
424. Alfaro E, Díaz-García E, García-Tovar S, Zamarrón E, Mangas A, Galera R, et al. Upregulated Proteasome Subunits in COVID-19 Patients: A Link with Hypoxemia, Lymphopenia and Inflammation. *Biomolecules*. 2022;12(3).
425. Tsalik EL, Fiorino C, Aqeel A, Liu Y, Henao R, Ko ER, et al. The Host Response to Viral Infections Reveals Common and Virus-Specific Signatures in the Peripheral Blood. *Frontiers in Immunology*. 2021;12:741837.
426. Seeland U, Coluzzi F, Simmaco M, Mura C, Bourne PE, Heiland M, et al. Evidence for treatment with estradiol for women with SARS-CoV-2 infection. *BMC Medicine*. 2020;18(1):369.
427. Hardie DL, Baldwin MJ, Naylor A, Haworth OJ, Hou TZ, Lax S, et al. The stromal cell antigen CD248 (endosialin) is expressed on naive CD8+ human T cells and regulates proliferation. *Immunology*. 2011;133(3):288-95.
428. Wu J, Lu G. Multiple functions of TBCK protein in neurodevelopment disorders and tumors. *Oncology Letters*. 2021;21(1):1-

429. Tintos-Hernández JA, Santana A, Keller KN, Ortiz-González XR. Lysosomal dysfunction impairs mitochondrial quality control and is associated with neurodegeneration in TBCK encephaloneuronopathy. *Brain Communications*. 2021;3(4).
430. Liu Y, Yan X, Zhou T. TBCK influences cell proliferation, cell size and mTOR signaling pathway. *PloS One*. 2013;8(8):e71349.
431. Papadopoli D, Boulay K, Kazak L, Pollak M, Mallette F, Topisirovic I, et al. mTOR as a central regulator of lifespan and aging. *F1000Res*. 2019;8.
432. Mannick JB, Del Giudice G, Lattanzi M, Valiante NM, Praestgaard J, Huang B, et al. mTOR inhibition improves immune function in the elderly. *Science Translational Medicine*. 2014;6(268):268ra179.
433. Porcù E, Benetton M, Bisio V, Da Ros A, Tregnago C, Borella G, et al. The long non-coding RNA CDK6-AS1 overexpression impacts on acute myeloid leukemia differentiation and mitochondrial dynamics. *iScience*. 2021;24(11):103350.
434. Hu MG, Deshpande A, Schlichting N, Hinds EA, Mao C, Dose M, et al. CDK6 kinase activity is required for thymocyte development. *Blood*. 2011;117(23):6120-31.
435. Thapa P, Farber DL. The Role of the Thymus in the Immune Response. *Thoracic Surgery Clinics*. 2019;29(2):123-31.
436. Thomas R, Wang W, Su D-M. Contributions of Age-Related Thymic Involution to Immunosenescence and Inflammaging. *Immunity & Ageing*. 2020;17(1):2.
437. Pack LR, Daigh LH, Chung M, Meyer T. Clinical CDK4/6 inhibitors induce selective and immediate dissociation of p21 from cyclin D-CDK4 to inhibit CDK2. *Nature Communications*. 2021;12(1):3356.
438. Kesheh MM, Mahmoudvand S, Shokri S. Long noncoding RNAs in respiratory viruses: A review. *Reviews in Medical Virology*. 2022;32(2):e2275.
439. Wang Y, Liu J, Huang BO, Xu YM, Li J, Huang LF, et al. Mechanism of alternative splicing and its regulation. *Biomedical Reports*. 2015;3(2):152-8.
440. Harries LW. Dysregulated RNA processing and metabolism: a new hallmark of ageing and provocation for cellular senescence. *The FEBS Journal*.n/a(n/a).
441. Nagai N, Kudo Y, Aki D, Nakagawa H, Taniguchi K. Immunomodulation by Inflammation during Liver and Gastrointestinal Tumorigenesis and Aging. *International Journal of Molecular Sciences*. 2021;22(5).
442. Du J, Wang Q, Ziegler SF, Zhou B. FOXP3 interacts with hnRNPF to modulate pre-mRNA alternative splicing. *Journal of Biological Chemistry*. 2018;293(26):10235-44.
443. Henderson AR. Testing experimental data for univariate normality. *Clinica Chimica Acta*. 2006;366(1):112-29.
444. Frasca D, Blomberg BB. Inflammaging decreases adaptive and innate immune responses in mice and humans. *Biogerontology*. 2016;17(1):7-19.
445. Rasa SMM, Annunziata F, Krepelova A, Nunna S, Omrani O, Gebert N, et al. Inflammaging is driven by upregulation of innate immune receptors and systemic interferon signaling and is ameliorated by dietary restriction. *Cell Reports*. 2022;39(13):111017.

446. Kordaß T, Osen W, Eichmüller SB. Controlling the Immune Suppressor: Transcription Factors and MicroRNAs Regulating CD73/NT5E. *Frontiers in Immunology*. 2018;9.
447. Dorneles GP, Teixeira PC, da Silva IM, Schipper LL, Santana Filho PC, Rodrigues Junior LC, et al. Alterations in CD39/CD73 axis of T cells associated with COVID-19 severity. *Journal of Cellular Physiology*.n/a(n/a).
448. Li X, Yan M, Chen J, Luo Y. The Potential of Mesenchymal Stem Cells for the Treatment of Cytokine Storm due to COVID-19. *BioMed Research International*. 2021;2021:3178796.
449. Pillalamarri N, Abdullah, Ren G, Khan L, Ullah A, Jonnakuti S, et al. Exploring the utility of extracellular vesicles in ameliorating viral infection-associated inflammation, cytokine storm and tissue damage. *Translational Oncology*. 2021;14(7):101095.
450. Fernandez RJ, 3rd, Johnson FB. A regulatory loop connecting WNT signaling and telomere capping: possible therapeutic implications for dyskeratosis congenita. *Annals of the New York Academy of Sciences*. 2018;1418(1):56-68.
451. Kared H, Tan SW, Lau MC, Chevrier M, Tan C, How W, et al. Immunological history governs human stem cell memory CD4 heterogeneity via the Wnt signaling pathway. *Nature Communications*. 2020;11.
452. Bam M, Yang X, Zumbun EE, Zhong Y, Zhou J, Ginsberg JP, et al. Dysregulated immune system networks in war veterans with PTSD is an outcome of altered miRNA expression and DNA methylation. *Scientific Reports*. 2016;6:31209.
453. Chaiworapongsa T, Romero R, Whitten A, Tarca AL, Bhatti G, Draghici S, et al. Differences and similarities in the transcriptional profile of peripheral whole blood in early and late-onset preeclampsia: insights into the molecular basis of the phenotype of preeclampsia. *Journal of Perinatal Medicine*. 2013;41(5):485-504.
454. Tong DL, Chen RG, Lu YL, Li WK, Zhang YF, Lin JK, et al. The critical role of ASD-related gene CNTNAP3 in regulating synaptic development and social behavior in mice. *Neurobiology of Disease*. 2019;130:104486.
455. Zhao B, Fan Q, Liu J, Yin A, Wang P, Zhang W. Identification of Key Modules and Genes Associated with Major Depressive Disorder in Adolescents. *Genes (Basel)*. 2022;13(3).
456. Qiao YQ, Huang ML, Zheng Q, Wang TR, Xu AT, Cao Y, et al. CNTNAP3 associated ATG16L1 expression and Crohn's disease. *Mediators of Inflammation*. 2015;2015:404185.
457. Holly AC, Melzer D, Pilling LC, Fellows AC, Tanaka T, Ferrucci L, et al. Changes in splicing factor expression are associated with advancing age in man. *Mechanisms of Ageing and Development*. 2013;134(9):356-66.
458. Lee BP, Smith M, Buffenstein R, Harries LW. Negligible senescence in naked mole rats may be a consequence of well-maintained splicing regulation. *GeroScience*. 2020;42(2):633-51.
459. Zhou P, Wu G, Zhang P, Xu R, Ge J, Fu Y, et al. SATB2-Nanog axis links age-related intrinsic changes of mesenchymal stem cells from craniofacial bone. *Aging*. 2016;8(9):2006-11.
460. Izgi H, Han D, Isildak U, Huang S, Kocabiyik E, Khaitovich P, et al. Inter-tissue convergence of gene expression during ageing suggests age-related loss of tissue and cellular identity. *Elife*. 2022;11.

461. Kan A, Hodgkin PD. Mechanisms of cell division as regulators of acute immune response. *Systems and Synthetic Biology*. 2014;8(3):215-21.
462. Bhowmik R, Pardasani M, Mahajan S, Magar R, Joshi SV, Nair GA, et al. Persistent olfactory learning deficits during and post-COVID-19 infection. *Current Research in Neurobiology*. 2023;4:100081.
463. Latorre E, Ostler EL, Faragher RGA, Harries LW. FOXO1 and ETV6 genes may represent novel regulators of splicing factor expression in cellular senescence. *FASEB Journal*. 2019;33(1):1086-97.
464. RUNX3 RUNX family transcription factor 3 [ Homo sapiens (human) ] [Internet]. 2022 [cited 29/07/2022]. Available from: <https://www.ncbi.nlm.nih.gov/gene/864>. Accessed on: 29/07/2022
465. Balogh P, Adelman ER, Pluvinage JV, Capaldo BJ, Freeman KC, Singh S, et al. RUNX3 levels in human hematopoietic progenitors are regulated by aging and dictate erythroid-myeloid balance. *Haematologica*. 2020;105(4):905-13.
466. Goronzy JJ, Weyand CM. Understanding immunosenescence to improve responses to vaccines. *Nature Immunology*. 2013;14(5):428-36.
467. Krishnan V, Ito Y. RUNX3 loss turns on the dark side of TGF-beta signaling. *Oncoscience*. 2017;4(11-12):156-7.
468. Roukens AH, Soonawala D, Joosten SA, de Visser AW, Jiang X, Dirksen K, et al. Elderly subjects have a delayed antibody response and prolonged viraemia following yellow fever vaccination: a prospective controlled cohort study. *PloS One*. 2011;6(12):e27753.
469. Crooke SN, Ovsyannikova IG, Poland GA, Kennedy RB. Immunosenescence and human vaccine immune responses. *Immunity & Ageing*. 2019;16:25.
470. Paz I, Kosti I, Ares M, Jr, Cline M, Mandel-Gutfreund Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research*. 2014;42(W1):W361-W7.
471. Risso D, Schwartz K, Sherlock G, Dudoit S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics*. 2011;12(1):480.