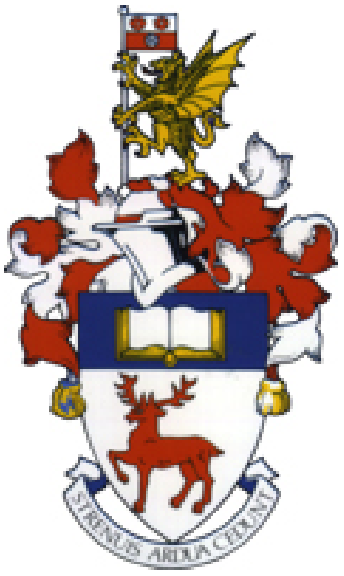


UNIVERSITY OF SOUTHAMPTON

Faculty of Medicine

Human Development & Health


# Advanced modelling of genomic data in Inflammatory Bowel Disease

by


Enrico Mossotto


A thesis submitted for the degree of Doctor of Philosophy

Supervisory Team:

Dr Benjamin MacArthur, Prof Jacek Brodzki & Prof Sarah Ennis

May, 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Human Genetics

Doctor of Philosophy

# ADVANCED MODELLING OF GENOMIC DATA IN INFLAMMATORY BOWEL DISEASE

by Enrico Mossotto

Advances in next generation sequencing technologies allow the collection of enormous volumes of genomic data on large patient cohorts. Concurrently, machine learning algorithms are rapidly evolving and, together, these technologies represent the new frontier of research and clinical management on a path leading toward personalised medicine.

The aims of this thesis are two. Firstly, to develop a mathematical framework for the analysis and integration of next generation sequencing data. Secondly, to model data from patients affected by inflammatory bowel disease (IBD), a common complex autoimmune condition with increasing incidence worldwide, by applying machine learning methodologies to clinical and transformed genomic data.

The analyses presented in this thesis are largely based on a cohort of paediatric IBD patients for which clinical data, immunology and whole exome sequencing data were available.

This research illustrates a supervised and unsupervised machine learning approach modelling histology and endoscopy data for assigning IBD patients with the correct CD/UC subtypes with superior accuracy.

Stratification and classification of IBD patients can be improved by layering genomic data on top of clinical evidence. This thesis also describes the development of GenePy, a mathematical model for transforming patients genomic data into a per-individual per-gene deleteriousness scoring system. GenePy is capable of

modelling and implementing important biological information from whole exome sequencing data from patient DNA. GenePy eases the analysis and interpretation of genomic data on an individual basis and concomitantly allows the comparison of genetic profiles across patients. GenePy gene scores can be further combined according to molecular processes or pathways.

This work describes eight novel immuno-genomic IBD sybtypes observed on a small cohort for which immune cytokine signalling and response cascades have been specifically profiled and GenePy scores obtained.

In addition, the GenePy algorithm is applied using both supervised and unsupervised approaches to classify IBD subtypes and to explore alternative disease classifications that discriminate molecular clinical subtypes that are clinically relevant for treatment and prognosis. This thesis reports the current highest performance in discriminating IBD subtypes using exome sequencing data and five novel genomic patient strata defined by different mutational burden of adaptive immune system genes.

This work demonstrates the power of integrating $21^{st}$ century high throughput digital data in machine learning frameworks and the potential to obtain clinically relevant strata for bench to bedside improvements in patient quality of life.

# List of Abbreviations

| | |
|---|---|
| AI | Artificial intelligence |
| ANN | Artificial Neural Network |
| AUC | Area under the curve |
| BED | Browser extensible data |
| BWA | Burrows-Wheeler Aligner |
| CAGI | Critical assessment of Genome Interpretation |
| CC | Correlation coefficient |
| CD | Crohn's disease |
| CNV | Copy number variation |
| CoV | Coefficient of variation |
| CV | Cross validation |
| DNA | Deoxyribonucleic acid |
| eQTL | expression quantitative loci |
| GA | Genetic algorithm |
| GATK | Genome analysis toolkit |
| GDI | Gene damage index |
| GI | Gastrointestine |
| GUI | Graphical user interface |
| GWAS | Genome wide association studies |
| HC | Hierarchical clustering |
| IBD | Inflammatory bowel disease |
| IBDU | Inflammatory bowel disease undetermined |
| IQR | Interquartile range |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| MAF | Minor allele frequency |
| MDS | Multidimensional scaling |
| ML | Machine learning |
| MSE | Mean standard error |
| NGS | Next Generation Sequencing |
| NLR | Nod-like receptor |

| | |
|---|---|
| NOD2 | Nucleotide-binding oligomerization domain-containing protein 2 |
| PCA | Principal component analysis |
| PCR | Polymerase chain reaction |
| PIBD | Paediatric inflammatory bowel disease |
| QC | Quality control |
| RFE-CV | Recursive feature elimination -cross validated |
| RNA | Ribonucleic acid |
| ROC | Receiver operating characteristic |
| SKAT | Sequence Kernel Association Test |
| SNP | Single nucleotide polymorphism |
| SNV | Single nucleotide variant |
| SVM | Support vector machine |
| TLR | Toll-like receptor |
| UC | Ulcerative colitis |
| VCF | Variant call format |
| WES | Whole Exome Sequencing |

# Contents

**Abstract** iii

**List of Abbreviations** v

**List of Figures** xiii

**List of Tables** xvii

**Published papers** xix

**Submitted Papers** xix

**Acknowledgements** xx

**Ethics approval** xxi

**Funders** xxi

**Declaration of authorship** xxiii

**1 Introduction** 1

1.1 The post-genomic era . . . . . 1

1.1.1 Next generation sequencing . . . . . 2
The human genome . . . . . 3
Whole Exome Sequencing . . . . . 6
1.2 Machine learning algorithms . . . . . 9
1.2.1 Types of learning . . . . . 13
1.2.2 Data sets and model validation . . . . . 14
1.2.3 Feature selection . . . . . 16
1.3 Machine learning applications to genomics . . . . . 17
1.4 Inflammatory Bowel Disease . . . . . 20
1.4.1 Disease classification . . . . . 22
1.4.2 Genetics of IBD . . . . . 24
Associated pathways . . . . . 27
1.4.3 Machine learning applications to IBD genetics . . . . . 30
1.5 Thesis outline, aims and contribution . . . . . 32

**2 Methods** **35**

2.1 Programming tools . . . . . 35
2.1.1 Iridis 4 . . . . . 35
2.1.2 Python . . . . . 36
Packages . . . . . 36
2.2 Bioinformatic tools . . . . . 36
2.2.1 BWA . . . . . 38

2.2.2 GATK . . . . . 39
2.2.3 ANNOVAR . . . . . 40
2.2.4 CADD . . . . . 42
2.2.5 MaxEnt . . . . . 43
2.3 Machine learning algorithms . . . . . 43
2.3.1 Supervised learning algorithms . . . . . 44
Support vector machine . . . . . 44
2.3.2 Unsupervised learning algorithms . . . . . 47
Principal components analysis . . . . . 48
Multidimensional scaling . . . . . 49
t-distributed SNE . . . . . 50
Hierarchical Clustering . . . . . 50
2.3.3 Resampling methods . . . . . 52
Permutation, bootstrapping and jackknife . . . . . 53
Cross-validation . . . . . 53
2.4 Feature selection algorithms . . . . . 56
2.4.1 Univariate feature selection . . . . . 56
2.4.2 Linear regression and lasso . . . . . 57

**3 Machine learning classification of inflammatory bowel disease patients using histopathology data 59**

3.1 Summary . . . . . 59
3.2 Introduction . . . . . 60

3.3 Methods . . . 62
3.3.1 Sample data . . . 62
3.3.2 Unsupervised machine learning . . . 62
3.3.3 Supervised machine learning . . . 63
3.4 Results . . . 66
3.4.1 Unsupervised clustering of CD and UC phenotypes . . . 66
3.4.2 Hierarchical clustering of PIBD subtypes . . . 68
3.4.3 Supervised classification of PIBD patients . . . 69
3.4.4 Model validation in an additional cohort . . . 72
3.4.5 IBDU reclassification . . . 72
3.5 Discussion . . . 73

**4 GenePy – a tool for estimating gene pathogenicity in individuals using next-generation sequencing data 77**

4.1 Summary . . . 77
4.2 Introduction . . . 78
4.3 Methods . . . 81
4.3.1 Sample data . . . 81
4.3.2 WES data processing . . . 82
4.3.3 Quality Control . . . 83
4.3.4 GenePy score . . . 84
4.3.5 Score validation . . . 85
4.4 Results . . . 87

4.4.1 QC results . . . 87
4.4.2 GenePy score behaviour – impact of allele frequency and zygosity . . . 88
4.4.3 GenePy score behaviour – impact of deleteriousness metric . 88
4.4.4 GenePy score testing . . . 93
4.5 Discussion . . . 98

**5 Stratification of paediatric patients using immunogenomic data 101**

5.1 Summary . . . 101
5.2 Introduction . . . 102
5.3 Methods . . . 104
5.3.1 Sample Data . . . 104
5.3.2 Immunological assay . . . 105
5.3.3 WES data processing . . . 105
5.3.4 Unsupervised stratification . . . 106
5.4 Results . . . 108
5.4.1 Cytokine responses of IBD patients . . . 108
5.4.2 Hierarchical clustering of immune-phenotypes . . . 109
5.4.3 Genomic interpretation of immuno-phenotypes . . . 113
5.5 Discussion . . . 115

**6 Machine learning modelling of genomic data of IBD 119**

6.1 Summary . . . 119

6.2 Introduction . . . . . 120

6.3 Methods . . . . . 122

6.3.1 Sample data . . . . . 122

6.3.2 Gene selection . . . . . 123

6.3.3 Supervised classification . . . . . 123

6.3.4 Unsupervised learning . . . . . 126

6.4 Results . . . . . 126

6.4.1 Supervised Classification of IBD subtypes . . . . . 126

6.4.2 Supervised Classification of IBD vs Control . . . . . 130

6.4.3 Gene enrichment . . . . . 132

6.4.4 Unsupervised Stratification of IBD patients using genomic data . . . . . 135

6.5 Discussion . . . . . 141

**7 Conclusions and Future work** **145**

**Bibliography** **156**

# List of Figures

1.1 **Historical trends in storage prices versus DNA sequencing costs and year** . . . 2

1.2 **DNA structure** . . . 3

1.3 **Whole exome sequencing workflow** . . . 6

1.4 **Single and paired end configuration** . . . 7

1.5 **Comparison of insert size distribution between WGS and WES** . . . 8

1.6 **Example of a 4 line FASTQ format** . . . 9

1.7 **Function fitting** . . . 12

1.8 **PCA clusters ethnic groups** . . . 14

1.9 **IBD inflammation patterns** . . . 23

1.10 **Genetics and environmental contribution to IBD onset** . . . 25

1.11 **Timeline of loci discovered to be associated with IBD phenotype** . . . 26

1.12 **NOD-like receptor signalling pathway** . . . 28

1.13 **IBD pathways** . . . 29

2.1 **Bioinformatic workflow for NGS data analysis** . . . 37

2.2 **Aligner accuracies** . . . . . 40

2.3 **Comparison of the gVCF format against the classic VCF** . 41

2.4 **Example of an hyperplane** . . . . . 45

2.5 **Margin maximisation** . . . . . 46

2.6 **PCA of Europe** . . . . . 49

2.7 **t-SNE applied on the hand writing digit recognition database** 51

2.8 **Example of a 5-fold cross-validation** . . . . . 55

2.9 **Univariate feature selection performance** . . . . . 57

3.1 **Model schematic and histopathology data processing** . . . . 64

3.2 **Dimensionality reduction approaches and hierarchical clustering of histopathology data** . . . . . 67

3.3 **Supervised classification performance and metrics using histopathology data** . . . . . 71

4.1 **Single variant GenePy score distribution under fixed deleteriousness values** . . . . . 89

4.2 **Median whole gene GenePy score profiles observed across the cohort of 508 patients with WES data for all sixteen metrics of deleteriousness** . . . . . 92

4.3 **IBD ethnicity imputation** . . . . . 94

4.4 **GenePy scores profiles for the *NOD2* gene in the CD and control groups for each of the sixteen implemented deleteriousness metrics** . . . . . 95

5.1 **Cytokine production in IBD** . . . . . 103

5.2 **Principal component analysis of immunoassay data** . . . . . 108

5.3 **Radar plot of immunoassay data** . . . . . 109

5.4 **Hierarchical clustering of immunoassay data from IBD patients** . . . . . 110

6.1 **CD *vs.* UC areas under the ROC curve for each IBD related pathway** . . . . . 128

6.2 **CD *vs.* UC areas under the ROC curve for each deleteriousness metric** . . . . . 130

6.3 **IBD *vs.* control areas under the ROC curve for each deleteriousness metric** . . . . . 132

6.4 **Protein-protein interaction network based on the 40 genes selected for the CD *vs.* UC classification** . . . . . 133

6.5 **Protein-protein interaction network based on the 28 genes selected for the IBD *vs.* controls classification** . . . . . 134

6.6 **Principal component analyses of IBD cases 16 different deleteriousness metrics** . . . . . 136

6.7 **Hierarchical clustering of IBD cases for all 16 deleteriousness metrics** . . . . . 137

6.8 **Hierarchical clustering of only IBD cases using M-CAP deleteriousness metric** . . . . . 140

7.1 **Schematic representation of phased data** . . . . . 149

7.2 **Median whole gene GenePyuncorrected score profiles observed across the cohort of 508 patients with WES data depicted separately for each of the sixteen deleteriousness metrics** . . . . . 156

7.3 **Median whole gene GenePy$_{cgl}$ score profiles observed across the cohort of 508 patients with WES data depicted separately for each of the sixteen deleteriousness metrics** . . . . 157

7.4 **Individual radar plots of cytokine responses per patient** . . 158

7.5 **Individual radar plots of cytokine responses per patient (2)** 159

# List of Tables

1.1 **Characteristic traits for UC and CD** . . . . . 21

3.1 **Preliminary assessment of linear and non-linear models** . . 69

3.2 **Performance of the three optimised supervised models** . . . 70

3.3 **Performance of the trained combined histopathology model over the validation set** . . . . . 72

4.1 **Pathogenicity scores for SNVs and their reported ranges in the dbsnfp database** . . . . . 86

4.2 **Statistical attributes of whole gene GenePy scores computed for sixteen deleteriousness metrics** . . . . . 91

4.3 ***NOD2* GenePy score statistics (maxima and means) and Mann-Whitney U tests across groups for all sixteen deleteriousness metrics** . . . . . 97

5.1 **Normalised cytokine response level for IBD cases and controls** . . . . . 112

5.2 **GenePy scores regression against immuno-phenotypes** . . . 114

6.1 **Pathways involved in IBD pathogenesis** . . . . . 124

6.2 **Performance of CD *vs.* UC classification for different ANOVA thresholds** . . . . . 127

6.3 **Supervised classification of CD vs UC using GenePy scores and CADD metric** . . . . . 129

6.4 **CD *vs.* UC classification performance for each deleteriousness metric** . . . . . 131

6.5 **Unsupervised IBD groups regression data** . . . . . 139

7.1 **Monogenic IBD genes** . . . . . 153

7.2 **All single nucleotide variants in the NOD2 gene used in GenePy validation.** . . . . . 154

7.3 **Genes driving the unsupervised clustering of IBD patients** 154

# Published papers

- JJ Ashton, Q Bonduelle, **E Mossotto**, T Coelho, A Batra, NA Afzal, B Vadgama, S Ennis, RM Beattie - Endoscopic and Histological Assessment of Paediatric Inflammatory Bowel Disease over a three year follow-up period. *J. Pediatr. Gastroenterol. Nutr.* 1 (2017)

- **E Mossotto**, JJ Ashton, T Coehlo, RM Beattie, B MacArthur, S Ennis . - Classification of Inflammatory Bowel Disease using Machine Learning, *Sci. Reports* 7, 2427 (2017)

- R J Pengelly, A A Gheyas, R Kuo, **E Mossotto**, E G Seaby, D W Burt, S Ennis and A Collins - Commercial chicken breeds exhibit highly divergent patterns of linkage disequilibrium. *Heredity*, 117 (5), 375-382 (2016)

# Submitted Papers

- **E Mossotto**, JJ Ashton, RJ Pengelly, RM Beattie, B MacArthur, S Ennis. - GenePy – a score for estimating gene pathogenicity in individuals using next-generation sequencing data, *Plos Comp. Bio.*

# Acknowledgements

I would mainly like to acknowledge Prof. Sarah Ennis for her valuable supervision. Prof. Ennis shared her expertise and provided a solid academic and moral support throughout all the years spent in completing this work.

Also, I would like to thank Dr. Benjamin MacArthur and Prof. Mark Beattie for sharing precious and insightful knowledge on their own specific field of expertise.

I am also grateful to all the families that took part in the Southampton genetics of IBD study and research nurse Rachel Haggarty. Heartfelt thanks to Crohn's in Childhood Research Association (CICRA) and the NIHR Southampton Biomedical Research Centre.

I can not exempt myself from thanking my family for supporting me until here and for being a stronghold in my life.

Finally I wish to thank friends and colleagues for all the quality time spent together.

# Ethics approval

This study was approved by the Southampton & South West Hampshire Research Ethics Committee (REC) (09/H0504/125).

# Funders

This studentship is supported by The Faculty of Medicine Doctoral Training Fund and the Institute for Life Sciences Hillary Marsden Scholarship of the University of Southampton.

# Declaration of authorship

I, **Enrico Mossotto** 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.

**Advanced modelling of genomic data in Inflammatory Bowel Disease**

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as:

   **E Mossotto**, JJ Ashton, T Coehlo, RM Beattie, B MacArthur, S Ennis. - Classification of Inflammatory Bowel Disease using Machine Learning, *Sci. Reports* 7, 2427 (2017)

Signed: .....

Date: .....

# Chapter 1

# Introduction

In this chapter I will cover the current state of the art useful to ease the interpretation of this thesis. The first section discusses functioning and features of next-generation sequencing technologies, then a second brief section introduces basic concepts of machine learning and their applications to genomics. The last section introduces inflammatory bowel disease, its forms, classification and genetics.

## 1.1 The post-genomic era

Genome sequencing has became trivial and is increasingly replacing many classical genetic approaches, defining a new epoch known as post-genomic era. More than 10 years have passed since the introduction of the first next-generation sequencing (NGS) techniques and the ability to sequence life forms constantly increased. NGS is one of the last achievements in sequencing technology commenced by Sanger [161] in 1977. With the advent of NGS the amount of collectable data by sequencing [13] quickly jumped from hundreds of mega bases to over one thousand billion bases (>1 tera bases) (Figure 1.1). Unfortunately the interpretation of these data did not experienced the same growth leaving a substantial gap between available data and our understanding of it. A similar escalation was also observed in other

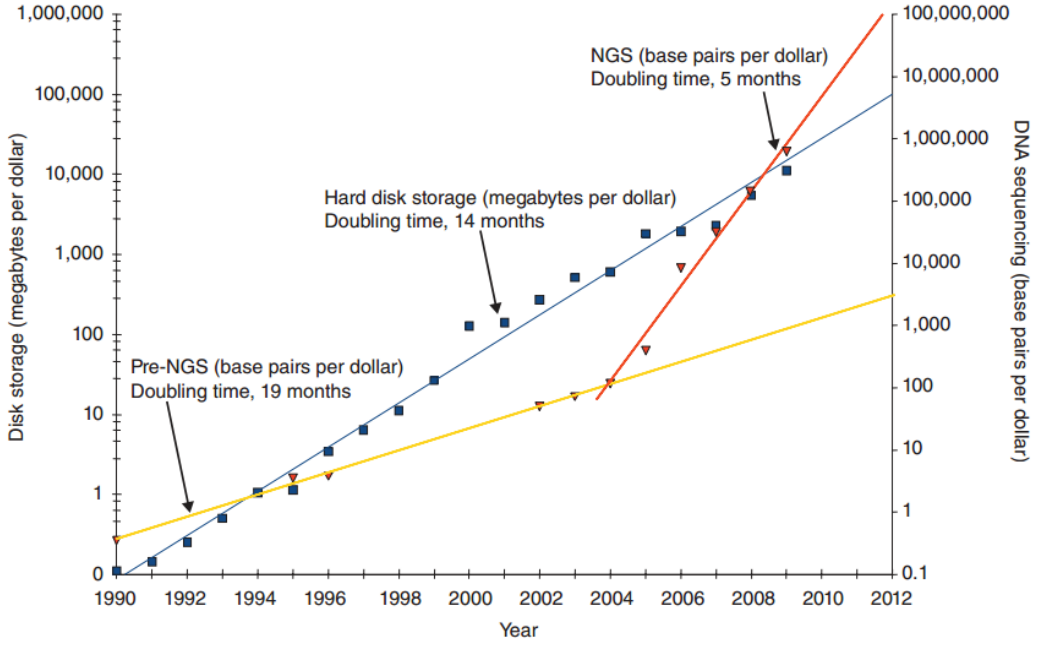


Figure 1.1: **Historical trends in storage prices versus DNA sequencing costs and year.** Image adapted from Stein *et al.* [178]

research branches generating a group of molecular data called "omics". Genomics, transcriptomics, proteomics, metabolomics, microbiomics, pharmacogenomics and so on, are the new frontiers of medicine and biology. Sensitive integration of all these data will provide a more accurate view of the mechanisms behind human traits and diseases. The application of such combined and extensive knowledge to medicine, known as precision medicine [60], aims to deliver targeted treatments which will results in a better and more efficient patient management.

### 1.1.1 Next generation sequencing

The advent of whole genome sequencing and whole exome sequencing made us aware that even the smallest and most subtle variation in our deoxyribonucleic acid sequence (DNA) may result in an adverse clinical manifestation. The main reason for investigating the human genome resides in the belief that it might hold most of the answers to almost every personal normal or pathogenic trait. As a consequence of NGS high-throughput we are now able to detect mutations within the human genome allowing more accurate investigations. Unfortunately,

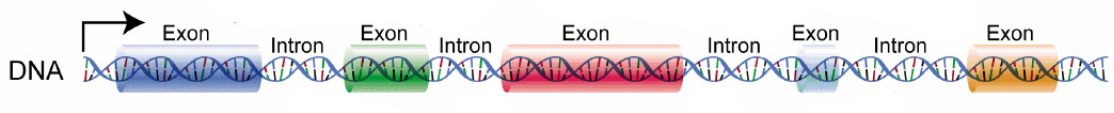


Figure 1.2: **Gene structure.** Coloured barrels represent gene exons.

the interpretation of these biological data is challenging and requires advanced analytical tools.

### The human genome

The human genome is a long sequence ($3.2 \cdot 10^9$) of repeated nitrogenous bases (nucleotides): adenine (A), guanine (G), thymine (T) and cytosine (C). This sequence is organised in 23 pairs of chromosomes and presents regions with a very specific roles and importance. One of the possible ways to organise the human genome is by the function of specific genomic region. DNA sequences used to generate proteins are defined as genes. Genes are as well divided in coding and non-coding regions that are respectively called exons and introns (Figure 1.2). The collection of all coding parts of the human genome is known as exome. When the cell triggers the production of a particular protein, the relative gene is transcribed in order to synthesise a messenger RNA (ribonucleic acid) sequence. RNA is a polymeric molecule similar to DNA that directs the synthesis of a specific protein.

After the transcription, introns (non coding parts within a gene) are removed from the mRNA sequence using a mechanism called splicing. Once introns are removed, the protein synthesis begins and dedicated machineries (ribosomes) translate the mRNA sequence into an amino acid sequence. This process is known as translation whereby triplets of nucleotides are translated into amino acids, the fundamental elements of a protein.

Only 2 percent of the human genome codes for proteins while the remaining 98 percent is involved in regulatory or spacing functions (e.g. telomers) [21]. In the late seventies, the attention was focused on coding regions and the remaining DNA

was labelled as "junk DNA" since it was apparently not involved in any function [58]. With the advent of sequencing technologies, regulatory regions, non-coding RNAs and other functional sequences were discovered in this non-coding part of the genome renouncing this previous label.

The availability of next generation sequencing techniques revealed the complexity of the human genome and made possible to generate a map of the human reference genome. The reference genome is not representative of a single individual but is made by collecting consensus sequences from hundreds of people. The better the NGS technique, higher the number of sequenced individuals, more accurate the reference genome [130]. Comparing any human genome to the reference [71], about 4 million discordant nucleotidic bases [91] can be detected using whole genome sequencing and around 25,000 [36] with whole exome sequencing. While whole exome sequencing uses some specific tools to capture only exonic sequences, whole genome sequencing can also map mutations within non-coding regions of the genome. Different types of mutations can be found in a genome and are classified into small and large scale variations. While large-scale mutations can change the structure of entire chromosomes, small-scale mutations involve a small number of nucleotides. Single nucleotide variation (SNV) is the term used to refer to mutations that afflict only one nucleotide and depending on the effect on the coding region it is possible to define the following types:

- Synonymous single nucleotide variation (SNV): a single base in the genome is mutated but does not affect the protein translation;
- Non-synonymous SNV: a different amino acid is coded and the protein sequence is altered;
- Stop gain SNV: the mutated base results in a premature stop of the amino acid sequence synthesis;
- Stop loss SNV: the mutated base results in a missing termination of the protein synthesis;

- Frameshift indel: the open reading frame (the way bases are read in triplets) is shifted due to new base(s) insertion or deletion;
- Non-frameshift indel: the insertion or deletion of bases does not change the open reading frame;
- Splicing-affecting SNV: the mutation falls within a region important for a correct splicing of introns and exons.

A specific nomenclature is also used to describe the frequency of a mutation in the human population. A mutation with a frequency greater than 1% among the population is called single nucleotide polymophism (SNP). Rare variant, instead, are defined as mutations that occur in less than 5% of a population. The International HapMap Project [57], is the reference consortium responsible for mapping human genetic variation.

As a consequence of the multitude of sequencing projects, it is now known that some regions of the genome are formed by a precise repetition of nucleotides (motifs). The alteration of the number of these repetitions represent a particular type of mutation called copy number variation (CNV). Unfortunately, the identification of copy number variations (CNVs) through exome sequencing is still a challenging task [81]. The limitation of exome sequencing in detecting CNVs resides in the insufficient length of the DNA fragments that are sequenced, conversely, whole genome sequencing is typically based on longer fragments allowing a precise detection of CNVs.

Humans are diploid organisms which means that every individual carry two copies of the same chromosome, paternal and maternal, and variants may be observed in one or both copies. If the mutation is observed in both chromosomes, the genotype is homozygous, while if the mutation is only in one chromosome, the genotype is heterozygous.

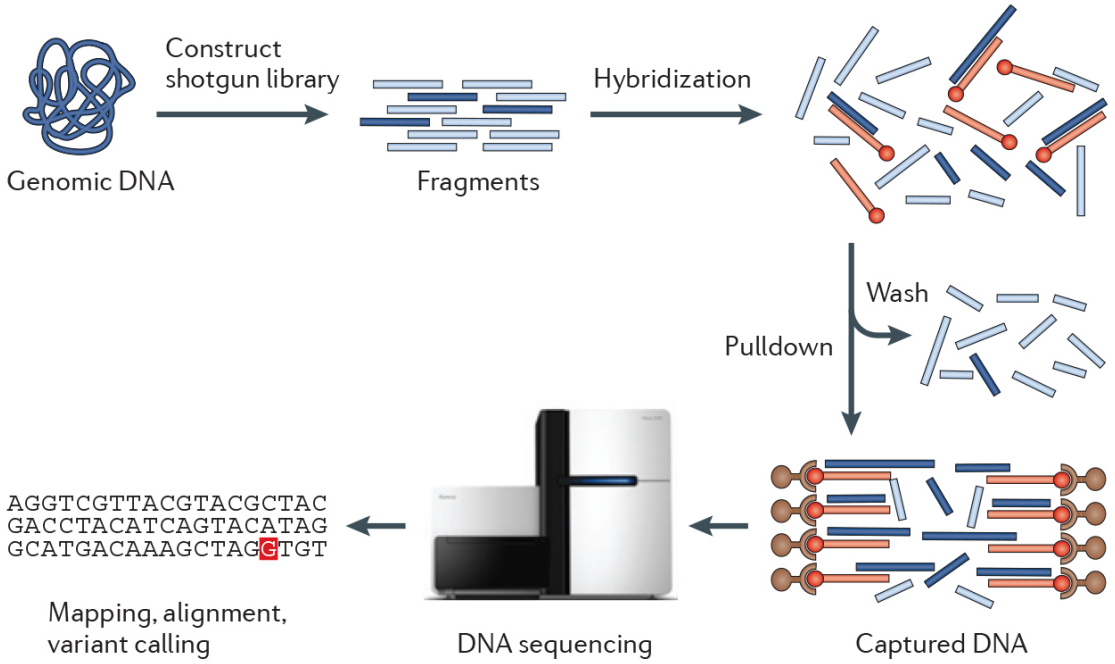


Figure 1.3: **Whole exome sequencing workflow.** Image adapted from Bamshad *et al.* [15].

### Whole Exome Sequencing

Whole exome sequencing (WES) is a next-generation sequencing technique that combines a powerful high-throughput approach and enrichment for selected coding regions of the genome. Steps required in a whole exome sequencing approach (Figure 1.3) are similar to ones performed during whole genome sequencing which consist of:

1. Library preparation;
2. Amplification and Enrichment;
3. Sequencing;
4. Data Analysis.

In the library preparation step, genomic DNA is fragmented to dimensions determined by the sequencing platform and also by the type of readings that is willing to produce. After the DNA fragmentation, linker sequences are bound to fragments with two possible approaches: single or paired end (Figure 1.4). With single-end

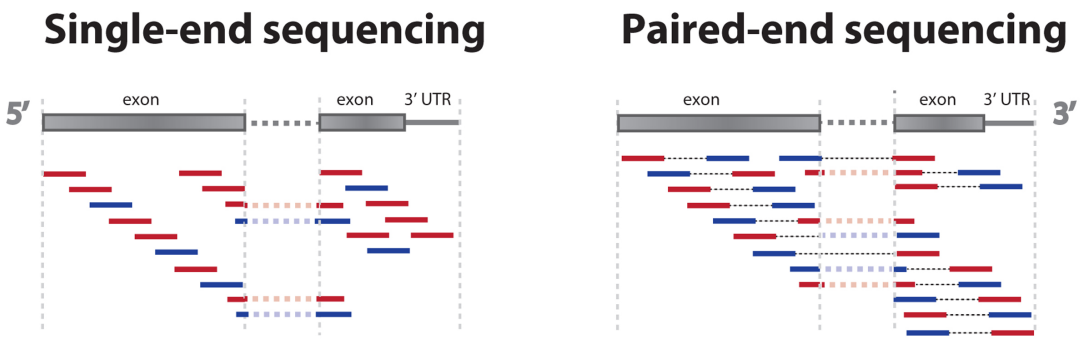


Figure 1.4: **Single and paired end configuration.** Image adapted from Zhernakova *et al.*[205]

linkers, only one end of the fragment is linked to a tag sequence and therefore sequenced along only one direction. With a paired-end approach, following the DNA shearing process, fragments with particular known length are selected and then linked to tags at both ends. With paired end tags both ends of a fragment are sequenced and, knowing the distance between tags increases the efficiency in mapping fragments on the correct genomic location of the reference genome. An accurate selection and shearing of fragments to the correct size is crucial for optimal mapping. Whole genome sequencing and whole exome sequencing usually requires different fragment sizes that are respectively 400 and 200 bases in length (Figure 1.5). If the distribution of insert sizes deviates too much from a normal distribution it will likely introduce bias in the alignment process. Observing figure 1.4 it is quite evident that in both cases (single or paired ends) the overlap of reads makes the sequencing process redundant. Indeed, the redundant alignment of multiple reads to the same genomic location increases the confidence and the quality of sequencing. The number of overlapping reads across the genome generates a measure that is known as coverage. In order to ensure that sequencing data reflect the true genomic sequence, the coverage must range between 10 to 30 reads per locus, with the latter representing the minimal requirement for clinical applications [182].

Following tags ligation, other adapters are linked to fragments providing the starting points for the amplification step. In the amplification step a polymerase chain reaction (PCR) is performed in order to increase the number of fragments that will be then sequenced, resulting in an increased signal strength.

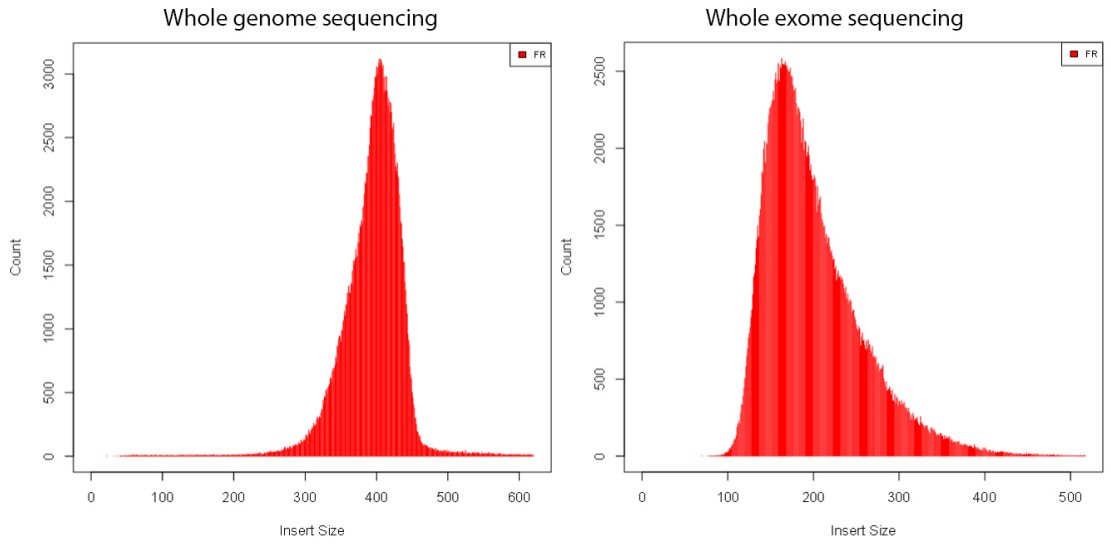


Figure 1.5: **Comparison of insert size distribution between WGS and WES.**

Following the PCR, the sequencing step begins. There are several methods for sequencing the DNA, however it is possible to group them in three classes: *(a)* sequencing by synthesis; *(b)* sequencing by ligation; *(c)* other minor methods (pyrosequencing, nanopore, ect...) In a sequencing by synthesis approach each DNA fragment is read by synthesising a complementary fragment using fluorescent modified nucleotides and a polymerase enzyme. Each fluorescent nucleotide corresponds to a different colour, by observing the colour sequence it is possible to determine the nucleotide sequence. Instead of synthesising nucleotides, the sequencing by ligation approach uses custom small nucleotidic (oligonucleotides) sequences and a ligase enzyme. Each oligonucleotide has a fluorescent element which is read after the ligation to the fragment.

Regardless of the method used, the functioning mechanism of the most common sequencing platforms consist of reading a different light signal (wave length) whenever a particular base is observed. However, some new platforms (e.g. Ion Torrent and Ion Proton) escape this canonical detection method by instead observing the ions released during the sequencing by synthesis process.

After completing the sequencing step, a file containing all the observed bases and the relative per base quality (the confidence of the reading), is obtained and ready to be analysed with bioinformatics tools. Depending on the sequencing

technology, files containing sequencing data may have different format. However, the most common FASTQ format [40] is becoming the standard output chosen by most companies. This format is made of 4 lines per each sequenced read:

1. read sequence ID preceded by a "@" character;
2. raw sequence in letters;
3. "+" as spacing character for optional notes;
4. quality values for sequence in line 2;

```
@HISEQ2500-09:25:C3F9YACXX:7:1101:8130:1970 1:N:0:ATCACG
NATCGCCGTTTGTTCTTTTCTTGGCTTTTCTTTCTGGAGGTCTAATGTTTCTTTCTCTTCTGCAGATATGCACCCAATGTCCAGGGAGCGTGCAAAATTT
+
#1:ADDDD@@;?:;4<EEDF@@4?B*?;CCDEDDD>D<D)9@@*?<BBDBDACDEIEIEEECDCICIACAAADCD(..7;>BAAA?(9;?38??>><3>A
```

Figure 1.6: **Example of a 4 line FASTQ format.**

The quality values are expressed through ASCII characters, each of which indicate a specific probability that the corresponding base call is incorrect. This score is also known as Phred quality score and the encoding character/score depends on the sequencing technology.

The aim of exome sequencing is to target and sequence specific regions coding for proteins [16], reducing both the volume of output data and analytical costs. In order to collect all the coding parts of the genome, several methods have been proposed [117]. The most common method applied to preferentially select exons is the *in-solution* enrichment. This approach uses a pool of custom probes that are hybridised in solution to the fragmented genomic DNA. Then, using capturing beads that recognise a specific element on the probes, exonic region are retained while non coding regions are washed out. Once the beads have been removed, coding fragments can be sequenced with the process just described (Figure 1.3).

## 1.2 Machine learning algorithms

When dealing with complexity, some problems cannot be solved only through classical experimental approaches. Thus, we need to appeal to other technologies such

as artificial intelligence (AI), software capable of analysing extensive datasets and to adapt their behaviour in order to keep performing properly as the data changes. AI, like sequencing, moved its first steps more than 80 years ago and was founded as an academic discipline in 1956 at the Dartmouth Summer Research Project on Artificial Intelligence [120]. Machine learning evolved from this field providing a new set of statistical and analytical tools for the analysis of complex data. Early machine learning algorithms date back to the early sixties when giant steps in artificial intelligence and computing were made. The aim of these algorithms is to find patterns within data and once found use those rules to predict or classify new data. Machine learning algorithms belong to data mining and computational statistics, fields where classic statistical methods meet the complex problem of extracting information from oversized dataset, also known as "Big Data". These algorithms were proved to be incredibly powerful when dealing with complex data, where hundreds of thousands variables have to be analysed simultaneously. Basic statistical tests could be still applied, however they have been overtaken in terms of speed and complexity. Due to their ability to work with complex dataset, made of multiple levels of information, machine learning algorithms were widely applied in biology and more generally in scientific research to extract patterns and to predict outcomes. Thanks to this polyhedric behaviour, many different algorithms have been developed, each of them with specific strengths and weaknesses. So far, the ultimate model that could be successfully applied in any scenario and on any type of data does not exist. For this reason, when approaching a machine learning problem it is recommended to test different models, acknowledging that it is impossible to predict which algorithm will perform optimally. Coupling this technologies with NGS can lead to new powerful ways to analyse complex data extracting more intelligible information.

The role of machine learning algorithms and, more in general, statistical learning is to find patterns in observed data and infer or predict new data. Considering a dataset made of hundreds of samples, represented by a multitude of variables, statistical learning tries to understand and predict mechanisms that correlate all those variables with the observed outcomes. The correlation, can be calculated

with different approaches whereby the first historical method formulated was defined by C.F. Gauss in 1795 as the *method of least squares* and then published by A.E. Legendre [180]. This approach was the first modern method for observing trends in big dataset. This revolutionary concept was based on the idea of searching for a function capable of describing the observed data and therefore used to predict future events (Figure 1.7). This is still today one of the pillars of machine learning algorithms. However, finding the perfect function that fits perfectly the observed data without any error, is impossible as infinite number of equations should be tested. The limits occurring when fitting a function to real data can be described by two types of errors: *reducible error* and *irreducible error*. While the function can be adjusted to decrease the reducible error choosing better parameters, it is impossible to lower the irreducible error. This limit is the reflection of the lack of knowledge about the dataset represented by unmeasured variables or variables that can not be physically measured. Suppose to observe a an output $Y$ and a set of variables $X_1, X_2, ..., X_n$ for each observation in a database. The aim of machine learning is to describe the relationship between $Y$ and $X$ using a function $f$. This can be generally written as:

$$Y = f(X) + \epsilon \tag{1.1}$$

Here $\epsilon$ denotes the error that separates the function $f$ from the perfect representation of the outcomes $Y$. A function with the mean of all the errors equal to zero has a perfect fitness.

The reasons of the interest in estimating a correct function are two: prediction and inference. In both cases the search space for functions capable of describing data is infinitely large. This concept can be expressed as the search of $\hat{f}$ that produces a set of predicted outcomes $\hat{Y}$ as close as possible to the set of true outcomes $Y$.

$$\hat{Y} = \hat{f}(X) \; with \; \hat{Y} \approx Y \tag{1.2}$$

The difference between prediction and inference is in the knowledge that we are trying to obtain from the model. When predicting an outcome, there is more

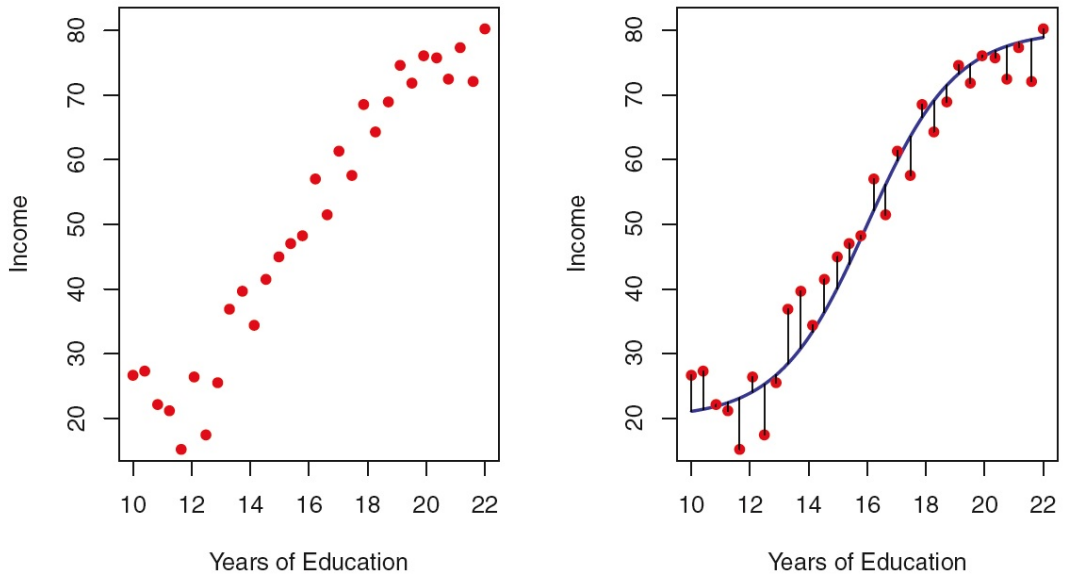


Figure 1.7: **Function fitting.** Example of fitting the correct function to the observed data. The red dots are observed incomes with respect to the years of education. The dataset was simulated, thus the best function (blue line) was known. Black lines represent the error $\epsilon$. Figure adapted from James *et al.*[76].

interest in obtaining the right prediction rather than understanding the mechanism that allowed the result. Conversely, the objective of inference approaches consist of achieving a higher understanding of the role of variables and their relationship to the outcome. In the first case, $\hat{f}$ can be treated as a *black box* while in the latter case, it is important to understand the functioning of it.

These concepts only describe the statistical side of a machine learning approach to data, whereas the rules for improving the fitness (learning) represent the core aspect of ML. By definition, a learning process is an automatic or semi automatic procedure in which the algorithm corrects its outputs, without external supervision or knowledge. *Learning* does not only refer to a memorizing process but it also involves plasticity, the ability to adapt to different inputs and then giving the best possible answer. As stated by I. H. Witten and E. Frank [200]:

'Things learn when they change their behaviour in a way that makes them perform better in the future'

A learning process is a set of rules that the algorithm continuously updates as it receives more data, the more training data the better the algorithm will be at learning.

### 1.2.1 Types of learning

The large variety of machine learning algorithms is a direct reflection of the broad applicability of such algorithms. ML algorithms can be divided in two broad classes: *supervised learning*, when the data we would like to model is labelled; and *unsupervised learning* when no labels are given and the model has to find an optimal clustering system.

**Supervised learning** Most of the background so far discussed refers to supervised algorithms which have a simpler interpretation and better understanding of the tasks that these algorithms try to accomplish. Supervised learning algorithms are principally devoted to prediction and their learning process is based on the comparison of the predicted labels to the provided one. This imply that input data come with a parallel information about true output of each element. Many algorithms have been developed to solve prediction and inference problems ranging from the classical linear regression to more modern methods like support vector machines, decisional trees and artificial neural networks. Supervised learning algorithms have been broadly applied to many fields including biology [171] and health care [114] [129]. The interest of these fields in machine learning approaches is recent and quickly scaled with the discovery of new high-throughput technologies delivering enormous amount of complex data. Correct predictions in biology and in healthcare could help scientists to treat patients better and quicker.

**Unsupervised learning** Unsupervised learning approaches are more challenging and harder to interpret. Since the real correct output of samples in the dataset is not known, the algorithm tries to represent and reduce the complexity of data by observing relationships between samples and variables. This process is known as *cluster analysis*, or clustering. Different unsupervised algorithms explore different mathematical approaches to the same problem. Using unsupervised learning algorithms made possible observing clusters in complex data like genome-wide polymorphism data. Figure 1.8 provides an example of the power of these mod-

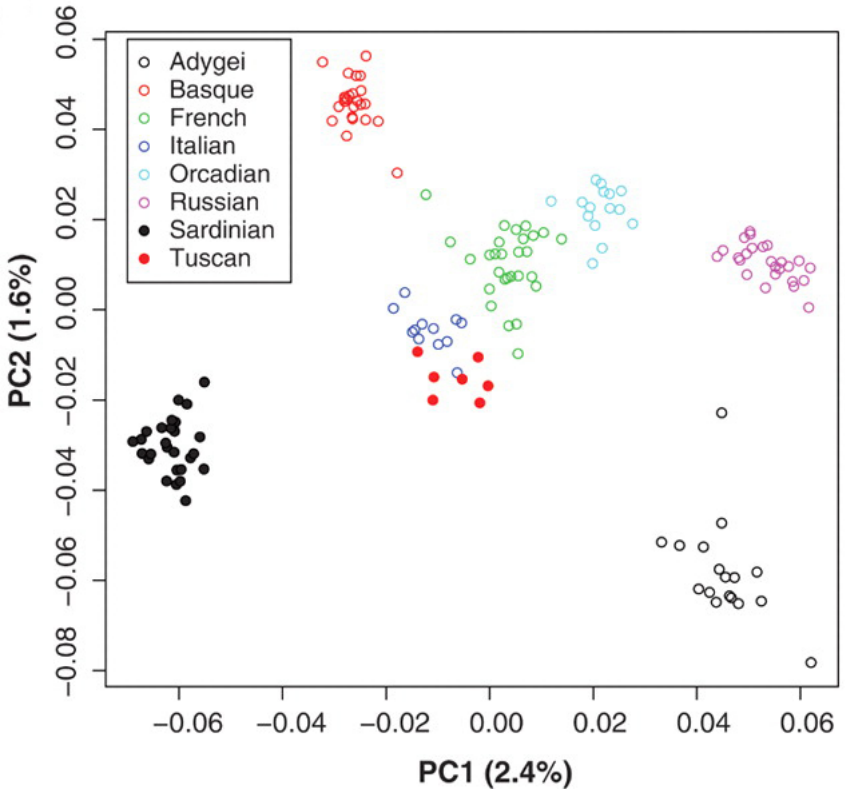


Figure 1.8: **PCA clusters ethnic groups.**Clustering of eight European populations using all autosomal SNPs. Figure adapted from Li *et al.*[107]

els in the identification of different ethnic groups using frequent single nucleotide variants (SNPs) [107]. In this model, the position of an individual is determined by the scores calculated by a principal component analysis (PCA) algorithm using more then 650,000 common SNPs. Clusters correlates with the ethnicity of the individuals and are highlighted with different colours.

### 1.2.2 Data sets and model validation

**Learning and testing** Data are the central elements of a machine learning analysis and the efficiency in extracting information might depends on data quality. "Garbage in, garbage out" is a common notion in machine learning to encourage people avoiding the analysis of faulty or not refined data in order to prevent obtaining results that are not reflecting the truth. Dataset preparation is therefore a fundamental step of any analysis. Depending on the selected method (supervised or unsupervised) the whole dataset bay be split in smaller partitions that will be

used during different steps of the machine learning approach. *Learning* and *testing* subsets are two fundamental partitions for a supervised approach. Using the learning subset, supervised methods fit data to identify a function optimally representing the input (fitting or learning). The algorithm ends the learning process after reaching the highest possible accuracy or after an arbitrary number of test-and-error corrections. Subsequently, the accuracy obtained during the learning process directly reflects algorithm's performance on fitting that specific dataset. In order to assess whether the model is capable to generalise its prediction over new data, a second step of validation is needed. Such validation is performed over the testing subset of the original dataset that was not used for training the model. If the starting database was too small to be divided into smaller subsets, the model can be still verified using a additional unseen data or validation techniques that involve different ways of testing during the learning step. On the other hand, unsupervised methods do not require labelled data and there is no prior knowledge to be used to verify the predicted output.

**Resampling methods** Fitting a machine learning model is usually the key step that determines whether the model will produce useful predictions or not. Among the large number of variables that may influence the model performance, variability of input data is the most important. In order to train a model that will correctly generalise on new data, many resampling methods have been proposed. Resampling methods consist of iterative fitting and testing on subsets of the original dataset. This approach is solid from a statistical point of view, however it is also computationally expensive since it entails fitting the model several times. Considering the advantages in terms of performance obtained by investing time and computational power, resampling methods are essential tools for any machine learning approach. Two of the most used methods are *cross-validation* and *bootstrap*. These cross-validation approaches will be better covered in the following chapter.

### 1.2.3 Feature selection

In ML approaches features are variables that can be used by an algorithm to make predictions or inference. Features can differ depending on the dataset used or on the algorithm applied. For example, when applying machine learning to clinical datasets, features are usually binary (0/1) variables that tells whether an individual have been treated with a particular drug or not. In some other cases, features can be continuous or discrete values. Not having limitations on the input data format, variables such blood test results, clinical history, genomic traits and many others, can be simultaneously implemented in ML models. This powerful adaptation to various inputs made machine learning algorithms perfect for biological problems, where multiple and different features have to be modelled at the same time. However, even if the variety of features offers a precise and detailed description of data, machine learning algorithms have been proved to perform better when provided with a pruned and selected number of features. The improvement in performance is a direct consequence of removing those features that were bringing no information other than noise. This noise can be attributed to multiple factors such as batch effect, human errors and random noise. This pruning process removes sources of meaningless variation from raw data and unmasks potential underlining biological patterns.

In ML approaches, features are usually implemented as dimensions although not all algorithms scales well to high-dimensional data, leading to exponential increase in computational time and biased data distribution over the space. This limit is called curse of dimensionality [20] and depends mainly on the algorithm used. The most important effect of curse dimensionality on data is the over spreading of data into many different space dimensions, nullifying any chance of observing patterns or cluster of similar objects. This problem does not only affect machine learning approaches but is also observed in any data analysis approach including database organisation, artificial intelligence and combinatorics. In machine learning approaches the effect of this problem, called the Hughes effect [141], results in the need of an enormous amount of training data to observe meaningful patterns.

With a fixed number of training examples the prediction power declines as the number of features increases. Whilst this limitation is not of particular concern for classification and regression problems, it represent a major issue in density prediction approaches. Several algorithms were developed in order to overcome the curse of dimensionality. A reduction/pruning of high-dimensional data can be performed through principal component analysis, multidimensional scaling or other algorithms that reduce the number of features by applying statistical tests. In machine learning, the ideal ratio between training examples and feature has been empirically derived and is usually around 5 examples per feature [176], informally named "*5 to 1 rule*". However, due to the nature of biological 'omics data, it is rare to observe applications following this empirical rule.

## 1.3 Machine learning applications to genomics

With the increased volume of biological data, machine learning (ML) is increasingly applied as a powerful tool for the analysis and interpretation of "Big Data". Its application ranges across different areas of biology such as cancer research [193] [31], drug discovery [109] [105] and genomics [125] [24] [118]. One of the first application of machine learning to biology dates back to 1986 with the work from Klein on predicting the secondary structure of proteins [89].

The ability of machine learning algorithms to look simultaneously at several interdependent variables make them appropriate for the description of biological systems, complex diseases and patient stratification. In order to better describe these aspects of research, it is important to collect longitudinal data using different techniques. Thanks to better technologies for collecting 'omics data (genomics, transcriptomics, proteomics, metabolomics, etc...), large longitudinal datasets representing different aspects of a disease, a patient's condition or a biological sample are becoming more accessible.

**Prediction of genomic elements** Machine learning algorithms have been extensively applied in genomics, in particular with the scope of predict and identify of protein coding regions in the human genome. Salzberg *et al.* [160] trained a tree classifier for the identification of new genes. This algorithm was trained and tested on known coding regions from the GenBank database and used 21 coding measures as features, some of which were nested. The overall accuracy in detecting coding and noncoding regions was 83.7% on DNA sequences 108bp long.

Machine learning algorithms have also been applied to detect splicing regions [33] and to predict splicing products [203]. A Bayesian deep learning model was trained to score the impact of mutations on splicing events. The resulting model was able to determine the deleteriousness of SNVs and to predict unexpected aberrant splicing leading to clinical conditions with a 94% accuracy (tested on causal splicing variants in spinal muscular atrophy and colorectal cancer genes).

RNA structure [24] [87] and expression quantitative trait loci (eQTL) have been also investigated using machine learning algorithms. Ackermann [2] proposed a model for a quicker identification of eQTL regions that can link genotypes and traits. In this supervised approach, random forest and LASSO machine learning algorithms were coupled performing better not only than their standalone version but also compared with other existing models.

Since 2010 the Critical Assessment of Genome Interpretation (CAGI) association has organised competitions for interpreting genomic data [28]. Whole exome sequencing data from several clinical conditions were collected and provided to competitors. The aim of CAGI competitions ranges from the discrimination of healthy and affected individuals to the identification of disease subtypes using WES data or novel genomic elements. While the identification of causal variants in individuals affected by rare diseases using WES is well established in clinics, the interpretation of large scale WES data in complex diseases (although less computational demanding than whole genome data) is still challenging. Using WES from hundreds of individuals, new computational and methodological limitations arise. For this reason, only few examples of WES data modelling using machine

learning models are available.

**Merging clinical and genomic data** Although machine learning approaches to genomic or clinical data are increasing in number, few studies apply ML using both type data. This approach requires complete and well curated clinical data, where for each individual, both clinical and genomic data are available. A recent study by Sio-Wee Chang et al. [34] provided a prognostic tool for oral cancer using both clinicopathological and genomic markers. The algorithm was challenged to predict the survival chances over three years for affected patients. Since the number of features did not fit the ideal 5:1 ratio, different machine learning algorithms were coupled with a range of feature selection algorithms. The starting dataset was composed of 31 oral cancer patients with almost complete clinicopathological description including social demographic data, clinical data and pathological data. Alongside this kind of information, immunohistochemistry was used to assess the activity of two genomic features, respectively TP53 and TP63 genes. The choice of this type of genomic data was made following literature search for the most commonly correlated genes with the oral cancer phenotype. The *TP53* and *TP63* genes are two of the most studied genes in cancer science since their role in the cell is to regulate the cell cycle and proliferation. They both fall into the category of tumour suppressor gene and negatively regulate the cell division by controlling a set of genes required for this process. Firstly, all the variables were converted to numeric values allowing an easier computation. In this study five feature selection methods were used: *(a)* Pearson's correlation coefficient (CC); *(b)* Relief-F; *(c)* genetic algorithm (GA); *(d)* CC combined with GA and *(e)* Relief-F combined with GA as hybrid approach. Using feature subsets selected by each of these approaches, 4 machine learning algorithms were taught: ANFIS [78], artificial neural network (ANN), SVM and logistic regression. Results indicated an overall improvement in the prediction accuracy when genetic variables are included in the feature subset regardless of the ML algorithm used. These results are just some few examples of the improvement that genomic data can provide in a classification problems.

## 1.4 Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is an umbrella term for the common gastrointestinal auto-immune diseases ulcerative colitis (UC) and Crohn's disease (CD). In 2013 the annual incidence of IBD in the United Kingdom was calculated as approximately 400 new cases for every 100,000 individuals [135]. As a consequence, the cost of IBD on health-care is expected to grow exponentially leading to increasing managing problems [85]. For these reasons it is important to understand the underlying mechanisms of this disease, in order to deliver a more targeted and effective treatment on a per-patient per-episode basis. Here, personalised medicine will not only be beneficial to patients but will also produce a positive economic impact on health-care systems.

Diagnosing IBD is difficult, the aetiology is not fully understood and the process for selecting the most correct treatment is protocol led but controversial particularly for complex cases . The main attribute of an inflammatory bowel disease phenotype is the chronic inflammation of the gastrointestinal (GI) trait, with localisation and severity being major factors in disease clinical presentation, disease classification and treatment decisions. Symptoms of IBD include diarrhoea, fever, abdominal pain, blood in the stools and weight loss [149]. Besides their medical relevance, symptoms of IBD have alters the normal social and working life of affected individuals, leading to a considerably lower the quality of life, higher health care, psychological and educational impact.

IBD subtypes, Crohn's disease (CD) and ulcerative colitis (UC), are distinguished though endoscopic and histology examinations. Endoscopic investigation consist in a macroscopic observation of the GI trait without requiring the collection of specimens and represents the standard examination for diagnosing a suspected IBD case. Unfortunately, it is not always possible to assign a diagnosis solely on endoscopic evidence and, therefore, a further histological exam is required. Opposite to endoscopy, histology requires the collection of multiple specimens from the patient intestine since it consist in a microscopic investigation of those tissue. Samples are stained and analysed, requiring more time and consumables compared

to an endoscopy exam. Thanks the higher resolution, histology can resolve the diagnosis for most of those patients where endoscopy results were inconclusive.

Crohn's disease is characterised by a non-localised inflammation of the gastrointestinal system, while the inflammation pattern in ulcerative colitis is more often continuous and restricted to the colorectal trait of the intestine (Figure 1.9) [194]. Alongside the inflammation localisation, other minor and sometimes subtle traits differentially characterise the two forms (Table 1.1). Despite these differences, clinical traits does overlap and increase the uncertainty of the diagnostic process. As mentioned before, endoscopy and histology offer two level of details, respectively macroscopic and microscopic.

Table 1.1: **Characteristic traits for UC and CD** [194].

| | **Ulcerative colitis** | **Crohn's disease** |
|---|---|---|
| Endoscopy | Ulcers | Ulcers |
| | Erythema | Cobblestoning |
| | Loss of vascular pattern | Skip lesions |
| | Granularity | Strictures |
| | Spontaneous bleeding | Fistulas |
| | Continuous distribution | Segmental distribution |
| Histology | Mucosal involvement | Submucosal or transmural involvement |
| | Crypt distortion | Crypt distortion |
| | Crypt abscess | Crypt abscess |
| | Goblet cell depletion | Granulomas |
| | Mucin granulomas | Focal changes |
| | Continuous distribution | Discontinuous distribution |

Currently there is no cure for IBD, but it is possible to treat the disease symptomatically. Treatments are decided depending on the symptoms and localisation of the disease according to the National Institute for Health and Care Excellence guidelines. These procedures can be ordered on a invasiveness scale. The nutritional approach is the least invasive and consist in changing to liquid diets based on specific formulae. These formulae force a controlled nutrition with low-complexity components. The nutritional approach is a valid alternative to steroid treatments which might lead to adverse effects. If the controlled nutrition is not sufficient to treat IBD symptoms, then a pharmacological approach is needed. Drugs used to treat IBD target the patient's immune system by reducing its response and,

depending on the severity of the condition, it is possible to choose between anti-inflammatory, antibiotics, steroids and antibody-based drugs. When drugs can not control IBD symptoms, a surgical intervention is needed. Usually, only the worst IBD cases undergoes surgery where the rectum and part of the colon are removed. Due to the scattered inflammation pattern of CD, surgery is less effective compared to the results obtained on UC patients.

### 1.4.1 Disease classification

Many classifications have been proposed in order to address clinicians to the most correct treatment, adapting the criteria according to new scientific breakthroughs. The first worldwide recognised criteria for classification for IBD was made in 1991 in Rome by the International Working Party and aimed to distinguish IBD subtypes. Crohn's disease (CD) and ulcerative colitis (UC) were the chosen classes and were discriminated depending on the anatomical distribution of the inflammation, clinical behaviour and operative history presented by the patient. However, the Rome classification presented different issues and was then revised in 1998 in Vienna[56] and then in 2003 in Montreal introducing age of onset as important variable alongside disease location and behaviour [163]. This classification was widely adopted but still not capturing the dynamic evolution of paediatric IBD. In order to overcome this problem, a new revision was proposed in Paris in 2011, distinguishing early-onset (less than 18 years old at diagnosis), very-early-onset (less than 6 years old) and infantile-onset (less than 1 years old) cases [99]. Unfortunately, it is not possible to take into account histological evidence due to the time required to perform those tests and invasiveness. Therefore, the Paris classification, like the previous classification, does not rely on histological evidence, which are frequently the key for an accurate diagnosis.

Although the classification system is continuously revised, the distinction between the two subtypes (Crohn's disease and ulcerative colitis) is often unclear and it is not always possible to assign a definitive diagnosis. Such cases are often referred as inflammatory bowel disease undetermined (IBDU). This uncertainty is mostly

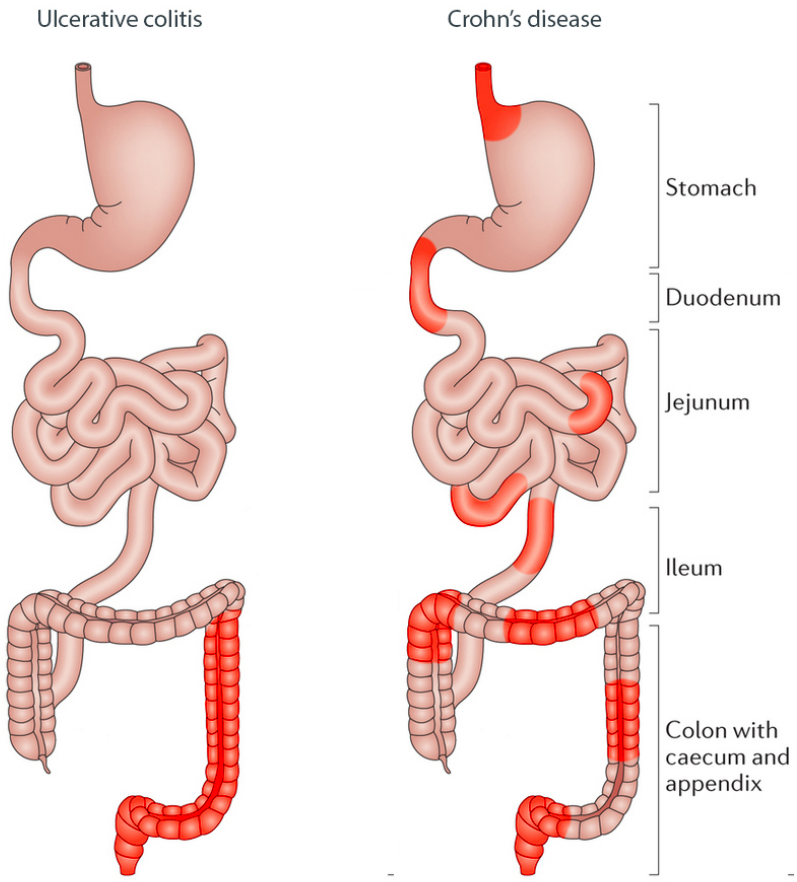


Figure 1.9: **IBD inflammation patterns.**Different inflammation patterns in ulcerative colitis (left) and Crohn's disease (right). Red highlighted areas represent possible localisation of the inflammation.

driven by the overlapping symptoms between subtypes and it is also not unusual to observe different subtypes within the same family pedigree.

IBD is a complex disorder and multiple factors are responsible for the overall phenotype. Thus, the pathogenesis of IBD can not be explained only by genetics and environmental factors may also alter the predisposition to disease in susceptible individuals [12]. Amongst these non-genetic elements we can include diet, smoking and the composition of the gastrointestinal microbiome (the bacteria naturally present in the human intestine). Assessing the effect of environment on IBD patients is challenging and the genetic component is still the main factor for disease characterisation. Thanks to the advent of new sequencing technologies, it was possible to identify IBD susceptibility loci explaining the positive familial history in 8% of IBD patients [25].

The need of considering the age of onset in the IBD classification is explained by the complex nature of the disease. Despite inflammatory bowel disease arises as a consequence of both genetic and environmental factors, individuals with rare and deleterious mutations are more likely to develop IBD in the first years of their life. In adults, instead, the IBD phenotype may be explained by multiple mild mutations and environmental factors. Moreover, this difference can also be explained by the different time that adult and young individuals are exposed to environmental hazards and unhealthy lifestyle. While adult forms of IBD is a concerto of genetic and epigenetic causes, paediatric IBD (PIBD) is mostly driven by genetics [194].

### 1.4.2 Genetics of IBD

In the last decades many disease revealed to be caused by multiple concurrent genetic mutations and environmental factors [72] being then labelled as complex diseases. These disorders do not have a clear inheritance pattern and escape classical Mendelian rules of inheritance. Unlike monogenic disorders where single genes are responsible for causing the phenotype, complex diseases are caused by multiple genetic factors. IBD is a complex polygenic disorder where the genetic component is accompanied by other risk factors such as immune dysregulation, altered microbial flora and a variety of environmental variables. Since all this elements play a role in defining the IBD phenotype, isolating the genetic variable and identifying causative or associated elements has been historically challenging.

Despite the lack of a clear understanding of the IBD aetiology, recent studies proved the genetic contribution to IBD being highly variable and having a direct impact on the disease onset [139]. The most accredited model depict the IBD phenotype as the interplay of genetics and the environment in respect to the age of a subject. This translates to a direct correlation between disease onset and genetics with very-early onset IBD showing the largest genetic component and the smallest environmental factor. Conversely, with adult onset IBD presentations, the genetic component has a minor role compared to the impact of environmental

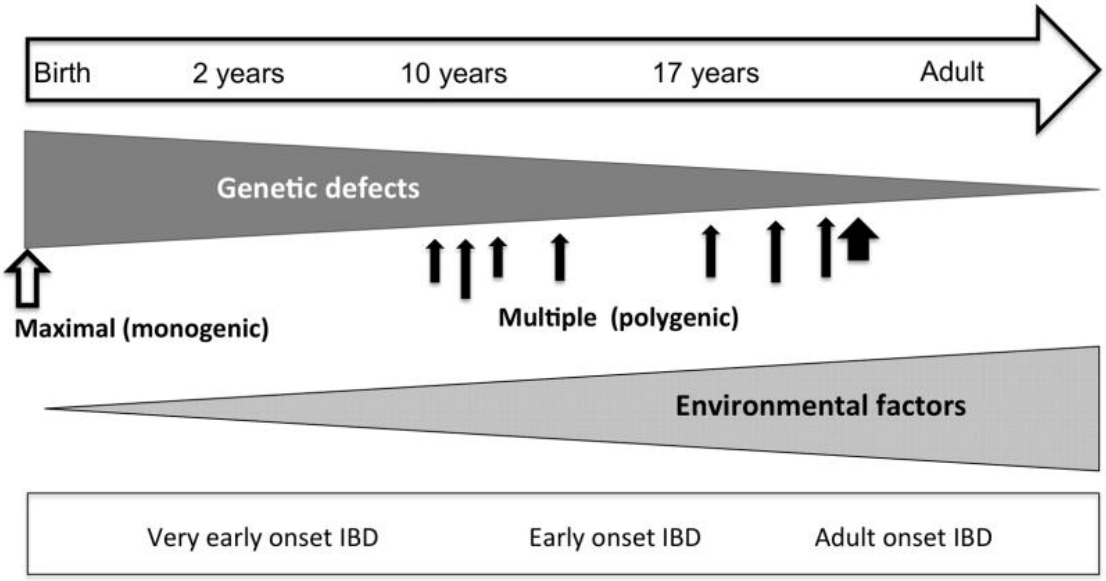


Figure 1.10: **Genetics and environmental contribution to IBD onset.** Very-early onset IBD presents more often as a monogenic condition whilst late onset IBD can be explained by a concurrence of mutations and environmental factors.

risk factors. Through this model is therefore possible to motivate both monogenic and polygenic IBD presentations (Figure 1.10).

Very-early onset cases of IBD deviate from the polygenic model by presenting a genetic profile much closer to a monogenic/oligogenic form. Here, highly deleterious mutations in one or few genes can trigger the IBD phenotype with stronger symptoms. However, treating the monogenic and oligogenic forms of IBD can be easier than treating its polygenic form thanks to its more predictable genetic behaviour. One of the most successful results in treating a monogenic and paediatric form of IBD was achieved by Worthey *et al.* [202] where a rare mutation in the X-Linked Inhibitor Of Apoptosis (*XIAP*) was first identified through the application of whole exome sequencing and then corrected with gene therapy.

Despite more than 50 genes were so far identified as causative of monogenic forms of IBD [192] (Supplementary table 7.1), the majority of IBD presentations cannot be explained by mutation(s) in a single gene. In order to detect genes with a more subtle contribution to the IBD phenotype, a range of approaches were developed in the last 20 years. Before the advent of NGS, linkage studies, consisting in the study of familial inheritance of specific genomic regions, were the standard analyses for the detection of susceptibility loci (Figure 1.11). Whenever a region was found

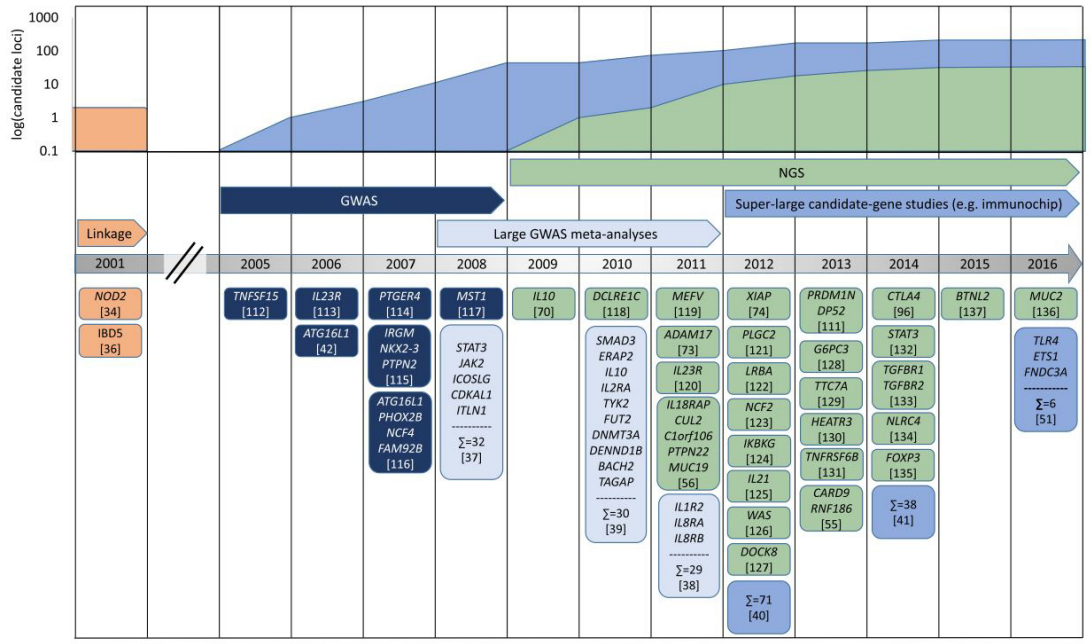


Figure 1.11: **Timeline of loci discovered to be associated with IBD phenotype.** Image adapted from Petersen *et al.*[146]

linked to a phenotype and recurring within the same pedigree, it was then possible to estimate the likelihood of that recurrence against a normal scenario. Though this method the Nucleotide-binding oligomerization domain-containing protein 2 (*NOD2*) gene was identified as the first susceptibility gene for Crohn's disease [70] [138].

With the introduction of SNP arrays, genotyping thousands of loci in large cohorts became affordable and started the so called "GWAS era". Instead of focusing on small pedigrees, GWAS involve large groups of unrelated individuals in order to observe mutations that are statistically more frequent in affected individuals and less in controls. Through Genome-wide association studies many new loci were associated to IBD [121][157][5][79][18]. Although more than 200 SNPs were identified so far[110], GWAS results can only explain 30% of IBD genetic component [79]. As mentioned before, association studies are based on SNP arrays and therefore are covering only a very limited percentage of the whole human variome (the whole set of human variations). This is motivated by the requirement of SNPs to have a moderate or high frequency in the population in order to be included in a SNP array, therefore ignoring all the rare (minor allele frequency <1%) and novel

variants.

Thanks to the advent of whole-exome sequencing and its drop in price, many new studies were then able to cover that part of rare human variation that were previously set aside. Mutations in the interleukin 10 gene (*IL10*) [90] and in the Baculoviral IAP Repeat Containing 2 and 3 (*BIRC2*, *BIRC3*) [6] are some successful examples of WES extrapolating new knowledge on IBD genetics. Figure 1.11 recapitulates the advances in discovering associated loci to IBD and the contribution of each technological breakthrough [146].

### Associated pathways

Due to autoimmune nature of IBD, many studies were focused on the analysis of immunological pathways in order to understand the mechanisms triggered in its pathogenesis. Particularly, the first extensively studied pathway was the cascade of events regulating the innate immune response. Since *NOD2* was the first gene associated to IBD, there was great interest in understanding how *NOD-like* (NLRs) and *Toll-like* (TLRs) receptors were regulating the inflammatory and apoptotic response and which elements were responsible of the dysregulation observed in IBD patients.

The *NOD* pathway (Figure 1.12), through the activation of NLRs is devoted to the detection of specific bacterial components (bacterial peptidoglycans from both Gram + and Gram - bacteria), triggering the innate immune response and maintaining the homoeostasis of intestinal microbiota [26]. The activation of such receptors causes the expression of pro-inflammatory cytokines and apoptosis respectively mediated by the nuclear factor kappa B (*NF-$\kappa$B*) and the mitogen activated protein kinase (*MAPK*) signalling pathways [73] [54]. Both signalling pathways have a well described role in inflammation, activation of stress responses, B-cell development, and lymphoid organogenesis [65].

Besides this cascade, other NLRs (*NLRP1*, *NLRP3*, *NLRP6*, *NLRP7*, *NLRP12*, *IPAF* and *NAIP*) have the ability to oligomerise forming a multiprotein known

as *inflammasome*. The inflammasome can activate the Caspase-1 (*CASP1*) which induces the production of other pro-inflammatory interleukins (*IL-1β* and *IL-18*) and pyroptosis leading to cell lysis and swelling.

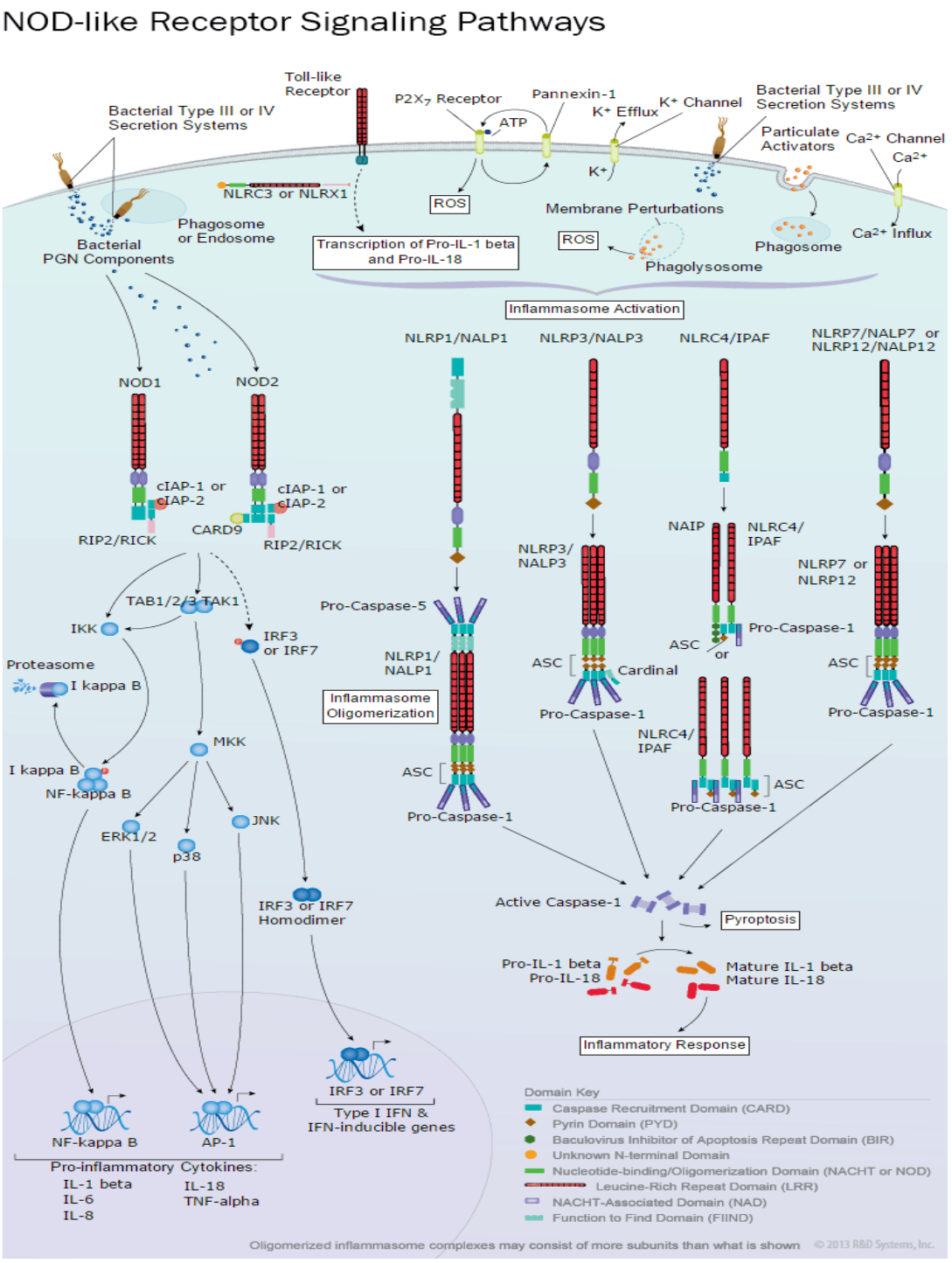


Figure 1.12: **NOD-like receptor signalling pathway.** The cascade on the left hand side reflects is the result of the activation of *NOD1* and *NOD2* genes. The cascade on the left hand side shows the inflammasome activation via NLRs. Image adapted from R&D systems repository[153]

By interrogating KEGG pathway (KEGG entry: hsa04621) [84], an on-line repos-

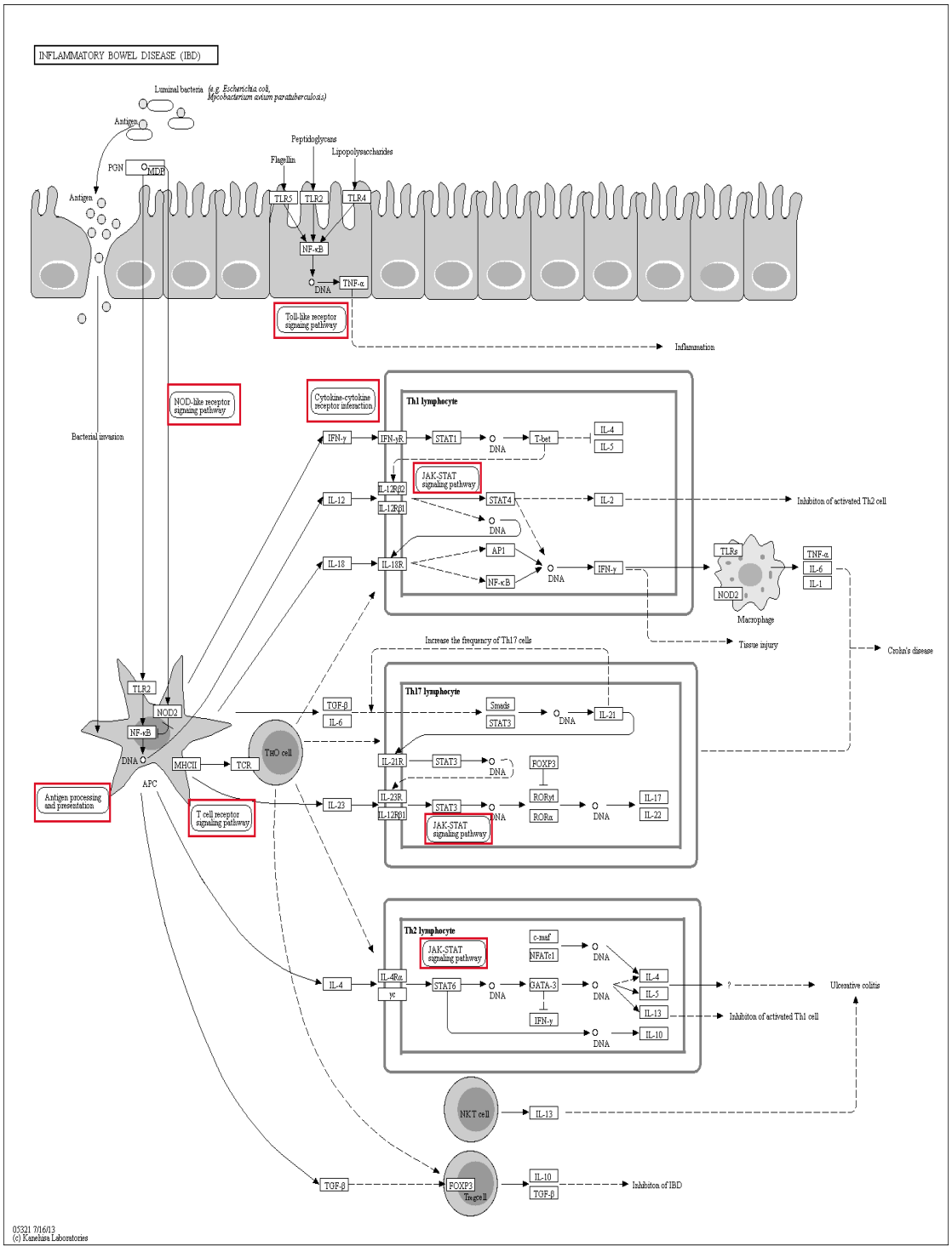


Figure 1.13: **IBD pathways.** Pathways and genes reported to be involved in IBD characterization by interrogating KEGG pathway (KEGG entry: hsa05321). Relevant signalling pathways are highlighted by a red box.

itory for multi-organism pathways, 170 genes are involved in the *NOD*-like signalling pathway but only 56 of them can be considered as forming the core *NOD* cascade.

Despite the distinctive autoimmune nature of IBD described by Targan *et al.*

[123] [168], there are many other pathways that are potentially involved in the IBD pathogenesis (Figure 1.13). Studies performed on murine models have shown the importance of G protein-coupled receptors (GPCRs), the regulation of innate and adaptive immunity (IL-10 signalling, Th17 differentiation programme, T and B cell signalling pathways) and the epithelial barrier function in the induction of IBD [86]. Moreover, a possible involvement of the interferon gamma (*IFN-*$\gamma$) and *TNF* signalling pathway was also postulated [10].

### 1.4.3 Machine learning applications to IBD genetics

Machine learning algorithms could be exploited to improve the diagnostic process in IBD. So far only few attempts were made, mostly because of the shortage of NGS database with matched clinical information.

Zhi Wei et al. [197] made use of thousands of single nucleotide polymorphisms (SNPs) to develop a model aimed at distinguishing Crohn's disease from ulcerative colitis. The original dataset was generated by the IBD Genetics Consortium's Immnunochip project and the initial 196,524 variants were filtered accordingly to standard association thresholds (p-value $< 10^{-4}$ and minor allele frequency $< 0.01$). To avoid overfitting the model with too many variables and few samples, and other previously illustrated modelling issues, more than 22 thousand control individuals were included in this study alongside 30 thousand IBD cases (17,000 CD and 13,000 UC). Subsequently, a supervised model based on a penalized logistic regression was trained with a 10-fold cross validation approach to ensure good data fitness and generalisation. The study was structured in order to obtain two different predictors, one for Crohn's disease and one for the ulcerative colitis. After the training, predictors performances were approximately $\sim 85\%$ accuracy in predicting CD and UC.

In terms of application of machine learning algorithms to IBD genomic data, the 2013 CAGI competition [28] challenged participants to identify Crohn's disease affected individuals amongst healthy ones using only whole exome sequencing data.

This challenge was based on the analysis of whole exome sequencing data, which includes more information respect to immunochip data presented by Wei *et al.*. The CAGI challenge provided whole exome sequencing data from 66 individuals, 51 with Crohn's disease and 15 without, including related individuals and two affected twins. Every submission was independently tested by the CAGI committee and the two best submissions (Tosatto and Radivojac) reached an area under the ROC curve of $\sim$ 87% [28]. As first step in both methods variants were filtered depending on the sequencing quality (higher than 30) and the minor allele frequency (MAF less than 2 percent). Both methods used an additive model, weighting SNVs according to their genotype: 1 for homozygous alternative SNVs, 0.5 for heterozygous SNVs, and 0 for homozygous reference SNVs. The submission from Tosatto was based on hierarchical clustering of variants found in genes potentially relevant to IBD. On a opposite path, the Radivojac submission exploited the MutPred score and the PhenoPred score to add additional weights respectively to variants and genes. Then, expecting a binary classification, they performed a k-means analysis scoring each sample depending on the distance from the two centroids. Both methods confirmed the additive model as the best choice for interpretation of SNVs and the need for a strong filtration on the input SNVs in order to remove the background noise. However, none of these models were validated on additional data.

## 1.5 Thesis outline, aims and contribution

The intention of the introduction chapter was to provide a brief background on the recent sequencing technologies and machine learning methodologies that will help the reader to contextualise what will be discussed in the following chapters. Since most of the results will be focused on the analysis of data describing inflammatory bowel disease, it was necessary describing the state of the art from a clinical and research perspective.

Following the introduction chapter, I provide more details on the bioinformatics tools and machine learning algorithms that will be extensively applied and cited across all result chapters. This method section is essential in order to ease the understanding of more complex mechanism which would move the focus of the reader if covered individually in each result chapter.

**Aim 1 - Machine learning classification of IBD patients using histopathology data** Chapter 3 shows the application of supervised and unsupervised machine learning methodologies to clinical data that is routinely collected and used on a daily basis to assign a diagnosis of either Crohn's disease or ulcerative colitis.

My contribution was to develop supervised machine learning models, analyse the histopathology data with unsupervised methods, investigate possible clustering strategies, performing statistical test and interpret results. The work was supervised by Prof Sarah Ennis and Dr Ben MacArthur from a research perspective and by Prof Mark Beattie from a clinical point of view.

**Aim 2 - Gene score development: GenePy** Chapter 4 shows the development and testing of a mathematical model capable of transforming NGS data in order to produce per-gene per-patient scores.

The work conducted in this chapter was conducted only by myself under the supervision of Prof Sarah Ennis and Dr Ben MacArthur. My contribution consisted in every step described in the chapter, including curation of the research database,

data quality control, data processing, pipeline and models development and application.

**Aim 3 - Stratification of paediatric patients using immunogenomic data** Chapter 5 investigates the opportunity of using immunological markers and GenePy-modelled whole exome data to stratify patients according to their immunological response.

Regarding immunological data, I was responsible of controlling data quality, normalisation and finally the application of clustering strategies. Concerning genomic data I was in charge of data processing, quality control and pipeline development. This work was conducted in collaboration with Dr. Tracy Coelho and supervised by Prof. Sarah Ennis and Prof Mark Beattie.

**Aim 4 - Classification of IBD using supervised machine learning and genomics data** Chapter 6 covers the application of supervised and unsupervised machine learning methodologies using uniquely genomic data to classify and stratify IBD patients.

My contribution consisted in every step described in the chapter, including curation of the research database, data quality control, data processing, pipeline and models development and their application. This work was supervised by Prof Sarah Ennis and Dr Ben MacArthur.

Finally, Chapter 7 summarises thesis findings and discusses future work.

# Chapter 2

# Methods

## 2.1 Programming tools

### 2.1.1 Iridis 4

Most of the work presented in this thesis was made using the computational power of Iridis 4. Iridis 4 is the computing cluster of the University of Southampton. In November 2015 was in the TOP500 list [47] of the most powerful computer around the world and is still one of the largest computational facilities in the United Kingdom. The cluster is made of:

- 750 computing nodes each with 16 CPUs and 64GB of memory;
- 4 high-memory nodes with two 32 cores and 256GB of RAM;
- 12,320 processors providing 250 TFlops peak.

Iridis 4 does not have a graphical user interface (GUI) and each operation or software has to be executed using the bash command line. Iridis 4 command line is based on the UNIX architecture.

### 2.1.2 Python

Python is a freely available programming language that in these recent years has become one of the most broadly used [112]. From games development to data analysis, its flexibility is the main reason why many people are choosing it as tool for developing software and pipelines. As many other languages, Python is structured with a set of standard functions that can be expanded with various packages. The version used in this dissertation is the 2.7.

#### Packages

**SciPy** SciPy is a library of programs for mathematics, science and engineering [140]. It includes fundamental tools like NumPy for array (matrix) calculations and Matplotlib for graphs and 2-3D plotting.

**Scikit-learn** Scikit-learn is the main package for statistical learning in python [144]. It includes algorithms for supervised and unsupervised machine learning nonetheless tools for choosing and validating parameters and models.

## 2.2 Bioinformatic tools

Next-generation sequencing data represents an important component of "big data" within the medical information field that requires advanced analytical tools. After data generation with NGS chemistries, a bioinformatic pipeline is required to detect and characterise variants. (Figure 2.1).

Firstly, raw NGS data has to be quality controlled, checking for errors generated by the sequencer or by a low quality sample. These QC checks usually consist in obtaining metrics about read number and mean coverage of regions targeted by capture kits.

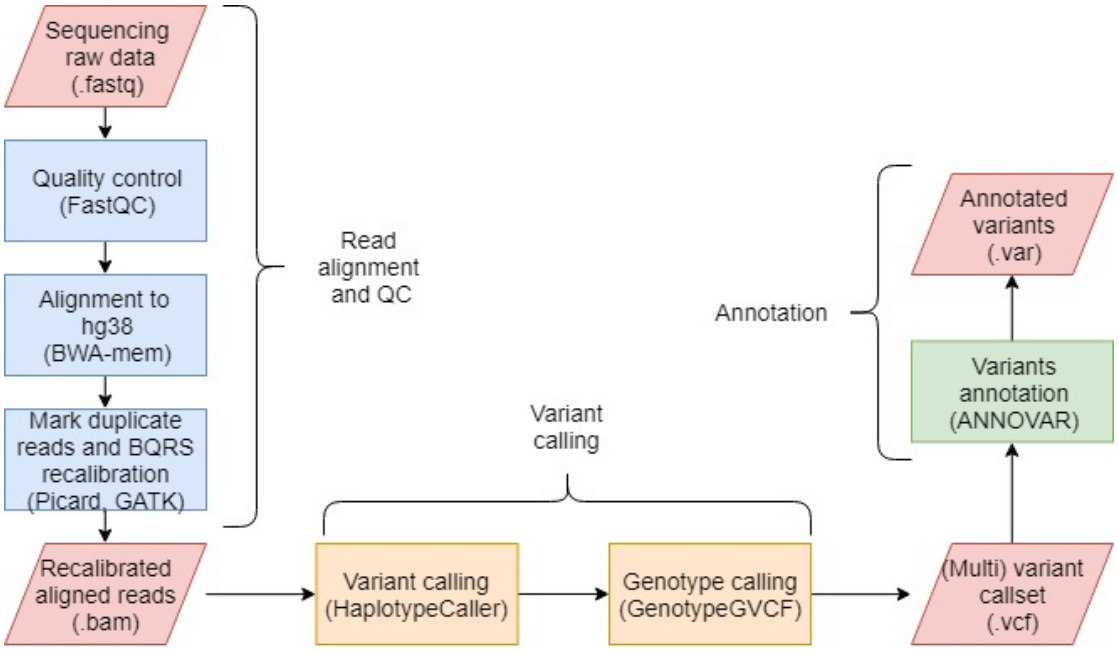


Figure 2.1: **Bioinformatic workflow for NGS data analysis.** The analysis consist in three major steps: read alignment (blue), variant calling (orange) and variant annotation (green). Parallelograms represent data (file format in parentheses) and rectangles denote processes (software used in parentheses).

Then, if the quality controls (QCs) are successfully passed, sequenced paired reads are aligned against a reference genome. The reference genome is a continuously updated reference sequence obtained by sequencing samples from several individuals. The version used in our analyses is the GRCh38 (hg38) released by the Genome Reference Consortium in 2013. Since chromosome coordinates and contexts change depending on the completeness of the sequence, in case of samples comparison it is important to consistently select the appropriate release. Our bioinformatic pipeline uses the Burrows-Wheeler Aligner (BWA) for the alignment step, set with a mean fragment length of 200 bp, a gap opening penalty of 65 and extension penalty of 7 as recommended by BWA guidelines .

Following the alignment, which generates a SAM file per individual that is converted into the BAM format, duplicate reads are marked using Picard. This tool identifies and tags duplicate reads originating from the same DNA fragment. An elevated percentage of duplicate reads is usually index of poor data quality. Once duplicate reads are removed, the BAM file is ordered by genomic coordinates and indexed. The resulting BAM file is then scanned by Picard for wrong mate-pair matched and fixed accordingly.

Prior to variant calling, base qualities have to be recalibrated in order to detect and correct for systematic errors. Variant calling efficiency is highly dependent on the quality scores assigned to each base pair reported in the BAM file. Due to systematic error by sequencing machines, it is important to correct avoiding over or under estimating sequencing quality. This step is performed using GATK's BQSR recalibrator which employs a machine learning algorithm to detect and adjust such discrepancies leading to more accurate variant calls.

Once the recalibration is complete, variants are called using GATK's HaplotypeCaller tool. This allows the simultaneous call of SNVs and indels and reports them in the so called gVCF format further described.

Following variant calling, gVCFs from multiple samples can be merged in a single VCF file containing the combined call set and specific genotype information. This step is performed with GATK's GenotypeGVCF tool.

Called variants are then annotated in order to obtain information regarding the observed frequency in the population (1000 Genomes Project, ExAc) and the induced protein alteration (GERP++, PolyPhen-2, ...). This step is performed using the ANNOVAR software which integrates and interrogates multiple repositories to produce a final report file (.var).

All the tools so far mentioned were utilised according to distributors guidelines when not stated differently. Figure 2.1 shows the steps of the bioinformatic pipeline just described.

### 2.2.1 BWA

BWA is an alignment tool for mapping sequence reads against a reference genome. This algorithm implement the theory of Burrows-Wheeler Transformation (BWT) to efficiently align short sequences allowing both mismatches and gaps.

Originally developed for compression purposes, the BWT algorithm is a fast algorithm for compressing data maintaining the reversibility without the need of

additional metadata. In the case of genomic data, characterised by a large volume of fragmented "text strings" BWA (thanks to BWT) can provide a reliable and quick alignment.

Despite being the second best aligner on the market in terms of number of mapped reads, its speed makes it the best tool for the analysis of large genomic data. As consequence of the constantly increasing throughput of sequencing technologies, trade-off between accuracy and computational time is continuously moving towards models that requires less running time rather than a complete alignment of input reads set.

By comparing the most recent aligners (Figure 2.2), Novoalign results being the most accurate in terms of mapping but also the most computational demanding. Implementing a modified version of the classic Smith-Waterman algorithm [174] used by BLAST, known as Needleman-Wunsh algorithm [132] despite the CPU vectorization to speed up the alignment process [104] Novoalign requires eight times more time to align paired-end data from a single sample. Moreover, Novoalign performance are strongly limited by the large memory required for the reference genome hashing, the first step of its algorithm. So far, BWA represent the best compromise for detecting structural variations (insertions, deletions, CNVs...) when using a long reference genome and short reads.

### 2.2.2 GATK

GATK, genome analysis toolkit (version 3.7), is an alternative library for reads alignment and variant calling developed by th Broad Institute [122]. This software is everyday becoming more popular thanks to is good scaling capability in both statistics calculation and variant calling. This makes GATK time-wise more efficient. The main tool in GATK for variant calling is the HaplotypeCaller.

HaplotypeCaller calls germline SNVs and indels re-assembling the reads whenever a mismatch is found. This approach increases the accuracy in calling variants close to each other and long indels. HaplotypeCaller was designed to analyse data with

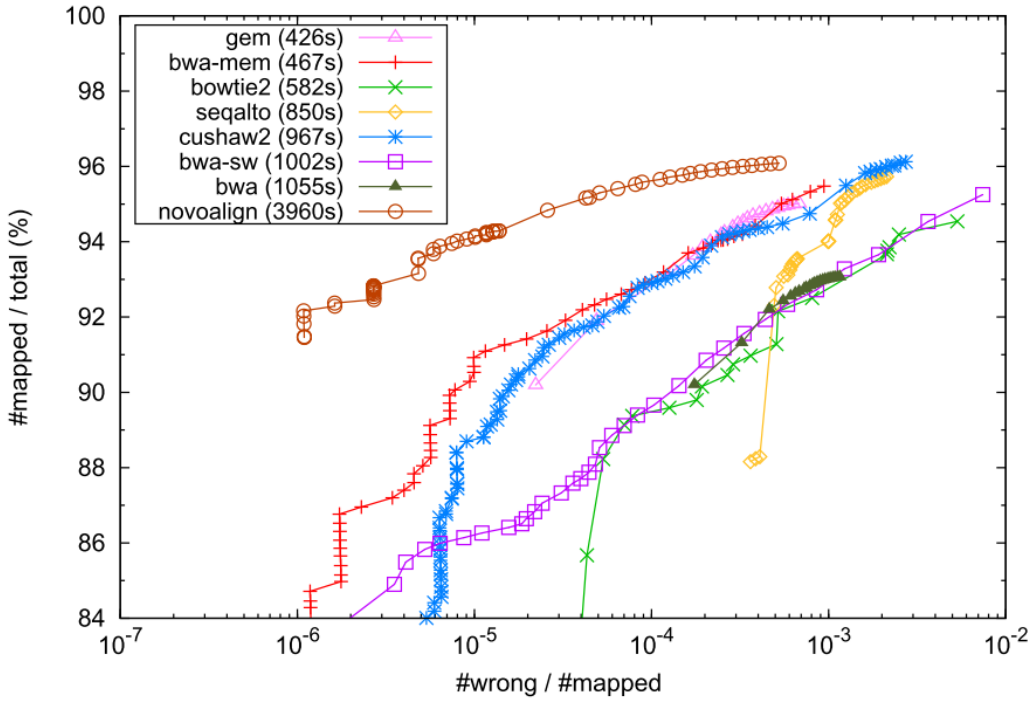


Figure 2.2: **Aligner accuracies.** Accuracy benchmark among the most known alignment tools. Image adapted from Li *et al.*[102]

a sample-by-sample approach, thanks to the gVCF format (figure 2.3) and can handle sample multicalling extracting information about homozygous reference calls. Due to its highly sensible algorithm for calculating the variant likelihood, HaplotypeCaller is not suited to the extreme allele frequencies observed in cancer samples. MuTect2 [39] is a modified version of HaplotypeCaller for calling somatic SNVs and indels.

### 2.2.3 ANNOVAR

ANNOVAR is a software for the annotation of variants [196]. This program has been specifically developed for the annotation of SNVs and insertions/deletions examining their functional consequences and frequency in general populations. As consequence of it simple architecture, ANNOVAR only requires flat files from any available annotation database and a VCF file to annotate. Amongst the large variety of annotations, this software can integrate the dbsnfp dataset, containing most of the known SVN deleteriousness metrics, and allele frequencies from the

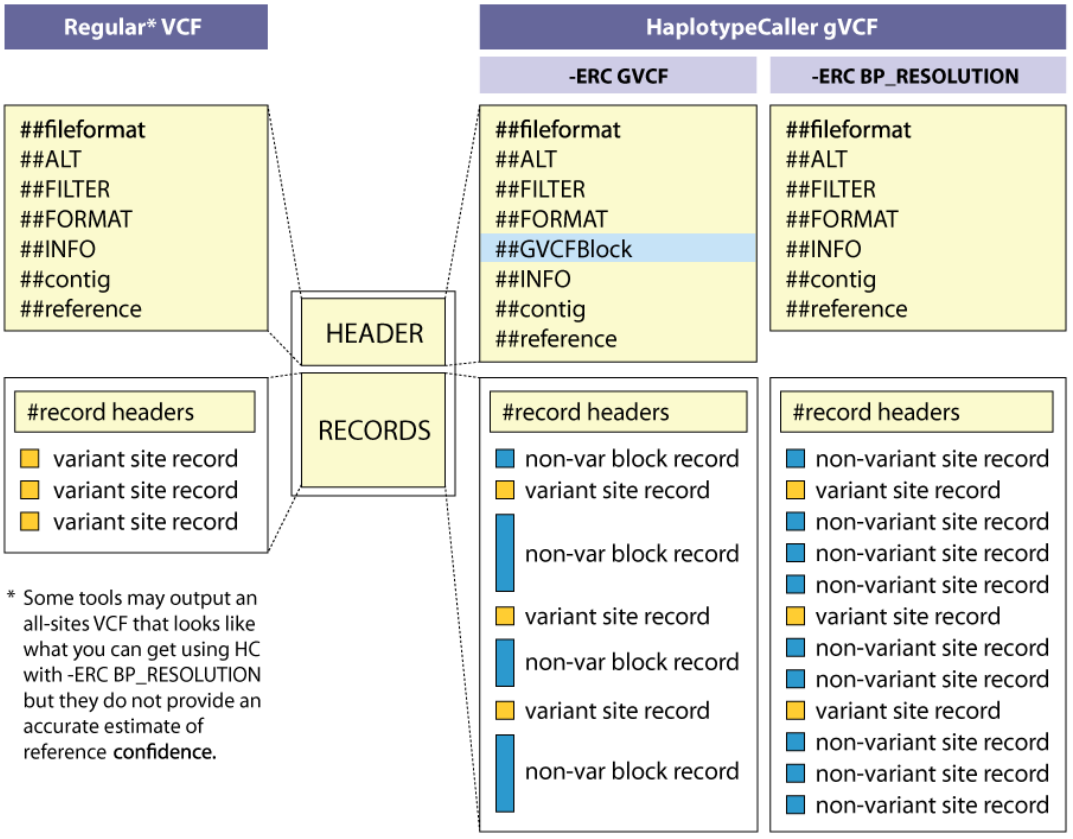


Figure 2.3: **Comparison of the gVCF format against the classic VCF.** Non-var blocks are regions known where sample's genotypes are homozygous as reference. These regions are not annotated in the normal VCF and makes therefore impossible a correct multicalling approach. Adapted from [122]

1000 Genomes Project [1] the OMIM dataset [63], the COSMIC repository [14], ExAc [96] and many others. Amongst the may deleteriousness metrics available trough ANNOVAR the following had been largely employed:

- Sift, sorting tolerant from intolerant, predicts whether a single amino acid substitution affects the protein function or not considering the degree of conservation for that site [173](0-1);
- Polyphen2, predicts the possible impact of a mutation on the structure and function of the encoded protein [3] (0-1);
- LRT, likelihood ratio test for deleteriousness using 32 vertebrates as reference [38] (0-1);
- MutationTaster, composite score that uses multiple statistics and a Bayes classifier for deleteriousness at the mRNA level [165] (0-1);

- MutationAssessor, asses the functional impact of a variant using conservation patterns and entropy formalisms [154] (0-1);
- FATHMM, functional analysis through Hidden Markov Models fitted for human variants [170] ($-\inf - +\inf$);
- VEST3, supervised machine learning classifier trained on 45,000 disease mutations annotated in the Human Gene Mutation Database (HGMD,[179]), [32] (0-1);
- PROVEAN, protein variation effect analyzer, specifically suited for filtering nonsynonymous variants and indels [37] (-14-+14);
- CADD, combined annotation dependent depletion,is a composite score of the deleteriousness of SNVs and indels [88] ($0 - +\inf$);
- DANN, is a composite score using the same training data as CADD but implemented in a deep neural network classifier[151] (0-1);
- GERP++, genomic evolutionary rate profiling, calculates the strength with which the genome rejects the variants due to the functional constraint [43] ($0 - +\inf$);
- phastCons7way, conservation score based on 7 vertebrates (0-1);
- SiPhy_29way, conservation score based on 29 mammals genomes [55] (0-37.97).

### 2.2.4 CADD

Combined Annotation Dependent Depletion (CADD) [88] is a composite score for both SNVs and indels. This combined annotation dependent depletion algorithm exploits a support vector machine (supervised machine learning algorithm) to differentiate about 15 million real human variants from other 15 million simulated variants. Moreover, scores were also calculated for all 8.6 billion possible

human SNVs and short indels. The aim of CADD is to distinguish between variants that are fixed or nearly fixed in the human genome and those simulated. Natural selection and more in general evolution should avoid the fixation of deleterious variants, therefore the closer a genotype is to a simulated scenario and far from being fixated, the higher the deleteriousness. To date, CADD is one of the most popular deleteriousness metrics and detain the highest AUC in classifying pathogenic variants in the ClinVar dataset.

### 2.2.5 MaxEnt

MaxEnt is a tool developed by the Massachusetts Institute of Technology for the evaluation of splicing motifs [204]. The original aim of this software was the identification of splicing sites (5' donor and 3' acceptor) in the human genome. The algorithm out-performed all the previous probabilistic models and become the best tool for assessing the presence of splicing motifs. The model was so good in this identification that was exploited to observe the impact of SNVs and indels on the splicing sequence. Worst the mutation higher the divergence, calculated as entropy, and therefore its effect on the phenotype.

## 2.3 Machine learning algorithms

Statistical learning, or machine learning, become popular a science when people started collecting data on a large scale. With massive datasets, classic statistical approaches did show limitations concerning result validation and the integration of multiple variables. Machine learning approaches can be divided in two large groups: supervised learning algorithms and unsupervised learning algorithms

### 2.3.1 Supervised learning algorithms

#### Support vector machine

Support vector machine (SVM) is a supervised algorithm for data prediction and regression analyses. Thanks to is almost ready-to-use approach and its good performance in a variety of scenarios, this model is currently one of the most applied. However, the easy application of this model should not be confused with simplicity, SMVs are elaborate algorithms developed starting from a simpler classifier called *maximal margin classifier*. This classifier applies the idea of using a hyperplane (that in 2 dimensions is a line) to separate data (Figure 2.4). It is important to define the concept of hyperplane as a subspace of dimension $p-1$. Considering a 2 dimensional space, any hyperplane of that space can be defined as:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{2.1}$$

where generalising to $p$ dimensions is easy as introducing more addends:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \tag{2.2}$$

From equation 2.1 defining parameters $\beta$ it is possible to plot a line that divides in half the 2D space (Figure 2.1). If a point $X$ does not satisfy the equation, that point can either fall above or below the hyperplane. Now, considering an imaginary dataset where each labelled sample is defined by 2 features that in the previous equation are represented by $X$ parameters, each point will be localised above, below or on the hyperplane. The idea of the *maximal margin classifier* is to use this mathematical concept to separate labelled data by correcting $\beta$ parameters. The logic used by this algorithm is to maximise the perpendicular distance between the closest points with opposite labels to the hyperplane. This set of points, given a hyperplane, defines the margin of the model. The classifier starts learning with a random set of $\beta$ values and, during the training, adjusts them according to the margin (Figure 2.5). The hyperplane margin does depend only on points close to it and not on other observations.

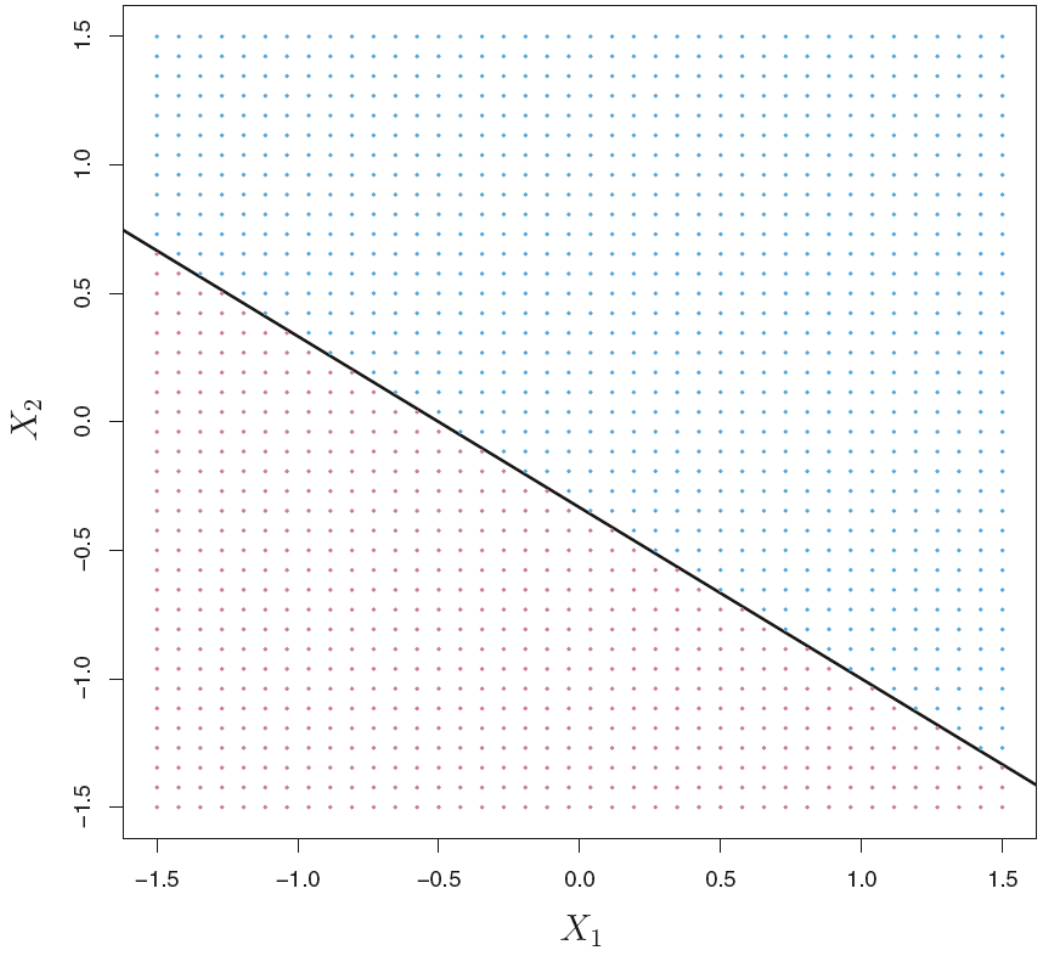


Figure 2.4: **Example of an hyperplane.**Hyperplane defined for $\beta_0 = 1, \beta_1 = 2, \beta_2 = 3$. Blue region made by points for which $1 + 2X_1 + 3X_2 > 0$, while purple region by points for which $1 + 2X_1 + 3X_2 < 0$. Image adapted from James *et al.*[76].

Once the model is trained with the training data and the coefficients are set, new data can be classified. Generally, the *maximal margin classifiers* as well as *support vector classifiers* and *support vector machines*, are extremely powerful when classifying data in two classes, but extensions to these approaches are available for the multi-class classification. Unfortunately, not every classification problem can be solved with the *maximal margin classifier*. Indeed, the linear separating hyperplane does not always exist. However, it is possible to generalise this model to overcome such limitation with *support vector classifier*.

**Support vector classifiers and machines** Since the solution obtained with the maximal margin hyperplane perfectly classify training data, it is extremely

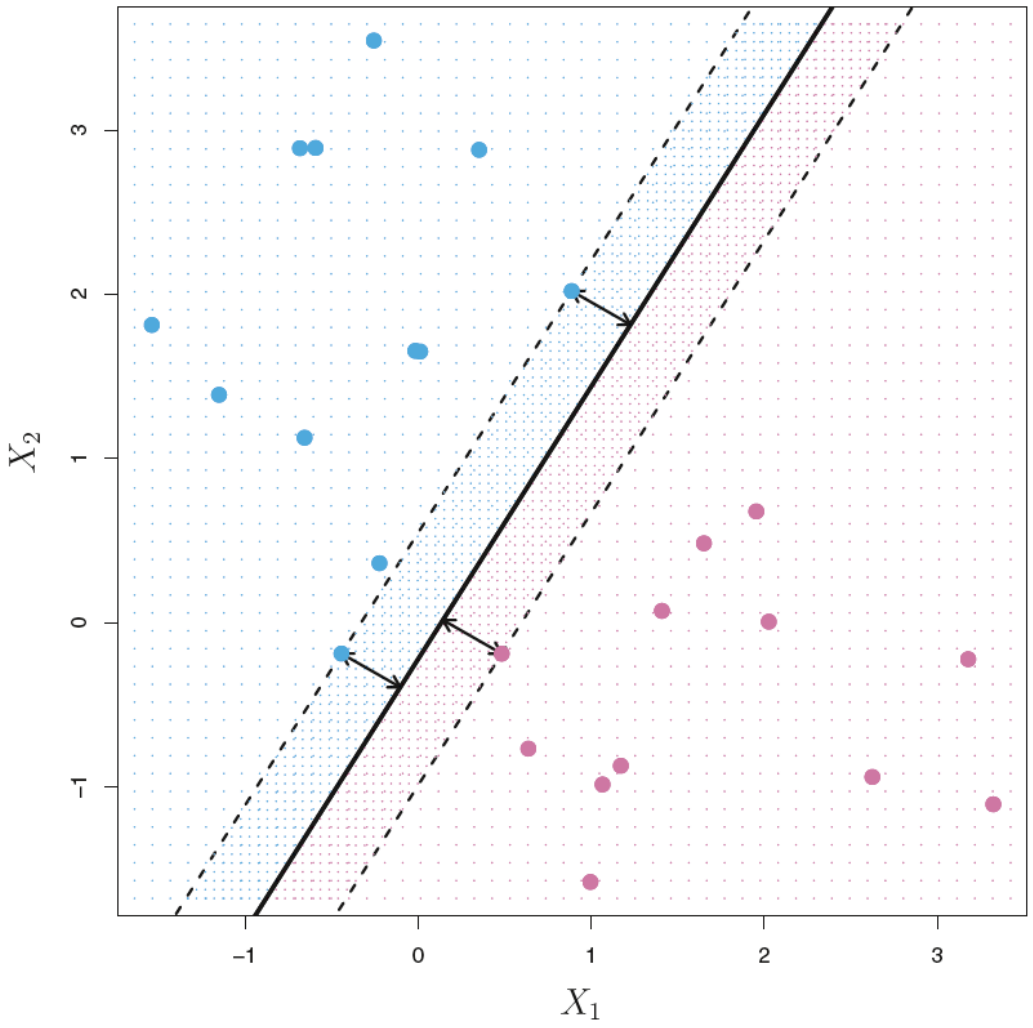


Figure 2.5: **Margin maximisation.**Example of an hyperplane maximising the margin for the classification of two different classes of observations. The two blue points and the purple one are the support vectors that define the margin. Figure from [76].

sensible to small changes in the input data, making overfitting a property of this model. In this case, support vector classifiers leave some imperfection during the model fitting, making it more robust to new individual observations and better in the classification of most of the training set. Support vector classifier introduces the concept of soft margin, where small misclassification are allowed. This allowance solves the problem where a linear hyperplane can not separate perfectly all the data.

If we state $M$ as a the solid margin that the maximal margin classifier aims to maximise, the new margin formulation for the support vector classifier is $M(1 - \epsilon_i)$ with $\epsilon_i > 0$ representing the error allowed in the classification of training

observations. If a point is located within the wrong margin $\epsilon_i$ is between 0 and 1 (violated margin), while if it is grater than 1, the point is in the wrong side of the hyperplane. The sum of all these errors have an upper boundary known as tuning parameter $C$. This variable has to be set before the training starts and reflects the tolerance of the model. A very low tolerance pushes the model towards a maximal-margin-like model with all the consequences. For $C > 0$ the maximum number of misclassified point can be no more than $C$: if the point is on the wrong side of the hyperplane $\epsilon_i > 1$. As $C$ increases, the margin does. The real key elements are then not the correctly classified observations but those that are within the margins or misclassified, those are the *support vectors*.

Support vector machines are an extension of support vector classifiers that includes *kernels*, non-linear representations of hyperplanes. SVMs, still use the same idea of soft margins but, instead of a linear relationship between $\beta$ terms and $M$, a non linear is used, allowing more complex separations. The two most used kernels are the polynomial, where the parameter $d$ defines the degree of the equation, and the radial (RBF, radial basis function) where the main parameter to be set is the positive constant $\gamma$. A direct consequence is the increased number of features that can be included in the model. To recap, to model data with SVMs it is important to select the correct tolerance $C$ and the shape of the kernel ($d$ or $\gamma$). This approach is known as *grid search* and is done testing for the best fitting several ranges of $C$'s given some kernels.

### 2.3.2 Unsupervised learning algorithms

Unsupervised learning algorithms are models used when the observations labelling is not available and the aim of the analysis is to understand the relationship between variables and samples. Most of the algorithms of this group are also known as class discovery approaches since they are powerful tools for clustering data. Moreover, it is also possible to apply these algorithms as methods for data visualisation or data pre-processing before applying other supervised techniques.

### Principal components analysis

Principal components analysis (PCA), is a linear class discovery and dimensionality reduction algorithm. PCA reduces the number of features to a fewer set of most representative variables. In a scenario where some data are represented by 5 variables, without using a dimensionality reduction algorithm like PCA would mean drawing 10 different 2D plots to represent all the information within that particular dataset. The principle of PCA is to find the lowest dimensional space to represent as much as possible of the original variation. This algorithm uses a linear combination of the original feature set to obtain a smaller number of dimensions, also known as components. The first principal component of the feature set $X_1, X_2, ..., X_p$ is their normalised linear combination [76] that maximise the variance

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + ... + \phi_{11} X_p \tag{2.3}$$

with $p$ the number of observations and $\phi$ known as loadings of the first principal component, meaning the weight that each feature is associated within the clustering process. With the normalisation, the sum of the loadings is equal to 1, avoiding bias in the variance comparison. Since the algorithm measures the variance of observations using a set of features, it is important to normalise all the values to the same scale. This avoids unrealistic results that are only dependent on the way a feature is reported. When fitting the model, the PCA algorithm starts assuming that all features are centred to have mean zero. Then, adjusting the loadings, it tries to maximise the linear combination 2.3 respecting the normalisation criteria where $\sum_{j=1}^{p} \phi_{j1}^2 = 1$. Using the same principle, the second and further components are calculated bearing in mind to avoid using linear correlations correlated to the one already used in previous components. Each resulting component will carry ( or explain) a percentage original variance and each observation will be associated with a coordinate ($X_p = (z_1, z_2, ..., zn)$, with $n$ as number of principal components). Since every additional component must be uncorrelated to the already computed, the explained variance will decrease after an empirically derived number of components. Then it will be possible to visualise data by plotting each

computed component against each other.

PCA has been demonstrated to be an efficient method for the analysis of genomic data [136]. Figure 2.6 shows the result of modelling more than half million DNA variable sites in more than 1,000 European individuals. PC1 and PC2 can recap both the ethnic and regional differences in Europeans. This result shows the potential of unsupervised machine learning algorithms in inference problems but also highlights the caution needed when mapping the genetics of a complex disease using samples with different origins.

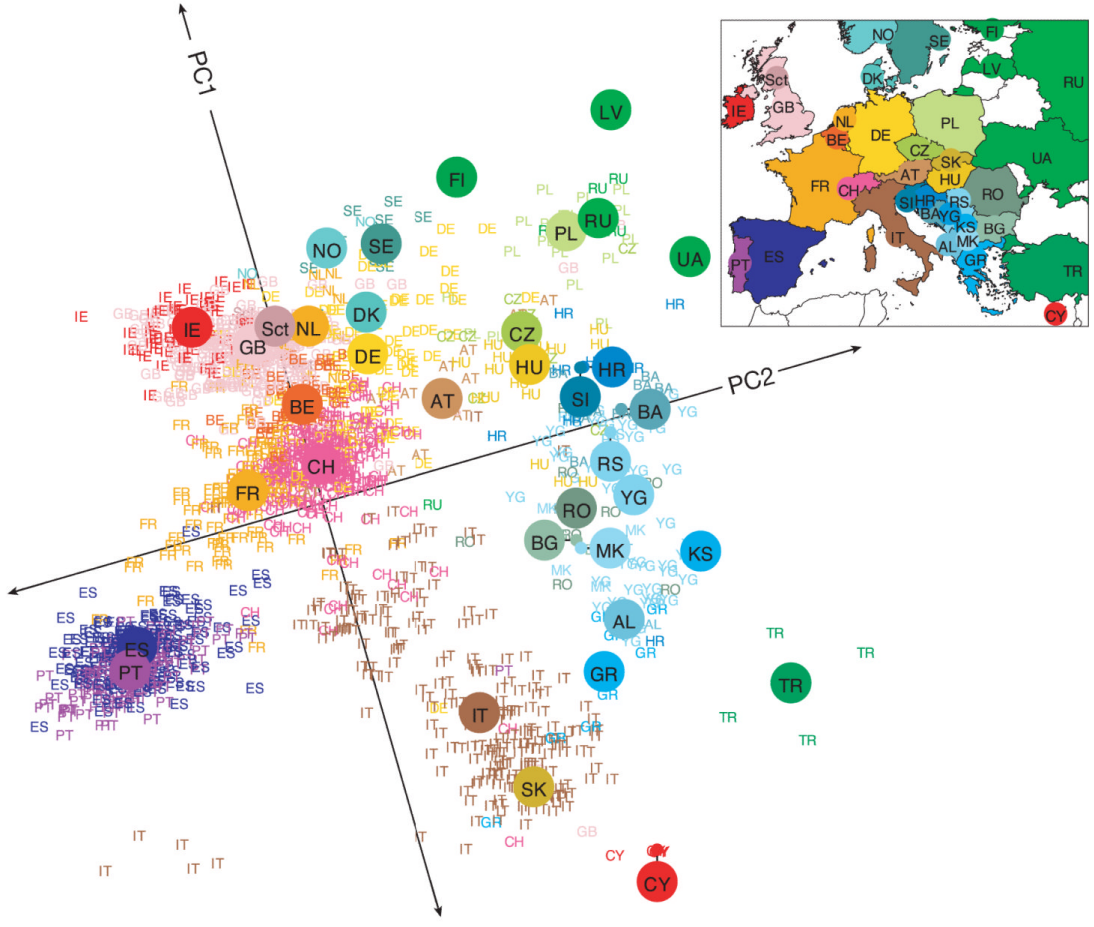


Figure 2.6: **PCA of Europe.** Principal component analysis of genomic data from European individuals. Image adapted from Novembre *et al.* [136].

## Multidimensional scaling

Like principal component analysis, multidimensional scaling (MDS) is a unsupervised learning algorithm for dimensionality reduction and data visualisation.

MDS, similarly to PCA, seeks to find a lower-dimensional representation of the data preserving, not the variance but, the pairwise distance between observations. Considering a set of observation $x_1, x_2, ..., x_N$ is possible to calculate the distance $d_{ij}$ between each couple of points. As distance, there are hundreds of possible choices but the more conventional is still the Euclidean distance $d_{ij} = \| x_i - x_j \|$. Moreover, there are many versions of MDS or methods that uses evolutions of the same algorithm. This led to MDS models that can explore solutions in the non linear space (local MDS) or that use graphs and geodesic distances (isometric feature mapping, ISOMAP) to reduce the dimensionality.

**t-distributed SNE**

The t-distributed stochastic neighbour embedding (t-SNE) [113] is a nonlinear dimensionality reduction algorithm. t-SNE converts a euclidean distance matrix into a probability distribution in a way where similar observations have a high probability, *vice versa* for dissimilar observations. A second probability distribution is also calculated over the original observations, not converted to euclidean distances. Then the Kullback-Leibler divergence is calculated between the two distributions. Once the divergence is minimised, the probability distribution done on the original observation (high dimensionality) can be approximated with the probability distribution converted to a lower dimensionality. The divergence is corrected step by step by changing the shapes of the t-distribution used to model the data. As any other unsupervised learning approach the performance of an algorithm depend only partially from the algorithm itself. Indeed, some datasets are better represented with some algorithm then others.

**Hierarchical Clustering**

Hierarchical clustering (HC) is an unsupervised clustering method that, using a matrix of distances can identify similarities between samples. Depending on the chosen criteria, HC can be labelled as bottom-up (or agglomerative) or top-down

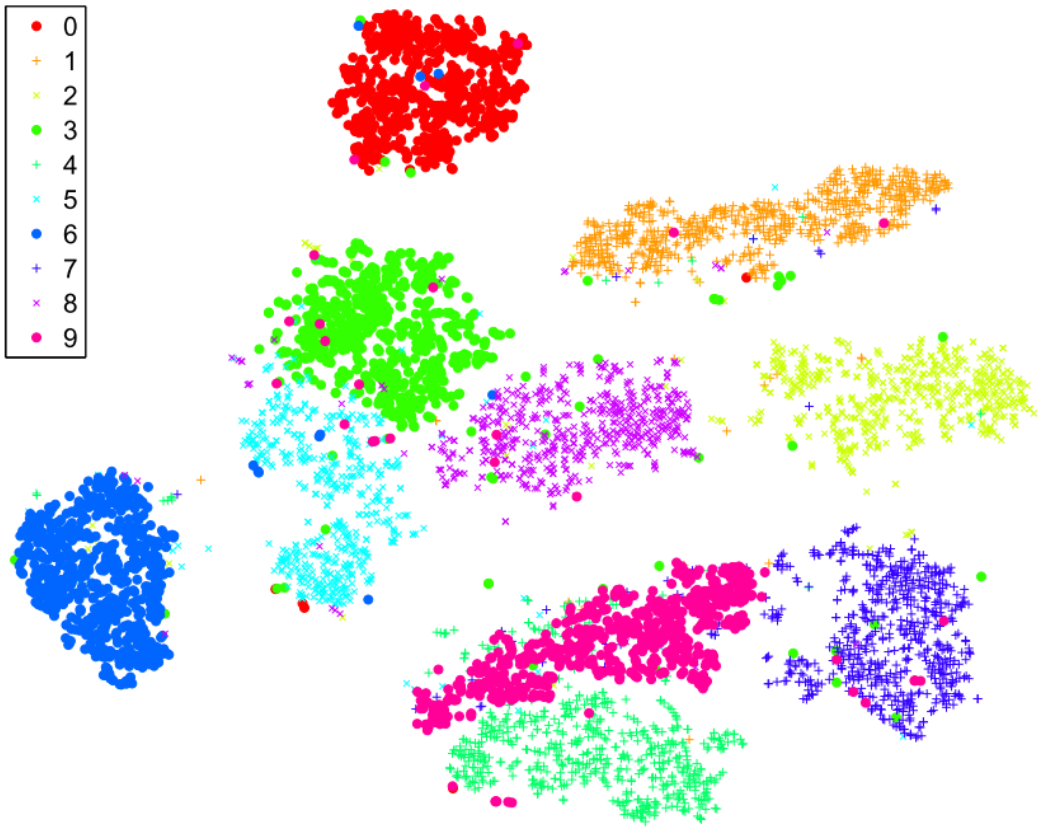


Figure 2.7: **t-SNE applied on the hand writing digit recognition database.** Each digit has been mapped with a different colour. Adapted from [113].

(or divisive) approach. With bottom-up, each sample start isolated and at each iteration pairs of clusters are then joined; vice versa, the in the top-down approach all samples initially belong to the same cluster and at each iteration are split.

The functioning of HC depends on two main parameters: the metric used to define how similar samples are, and the linkage criteria which defines the rule for merging (or splitting) clusters.

The distance metric determines the similarity of two samples where larger values represent a higher difference between the two. HC can implement any distance measure as long as it respect the rules defining a metric. Such conditions are the non-negativity ($d(a, b) \geq 0$), the identity (if $d(a, b) = 0$ then $a = b$), the symmetry ($d(a, b)$ is the same as $d(b, a)$) and the triangle inequality ($d(a, c)+d(b, c) \geq d(a, b)$). Depending on the selected metric, different clustering can be obtained, however such choice is not regulated and is highly dependent on the data that is analysed.

Some of the most applied metrics for HC are the Euclidean and its squared version, Manhattan, Maximum, Hamming and Levenshtein. Usually, Euclidean distances are the most applied in biology.

The linkage criteria is a set of rules for merging or splitting samples according to their distances. As per distance metrics, there is a wide range of linkage criteria available (complete linkage, single linkage, average linkage, entropy based linkage, Ward linkage) and the choice of the optimal is purely data-driven.

Once these two parameters are set, the HC can produce a dendrogram, a tree representation, of the similarities and clusters observed in the data. Similarly to a family tree, greater the distance between two clusters the the oldest is the common ancestor. Given such representation, it is possible to identify a variable number of clusters depending on the depth (distance) at which the dendrogram is interpreted. Usually, the number of clusters observed increase as the distance between samples reduces.

Choosing the right depth at which interpreting the result of HC frequently depends on the user intuition or some biological prior knowledge. However, as it is going to be discussed in following sections, there are methods to randomise the initial data up to a certain number of time in order to observe the rarity and therefore statistical validity of selected clusters. Such methods are known as resampling techniques.

### 2.3.3 Resampling methods

Resampling methods are tools becoming essential as data volume increase. They provide and efficient way to test the model fitness and obtain statistics on the model performance. Generally these approaches could be computationally expensive and time consuming, however the information obtained is essential.

### Permutation, bootstrapping and jackknife

Permutation, bootstrapping and jackknifes are three commonly used resampling methods used to test the validity of a null-hypothesis. Especially in clustering approaches, this null-hypothesis is the hypothesis that the detected clusters (or some classification accuracies) were obtained purely by chance.

In a permutation approach, the labels assigned to individual samples following clustering or classification, are randomised the same number of times as the desired level of significance. For example, if four clusters were identified by hierarchical clustering at a specific distance, this approach tests how many times in the permutation process we observe the same number of clusters.

The bootstrap approach is instead based on the idea of resampling as selection of subsets from the original dataset. This methodology test how close the results observed within the subset are to the one obtained when using the complete data. The subsampling performed by a bootstrap approach is a random sampling with replacement.

The bootstrap methodology was inspired by the jackknife approach. The jackknife consist in calculating statistics (or observing clusters) repeatedly by leaving one or more samples out during the iterations. As per other resampling methods, the numbers of iterations define the level of confidence of a measure.

### Cross-validation

Cross-validation is a method to assess the test error rate, the error that a model is producing after the learning step. This check is done by saving part of the dataset for testing and not using for fitting the model. Depending on the size and the number of test sets different cross-validation are performed.

**Validation set** With this method the dataset is divided randomly in half creating the validation (test) set. Ideally, while the model is fitted on the training

set, if not overfitted it will perform with similar accuracy and sensitivity also on the testing set. If this does not happen, the model has to be refitted. The error observed when modelling testing set is usually assessed through the mean standard error (MSE). The heterogeneity of both sets is essential, otherwise the model will learn a pattern that is present only in the training set. Since there is not possible to select a perfect heterogeneity, this is one major drawback of using a single validation set. Another issue that might rise is due to the limited number of observations that the model will use to train and is well known that fewer the examples worst the performance.

**Leave-one-out cross-validation** Leave-one-out cross-validation (LOOCV) approach is similar to the validation set technique but is aimed to solve those previously explained issues. The LOOCV logic consists in train and test the model multiple times where the test set is made of a different single observation for each iteration. Thus for each iteration the training is performed on the $n-1$ set and the remaining observation $(x_1, y_1)$ is tested. This will return a MSE relative to that particular observation. To estimate the overall performance of the model, the MSE is the average of all the standard errors:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{2.4}$$

This method will solve both issues observed in the normal validation, moreover since the whole dataset is used to train the model, there will not be any sampling problem returning always the same result.

**k-Fold cross-validation** Similarly to the LOOCV, the k-fold cross-validation involves multiple iterations of training and testing. Here, instead of holding out a single observation, the whole dataset is split in $k$ groups where one group is alternatively left out (Figure 2.8). Like in the LOOCV, the MSE will be calculated

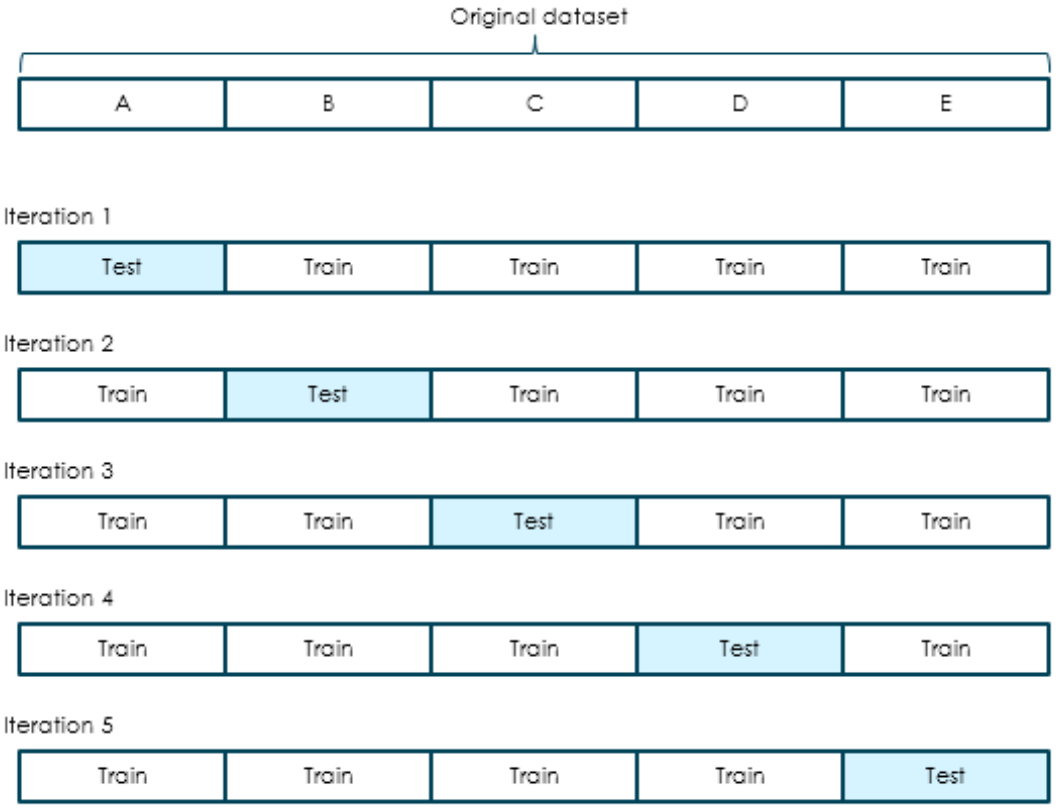


Figure 2.8: **Example of a 5-fold cross-validation.** After partitioning the original dataset in 5 subsets ($A$,$B$,$C$,$D$,$E$), 4 subsets are used as training set while the remaining one is used as testing set. This process is iterated until each subset has been used as testing set.

as the average MSE over the number of $k$ iterations:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{2.5}$$

The k-fold cross-validation adds to the LOOCV the advantage of speed in the validation process. Indeed, this method requires a limited number of iterations. This might led to the similar error rate overestimation observed in the standard validation where part of the dataset was not represented during the training, making the LOOCV the preferred method. However, considering the number of iteration done with a LOOCV logic, the variance observed in $n$ MSE is much higher than the one observed in $k$ MSE with a k-fol logic. The trade-off between bias and variance is then associated with the choice of $k$ and it has empirically demonstrated that a 5 to 10 fold cross-validation is the optimal choice for both variance and bias.

## 2.4 Feature selection algorithms

The big data term does not only refer to datasets with a great number of observations but also to those datasets where each element is observed from multiple points of view. The increasing number of features pushed the statistical learning experts to develop new methodologies to select features to be included in the analysis. The need of shortening the list of features is mainly aimed to avoid the curse of dimensionality and overfitting the model. From simple scoring systems to more complex selection criteria, many algorithms are available performing differently and promoting different aspects of data.

### 2.4.1 Univariate feature selection

Univariate feature selection is the most direct approach to reduce the number of dimensions without requiring fitting tortuous models and optimising complex functions (Figure 2.9). The univariate selection consist in exploiting univariate statistical test to rank and then select features.

$\chi^2$ and $F$ regression tests represent the most common models for testing the contribution of each feature. While $\chi^2$ and ANOVA $F$ tests are appropriate for classification problems (e.g. cases *vs* controls), the univariate linear regression test is better suited to solve regression problems.

In both approaches, the statistical test associate a p-value to each feature and subsequently ranks according to significance. When the number of tested features is much grater than the number of samples, a Bonferroni correction can be applied in order to correct for false positives. Bonferroni correction consist in multiplying each p-value for the number of tested features.

These methods are the fastest option for performing a feature selection. Although these good performance in terms of computational time, univariate feature selection models test one feature at time and do not take into consideration the combined effect of multiple variants. The effect of this limitation on the regression

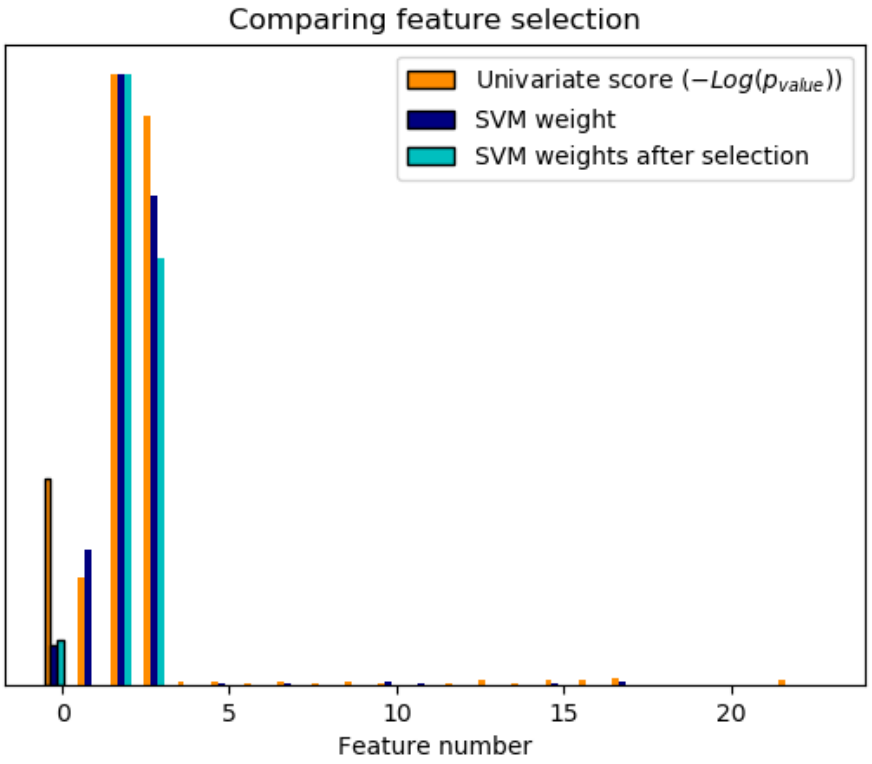


Figure 2.9: **Univariate feature selection performance.** Simulation of a feature selection problem where the first four feature are the only one significant. Blue bars show the feature weights assigned by a SVM before and after the feature selection. By applying such filtering approach SVM does not select uninformative features compared to the same model without selection. Features are shown on the x-axis. Adapted from [144]

or classification can not be predicted. Therefore, if the interaction of features is an important aspect in the approach, it is better to focus the attention on more elaborate feature selection models.

### 2.4.2 Linear regression and lasso

Linear regression considers the combined effect of multiple variants. In this case, the linear regression is called multiple linear regression in order to distinguish from its univariate form.

Multiple linear regression models the relationship between the supervised labelling and the feature set and is described by a function that minimise the fitting error, usually calculated with the least squares approach. This cost function can be substituted with other norms that penalise the common least square method as

the lasso and the ridge regression.

The most applied regularisation is the lasso L1-penalty, developed by Robert Tibshirani in order to increase the accuracy of regression models [187]. The aim of lasso is to reduce the overfitting of a model by forcing the regression coefficients (the weight of each feature) to be less than a threshold value. This regularisation induces certain feature coefficients to be shrunk to zero and therefore excluded during the selection process. An important limitation of this method is its sensitivity to the size of the input dataset. With a limited number of sample to be fitted on, lasso can not select the best features and performs at random.

$$\min_{w} \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \alpha \|w\|_1 \tag{2.6}$$

Equation 2.6 shows the objective function minimised by the model.The only parameter to be tuned in the lasso model is alpha, the penalising factor responsible for the regression coefficient shrinkage. When alpha is equal to zero, then the model performs like a classical linear regression model, while increasing alpha the model forces more coefficients to take values closer to zero. The tuning of alpha is usually an empirical process and for best results is performed via cross-validation tests.

# Chapter 3

# Machine learning classification of inflammatory bowel disease patients using histopathology data

## 3.1 Summary

In this chapter I describe the application of machine learning approaches to classify IBD patients using endoscopic and histological data. Unsupervised approaches, such as PCA and MDS, revealed a substantial overlap of CD and UC with broad clustering but no clear subtype delineation reflecting clinical complexity in distinguishing IBD subtypes. Hierarchical clustering of endoscopic and histological data identified four novel patient subgroups characterised by differing colonic involvement. Three supervised machine learning classifiers were developed utilising endoscopic only, histological only and combining endoscopic/histological data to yield classification accuracy of 71.0%, 76.9% and 82.7% respectively. The optimal combined model was tested on a statistically independent cohort of 48 additional PIBD patients and accurately classified 83.3% of patients. IBDU patients were

then reclassified by the combined model and for seventeen of them it was possible to assign a subtype diagnosis with a posterior probability greater than 80%.

Whilst Dr. James Ashton was responsible for the collection of endoscopy and histological data trough clinical notes, I was responsible for all analyses presented in this chapter.

## 3.2 Introduction

The incidence of paediatric inflammatory bowel disease, comprising Crohn's disease, Ulcerative Colitis and Inflammatory bowel disease unclassified (IBDU), has increased significantly over the last 30 years with a 46.6% increase only in England [67, 8, 68]. IBD is diagnosed through endoscopic and histological examination which inform the location and extent of the inflammation of the gastrointestinal system. As described in Chapter 1.4.1 CD and UC are distinct forms of IBD despite a substantial overlap of symptoms. Endoscopic investigation of disease is macroscopic and typically determines initial treatment and provisional diagnosis. However, the endoscopic assessment of the gastrointestinal system is not always sufficient for diagnosis and histological (microscopic) examination of biopsies from the upper and lower GI tracts is vital to determine disease extent and confirm diagnosis.

There is a well-established discordance between endoscopic and histological disease extent [51, 9, 191] with mucosal healing are frequently cited as the best marker of disease remission. Despite this, the Paris classification of PIBD (Section 1.4.1) is based exclusively on endoscopic and radiological disease extent [162, 127, 98]. Previous data has already indicated histological disease extent to be significantly greater than endoscopic disease extent, at both diagnosis and follow-up [51, 9]. This raises the possibility of a modification to the current classification to account for histological evidence as an additional measure of disease extent. However, the current endoscopic Paris classification remains a validated tool to guide diagnosis and treatment [191, 22].

The accuracy of diagnosis in PIBD is key to prompt and effective treatment. Uncertainty in the classification or the severity and extent of disease can lead to delays or inappropriate treatment [98]. Tools to assist clinicians in making a more accurate diagnosis are therefore attractive and might provide a better categorisation of disease into novel specific phenotypes with implications for how best to treat. Plevy *et al.* previously developed a multi-component machine learning model, based on serological and genetic markers, in adult IBD to discriminate current CD and UC subtypes [148]. However, genetic markers are expensive, slow and not routinely available in most hospitals. To date there are no mathematical models based solely on routinely collected clinical data to assist with diagnosis and classification.

Machine learning is a branch of artificial intelligence particularly well suited for analysis of complex data. As described in Section 1.2 machine learning algorithms aim to find patterns within data and use them to make predictions and classifications or infer new knowledge.

In this chapter we are going to utilise unsupervised models to examine the evidence for clearly distinguishable IBD strata identifiable through endoscopic and histopathological data. Potential novel grouping are then examined and regressed against main clinical features. Following this approach, we then investigate supervised support vector machine (SVM) as model for classify patient samples with established diagnoses of either CD or UC. The resulting model is tested for accuracy and its validity assessed on an unseen validation cohort. Such methodology has been used successfully in medicine and biology for cancer subtype classification, novel drug discovery and genomics [193, 104, 109, 118]. Here we use paediatric patient endoscopic and histological data to assess the utility of such approaches for the diagnosis and management of this complex disease.

## 3.3 Methods

### 3.3.1 Sample data

Patients were recruited from the Wessex Paediatric Inflammatory Bowel Disease Clinic in the genetics of paediatric inflammatory bowel disease study at Southampton Children's Hospital. Data were collected from prospectively entered electronic clinical records using a standardised proforma. Fully anonymised patient data were from endoscopy and histology reports at initial diagnosis. All patients were diagnosed according to the revised Porto criteria [100]. The dataset comprised manually collected data from 287 patients, 178 with Crohn's disease, 80 with ulcerative colitis and 29 with inflammatory bowel disease unclassified. The ratio of CD to UC is typical of paediatric onset disease.

Ten gastrointestinal (GI) locations were investigated for the presence of macroscopic and microscopic evidence of disease: mouth, oesophagus, stomach, duodenum, ileum, ascending colon, transverse colon, descending colon, rectum and perianal. Clinical observations were converted into numerical variables [-1, 0, +1] depending on tissue abnormalities. At each location, abnormal tissues observations were coded as +1 and normal were coded as -1. Null values (0) were assigned for missing data such as in the case of restriction at endoscopy. Mouth and perianal locations are not typically biopsied for histology, therefore these features were excluded in the unsupervised approach and automatically excluded in the supervised approach.

### 3.3.2 Unsupervised machine learning

In order to observe whether clinical features can induce the formation of the two clusters representing CD and UC, data were modelled using principal component analysis (PCA) and multidimensional scaling (MDS) algorithms as unsupervised machine learning approaches. As explained in Section 1.2, in unsupervised machine learning the diagnosis of CD, UC or IBDU is hidden from the model, leaving

the algorithm to return the most relevant strata. Both PCA and MDS are dimensionality reduction algorithms that convert a high dimensional space (here each dimension corresponds to a measured traits), to a lower dimensional space (usually 2D or 3D). The main difference between PCA and MDS is the search space of those two algorithms. While PCA investigates linear feature associations, MDS can also uncover non-linear associations. However, if the associations between features are essentially linear then multidimensional scaling will provide a similar representation to that of PCA.

To better visualise the relationship between patients and traits, hierarchical clustering with Hamming distance[62] and average linkage[175] was performed. Groups identified by hierarchical clustering were assessed with respect to: age of onset and C-reactive protein levels at diagnosis, using ANOVA disease subtype, gender, family history and personal history of autoimmune disease using $\chi^2$. Statistic were performed applying Python SciPy package[140].

### 3.3.3 Supervised machine learning

In order to discriminate CD and UC patients, a model was assembled utilising different techniques of supervised machine learning. We applied a supervised machine learning model where the diagnosis of CD and UC was seen by the model. In order to isolate the key histological and endoscopic features that determined diagnostic subgrouping we tested a range of classification strategies including ensemble learners (Boosted and Bagged Trees), linear discriminant analysis and support vector machines (SVMs) with a variety of different kernels[64, 45].

Data were split in order to construct and then validate the model, 210 patients (CD=143) and (UC=67) patients were included in the model construction step. Forty-eight patients (CD=35, UC=13) were set aside to validate the model on unseen data. Data from IBDU patients (n=29) were used only for a final reclassification. Figure 3.1 is a schematic representation of the model and shows the usage of the different subsets.

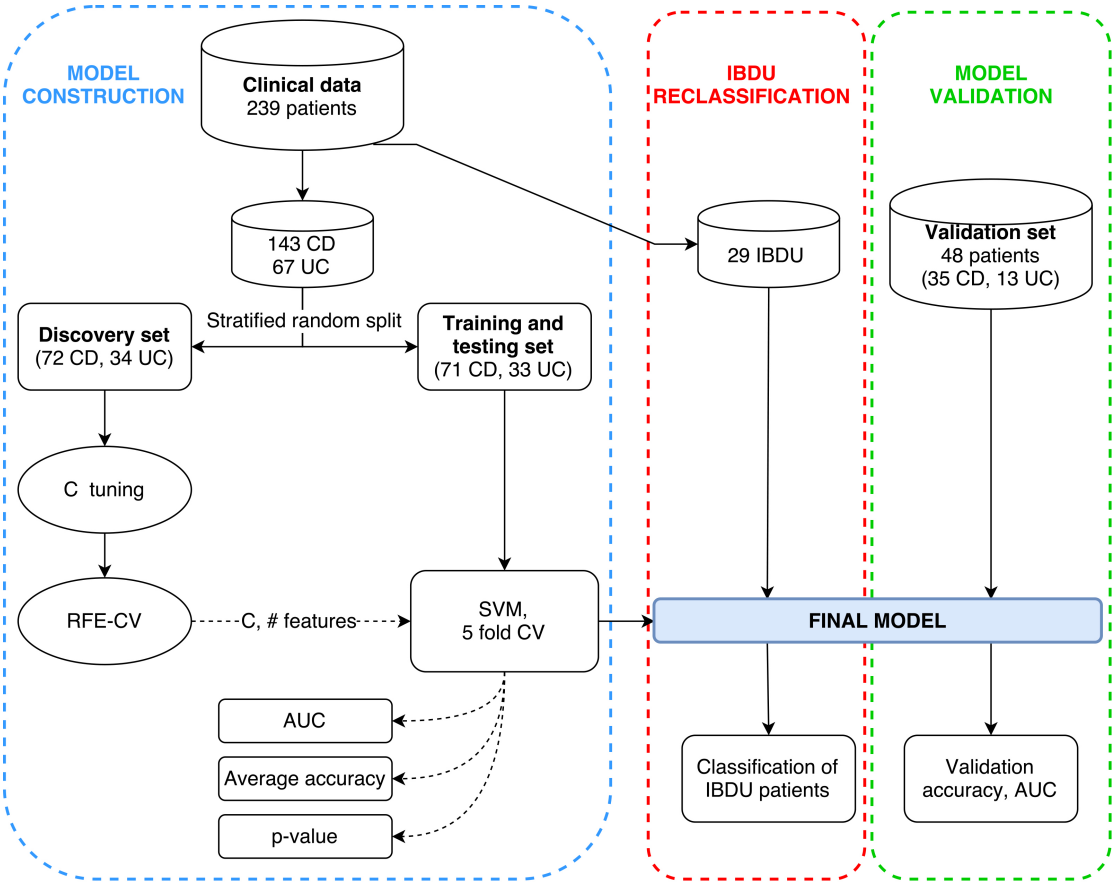


Figure 3.1: **Model schematic and histopathology data processing.** Schematic representation of the model construction (blue section), validation (green section) and IBDU reclassification (red section) phases. Solid arrows represent data stream while dashed arrows represent parameters or metrics stream. The discovery set was used to identify the optimal penalty parameter (C) and number of features using the recursive feature elimination with cross validation algorithm (RFE-CV). These two elements were then passed to the training and testing set which was then modelled using a support vector machine (SVM). Three metrics were collected: area under the ROC curve (AUC); accuracy over the 5 folds and; a permutation-generated p-value.

To create a model robust to unseen data, the 210 CD and UC samples were randomly split in two subsets preserving the original disease subtype ratio. The first data subset was used for searching the best parameters for the CD versus UC classification (discovery set). The second data subset was used for training and testing the model according to the parameters determined during the discovery phase. After assessing the performance of the final model, data from IBDU patients were passed to the model in order to classify them as either CD or UC.

Construction of optimal model utilised a linear support vector machine, allowing for regression of weights for each feature and assessment of the relative importance of each variable. Additionally, linear SVMs require estimation of a single penalty parameter (C) that allows for misclassification within the training set. In an

attempt to improve model performance when optimizing the classifier we allowed the search space for C values to range from $1 \cdot 10^{-3}$ to $1 \cdot 10^{2}$. Large values of C are less prone to misclassify data points, but perform suboptimally when classifying outliers in unseen data. Small C values generate models that are more robust to outliers by allowing more misclassified data points at the expense of the training accuracy.

Machine learning approaches are weakened by the inclusion of features that are not relevant to the classification problem (confounding factors or 'noise') and reduce model performance. In order to minimise noise from non-informative features, we applied a recursive feature elimination algorithm combined with a 5-fold cross validation scheme (RFE-CV) selecting pertinent features as described by Guyon et al. [61]. Including a 5-fold cross validation avoids overfitting the model to the discovery set by selecting parameters and features that are specific to this set but do not generalize well, and therefore perform poorly on the test subset. The selection of the best feature subset and optimal C were chosen to maximise the classification accuracy over the discovery set.

Following the identification of the optimal C and set of features, we trained a new support vector machine and tested its efficiency (Figure 3.1). With a 5-fold cross-validation scheme the algorithm repeatedly fitted and tested data from the training/testing set, providing the average accuracy in the CD vs. UC classification. The area under the receiver operating characteristic curve (AUC) was used to assess model efficiency. Statistical significance of the observed accuracy was determined through permutation testing of 1,000,000 randomly generated models in which sample labels were shuffled. The p-value was then determined by calculating the frequency at which the observed accuracy was replicated by the random models. Finally, the overall performance of the model was verified by classifying unlabelled data from the validation dataset of 48 patients.

Once the model had been fully trained and validated, it was used to classify IBDU patients and posterior probabilities for membership to both the UC and CD classes were obtained. These probabilities depend on the distance between an observation

and the decision function that SVM uses in order to discriminate between the two groups. The uncertainty in the classification of an individual increases as its profile is closer to the decision boundary (which is defined by the SVM decision function).

Data manipulation and modelling was performed using Matlab24 (R2016b), Python (2.7) and the Scikit-Learn27 (0.17.1) package.

## 3.4 Results

Endoscopic and histological data were collected for 287 patients; 178 patients with Crohn's disease, 80 with ulcerative colitis and 29 patients with inflammatory bowel disease unclassified. Machine learning was applied to 239 patients (CD=143, UC=97, IBDU=29). Females account for 37% (107) of the individuals in the dataset. Average age of onset was 11.5 years (range 1.6 to 17.6 years). Twenty-six (9%) of patients were diagnosed below 6 years of age (very-early onset IBD). The remaining 48 patients (CD=35, UC=13, average age of onset 13.2 years) were used to validate the model.

### 3.4.1 Unsupervised clustering of CD and UC phenotypes

Endoscopic and histological data underwent principal component analysis with the first three components being representative of 52.2% of the total variance of data. According to both PCA and multidimensional scaling, there was no clear separation of Crohn's disease and ulcerative colitis (Figure 3.2 A, B).

Despite the lack of distinct clusters, CD and UC individuals are differently distributed across the 3D space with regions predominantly populated by one or the other class. As anticipated, IBDU patients were distributed uniformly throughout the CD and UC data. The same clustering pattern was observed with MDS (Figure 3.2 B) strongly suggesting linear relationships between the measured features. The lack of clear clusters confirms the complexity in distinguishing CD and UC

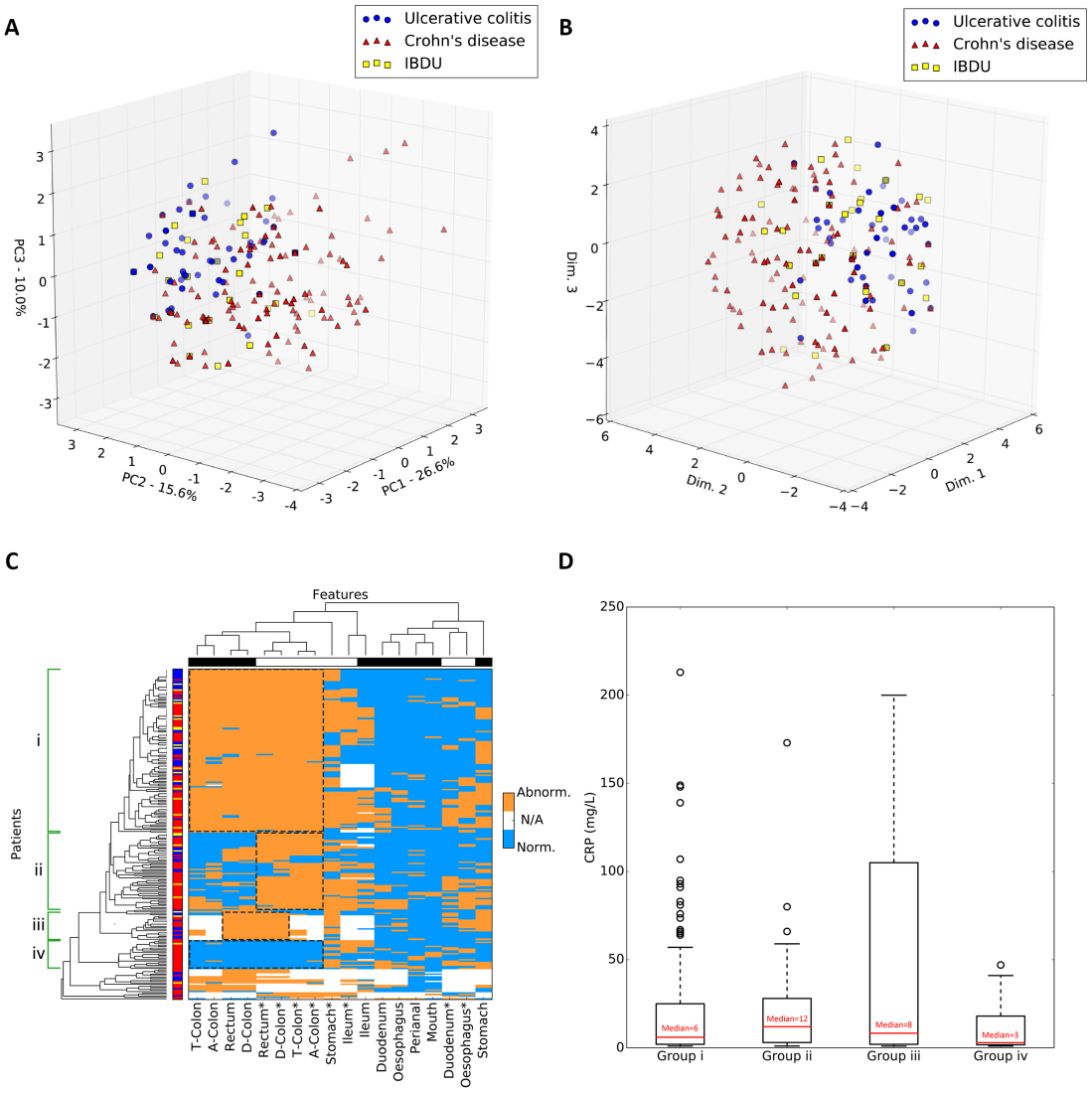


Figure 3.2: **Dimensionality reduction approaches and hierarchical clustering of histopathology data. A and B** - Principal component analysis (A) and multidimensional scaling (B) of clinical data from 239 PIBD patients. The first three PCA components account for 52.2% of the total variance. Important note – UC/CD/IBDU diagnoses were used only to retrospectively colour data points and were not included in actual modelling. **C** - Heatmap of endoscopic and histological tissue abnormalities in PIBD patients. Abnormal manifestations are shown in orange, normal in light blue and missing data in white. Asterisks indicate histology features. Ascending colon, transverse colon and descending colon labels were shortened to A-Colon, T-Colon and D-Colon respectively. Left hand side bar shows the referred diagnosis: CD in red, UC in blue, IBDU in yellow. Again, UC/CD/IBDU diagnoses were not used to model data but only to retrospectively colour each element. The top bar shows the type of investigation: histology in white, endoscopy in black. Identified colorectal groups are shown by dashed boxes and labelled from one (i) to four (iv). **D** - Box and whisker plot depicting C-reactive protein (CRP) levels recorded at diagnosis across the four identified groups. Each box represents data from the first (bottom edge) and the third (top edge) quartile. Red bars and numbers are the median CRP level. Dashed whiskers show the lowest and highest CRP within each group. Black circles are outlier data points.

phenotypes from microscopic and macroscopic observations.

### 3.4.2 Hierarchical clustering of PIBD subtypes

In accordance with PCA and MDS analyses, hierarchical clustering did not stratify patients according to CD, UC and IBDU diagnosis (Figure 3.2 C). However, it did reveal the presence of distinct subgroups of patients, corresponding to complex patterns of abnormalities. As expected, most of the macroscopic and microscopic dysregulations were observed in the colorectal region. Considering only the colorectal region, it is possible to observe four distinct groups (Figure 3.2C, i-iv). In the first group (i) patients exhibit tissue abnormalities identified by both endoscopy and histology. The second group (ii) shows colorectal abnormalities only after a microscopic investigation. Patients belonging to the third group (iii) present with inflammation of the rectum and the descending colon. Finally, the fourth group (iv) does not show any disruption of the colorectal region. Some patients are not placed within any of these four groups since they do not show any clear colorectal pattern. These patients have higher numbers of disease locations with null values (reflecting restriction at endoscopy). The ileum exhibited an inconsistent pattern of disruption, acting as interface between mostly-abnormal and mostly-normal regions (left hand side vs. right hand side of Figure 3.2C). Additionally, endoscopic or histological abnormalities in the upper GI tract are less frequent compared to lower GI tract abnormalities, this is equally applicable to all patients, regardless of their diagnosis (of CD or UC).

The four groups were analysed for any difference in their composition of patients with: a diagnosis of CD or UC; gender; positive or negative family history and clinical diagnosis of any other personal autoimmune disease. There was no significant difference between the groups with regard to any of these variables with the exception of diagnosis. Group iii (inflammation of the rectum and the descending colon) was significantly enriched for patients with ulcerative colitis ($p = 0.046$) and group iv (no colorectal involvement) was significantly enriched for patients with Crohn's disease ($p = 0.007$). Groups i and ii were not significantly enriched either for CD or UC indicating presence of both disease types.

Regression analysis of the four groups identified a significant ($p = 0.003$) increase

Table 3.1: **Preliminary assessment of linear and non-linear models.** Linear support vector machine (SVM) was the selected model.

| **Method** | **Accuracy ($\sigma$)** |
|---|---|
| Simple Tree (4 splits) | 78.1 (±1.3%) |
| Medium Tree (20 splits) | 75.2 (±1.1%) |
| Complex Tree (100 splits) | 76.7 (±2.1%) |
| Linear discriminant | 81.0 (±0.6%) |
| Linear SVM | 80.5 (±1.4%) |
| Quadratic SVM | 78.1 (±1.6%) |
| Cubic SVM | 73.8 (±0.4%) |
| Boosted Trees | 74.8 (±1.2%) |
| Bagged Trees | 77.6 (±1.5%) |

in CRP for patients in group iii compared to the other groups (Figure 3.2 D). There was no significant difference in age of diagnosis across groups.

### 3.4.3 Supervised classification of PIBD patients

Model selection was based by testing a range of different algorithms and kernels. Table 3.1 reports classification accuracies obtained fitting and testing models on the whole dataset excluding IBDU patients and the validation cohort. Reported accuracies are only informative in terms of comparing different models and were not validated on external dataset. Linear discriminant and linear support vector machine outperformed other tested algorithms. Linear models performed better than Tree-based model and non-linear SVMs. Although 0.5% less accurate compared to a linear discriminant model, linear SVM has a larger standard deviation, allowing potential better result. Moreover, linear SVMs represent the best choice in terms of adaptability and interpretation. Linear discriminant models assume data have the same covariance and a normal distribution, while SVMs does not have such requirements and is better suited for discriminative tasks [134]. Therefore, an SVM with a linear kernel was used as core classifier in our model.

In order to elucidate which observations are needed for optimal disease classification of patients, three supervised models were generated implementing ten endoscopic features, ten histological features and both endoscopic and histological features. The combined model outperforms the other two models achieving the highest accuracy; the model correctly assigns the diagnosis of CD or UC to a patient in 82.7% of cases (Table 3.2). All metrics that assess model performance agree

Table 3.2: **Performance of the three optimised supervised models.** Asterisks indicate histological features. All metrics represent the average over the 5-folds of the cross validation.

| Input | Accuracy | AUC | Precision | Recall | F1-score | (#) Features |
|---|---|---|---|---|---|---|
| Endoscopy | 71.0% | 0.78 | 0.89 | 0.68 | 0.75 | (5) Duodenum, Ileum, D-Colon, Rectum, Perianal |
| Histology | 76.9% | 0.82 | 0.81 | 0.86 | 0.83 | (1) Ileum |
| Combined (Endoscopy + Histology) | 82.7% | 0.87 | 0.91 | 0.83 | 0.87 | (8) Duodenum, Ileum, D-Colon, Rectum, Perianal, Oesophagus.*, Ileum*, A-Colon* |

in the superior efficiency when using combined endoscopy and histology data. The combined model shows the highest accuracy, precision and F1-score; recall is close to that observed in the histological model. The endoscopy model performs well in terms of precision but is poorer in recall. Conversely, the histological model has the lowest precision but highest recall. This indicates that using endoscopy data the model is highly precise in identifying most of individuals from both classes (CD and UC). However, the endoscopy model is prone to produce more false negatives (recall) compared to the histology model. Both the accuracy and the F1 score, which combines precision and recall metrics, indicate that histology model is superior to the endoscopy model although having a lower precision.

Moreover, the combined model selects all the features selected by the endoscopy and histology models plus two additional histological features (oesophagus and ascending colon). As expected, the ileum location appears to be consistently informative for the discrimination of CD and UC patients in every model, and in the histological model is sufficient to diagnose CD or UC in 76.9% of cases. Features with similar observations in both CD and UC patients are not informative for the classification while locations with a more variable manifestation of tissue damage were typically selected in the RFE-CV selection.

The greatest area under the curve (AUC) was observed in the combined model (0.87) followed by the histology (0.82) model and then the endoscopic model (0.78) (Figure 3.3 A). The endoscopic, the histological and the combined models showed a statistical significance of p= $3 \cdot 10^{-3}$, p = $5 \cdot 10^{-6}$ and p = $1 \cdot 10^{-6}$ respectively (Figure 3.3 B).

For each training fold of the combined model, the observed accuracies (in decimals)

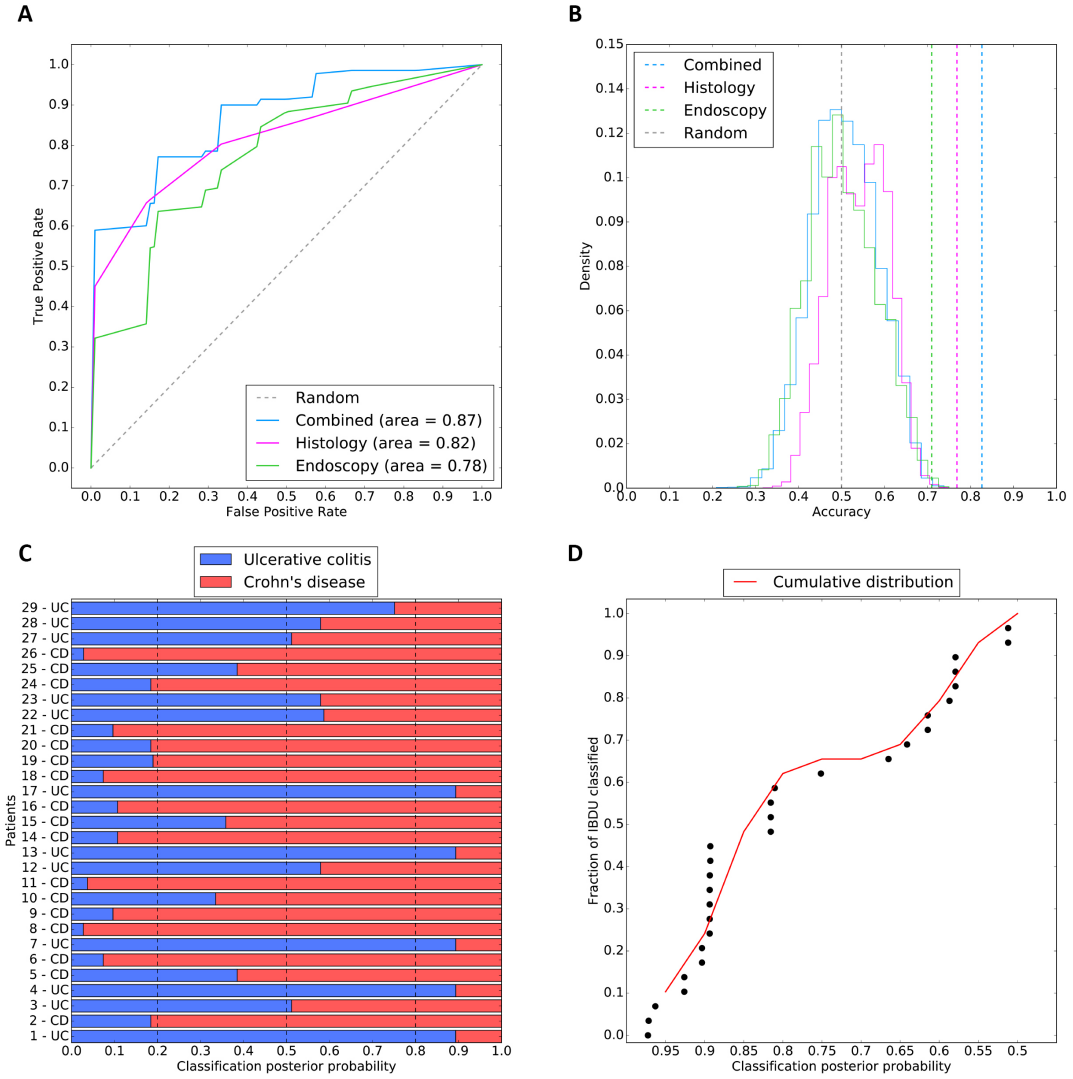


Figure 3.3: **Supervised classification performance and metrics using histopathology data. A** - Receiver operating characteristic of the combined (light blue), histology (purple) and endoscopy (green) models. The grey dashed line represents the expected performance of a random model. **B** - Permutation tests of models: dashed lines represent the observed accuracy of the combined (light blue), histology (purple) and endoscopy (green) models. The endoscopic, histological and combined models have a p-value of p=3x10-3, p=5x10-6 and p=1x10-6 respectively. The grey dashed line represents the average expected performance of random model. Solid coloured lines show the distribution of random permutations for each model. **C** - Classification of IBDU patients with the combined model in Crohn's disease (red) or ulcerative colitis (blue) subtypes. The classification posterior probability indicates the confidence of the model in assigning UC or CD labels. **D** – Cumulative confidence in IBDU reclassification represented as cumulative density function (red line) of posterior probabilities for 29 IBDU patients. Each dot represents an IBDU patient.

were 0.86, 0.67, 0.95, 0.85 and 0.80 respectively. Overall, the mean accuracy was 0.83, the median 0.85, the standard deviation 0.09 and the standard error 0.05. Over the 1,000,000 label permutations, none of the randomised models achieved an accuracy equal or greater than the observed ($p = 1 \cdot 10^{-6}$). These metrics indicate good overall performance and no overfitting of the model.

Table 3.3: **Performance of the trained combined histopathology model over the validation set.**

| Validation set | Accuracy % | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| UC | - | 0.65 | 0.85 | 0.73 | 13 |
| CD | - | 0.94 | 0.83 | 0.88 | 35 |
| Average/Total | 83.3% | 0.86 | 0.83 | 0.84 | 48 |

### 3.4.4 Model validation in an additional cohort

In order to further validate the combined histological and endoscopic model (using the 8 features shown in Table 3.2) we applied it to classify 48 anonymised PIBD patients (validation set, Figure 3.1). These data had not been used in the optimisation or training of the model. The model was accurate in classifying this additional cohort, correctly assigned the diagnosis of CD or UC in 83.3% of cases (Table 3.3). The performance metrics calculated on the validation set confirm the previous results in terms of accuracy and recall. However, precision, and consequently the F1-score, are lower when compared to the performance calculated over the test set. F1-score of the validation set is still higher than the histology and endoscopy only models.

Since the validation set never took part in any phase of the model generation, and since the model was already trained and tested avoiding overfitting, the accuracy over the validation set did not required any additional shuffling.

### 3.4.5 IBDU reclassification

The combined model was used to attempt to classify the 29 IBDU patients by assigning them to either a CD or UC subtype and computing the posterior probability of belonging to each class (Figure 3.3 C). It should be noted that the model was not trained to classify IBDU therefore patterns restricted to this class were not learnt by the algorithm. Instead the model aims to identify patterns learnt from UC and CD data in these previously unseen IBDU cases. When applied to the 29 IBDU patients, 17 patients were assigned as Crohn's disease and 12 as ulcerative colitis. In 17 of these patients the IBD subtype classification was estimated with a probability greater than 80% (Figure 3.3 C). Exploring the distribution of the

posterior probabilities (Figure 33.3 D), patients are not equally distributed across the entire probability range. The sigmoidal distribution reflects higher certainty of the model predication where patients present with a pattern learnt during the construction step but prediction accuracy declines rapidly for patients exhibiting previously unseen patterns.

## 3.5 Discussion

In this chapter we have applied machine learning algorithms to endoscopic and histological data in order to aid with classification of IBD diagnosis in paediatric patients. The resulting model demonstrates high accuracy in discriminating CD and UC patients and also provides an effective visualization of the complex overlap of these two disease subtypes.

Interpretation of the unsupervised models confirms uncertainty in discriminating CD and UC subtypes with overlapping and undefined clusters based only on disease location. We observed a limited separation of Crohn's disease and ulcerative colitis patients, with UC presenting less variance than CD cases. Based on the endoscopic and histological disease location the unsupervised models did not classify disease into distinct CD/UC subtypes, instead four distinct groups of patients were characterised by different colorectal involvement. The hierarchical clustering was not able to fit some individuals in those previously described groups. There are clear challenges in diagnostic categorisation based solely on disease location, however this model points to further subcategorization of disease, with significant overlap between UC and CD in groups i and ii. Whilst group iv is almost exclusively CD all colonic involvement has some overlap between disease types suggesting sub-classification of disease may be useful in distinguishing subtypes of CD or UC, potentially with impacts on management decisions. This theory has been raised previously through mathematical modelling of complex IBD data including serological and genetic markers. Regression analysis of CRP level at diagnosis with groups i-iv indicates a statistically significant increase in CRP in

group iii, whilst the reason behind this are uncertain there is a need to identify patients with increased systemic inflammation in order to optimise treatment. Here we provide potential evidence of the need for further subcategorization of disease based on solely on clinical parameters used in standard practice.

It is well established that ileal inflammation is key to diagnosis of Crohn's disease. Here we found that ileal inflammation (endoscopic or histological) is the only feature selected as important in all the models we constructed, providing evidence that ileal disease is the single most important factor for disease classification. Additionally, whilst colonic inflammation is important in paediatric UC, we find that it is also frequently present in CD with significant overlap between the two diseases.

Our machine learning models have been utilised for solving a classification problem (CD vs UC) and additionally to observe data structure and complexity with a view to improvement of current classification. Through the application of machine learning to these data we confirmed the higher accuracy of histological over endoscopic data if used in isolation. We also demonstrated that both investigations are needed for an optimal classification, although the current Paris classification only accounts for endoscopic disease location. Recently there has been interest in discrepancies between endoscopic and histological disease extent, with some calls to review the Paris classification of paediatric IBD to incorporate an additional histological score (Section 1.4.1). This model provides further evidence to suggest that there are significant differences between endoscopic and histological disease extent, with notable differences seen in figure 3.2C. Additionally the classification accuracy of the model of endoscopic disease alone is less than a combined model, further raising the need to discuss a modification to the Paris classification. The potential clinical utility of machine learning models such as the one we have developed are significant. By placing these basic data into the model a clinician will get a disease probability score. The model is open to incorporating additional data coming from independent clinics, leading to increasing accuracy over time.

IBDU presents an ongoing challenge to clinicians. There is broad guidance on

treatment but increasingly there is uncertainty with the diagnosis and reclassification of disease at a later stage [100]. The model described here has been developed in an attempt to classify Crohn's disease and Ulcerative Colitis at diagnosis, and not to reclassify IBDU based on disease location. Despite this, IBDU patients appear throughout the PCA/MDS plots and do not cluster, indicating a heterogeneous disease phenotype. We applied the model to 29 patients diagnosed with IBDU at initial endoscopy, 17 of these patients were assigned a probability of greater than 80% to either CD or UC based on their disease location. Posterior probabilities obtained from the classification of IBDU patients as either CD or UC, resulted in either high ($p>0.85$, n=14) or low ($p<0.65$, n=10) values, with few (n=5) exceptions. This distribution suggests the presence of at least two subgroups within IBDU patients. The first, where the model assigns the CD/UC label with high confidence, might represent a subset of patients with a clinical presentation similar to those already observed and learnt in CD and UC cases. The second subgroup, labelled with low confidence, might instead reflect a distinct clinical presentation that does not fit in the current classification criteria. Support from ML modelling may be particularly attractive for IBDU cases.

The strengths of this study lie in the robust nature of data collection. Patients recruited to this study were diagnosed by 4 different clinicians from Southampton Children's Hospital, therefore the pattern discovered by the model is not that of a single gastroenterologist. The supervised model combines different machine learning elements, but its relative simplicity makes it quick and easily interpretable. The feature selection step (RFE-CV) implicated the most informative GI locations for diagnosing IBD subtypes.

Through this model we report a diagnostic accuracy of 82.7% with an area under the ROC curve of 0.87, although for clinical application this would need to be increased to exceed 0.95 [155]. Comparing the metrics of the tested model with the performance over the validation set we conclude that: 1) the combined model performs better than individual histology or endoscopy models; 2) that both endoscopic and histological evidences are needed for an optimal classification of PIBD and 3) performance over the validation set is similar to that observed over the

test set, confirming the absence of overfitting and good generalisation. Moreover, performance metrics seen in the validation set, suggest that classification of UC patients is much more complex than for CD patients, reflecting the uncertainty observed in clinics. In total, 94% of Crohn's disease patients were successfully labelled as CD while only 65% of UCs were correctly labelled in the model. In conclusion, the missing 17% percent in accuracy can be mostly attributed to a lower discriminability of patients affected by UC. Additionally, this work can be seen as a blueprint for improvement of IBD categorisation in the future, through modelling of additional data, such as variants from whole-exome sequencing, transcriptome profiles and microbiome signatures it may be possible to gain further, clinically relevant, disease groups [198]. In the future this may aid with treatment selection, prognostication and ongoing management.

In conclusion, we presented a mathematical model of histological and endoscopic data within IBD; it provides a model with high diagnostic accuracy on unseen data (83.3%). We present 4 novel subgroups of disease identified by unsupervised machine learning based on colonic disease.

The purpose of this chapter was two-fold, to better understand disease aetiology, heterogeneity and classification and to understand the potential for machine learning to assist with disease classification using solely clinical data. Through further work machine learning can aid clinicians to accurately subtype disease and personalise treatment. Additionally this may help with classification of IBDU. Whilst existing methods for diagnosis appear robust, the opportunity to improve and personalise therapy for patients through new and more accurate subtyping of disease is exciting and increasingly tangible.

# Chapter 4

# GenePy – a tool for estimating gene pathogenicity in individuals using next-generation sequencing data

## 4.1 Summary

In this chapter I describe the design, development and testing of GenePy, a gene score to transform NGS data that preserves biological information. GenePy aims to fill the lack of approaches for annotation and interpretation of genomic sequencing data for complex diseases. The key aspect of GenePy is the gene-based approach capable of assigning a score of deleteriousness on a per-patient basis over sub-genomic regions such as genes. Implementing known deleteriousness metrics in addition to incorporating allele frequency and zygosity information, GenePy improves the modelling of biological information brought by NGS data. Scores were generated for 15,000 genes across 508 individuals using whole exome sequencing data.

Despite relatively modest sample sizes, typical for NGS data, when assessing

GenePy scores for *NOD2*, a well established IBD associated gene, Crohn's disease patients exhibit a higher level of deleteriousness compared to controls ($p = 1.37 \cdot 10^{-4}$). GenePy scores demonstrate increased power to significantly discriminate CD patients from controls than SKAT-O, the most popular test used for assessing the combined effects of common and rare variation. This chapter additionally describes the potential of GenePy as per-gene/per-individual score to facilitate downstream integration of NGS data into machine learning, network and topological analyses.

All the work presented in this chapter, from the NGS data processing to the model development and testing, was conducted by myself.

## 4.2 Introduction

As result of price reduction and increased throughput, next-generation sequencing (NGS) has emerged as an effective tool for detecting single nucleotide variants (SNVs) causing rare Mendelian conditions [189]. This resulted in an increased application of whole exome sequencing in clinical framework, increasing the diagnostic yield of rare diseases by 25-31% [77, 167]. Through comparison against human genome reference sequence, it is possible to identify in excess of 30,000 variants when based on whole exome data that captures all the coding regions of the genome. As explained in Chapter 1.1.1, the number of identifiable variants scales quickly when sequencing is performed on the entirety of human genome.

The sole identification of very rare variants in empirically implicated candidate genes related to the phenotype of interest is not sufficient to imply causality. Further exclusion/filtering criteria has to be applied in order to remove variants which might not have an impact on protein amino acid sequence or that occur more frequently than the disease of interest. These steps can reduce the search space for causal variation by orders of magnitude to smaller sets of hundreds or even tens of genetic alterations that are then prioritised by in silico methods [59].

Many prediction tools have been developed in order to estimate the potential impact of genetic variants on gene/protein function. Predicting pathogenicity or deleterious impact can be achieved through a variety of algorithms that focus on one or more specific biological aspect(s). Three broad classes of deleteriousness prediction metrics are: (i) conservation metrics, (ii) function alteration metrics and (iii) composite scores. Conservation metrics such as GERP++ [41], phastCons [172] and phyloP [150] assign a high deleteriousness to variants where the homologous position in other species has remained constrained over evolutionary history. Scores focused on predicting the potential disruption of protein functionality, for example through alteration of resultant protein amino acid sequence, include SIFT[173], FATHMM [170], fathmm-MKL [169], PolyPhen2 [3], MutationTaster [166], PROVEAN [37] and VEST3 [32].

To date, no single metric has proven unilateral superiority in estimating consequent severity, despite an expanding list of metrics based on subtly different foundations and assumptions [30]. While individual metrics have the ability to perform well in isolation, discordant evidence when assessing the same data with multiple metrics has led to increased uncertainty in choice of prediction tool [38]. This in turn has led to the development of a range of composite prediction tools applying statistical and machine learning methodologies that combine metrics assessing both conservation and functionality in order to obtain higher accuracy [186]. Amongst the most utilised composite scores there are CADD [88], MetaSVM and MetaLR [46], M-CAP [75] and DANN [151]. CADD currently detains the highest AUC in detecting pathogenic variants in the ClinVar database whilst DANN is its evolution based on artificial neural network. Despite so, no one method emerged as optimal [116]. For this reason, when assessing variant deleteriousness it is still necessary to observe consensus prediction based on multiple scoring metrics rather than focusing on any single score [106]. This remains the case when studying rare Mendelian disease where single gene mutations imparting severe consequence are expected to represent the most extreme set of deleterious variants.

In contrast to rare diseases, common genetic diseases such as ischemic heart disease, asthma, inflammatory bowel disease (IBD) and Alzheimer's disease are

caused by the combined action of multiple genetic variants working in combination with environmental factors [49]. Collectively, common diseases impose an enormous economic burden and arguably have the greatest unmet need for diagnosis and stratified treatment [164]. Despite the same clinical presentation, the set of genes and variants inducing the phenotype varies from patient to patient. This large heterogeneity is one of the key elements of complex diseases.

Prior to the advent of NGS, genome-wide association studies (GWAS) were the most fruitful approaches for linking genetics to the molecular bases of complex diseases. These studies typically reports more than a million common single nucleotide variants across the genome and identified statistically significant associations of bialleleic markers with very large cohorts of independent patients compared to ethnically match controls. Genetic regions implicated by GWAS were assumed to harbour common variants in genes or regulatory elements underpinning the disease of interest. However, since these genetic breakthroughs were achieved using necessarily huge cohorts of patients and controls, they were largely uninformative on an individual patient basis. The variants identified simply associate genomic regions without necessarily being causal. Importantly, the relevance and value of GWAS findings to individual patients cannot be translated to clinical practice in terms of either diagnosis or treatment.

The application of NGS to improve our understanding of common complex diseases has been largely limited to burden tests, consisting of combined association tests integrating information from common and rare variation across defined genomic regions such as genes. While this approach broadens the search space by including rare variants, detectable by sequencing approaches but not through GWAS, they are most often implemented through collapsing multiple variants into a single value for univariate analysis. The modest success of these approaches may be partly attributable to their intrinsic lack of biological information and inclusion of both causal and benign genetic variation [101, 131]. In order to address this limitation, Neale *et al.* developed the C-alpha test, correcting for both protective and deleterious variants but at the cost of losing statistical power. Currently, SKAT (and SKAT-O optimised for small sample size) [95] represents the most

sensitive approach to test for association between a genomic region and a phenotype. SKAT jointly assesses both rare and common variants maximising the statistical power and representing a new class of analysis lying between burden and association tests and has been successfully applied to a large variety of complex diseases [195, 158, 159, 184, 185]. While NGS is proving a revolutionary technology for the diagnosis and treatment of rare diseases, its relatively modest application in common diseases is limited by the lack of analytical approaches incorporating individual profiles of genetic variation annotated with biologically meaningful information. Instead of variant-level approaches, typical for rare disease or large cohort approaches (e.g. GWAS), contemporary analyses of complex polygenic disorders requires the development of tools that combine both mutational burden and biological impact of a personalised set of mutations into single scores for discrete genes. This chapter, describes the development and implementation of GenePy, a novel gene-level scoring system for integration and analysis of next-generation sequencing data on a per-individual basis. GenePy incorporates variant pathogenicity scores, allele frequency and zygosity and sums across all variants within a gene for each patient. As consequence of a standardised approach for collecting GenePy scores, single gene values can be compared between individuals. Following correction for gene size, all gene scores or subsets reflecting pathways, can be implemented in downstream network analyses or used as input for machine learning to stratify or classify disease subtypes. We validate GenePy performance by comparing the genes scores for a cohort of paediatric IBD patients against a non-IBD cohort for the *NOD2* gene – a widely, accepted 'positive control' gene for causality in complex IBD (Section 1.4.2).

## 4.3 Methods

### 4.3.1 Sample data

Whole exome sequencing (WES) data were derived from two sources. This first group comprised 309 patients diagnosed in childhood with IBD. This cohort (fur-

ther described in chapter 3) includes unrelated, Caucasian patients ascertained and recruited through Southampton Children's Hospital who were diagnosed under the age of 18 years according to the modified Porto criteria [97]. Additional WES data from a cohort of 199 anonymised individuals diagnosed with an infectious disease but unselected for any form of autoimmune disease were also used to give a total cohort size of 508 individuals with WES data.

Genomic DNA was extracted from peripheral venous blood using the salting out method [124]. DNA concentration was estimated using the Qubit 2.0 Fluorometer and the 260:280 ratio calculated using a nanodrop spectrophometer. Fragmented DNA was subjected to adaptor ligation and exome library enrichment using the Agilent SureSelect All Exon capture kit versions 4, 5 and 6. Enriched libraries were sequenced on Illumina HiSeq systems.

### 4.3.2 WES data processing

WES data from the IBD and control cohorts were processed with the Southampton custom pipeline (Section 2.2). VerifyBamID [80] was utilised to check the presence of DNA contamination across our cohort of 508 individuals.

Alignment was performed against the human reference genome (GRCh38/hg38 Dec. 2013 assembly) using BWA[103] (version 0.7.12). Aligned BAM files were sorted and duplicate reads were marked using Picard[147] (version 1.97). Following GATK recommendations [44], base qualities were recalibrated in order to correct for systematic errors produced during sequencing. Finally, variants were called using GATK[122] (version 3.7) HaplotypeCaller producing a gVCF file for each sample. Samples were processed on IRIDIS4, the University of Southampton computing cluster, and required on average 4 hours of running time on a 16 processors node per sample. The bioinformatic pipeline is further detailed in Section 2.2.

While the standard VCF format reports only alternative calls, the gVCF format records also regions where variants were not observed, also known as non-variant

blocks. This difference enables calling of homozygous reference loci when combining the call sets from IBD and controls cohorts. The multi-sample variant calling was performed by calling separately each sample and then merging all gVCFs using GATK GenotypeGVCFs. This procedure ensures the accurate calling of homozygous reference genotypes which are otherwise ignored and reported as missing genotypes (./.). Since directly genotyping 508 samples would result in weeks of computational time, samples were first combined in six batches using GATK's CombineGVCFs (approx. 6 hours/batch on a 16 proc. node) and then gennotyped with GenotypeGVCFs (approx 1h on a 16 proc. node).

Variant annotation was performed on the genotyped VCF files obtained from the previous step. Using Annovar (version 2016Feb01) variants were annotated against: refSeq gene transcripts (refGene), deleteriousness scores databases (dbnsfp33a) and dbSNP147. Variants allele frequencies were obtained through Annovar (ExAc03) and the ensembl human variation API [52].

### 4.3.3 Quality Control

In order to reduce heterogeneity, it is necessary to control for bias encountered due to alternative capture kit versions and variant quality. For the entire cohort of 508 samples, exon enrichment was performed using Agilent SureSelect capture kits but at different time-points. For this reason, there is inter-capture kit variability across the 508 cohort with kit versions 4, 5 and 6 being applied. To correct for disparity in the regions targeted by respective versions, all downstream analyses were restricted to the set of overlapping targeted genomic locations (as defined by respective kit BED files) using BEDtools v2.17 [152]. Following GATK best practice guidelines, HaplotypeCaller default settings were utilised, implying that only variants with a minimum Phred base quality score of 20 were called.

### 4.3.4 GenePy score

Individuals often have more than one variant in a gene making the interpretation of their combined effect a challenging task. In order to quantify the contribution of multiple variants within a gene in defining the IBD phenotype, we developed a gene-level score that considers the genotype of observed variations, their frequencies in the general population and the estimated deleteriousness assessed using a variety of existing deleteriousness metrics.

We hypothesised that for each individual sample $h$ within our cohort $H = \{h_1, h_2, \ldots, h_n\}$ the loss of integrity of any given gene $g$ in the refGene database $G = \{g_1, g_2, \ldots g_m\}$ can be quantified as the sum of the effect of all $(k)$ variants within its coding region observed in that sample, where each biallelic mutated locus $(i)$ in a gene is weighted according to its predicted allele deleteriousness $(D_i)$, zygosity and allelic frequency $(f_i)$. The GenePy score $S_{gh}$ for a given gene $(g)$ in individual $(h)$ is

$$S_{gh} = -\sum_{i=1}^{k}(D_i \log_{10}(f_{i1} \cdot f_{i2}))) \tag{4.1}$$

Importantly, the choice of variant deleteriousness score is user-defined, and therefore the GenePy score is able to take into account different definitions of pathogenicity depending on context. Herein we examine the relative attributes of using any one of sixteen of the most commonly applied scores (Table 4.1). At any one variant locus (i), we represent both parental alleles using $f_{i1}$ and $f_{i2}$ to embed the population frequency of $\text{allele}_1$ and $\text{allele}_2$ and in doing so model observed biological information on both frequency and zygosity. Any homozygous genotype therefore is simply the observed allele frequency squared whereas the product of each of the observed alleles is calculated for heterozygous genotypes. The latter can therefore accommodate variant sites with multiple alleles in addition to the typically encountered bialleleic single nucleotide polymorphisms (SNPs). Hemizygotic variation from male X-chromosomes are treated as homozygotic. Where a variant may be novel to an individual or absent from reference databases, we impose a lower frequency limit of 0.00001. This lower limit is arbitrarily set to con-

servatively reflect the lowest frequency that can be observed in the largest current repositories of human variation (ExAc03). The log function is applied to upweight the biological importance of rare variation.

Deleteriousness metrics were developed to assess damage induced by nonsynonymous variation, therefore structural variants such as frameshifts or stop mutations that truncate proteins are not routinely assigned deleteriousness values. Due to their highly detrimental impact to function we assign all protein truncating mutations the maximal deleteriousness value of 1. Synonymous and splicing variants are not routinely annotated by ANNOVAR and were not included in the current assessment.

Sixteen of the most common deleteriousness ($D$) metrics were selected for implementation within the GenePy algorithm (Table 4.1). Five of these metrics (shown in bold) are unbounded. In order to implement unbounded metrics in GenePy it was necessary to impose lower and upper limits by applying the respective minimum and maximum values observed in the dbnsfp33a database of 83,422,341 known SNV mutations. These limits were used to transform observed values in our cohort scaled to 0-1.

As a function of their size alone, larger genes have greater opportunity to accumulate higher deleterious GenePy scores through having a greater number of variants thus inflating GenePy scores. We therefore generated GenePy scores corrected for gene length ($\text{GenePy}_{cgl}$) by dividing the GenePy score by the targeted length in base pairs and then multiplying by the median observed targeted gene length in our data (1461 base pairs). A final set of 16 deleteriousness metrics, each with a range of 0-1 where highest values were most deleterious, were individually implemented in the model.

### 4.3.5 Score validation

In the absence of any comparable gene based scoring system for individuals, GenePy performance was benchmarked by assessing its power to determine sig-

Table 4.1: **Pathogenicity scores for SNVs and their reported ranges in the dbsnfp database.** § In order to maintain uniform directionality, the complement (1 – score) of a value was taken so that across scores, a value of 0 consistently indicated benign variation and a value of 1 inferred maximal pathogenicity.

| Metric | Type | Implementation | Actual range | Imposed range for transformation |
|---|---|---|---|---|
| **CADD** | Composite | Score | -∞ to +∞ | -7.53 to 35.79 |
| DANN | Composite | Score | 0 to 1 | - |
| **FATHMM§** | Functionality | 1-Score | -∞ to +∞ | -16.13 to 10.64 |
| fathmm-MKL | Composite | Score | 0 to 1 | - |
| **GERP++_RS** | Conservation | Score | -∞ to +∞ | -12.3 to 6.17 |
| M-CAP | Composite | Score | 0 to 1 | - |
| MetaLR | Composite | Score | 0 to 1 | – |
| **MetaSVM** | Composite | Score | -∞ to +∞ | -2 to 3 |
| MutationTaster§ | Functionality | 1-Score if N/P; Score if A/D | 0 to 1 | - |
| phastCons | Conservation | Score | 0 to 1 | - |
| **phyloP** | Conservation | Score | -∞ to +∞ | -13.28 to 1.2 |
| Polyphen2_HDIV | Functionality | Score | 0 to 1 | - |
| Polyphen2_HVAR | Functionality | Score | 0 to 1 | - |
| PROVEAN§ | Functionality | 1-Score | -14 to 14 | - |
| SIFT§ | Functionality | 1-Score | 0 to 1 | - |
| VEST3 | Functionality | Score | 0 to 1 | - |

nificantly different score distributions in disease cases compared to controls for a known causal gene and using the same variant data, comparing GenePy results against that of SKAT-O, the most commonly applied gene level association test. The cohort comprised 309 individuals diagnosed with inflammatory bowel disease (IBD) and 199 controls unselected for autoimmune conditions. The analysis focussed only on the *NOD2* gene which represents the most strongly and repeatedly associated common disease gene conferring strong association specifically with the Crohn's disease (CD) subtype of IBD [70, 108, 93]. NOD2 was selected as a positive control gene, whereby evidence for increased burden of deleterious mutation encoded in CD patient DNA compared to either ulcerative colitis or control DNA is expected.

The matrix of *NOD2* GenePy scores calculated for all 508 samples was split into controls and cases with the latter further divided into UC and CD subtypes. Statistical significance of GenePy score distribution difference between groups was calculated using the Mann Whitney U test for unpaired data. Using the same variant input data, the SKAT-O gene based test for association was performed twice using default settings: firstly by considering all variants called within *NOD2* and secondly including only rare variants (MAF<0.05) as per developer recommendations [95].

Association tests are highly sensitive to false positive results due to spurious association brought about by population stratification or systematic differences in case versus control data. By analysing data from multiple populations the variability that comes with different ethnicities can introduce noise and might mask other possible stratifications. Since the Caucasian ethnic group is the most represented in our cohorts, we excluded non-Caucasian individuals identified through comparison against the 1000 Genomes Project [1] using Peddy software [143] for ethnic imputation. Due to compatibility requirement the multi-sample VCF for our cohort was lifted-over to hg19 reference genome build from the original hg38. Since the lift-over was necessary solely for ethnicity imputation, we did not investigated the performance of this step in terms of contigs that were not mapped to the previous build. The power to call genetic variants is dependent on the quality of sequencing data [4]. Ajay *et al.* demonstrated the efficiency of genotype calling as a function of average depth of coverage and reported that the variant detection reaches saturation with an average coverage between 40X to 45X. An average coverage of 50X therefore ensures the detection of 95% of variants observed at the ideal coverage of 100X. To ensure all samples were of comparable quality, those showing an average coverage less than 50X per each investigated gene were not included in downstream analyses.

## 4.4 Results

### 4.4.1 QC results

All WES data (N = 508; $N_{ibd}$ = 309; $N_{ctrl}$ = 199) underwent quality control assessment for contamination using VerifyBamID and were confirmed free of contamination (free-mix statistic <0.01). Out of 508 individuals, we identified three pairs of first degree relatives, one set of monozygotic twins and one mother-father-child trio. In order to correct for relatedness (which would bias association tests) for each related pair, the sample with poorest coverage data was excluded. For the trio, the child data were excluded and unrelated parents retained. The combined

genomic dataset (multicalling VCF) containing all the variants called across 508 individuals consisted of 381,451 unique variants.

### 4.4.2 GenePy score behaviour – impact of allele frequency and zygosity

Simulated GenePy score (y-axis) were calculated using a range of deleterious metric scores (0.1, 0.5, 0.75, 0.9, 0.95, 0.99) and varying minor allele frequency (x-axis) (Figure 4.1). The resulting distributions demonstrate the impact of deleteriousness and frequency as well as heterozygote versus homozygote states. The plot reveals the logarithmic nature of GenePy scores for a single variant only (whereas for any individual, their per gene GenePy score is weighted sum of all variant scores observed in that individual across that gene). For any single variant, the theoretical maximum observable GenePy value of ten occurs only with highest deleteriousness value ($D$), the lowest minor allele frequency ( MAF = 0.00001) and in the homozygous state whereas the upper limit for a heterozygote with the same deleteriousness and frequency settings is five. The logarithmic scale implemented in GenePy algorithm confers rapidly increasing scores as the MAF approaches novelty.

### 4.4.3 GenePy score behaviour – impact of deleteriousness metric

While there are 27,238 genes annotated in RefSeq, we aimed to generate GenePy scores only for the overlapping subset of 21,577 target genes captured by all versions of the Agilent SureSelect capture kits applied. The GenePy scoring algorithm was executed for each of sixteen commonly applied metrics (Table 4.1). There is fluctuation in the number of genes for which variants were annotated with deleteriousness metric data using ANNOVAR ranging from 12,921 for M-CAP (one of the most recently released scores) to 14,745 genes annotated scores for Polyphen2_HDIV (one of the earliest developed deleteriousness scores) (Table

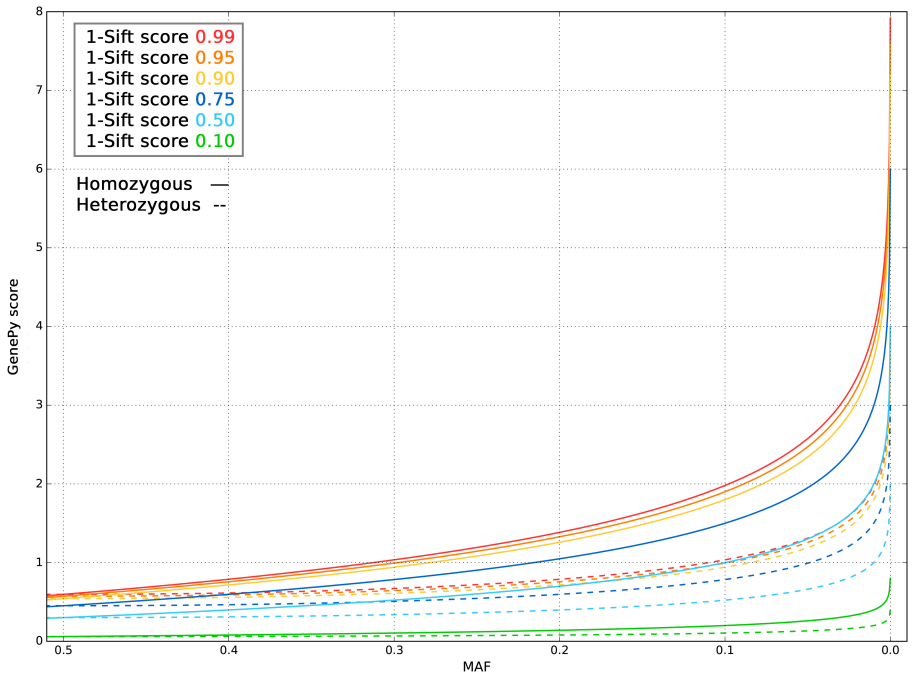


Figure 4.1: **Single variant GenePy score distribution under fixed deleteriousness values.**

4.2). Among the 508 individuals that underwent GenePy scoring of exome data, the majority of genes are invariant within any one individual (e.g. median 9917 for CADD metric). This is expected for intrinsically sparse genomic data. However, across the cohort, no single gene returns a GenePy score of zero in all individuals indicating all genes have at least one variant observed amongst the 508 individuals. The vast majority of genes are scored with GenePy values of less than 0.01 and correction for gene length marginally increases the number of genes achieving lowest scores. More than 97% of genes achieve a score of less than 0.01 when the M-CAP metric is used whereas FATHMM scores approximately 65% of genes in the 0 – 0.01 range. The inflated percentage of invariant genes observed when implementing M-CAP is explained by its tendency to depress weight for benign variants compared to other tested metrics[75].

Across the $14,\tilde{0}00$ genes achieving GenePy scores, the observed score mean (uncorrected for length) in our cohort of 508 samples ranges from 0.02 to 0.40 depending on the applied deleteriousness metric. There is only modest effect on the range of the mean scores observed after correction for gene length (0.02 – 0.31). However,

the gene length correction causes increased spread of the data reflected by an approximate two-fold increase in the coefficient of variation (CoV) for GenePy scores generated that is consistent across all sixteen deleteriousness metrics. GenePy scores generated with M-CAP are least impacted by gene length correction but maintain the largest CoV despite this score demonstrating the lowest maximum value. Moreover, following the correction, the maxima across deleteriousness metrics increase by approximately three to four folds.

Table 4.2: **Statistical attributes of whole gene GenePy scores computed for sixteen deleteriousness metrics.** Number of genes for which GenePy scores were calculated, median number of non-variant genes (GenePy=0), mean GenePy scores, mean and standard deviation across our cohort (n=508), coefficient of variation (CoV, defined as $\sigma/\mu$) and the median number of genes with a GenePy score <0,01 as percentage of the total number of genes. The same information is reported for GenePy$_{cgl}$. § Across the cohort of 508 individuals assessed, individual samples have a very high median number of invariant genes resulting on GenePy scores of zero.

| Metric | Gene scores calculated | §Median no. of genes with GenePy=0 within individuals (%) | Max GenePy | Mean GenePy | CoV uncorrected | Median no. of genes with GenePy<0.01 within individuals (%) | Max GenePy$_{cgl}$ | Mean GenePy$_{cgl}$ | CoV$_{cgl}$ corrected | Median no. of genes with Genepy$_{cgl}$<0.01 ($\delta$%) |
|---|---|---|---|---|---|---|---|---|---|---|
| CADD | 14184 | 9917 (69.92%) | 32.15 | 0.10 | 3.81 | 10231 (72.13%) | 74.19 | 0.08 | 8.09 | 10304 (0.51%) |
| DANN | 14184 | 9917 (69.92%) | 110.48 | 0.33 | 3.37 | 10153 (71.58%) | 304.15 | 0.25 | 6.96 | 10196 (0.30%) |
| FATHMM | 13143 | 9981 (75.94%) | 72.73 | 0.16 | 4.15 | 10923 (83.11%) | 269.62 | 0.11 | 6.42 | 11092 (1.29%) |
| fathmm-MKL | 14178 | 9039 (63.75%) | 50.10 | 0.16 | 3.29 | 9282 (65.48%) | 131.34 | 0.12 | 7.55 | 9332 (0.36%) |
| GERP++_RS | 14197 | 9910 (69.80%) | 100.44 | 0.32 | 3.35 | 10116 (71.25%) | 283.69 | 0.24 | 6.47 | 10143 (0.19%) |
| M-CAP | 12921 | 12577 (97.34%) | 24.52 | 0.02 | 12.65 | 12596 (97.48%) | 59.88 | 0.02 | 19.05 | 12630 (0.26%) |
| MetaLR | 14063 | 12752 (90.68%) | 38.14 | 0.04 | 8.77 | 13146 (93.48%) | 87.80 | 0.04 | 16.14 | 13253 (0.76%) |
| MetaSVM | 14076 | 9845 (69.94%) | 36.76 | 0.10 | 3.95 | 10141 (72.04%) | 99.44 | 0.08 | 8.94 | 10207 (0.47%) |
| MutationTaster | 14039 | 12161 (86.62%) | 90.86 | 0.13 | 5.24 | 12521 (89.19%) | 332.05 | 0.09 | 9.02 | 12579 (0.41%) |
| phastCons | 14197 | 10217 (71.97%) | 100.64 | 0.21 | 3.79 | 11018 (77.60%) | 324.41 | 0.14 | 5.76 | 11116 (0.69%) |
| phyloP | 14202 | 9910 (69.78%) | 118.81 | 0.40 | 3.31 | 10107 (71.17%) | 332.05 | 0.31 | 7.15 | 10131 (0.17%) |
| Polyphen2_HDIV | 14745 | 11824 (80.19%) | 65.48 | 0.14 | 4.89 | 12558 (85.16%) | 257.00 | 0.12 | 12.08 | 12658 (0.68%) |
| Polyphen2_HVAR | 14741 | 11470 (77.81%) | 59.67 | 0.11 | 5.47 | 12621 (85.62%) | 239.71 | 0.09 | 14.03 | 12778 (1.07%) |
| PROVEAN | 13888 | 9733 (70.08%) | 74.16 | 0.23 | 3.37 | 9958 (71.70%) | 219.39 | 0.17 | 7.93 | 10003 (0.32%) |
| SIFT | 14561 | 11088 (76.15%) | 99.69 | 0.25 | 3.69 | 11224 (77.08%) | 265.64 | 0.20 | 7.04 | 11257 (0.23%) |
| VEST3 | 14170 | 9919 (70.00%) | 53.36 | 0.09 | 5.69 | 10528 (74.29%) | 136.56 | 0.08 | 12.56 | 10821 (2.07%) |

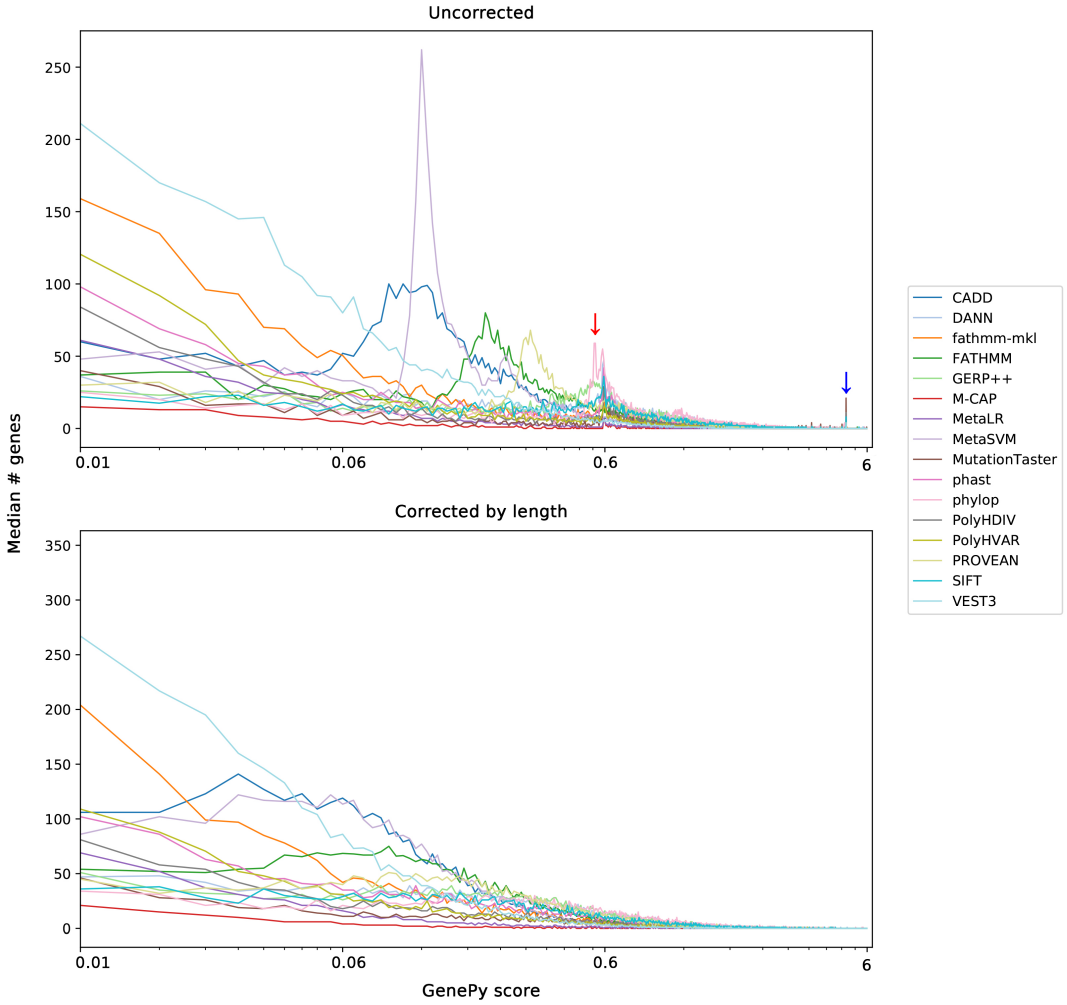


Figure 4.2: **Median whole gene GenePy score profiles observed across the cohort of 508 patients with WES data for all sixteen metrics of deleteriousness.** Uncorrected GenePy scores (upper panel) exhibit characteristic spikes reflecting gene scores strongly influenced by the effect of: single highly deleterious (D = 1) common homozygous variants (red arrow) or; single highly deleterious very rare/novel variants (MAF = 0.00001) (blue arrwiw). GenePy$_{cgl}$ score profiles (lower panel) do not display these spikes. Invariant genes conferring a GenePy score <0.01 are overrepresented and not shown here by commencing the x-axis with the 0.01-0.02 bin. All sixteen versions of the GenePy score exhibit long tails in the GenePy score distribution truncated here at a score of six. X-axis was log transformed to improve interpretation of scores <0.6.

In order to further investigate the behaviour of GenePy scores across genes, we calculated the median number of genes exhibiting scores falling within non-overlapping bins across the entire cohort. Figure 4.2 (and Supplementary Figure 7.2) shows the profiles for the 0.01 to 6 range of GenePy scores and a bin size of 0.01. Genes with scores <0.01 are overrepresented (Table 4.2) and not shown. Across most of the sixteen metrics, a distinct pattern characterised by two spikes around uncorrected GenePy scores of 0.6 and 5 represent genes strongly influenced by a single highly deleterious common homozygous variants (D=1, MAF=0.5) or a single highly deleterious very rare heterozygous variant (D=1, MAF=0.00001) respectively. This

profile was apparent for most deleteriousness metrics (except CADD, FATHMM, MetaSVM and VEST3, see Supplementary Figure 7.2). These two distinctive spikes are not observable once GenePy scores are corrected for the targeted gene length (Figure 4.2, lower panel and Supplementary Figure 7.3 ). We did not observe further spikes or other anomalies in the long right tail of the distribution of scores greater than 6.

### 4.4.4 GenePy score testing

When testing for association, it is necessary to remove sources of variation that would bias the statistical test. In this framework, bias conferred by uneven *NOD2* gene coverage, related samples and non-Caucasian ethnicity was removed from all IBD cases and non-IBD control samples respectively. Six IBD cases were omitted from further analyses due to sufficient coverage (<50x). One sample was removed due to a inflated number of shared alleles (>7000) indicating a first degree relation to another case of the cohort. Twenty IBD samples were predicted as non-Caucasian. Sixteen control samples shown an insufficient coverage, four were related and thirteen were non-Caucasian. If not controlled, these three characteristics are known source of bias when testing for association, in particular, false positives due to ethnic differences are well known and described in the current literature [92, 111]. Figure 4.3 shows the PCA obtained when modelling the genomic variation of all 508 individuals using the known ethnic background from the 1000 Genomes Project.

There remained 282 IBD cases for analysis of which 172 were diagnosed with Crohn's disease (CD), 100 with ulcerative colitis (UC) and a further 10 patients had a diagnosis of IBD undetermined (IBDU). There was a corresponding number of 166 controls. The *NOD2* GenePy scores for the 282 IBD and 166 control individuals were calculated using all sixteen deleteriousness metrics. Given *NOD2* gene variant association is specific to the CD subtype of IBD, we calculated GenePy scores for both subtypes grouped separately. By observing the distribution of such scores between cases and controls, it is possible to observe that *NOD2* GenePy

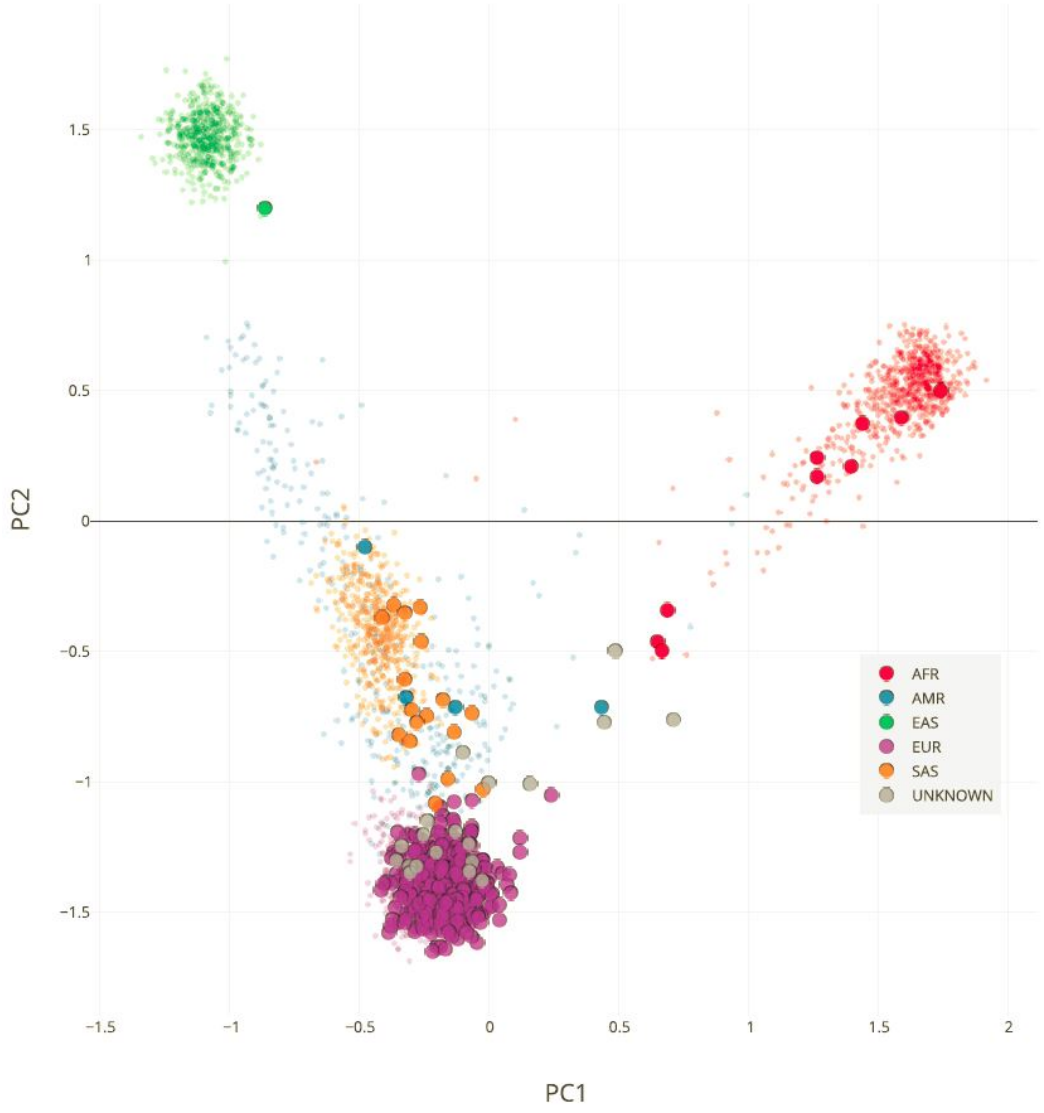


Figure 4.3: **IBD ethnicity imputation.** Principal component analysis for the imputation of sample ethnicity. Small dots represent individuals from the 1000 Genomes project used as background. Larger solid dots represent individuals in this study coloured according to the imputation result. In order to prevent selection bias, downstream analyses were restricted to Caucasian individuals only. The assessment of the ethnicity was performed modelling 2504 individuals from the 1000 Genomes Project alongside 508 individuals from this study. The multi-sample VCF for our cohort was lifted-over to hg19 reference genome build and then analysed using the Peddy software for ethnic imputation. Through Peddy, it was possible to calculate the identity-by-state of all possible sample pairs to impute their ethnicity.

scores do not follow a normal distribution (Figure 4.4). This made necessary the use of a non-parametric test for assessing statistical differences between subtypes.

The Mann-Whitney U test comparison of the distribution of *NOD2* GenePy scores between all IBD, CD and UC subtypes against controls identified statistically significant differences (Table 4.3). Modestly significant differences were observed for three of the implemented deleteriousness metrics (M-CAP, fathmm-mkl and MutTaster) were observed comparing all IBD against controls in this relatively

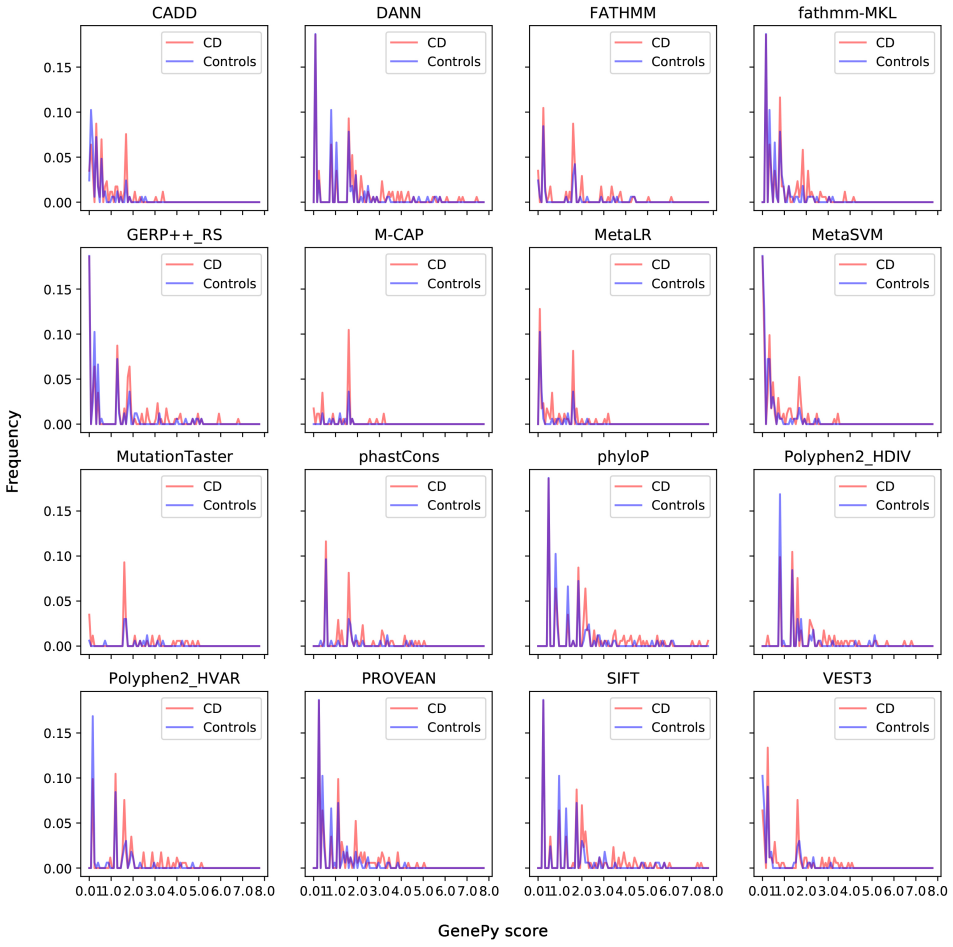


Figure 4.4: **GenePy scores profiles for the *NOD2* gene in the CD and control groups for each of the sixteen implemented deleteriousness metrics.** X-axis indicates GenePy scores grouped in bins of size 0.01 with the first bin shown 0.01-0.02. The y-axis shows the observed frequency of GenePy scores across the CD and control cohorts.

small sample. When the cases were stratified by disease subtype, UC samples had significantly lower GenePy scores compared to controls for two of the implemented deleteriousness metrics (MetaLR, phastCons). As expected, the most significant difference in *NOD2* score distribution was observed when comparing CD patients only against controls. Without exception, a highly significant difference was observed using every deleteriousness metric with M-CAP the most significant ($p = 1.37 \cdot 10^{-4}$) all of which would withstand correction for the three independent tests performed. Regardless of which deleteriousness metric is used, the mean GenePy score is consistently higher in CD patients when compared with

controls. Interestingly, similar results were observed for the SKAT-O gene test of association when using all variant frequency data but not when restricting to rare variation (MAF <0.05). Importantly, the magnitude of the SKAT-O difference between CD patients and control groups was statistically weaker ($p = 0.0346$) and less robust to correction for multiple testing. Although not the purpose of this comparison, we confirmed GenePy whole gene comparison provided statistical evidence two orders of magnitude greater than any single variant association result (Supplementary Table 7.2).

Table 4.3: ***NOD2* GenePy score statistics (maxima and means) and Mann-Whitney U tests across groups for all sixteen deleteriousness metrics.** p-values smaller than $1 \cdot 10^{-2}$ or smaller than $5 \cdot 10^{-2}$ are highlighted in red and green respectively. SKAT-O gene association results comparing patient groups against controls provided below thick line.

| **Metric** | **Controls (n = 166)** | | **All IBD (n = 282)** | | | **UC (n = 100)** | | | **CD (n = 172)** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *max* | *mean* | *max* | *mean* | *Mann-Whitney U comparison against controls* | *max* | *mean* | *Mann-Whitney U comparison against controls* | *max* | *mean* | *Mann-Whitney U comparison against controls* |
| CADD | 2.71 | 0.28 | 3.52 | 0.40 | $1.04 \cdot 10^{-1}$ | 2.66 | 0.20 | $1.38 \cdot 10^{-1}$ | 3.52 | 0.54 | $4.62 \cdot 10^{-4}$ |
| DANN | 5.92 | 0.84 | 7.62 | 1.06 | $1.36 \cdot 10^{-1}$ | 5.62 | 0.57 | $1.22 \cdot 10^{-1}$ | 7.62 | 1.38 | $8.16 \cdot 10^{-4}$ |
| FATHMM | 3.33 | 0.49 | 4.34 | 0.66 | $1.04 \cdot 10^{-1}$ | 3.14 | 0.38 | $1.47 \cdot 10^{-1}$ | 4.34 | 0.84 | $4.84 \cdot 10^{-4}$ |
| fathmm-MKL | 4.53 | 0.37 | 6.24 | 0.55 | $4.54 \cdot 10^{-2}$ | 3.78 | 0.25 | $3.15 \cdot 10^{-1}$ | 6.24 | 0.76 | $1.79 \cdot 10^{-4}$ |
| GERP++_RS | 5.30 | 0.64 | 7.00 | 0.87 | $1.26 \cdot 10^{-1}$ | 4.95 | 0.42 | $1.27 \cdot 10^{-1}$ | 7.00 | 1.17 | $6.95 \cdot 10^{-4}$ |
| M-CAP | 1.87 | 0.12 | 3.39 | 0.22 | $1.58 \cdot 10^{-2}$ | 1.73 | 0.08 | $4.62 \cdot 10^{-1}$ | 3.39 | 0.32 | $1.37 \cdot 10^{-4}$ |
| MetaLR | 2.42 | 0.16 | 3.39 | 0.29 | $2.71 \cdot 10^{-1}$ | 1.81 | 0.10 | $2.34 \cdot 10^{-2}$ | 3.39 | 0.42 | $1.63 \cdot 10^{-3}$ |
| MetaSVM | 2.67 | 0.30 | 3.61 | 0.43 | $9.88 \cdot 10^{-2}$ | 2.50 | 0.22 | $1.50 \cdot 10^{-1}$ | 3.61 | 0.57 | $4.39 \cdot 10^{-4}$ |
| MutationTaster | 4.38 | 0.26 | 5.10 | 0.39 | $4.48 \cdot 10^{-2}$ | 2.65 | 0.13 | $4.37 \cdot 10^{-1}$ | 5.10 | 0.57 | $7.47 \cdot 10^{-4}$ |
| phastCons | 4.66 | 0.35 | 5.24 | 0.56 | $2.86 \cdot 10^{-1}$ | 3.54 | 0.24 | $2.70 \cdot 10^{-2}$ | 5.24 | 0.77 | $2.16 \cdot 10^{-3}$ |
| phyloP | 6.32 | 1.02 | 7.93 | 1.27 | $1.23 \cdot 10^{-1}$ | 5.92 | 0.75 | $1.38 \cdot 10^{-1}$ | 7.93 | 1.62 | $7.09 \cdot 10^{-4}$ |
| Polyphen2_HDIV | 5.32 | 0.68 | 7.03 | 0.82 | $2.02 \cdot 10^{-1}$ | 2.30 | 0.33 | $6.22 \cdot 10^{-2}$ | 7.03 | 1.13 | $1.20 \cdot 10^{-3}$ |
| Polyphen2_HVAR | 4.86 | 0.46 | 5.31 | 0.64 | $1.65 \cdot 10^{-1}$ | 2.07 | 0.21 | $7.22 \cdot 10^{-2}$ | 5.31 | 0.92 | $7.90 \cdot 10^{-4}$ |
| PROVEAN | 4.33 | 0.66 | 5.23 | 0.86 | $1.04 \cdot 10^{-1}$ | 4.08 | 0.49 | $1.45 \cdot 10^{-1}$ | 5.23 | 1.10 | $4.84 \cdot 10^{-4}$ |
| SIFT | 5.91 | 0.95 | 7.61 | 1.14 | $1.47 \cdot 10^{-1}$ | 5.43 | 0.64 | $1.16 \cdot 10^{-1}$ | 7.61 | 1.47 | $9.64 \cdot 10^{-4}$ |
| VEST3 | 3.28 | 0.30 | 4.21 | 0.44 | $1.36 \cdot 10^{-1}$ | 2.24 | 0.17 | $1.13 \cdot 10^{-1}$ | 4.21 | 0.62 | $7.48 \cdot 10^{-4}$ |
| SKAT-O (all variants) | - | - | | | $5.41 \cdot 10^{-1}$ | | | $9.76 \cdot 10^{-2}$ | | | $3.46 \cdot 10^{-2}$ |
| SKAT-O (MAF<0.05) | - | - | | | $4.63 \cdot 10^{-1}$ | | | $1.37 \cdot 10^{-1}$ | | | $5.02 \cdot 10^{-2}$ |

## 4.5 Discussion

Multiple metrics have been recently developed aim to annotate individual mutations in order to sensitively discriminate causal versus non-causal variation. However, for common complex diseases where the action of multiple variants converge and combines to the one brought by environmental factors, the assessment of disease susceptibility through individual mutation profiles is necessarily. Furthermore, in order to interpret and translate genomic data into clinical management, it is important that novel methodologies provide metrics and evidence for individual patients and not just indications of modest genetic effects across large cohorts.

Herein, we describe the implementation of GenePy representing a novel model to establish the genetic burden as direct measure of the combined effect of mutations across each gene for each individual. The scoring system permits end-users freedom of choice of appropriate/preferred variant deleteriousness metric. GenePy should not be interpreted as one of the many new models that try to integrate multiple deleteriousness metrics that relate to variants only but rather as a novel method for scoring a whole gene in an individual. By summing across genes, GenePy further integrates biological information on frequency and zygosity and when being used to examine all genes or subsets thereof, can be corrected for gene length.

The analysis of GenePy profiles reveals the high variance necessary to distinguish mutational burden. As a consequence of the logarithmic implementation of allelic frequencies, GenePy up weights rare pathogenic variants making the additive score across a gene theoretically limited only by the number of variant sites within that gene. The majority of genes return a GenePy score of zero for any one individual but as most coding variation is rare, the likelihood of observing variation in any one gene is positively correlated with cohort size.

We provide proof of principle that the GenePy improves detectability of clinically meaningful gene perturbations. GenePy performance compares favourably against the most commonly applied gene based association test optimised for small data sets (SKAT-O). Such superiority to detect the subtle effects of genes in com-

plex disease is likely attributable to the additional modelling of innate biological features of mutations. Power to determine significant GenePy score differences between patient and control groups was consistent across sixteen different metrics of variant deleteriousness. Despite differing underlying principles, all metrics performed concordantly reporting a similar level of significance, with GenePy scores generated using M-CAP metric returning the most significant difference in CD patients compared to controls. Despite this, it is likely that no metric will prove optimal in all situations or conditions. The GenePy scoring system can simply implement new and improved variant deleteriousness metrics that are constantly evolving with improved interpretation of NGS data.

As with all large-scale data, GenePy scoring is dependent upon data integrity and elimination of systematic bias or technical artefacts. High quality individual DNA samples must be sequenced to sufficient depth of coverage to return confident variant calls. Particularly for larger scale analyses using multiple samples, it is essential to employ concordant capture kits, sequencing platforms and informatic pipelines or correct raw data accordingly. While these pre-processing quality control steps and generation of the multi-calling VCF file represent the highest computational burden, GenePy score calculation is amenable to batching and computationally trivial.

Many of the currently available deleteriousness scores implemented in GenePy fail to annotate synonymous, splicing or protein truncating variation. While we arbitrarily imposed maximum deleteriousness scores to protein truncating mutations, we standardised the set of variants examined across metrics by excluding synonymous and splicing variants from this analysis. Deleteriousness metrics based on solely on conservation could be calculated for all genomic locations and implemented for the assessment of non-coding regions derived from whole genome sequencing. Due to association testing in Caucasian samples only, we restricted allele frequency annotation to that ethnic group. Arguably, there is merit in implementation of global allele frequency estimates or those from more ancestrally diverse populations.

Further versions of the GenePy scoring system might see the integration of gene essentiality [145] (and conversely gene redundancy) or gene damage indices (GDI) [74] to improve the amount of biological information modelled. Similarly, the frequency of synonymous variants, so far ignored my our model, or the presence of GpC rich regions can be exploited to calculate the mutability rate on a per-gene basis and therefore correct GenePy scores accordingly. Since IBD is a condition restricted to the gastrointestinal system, gene expression rates (obtained from the GTEx database) could be integrated into GenePy to provide IBD-tailored scores that would take into consideration tissue-specific effects. On a more challenging level, long read NGS data enabling the discrimination of gametic phase would substantially advantage integration of inheritance models and haploinsufficiency. However, the technology to produce such data is still in development or not as cost effective as short read WES.

The key advantage of GenePy is its provision of a continuous quantitative measure of biological integrity of a gene within individuals, resulting in a score that is easily integrated into downstream analyses. GenePy scores are not dependent on cohort size and can be calculated and assessed on per-patient patient basis. GenePy scores are suited to pathway analyses where scores can be summed across defined molecular cascades. For the particular assessment of complex disease, machine learning tools that integrate multi-omic and extensive 'Big Data' to determine ambiguous patterns are increasingly applied. The ability to input biologically rich information at the gene and individual level represents an important step change from the more traditional methods of assessing genetic data at the variant and cohort level.

# Chapter 5

# Stratification of paediatric patients using immunogenomic data

## 5.1 Summary

In this chapter we investigate novel approaches to stratify paediatric IBD patients based on the integrity of their innate immune response. Through the application of unsupervised approaches cytokine response levels were modelled to identify eight novel immuno-phenotypes. Based on this information, we applied GenePy (Chapter 4) in order to collect gene scores from individuals for whom immunology data was obtained. GenePy scores from specific cytokine-related genes provided evidence of the exact molecular levels involved in characterising one of the novel immuno-phenotypes observed in our cohort. Moreover, this work represent a further proof of concept for modelling of WES data performed by GenePy.

Whilst the recruitment, immunoassay design and data collection was performed by Dr. Tracy Coelho, I was responsible for normalization and modelling of cytokine response data. I was also responsible for the collection, quality control, processing, structured analysis, modelling and all the analyses involving WES data ang

GenePy scoring.

## 5.2 Introduction

The principal symptom of inflammatory bowel disease is recurrent inflammation of the gastrointestinal system, that alternates between acute phase to remission of disease. According to the current literature, the chronic activation of the innate immune response through various signalling pathways represent one of the main mechanisms of inflammation involved in the pathogenesis of IBD [133]. Cytokines have been identified as key players in regulating such activity through anti- and pro- inflammatory effects. Subsequent to a plethora of upstream signals, cytokines are produced by a wide range of immune cells. Cytokines represent a broad class of small proteins divided in chemokines, interferons and tumour necrosis factors (TNFs). From a molecular perspective, pathways are usually modulated by cell-surface receptors that are triggered by specific cytokines. Figure 5.1 shows the mechanism of cytokine production from immune cells upon bacterial invasion. In addition to the crucial role of maintaining a correct mucosal homoeostasis, cytokines are important proliferation regulators, therefore a disregulated expression of such proteins can lead to IBD-associated intestinal cancer.

The production level of cytokines is therefore strongly regulated by multiple pathways all interacting through the *NF-kB* pathway, which represents the main hub were stimuli are integrated and modulated. Together with mitogen-activated protein kinases (*MAPK*) pathway, the *NF-kB* pathway controls the inflammatory response in IBD by promoting the expression of pro-inflammatory genes [10]. Due to the hub role of *NF-kB*, a large variety of genes are differentially regulated between different cell types. Despite this complexity, an increased expression of *NF-kB* in macrophages was observed IBD patients with the subsequent augmented secretion of TNF-$\alpha$, IL-1 and IL-6 cytokines. The increased production of these cytokines directly reflects the mucosal tissue damage observed in IBD [133].

Due to the powerful role of cytokine as molecular modulators, many drugs has

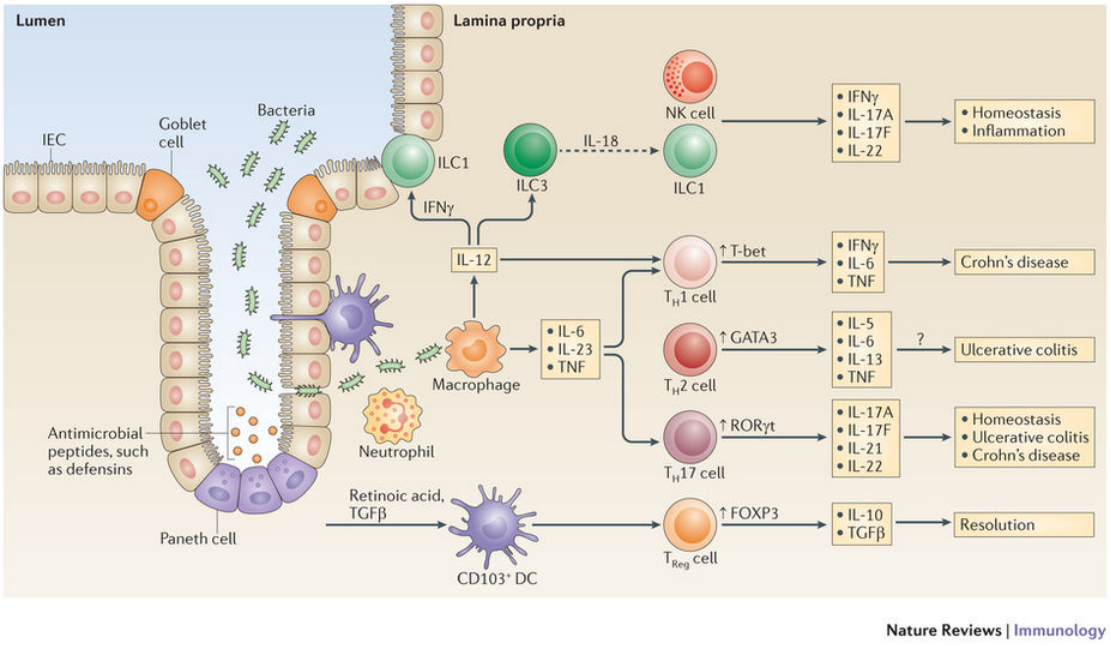


Figure 5.1: **Cytokine production in IBD.** Cytokines and transcription factors are produced by immune cells to control homoeostasis and inflammation. An unbalanced production of such proteins can lead to both ulcerative colitis and Crohn's disease. Image from [133].

been developed either mimicking or inhibiting their activity. Drugs that interfere with cytokine activity have become of common use in the treatment of paediatric and adult IBD cases with the TNF-$\alpha$ blocker one of the most employed [86].

The aim of this study was to identify novel IBD patient strata through the analysis and modelling of cytokines levels. Due to the high impact of canonical treatments on cytokine levels, data must be collected uniquely from treatment naïve patients which would have an unbiased expression level [48, 201].

Of the many pathways involved in IBD pathogenesis, the *NOD2*, *TLR1-2* and *TLR4* pathways are the most closely related to the cytokine regulations [133]. In order to trigger these pathways and therefore evaluate the integrity of the *NF-kB* hub in terms of cytokine production, peripheral blood mononuclear cells isolated from paediatric IBD patients were induced using three specific ligands.

Whilst simple statistical approaches are the gold standard for evaluating difference between samples, unsupervised machine learning algorithms (Section 6) can provide accurate patient stratification. Methods such as principal component analysis, t-SNE or more standard clustering algorithms can be used to identify patterns

in complex datasets. The advantage of unsupervised approaches is their complete blindness towards any clinical label that might bias stratification. Conversely, this lack of *a priori* knowledge ensures that obtained clusters, if any, are uniquely data-driven.

Due to the tight relationship between cytokine/regulatory pathways and genetics, next generation sequencing data might be coupled to cytokine expression levels leading to a better understanding of the molecular strata in paediatric IBD patients.

To summarise, this chapter focus on the identification of induced immune responses patterns in treatment naïve paediatric patients with IBD and the coupling of immunoassay data together with whole exome sequencing data. By merging these two data types we expect to observe novel IBD immuno-phenotype supported by both immune and genetic components.

## 5.3 Methods

### 5.3.1 Sample Data

Patients were recruited through the genetics of paediatric inflammatory bowel disease study at Southampton Children's Hospital. Patients were diagnosed under the age of 18 years according to the modified Porto criteria [97]. The cohort used in these analysis comprised 22 treatment naïve patients suspected to have inflammatory bowel disease, of which 14 were subsequently diagnosed with Crohn's disease (CD) and 8 with ulcerative colitis (UC). Ten additional patients for which a suspected IBD diagnosis was formulated and then not confirmed were included in the analyses as controls.

Genomic DNA was extracted from peripheral venous blood and fragmented DNA subjected to adaptor ligation and exome library enrichment using the Agilent SureSelect All Exon capture kit version 6. Enriched libraries were sequenced on a

Illumina HiSeq 2500 machine.

### 5.3.2 Immunological assay

Peripheral blood mononuclear cells (PBMCs) were extracted from blood samples obtained from both IBD and control individuals. In order to trigger specific immune response pathway in PMBCs, cells were activated using muramyldipeptide (MDP, *NOD2* agonist); Pam3CysSerLys4, a synthetic tri-palmitoylated lipopeptide (Pam3CSK4, *TLR1-2* agonist); and lipopolysaccharide (LPS, *TLR4* agonist). These three stimulating ligands were specifically selected to activate innate signalling pathways known to be involved in the pathogenesis of IBD.

Immunological assay was used to simultaneously measure the concentrations of 4 cytokines including *IL-10*, *IL-6*, *IL-1*$\beta$ and *TNF-*$\alpha$ with two technical replicates. Due to the high variability in monocyte counts in a healthy population, cytokines concentrations required normalisation according to each patients PBMC readouts. Monocyte counts were obtained through flow-cytomoetry analysis.

### 5.3.3 WES data processing

Whole exome sequencing data was obtained for 21 out of 22 IBD patients with raw sequencing data processed as previously described in Section 4.3.2. By applying GenePy (Chapter 4) gene scores were calculated for all available genes for each of the 21 IBD patients. GenePy is a per-patient gene scoring algorithm for integrating next generation sequencing data preserving important biological information such as variant deleteriousness, zygoisity and rarity. GenePy provides continuous scores that can be compared across individuals or genes. Since genes have different length, GenePy scores were normalised by dividing each gene for the size (expressed in bases) of the region captured by the exon enrichment kit. GenePy can produce gene scores implementing a wide range of deleteriousness metrics ranging from conservation to composite scores. In the following analyses, GenePy

scores were calculated implementing CADD [88], the best established composite score currently available which still detain the highest AUC in discriminating pathogenic vs. benign variants in the ClinVar dataset.

### 5.3.4 Unsupervised stratification

Raw immunoassay data obtained from patients and controls were obtained in five distinct batches including both cases and controls.

**Immunoassay data normalisation** In order to compare IBD patients cytokine responses with those observed in control samples and to detect specific signatures of individual IBD patients, cytokine levels required a normalisation within a reference range. Mean and standard deviation were calculated for the paediatric controls (n = 10) and used to scale IBD sample readouts. This transformation ensures that each ligand-cytokine feature is centred on a mean of zero and scaled to unit variance. The scaling, also known as standard scaling, was applied to each value $X = \{x_{1c}, x_{2c}, ..., x_{ic}\}$ for $i$ representing each individual in the IBD cohort and $c$ a specific ligand-cytokine combination and is defined as follows:

$$X'_c = \frac{X_c - \mu_{controls,c}}{\sigma_{controls,c}} \tag{5.1}$$

We herein define hypo-inflammatory or hyper-inflammatory states where IBD patient levels deviate respectively more than twice the standard deviation ($2\sigma$) below or above that observed in controls. Cytokine responses within the $\pm 2\sigma$ range were considered normal. The divergence of more than two standard deviations from normality is a well established and accepted threshold in the analysis of immunologic data [142]. Standard scaling was performed using the Python v2.7 Scikit-learn v0.19 package.

For each patient, standardised response values were represented through a 12 spokes radial plot. Each of the spokes represents one of the 12 ligand-cytokine combination. Individual cytokine response profiles were represented by joining

the data points on each spoke.

**Principal component analysis** Principal component analysis (PCA) was used to cluster samples using immunoassay data and to test for any batch effect. PCA, is an unsupervised machine learning algorithm capable of deconvoluting a multi-dimensional dataset into a selected lower number of dimensions (components) by linearly combining the original features (Section 2.3.2). The PCA algorithm transforms the data according to the variance observed in the original dataset with first component expressing the largest explained variance. As requirement of the PCA algorithm, data have to be scaled so that each feature has mean zero and each data point scaled to unit variance. Since PCA is based on the analysis of variance it is extremely sensitive to unit measure differences (Section 2.3.2). This transformation was performed using the equation 5.1.

**Hierarchical clustering** Hierarchical clustering (HC) is an unsupervised clustering algorithm that agglomerates samples according to their similarity (Section 2.3.2). In order to decide whether two samples should be linked, the algorithm uses a distance metric and a linkage method. Whilst the distance metric is a measure of similarity of samples, the linkage method instructs the clustering algorithm on grouping criteria. In this analysis, euclidean distances and average linkage were used to compute clusters. Prior to modelling data through HC, raw cytokine values were normalised using control sample readout by subtracting the median and then scaling according to the Inter Quartile Range (IQR, range between the 1st and the 3rd quartile). Despite not being compatible with PCA modelling, this latter scaling is more robust against extreme outliers. Based on the clustering provided by HC, samples where grouped into functional clusters.

## 5.4 Results

The mean age at diagnosis observed in our cohort of 22 IBD patients was 11.9 years and the male to female ratio is 1.20. In order to exclude batch effect, a PCA was performed on IBD cases and controls (Figure 5.2). The two first components explain 48.5% and 16.7% of the original variance respectively. We do not observe clustering of samples by batch. Moreover, there is a visibly larger spreading of control samples compared to IBD cases indicating greater variability in the non-affected individuals. Despite the lack of clear clustering between cases and controls, by observing the vector loadings of the PCA it is possible to observe an opposite directionality of IL-1$\beta$ and IL-6 regardless of the used stimulating ligand. This implies that, according to the second principal component, a sample with high IL1$\beta$ usually present low IL-6 response and *vice versa*.

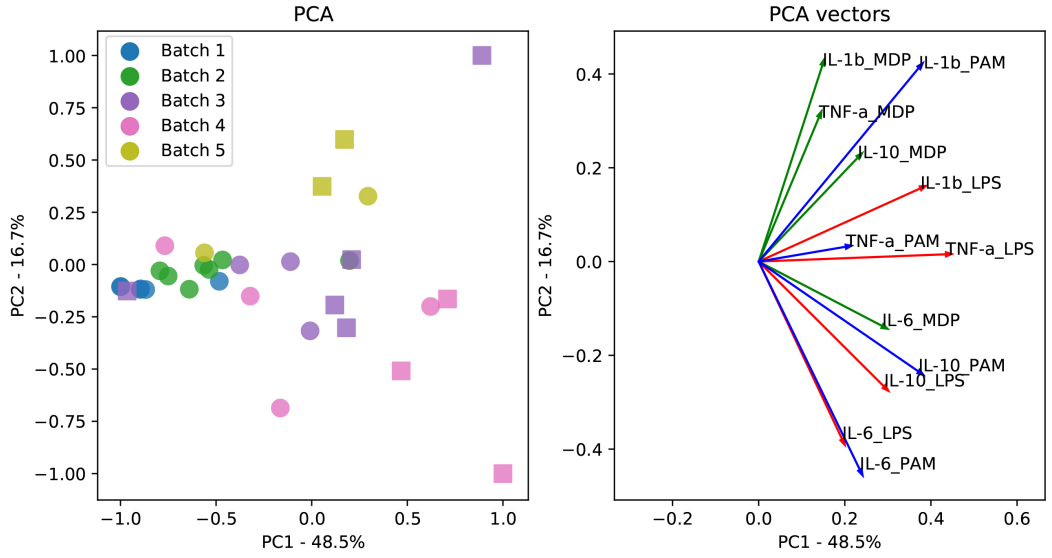


Figure 5.2: **Principal component analysis of immunoassay data.** Left panel shows the PCA of cases and controls represented as circles and squares respectively. Each batch (1-5) is represented by a unique colour. Right panel shows the vector loadings for each of the 12 possible ligand-cytokine combination of the same PCA. Colours match the stimulating ligand respectively: MDP in green, PAM in blue and LPS in red.

### 5.4.1 Cytokine responses of IBD patients

In order to detect samples with abnormal cytokine responses, IBD sample data was normalised according to control readouts and plotted on a radar plot (Figure

5.3). Following normalization through standard scaling, none of the IBD affected individuals show a hyper-inflammatory response, deviating more than $+2\sigma$ from the mean. However, it is possible to observe that, regardless of the ligand-cytokine combination, samples are predominantly distributed in between the mean of normality and $-2\sigma$ suggesting a tendency to cluster towards the hypo-inflammatory direction. Only one patient shown a hypo-inflammatory ($< -2\sigma$) IL-1$\beta$ response when stimulated with LPS. Individual radar plots from each IBD patient (Supplementary Figure 7.4 and 7.5) identified recurrent patterns.

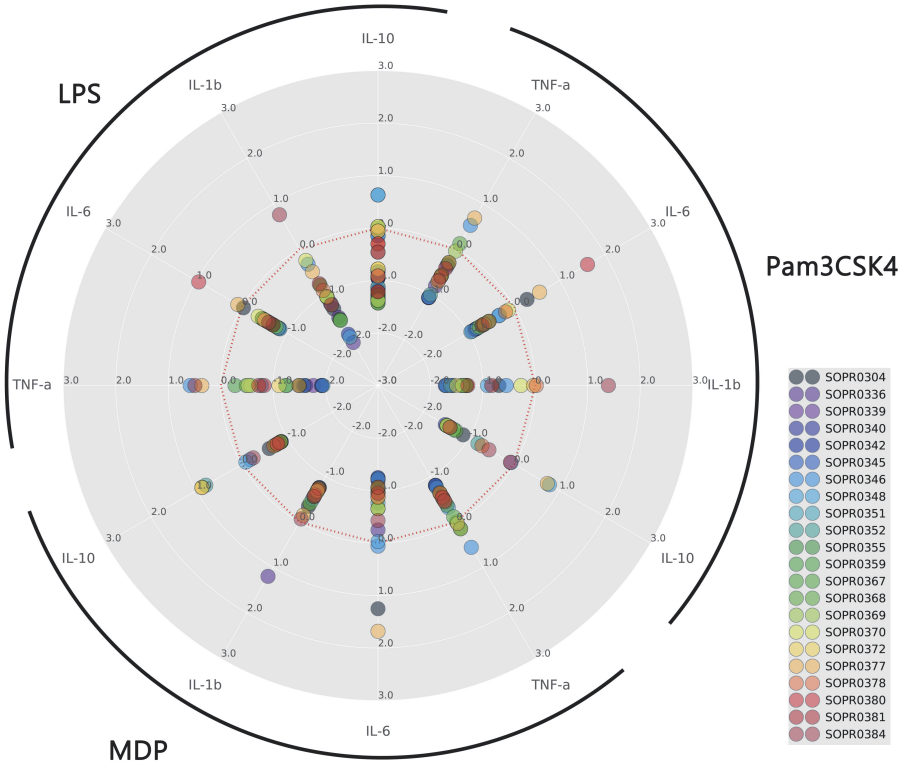


Figure 5.3: **Radar plot of immunoassay data.** Each spoke of the radar represents a ligand-cytokine combination with the red dashed line indicating the mean response of the control cohort. Each individual is is coloured differently.

### 5.4.2 Hierarchical clustering of immune-phenotypes

Immunoarray data normalisation, euclidean distances were calculated for each IBD case pair and used to perform hierarchical clustering. This unsupervised approach identified eight immuno-subtypes that might represent novel strata for IBD immunology (Figure 5.4). Of the eight patterns identified, three are represented by

single individuals (clusters 3, 5 and 7). The highest similarity is observed within individual belonging to clusters 1, 2 and 4. Clusters one and six are the most represented with five individuals each.

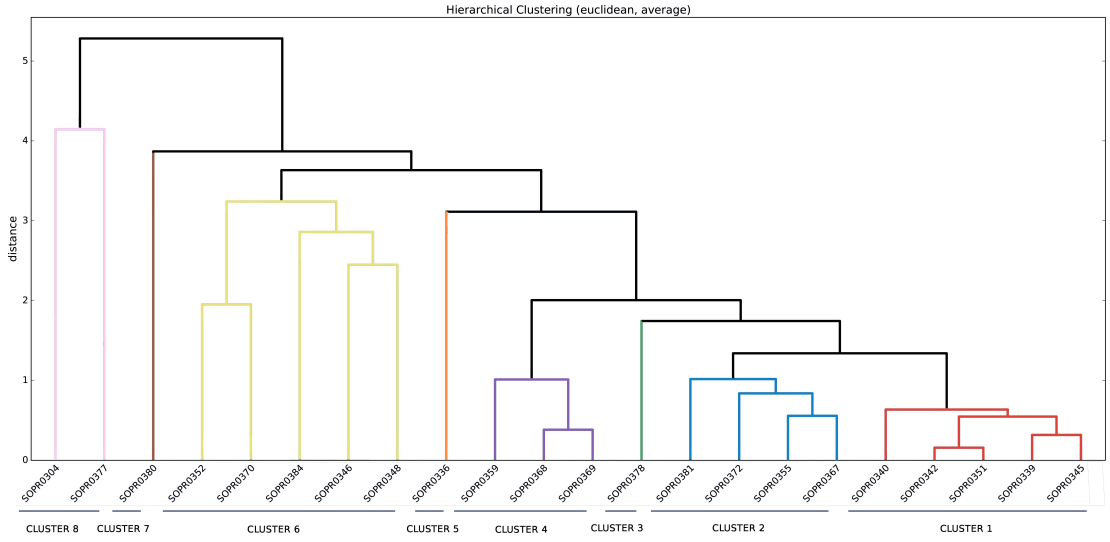


Figure 5.4: **Hierarchical clustering of immunoassay data from IBD patients.** The dendrogram was obtained performing hierarchical clustering with euclidean distances and average linkage strategy. Each leaf represent one of the 22 IBD patients and colours indicate the selected clusters. The vertical axis represent the distance between each node of the tree.

In order to test for significant differences between cases and controls immune responses, a two-tailed t-test was performed for each ligand-cytokine combination (Table 5.1). Statistical significance was observed in the levels of *IL10* when stimulated with LPS ($p = 0.045$) or Pam3CSK4 ($p = 0.018$); IL-1$\beta$ either stimulated with with LPS ($p = 0.010$) or Pam3CSK4 ($p = 0.015$) and TNF-$\alpha$ stimulated with LPS ($p = 0.018$). The stimulation of the *NOD2* pathway with MDP did not induced any statistically significant differential response between IBD patients and controls. Similarly, IL-6 cytokine levels were always comparable between cases and controls regardless of the stimulus used.

The cumulative effect of all the 12 conditions, obtained by summing the cytokine responses across stimuli, identified a trend of hypo-inflammatory response with individuals belonging to cluster one with the most negative values (Table 5.1). Conversely individuals from cluster six to eight exhibit the least negative, therefore close to normality, immune response. Despite this, the overall summative effect of

cytokine responses were significantly higher in IBD patients compared to controls ($p = 0.0025$).

Table 5.1: **Normalised cytokine response level for IBD cases and controls.** Samples are ordered according to clusters obtained through HC. The summative effect consits in the sum of all 12 conditions in each patient. Statistically significant differences between cases and controls were assessed performing a Student t-test per each cytokine-ligand combination (columns). p-values smaller than 0.05 are highlighted in red.

| **Sample** | **Cluster** | **TLR4 inuction (with LPS)** | | | | **NOD2 induction (with MDP)** | | | | **TLR1-2 induction (with PAM3CSK4)** | | | | **Summative effect** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *IL-10* | *IL-1β* | *IL6* | *TNF-α* | *IL10* | *IL-1β* | *IL-6* | *TNF-α* | *IL-10* | *IL-1β* | *IL-6* | *TNF-α* | |
| SOPR0339 | 1 | -0.95 | -1.50 | -0.69 | -1.77 | -0.86 | -0.75 | -1.24 | -0.81 | -1.46 | -1.60 | -0.85 | -1.06 | -13.55 |
| SOPR0340 | 1 | -0.94 | -1.18 | -0.55 | -1.92 | -0.84 | -0.73 | -1.22 | -0.81 | -1.52 | -1.72 | -0.93 | -1.07 | -13.44 |
| SOPR0342 | 1 | -1.26 | -1.87 | -0.83 | -1.93 | -0.86 | -0.75 | -1.20 | -0.80 | -1.48 | -1.69 | -0.86 | -1.06 | -14.59 |
| SOPR0345 | 1 | -0.83 | -1.56 | -0.80 | -1.58 | -0.87 | -0.76 | -1.25 | -0.81 | -1.41 | -1.57 | -0.84 | -1.07 | -13.35 |
| SOPR0351 | 1 | -1.12 | -1.93 | -0.84 | -1.95 | -0.83 | -0.76 | -1.22 | -0.76 | -1.43 | -1.70 | -0.95 | -1.05 | -14.55 |
| SOPR0355 | 2 | -1.43 | -1.54 | -0.68 | -1.50 | -0.84 | -0.70 | -0.96 | -0.37 | -1.46 | -1.31 | -0.69 | -0.61 | -12.09 |
| SOPR0367 | 2 | -1.29 | -1.57 | -0.75 | -1.24 | -0.86 | -0.74 | -1.04 | -0.63 | -1.34 | -1.32 | -0.77 | -0.30 | -11.84 |
| SOPR0372 | 2 | -0.78 | -1.04 | -0.49 | -1.11 | -0.78 | -0.68 | -0.88 | -0.53 | -1.41 | -1.50 | -0.66 | -0.72 | -10.59 |
| SOPR0381 | 2 | -1.21 | -0.76 | -0.69 | -0.83 | -0.83 | -0.56 | -1.07 | -0.55 | -1.42 | -0.83 | -0.71 | -0.61 | -10.06 |
| SOPR0378 | 3 | -0.91 | -0.91 | -0.51 | -1.49 | -0.71 | -0.59 | -0.87 | -0.45 | -0.71 | 0.02 | -0.19 | -0.29 | -7.62 |
| SOPR0359 | 4 | -1.29 | -0.79 | -0.56 | -0.27 | -0.74 | -0.42 | -0.66 | 0.00 | -1.29 | -1.39 | -0.66 | -0.57 | -8.65 |
| SOPR0368 | 4 | -1.35 | -1.06 | -0.46 | -0.49 | -0.86 | -0.71 | -0.98 | -0.12 | -1.46 | -1.32 | -0.57 | 0.12 | -9.24 |
| SOPR0369 | 4 | -1.38 | -1.06 | -0.46 | -0.58 | -0.85 | -0.72 | -1.06 | 0.15 | -1.48 | -1.39 | -0.55 | -0.04 | -9.41 |
| SOPR0336 | 5 | -1.17 | -2.05 | -0.63 | -1.93 | -0.87 | 1.20 | -0.25 | -0.53 | -1.48 | -1.40 | -0.66 | -0.81 | -10.59 |
| SOPR0348 | 6 | 0.63 | -1.04 | -0.45 | -0.78 | -0.09 | -0.67 | -0.03 | -0.32 | 0.77 | -0.91 | -0.32 | -0.38 | -3.58 |
| SOPR0352 | 6 | 0.00 | -1.23 | -0.80 | -1.26 | 0.79 | -0.56 | -0.76 | -0.52 | -0.80 | -1.62 | -0.81 | -1.00 | -8.57 |
| SOPR0370 | 6 | 0.03 | -0.25 | -0.36 | -0.52 | 0.87 | -0.69 | -0.66 | -0.64 | -1.48 | -0.27 | -0.13 | -0.58 | -4.70 |
| SOPR0346 | 6 | -0.15 | -0.33 | -0.80 | 0.58 | -0.18 | -0.50 | 0.06 | 0.56 | -0.07 | -0.55 | -0.34 | 0.52 | -1.22 |
| SOPR0384 | 6 | -0.46 | 0.76 | -0.60 | 0.49 | -0.25 | -0.07 | -0.43 | -0.47 | -0.56 | 1.39 | -0.55 | -0.69 | -1.43 |
| SOPR0380 | 7 | -0.30 | -1.22 | 0.95 | -0.73 | -0.78 | -0.74 | -0.95 | -0.63 | -0.08 | -1.29 | 1.61 | -0.43 | -4.59 |
| SOPR0304 | 8 | -1.35 | -1.31 | -0.04 | -1.62 | -0.61 | -0.35 | 1.25 | 0.14 | -1.13 | -0.69 | 0.28 | -0.34 | -5.76 |
| SOPR0377 | 8 | -0.06 | -0.50 | 0.09 | 0.36 | 0.88 | -0.15 | 1.68 | 0.03 | 0.73 | -0.04 | 0.56 | 0.69 | 4.27 |
| SOPR0330 | Control | -0.69 | -0.03 | 0.43 | -0.18 | -0.12 | 0.03 | 2.46 | 0.30 | 0.96 | 0.87 | 0.82 | 0.42 | 5.26 |
| SOPR0364 | Control | 0.26 | 0.95 | -0.53 | 0.87 | 2.58 | 2.68 | -0.94 | 0.92 | 0.34 | 1.96 | -1.02 | -0.42 | 7.66 |
| SOPR0365 | Control | 1.20 | -0.41 | -0.35 | -0.80 | 1.05 | -0.53 | -0.07 | -0.64 | 0.89 | -0.83 | -0.16 | -0.86 | -1.50 |
| SOPR0371 | Control | 0.79 | -0.64 | -0.59 | -0.37 | -0.33 | -0.61 | -0.52 | -0.76 | 1.05 | 0.14 | -0.53 | -0.86 | -3.23 |
| SOPR0374 | Control | -1.40 | -1.75 | -0.46 | -1.86 | -0.83 | -0.75 | -1.15 | -0.76 | -1.50 | -1.74 | -0.88 | -0.95 | -14.04 |
| SOPR0375 | Control | -0.47 | -0.82 | -0.34 | -0.08 | -0.39 | -0.52 | 0.14 | 0.03 | -0.34 | 0.40 | -0.15 | 2.06 | -0.46 |
| SOPR0376 | Control | 0.93 | -0.40 | 0.24 | 0.20 | -0.01 | -0.49 | 1.06 | -0.55 | 1.00 | -0.50 | 0.63 | -0.51 | 1.59 |
| SOPR0379 | Control | 1.45 | 1.50 | 2.82 | 2.15 | -0.75 | -0.75 | -0.24 | -0.62 | 0.28 | -0.94 | 2.47 | 0.56 | 7.91 |
| SOPR0382 | Control | -1.20 | 0.07 | -0.63 | 0.43 | -0.65 | 0.66 | -0.37 | 2.57 | -1.28 | 0.04 | -0.60 | 1.35 | 0.38 |
| SOPR0383 | Control | -0.87 | 1.54 | -0.58 | -0.35 | -0.54 | 0.27 | -0.38 | -0.49 | -1.41 | 0.61 | -0.58 | -0.78 | -3.57 |
| Cases vs Controls P | | 0.045 | 0.010 | 0.178 | 0.018 | 0.188 | 0.164 | 0.116 | 0.267 | 0.018 | 0.015 | 0.245 | 0.169 | 0.0025 |

In order to assess the contribution of each cluster to the hypo-immune response of IBD patients, normalised cytokine levels from individual cluster were tested against controls. Statistically significant differences were observed in cluster 1 and cluster 2 compared to controls. Cluster 1 shown a significant hypo-immune response in IL-1$\beta$ (stimulated with LPS (p = 0.006) or Pam3CSK4 (p = 0.004)), TNF-$\alpha$ (stimulated by LPS (p = 0.002)) and IL-10 (stimulated by Pam3CSK4 (p = 0.009)). Cluster 2 shown the same pattern of significance except for TNF-$\alpha$. All the remaining clusters were not showing significant different levels of cytokines compared to the control group. Statistical tests were not performed on clusters populated by a single individual.

### 5.4.3 Genomic interpretation of immuno-phenotypes

Through hierarchical clustering of immunological assay data, eight novel immuno-phenotypes were identified with a specific involvement of the *TLR4* and the *TLR1/TLR2* signalling pathways. In order to investigate a potential link between individuals genetics and immune responses, whole exome sequencing data was transformed with GenePy and then compared against immuno-phenotypes.

Although GenePy scores were available for more than 14,000 genes, the analysis was restricted to genes belonging only to the *TLR4* and *TLR1/TLR2* signalling cascades. Three separate gene panels were therefore defined in order to represent three molecular strata: i) the receptor level; ii) the downstream signal modulation and; iii) the cytokine level. The receptor level gene panel was obtained by interrogating the PathCards database for *TLR4* and *TLR1/TLR2* pathways (n = 127 genes), the signal modulation panel consisted of genes involved in the *MAPK*, *NF-$\kappa$B* and inflammasome pathways (n = 173 genes), and the lower cytokine level panel consisted of the *IL-10* and *TNF-$\alpha$* pathways (n = 77 genes).

Cumulative GenePy scores for each individual were obtained by summing all the gene scores belonging to each gene panel. Subsequently, statistical significance was assessed for differences between cluster 1 or cluster 2 against clusters 3 to

8. Table 5.2 reports the p-values obtained following this approach. A significant difference (withstanding Bonferroni correction) between cluster 1 and clusters 3 to 8 was observed at the cytokine level underpinning a strong association between the immune response of those individuals and their genetic asset.

Table 5.2: **GenePy scores regression against immuno-phenotypes.** Cumulative GenePy scores from three selected gene panels were used to test significant differences between clusters. Significant p-values are highlighted in red.

| Gene panel | Genes | GenePy scores available | t-test | p-value | Bonferroni corrected p-value |
|---|---|---|---|---|---|
| Receptor level (TLR1-2, TLR4) | 127 | 74 | Cluster 1 vs 3-8 | 0.938 | 5.630 |
| | | | Cluster 2 vs 3-8 | 0.482 | 2.891 |
| Signal modulation level (MAPK, NF-kB, Inflam.) | 173 | 104 | Cluster 1 vs 3-8 | 0.283 | 1.699 |
| | | | Cluster 2 vs 3-8 | 0.629 | 3.773 |
| Cytokine level (Il-10, TNF-a) | 77 | 52 | Cluster 1 vs 3-8 | 0.002 | 0.012 |
| | | | Cluster 2 vs 3-8 | 0.163 | 0.977 |

## 5.5 Discussion

Inflammatory bowel disease is common autoimmune condition driven by a chronic activation of innate immune response leading to inflammation of the gastrointestinal system. Cytokine based inflammation has been directly implicated as main mechanism of bowel mucosa alteration, causing ulceration and manifestations of the disease. Current literature well describes the increased production of pro-inflammatory cytokines such as TNF-$\alpha$ and IL-6, by the immune cells in the bowel mucosa of IBD patients [82, 181]. In order to reduce the inflammation level, anti-cytokine drugs have been developed and are currently used in routine treatment of IBD [11, 128]. This chapter was therefore focused on modelling and stratifying paediatric IBD patients according to their cytokine response levels [137].

The design of immunoassay used in this study was aimed to test the functional integrity of the *NF-kB* pathway, a critical hub for other inflammatory pathways implicated in IBD such as the *NOD2* signaling, TLR signaling and inflammasome activation. The activation of these signalling pathways induce the production of various proteins, in particular pro-inflammatory cytokines.

Through the comparison of cytokine production following stimulation between paediatric IBD cases and controls, we observed an unprecedented reduction in cytokine levels in cases. This unexpected hypo-inflammatory response of IBD patients might be explained by a strong genetic component affecting and disrupting the key signalling pathways controlling the innate immune response. It is therefore expected that, upon invasion from exogenous bacteria, the inability of a correct response and clearance might induce a persistent or chronic bowel inflammation in IBD patients.

In order to compare IBD responses to the expected normality, normal values were obtained from a modest cohort of paediatric controls. Since reference normal values are not unique to IBD, we believe such information will be useful and applicable in the investigation of other autoimmune conditions.

After determining the normal ranges of responses in the paediatric control cohort,

IBD cases readings were normalised and analysed through hierarchical clustering for stratification purposes. Through this approach we identified 8 novel functional clusters, based solely on the cytokine production levels in the 12 assay conditions. Hierarchical clustering also highlighted a consistent gradient from hypo- to normal functionality with the most divergent group presenting the better characterised pattern. Despite this, only one individual exhibited a significant divergence from normality ($>\pm 2\sigma$) with the majority of individuals presenting a statistically significant sub hypo-immune response ($p = 0.0025$).

Subsequently, we identified the TLR4 and TLR1-2 as the main pathways responsible for the altered immune response with clusters 1 and cluster 2 particularly affected. In order to establish a causal link between observed abnormal immune profiles and individuals genetic assets, we analysed exome data through GenePy, a per-patient gene score capable of integrating most of the biological information available through next generation sequencing data. GenePy can provide a direct measure of gene deleteriousness by weighting and summing gene-specific single nucleotide variants according to their zygosity, rarity and predicted deleteriousness. Thanks to the robust nature of GenePy scores, it was also possible to assess the deleteriousness of entire pathways or gene panels by simply summing the effect of involved genes.

We therefore assessed whether patients with an altered immune response present also a high degree of genetic disregulation. By testing gene score differences between cytokine-defined clusters and the three main gene panels dissecting the TLR cascade, we observed statistically significant results when comparing cluster 1 against clusters from 3 to 8 using genes involved in the cytokine level part of the cascade.

In conclusion, this study provides further insights on the role of cytokines in defining a wide range of immuno-phenotypes of IBD, which might be further investigated for association to clinical traits. Unsupervised machine learning approaches resulted fundamental for data interpretation and for patient stratification, identifying novel strata with a potential clinical relevance. Moreover, we also highlighted

the importance of collecting genomic data and their value in deciphering complex immune condition. Through this study we also demonstrated the potential of GenePy in modelling NGS data and its effortless integration with other data types, such as immunoassay results.

# Chapter 6

# Machine learning modelling of genomic data of IBD

## 6.1 Summary

In this chapter we investigate supervised and unsupervised approaches to classify and stratify paediatric IBD patients based on genomic data from selected IBD-related genes.

We report the current highest AUC for the classification of CD and UC patients (AUC = 91%) and the highest performance in discriminating IBD patients from controls (AUC = 85%). Following an enrichment analysis we observe that the *FGFR3* signalling pathway is significantly contributing to the CD/UC discrimination whilst cytokine signalling discriminates IBD patients from controls. GenePy scores proved valuable in correctly modelling complex NGS data. Through the application of unsupervised models, we identify five novel IBD strata uniquely driven by genomic data.

Concerning the results presented in this chapter, I was responsible for quality control, processing, modelling and all the analyses involving genomic data.

## 6.2 Introduction

Inflammatory bowel disease (IBD) is a complex condition where genetics and environmental factors contribute to the final phenotype. Depending on age of onset, the balance between these to components can be shifted from a predominantly genetic-driven (very early onset IBD) to a predominantly environmental-driven condition in late adulthood. In between these extremes sits a spectrum of IBD manifestations where genetics and environment variably contribute (Section 1.4.2) to determining the adverse phenotype.

Similar to other complex diseases such as asthma [23] or cancer [27], this multitude of IBD manifestations intuitively suggest the presence of multiple disease subtypes rather than one homogeneous phenotype. As already covered in Section 1.4.1, IBD is currently classified in two main subtypes characterised by different localisation and extent of inflammation, the main symptom of this autoimmune condition. Although Crohn's disease (CD) and ulcerative colitis (UC) forms of IBD were described and applied in the mid twentieth century, a formal worldwide classification was made only in 1991 [56]. Since then, this original classification underwent several revisions in order to match new clinical and molecular discoveries.

The diagnosis of CD or UC has a direct impact on the clinical approach in treatment. For example, proctocolectomy, which consist in the surgical resection of the rectum an all or a part of the colon completely removes the disease in UC patients where the inflammation is localised in those GI tracts [17]. However, the same approach is not curative in CD patients where the inflammation is instead scattered across the entire GI system. Specific surgical procedures and a wide range of drugs are currently available for treating IBD symptoms, however there is still no cure for such condition. As a consequence, a precise and prompt diagnosis is crucial to avoid delivering ineffective treatments to patients.

Despite the great advances in diagnostic tools and procedures, assigning IBD patients with a specific CD or UC subtype remains challenging and not always successful. This leads a substantial percentage of patients, especially in paediatric

age, being diagnosed with the IBD undetermined (IBDU) form. The complexity in assigning a specific diagnosis is mostly attributable to the high overlap of symptoms between CD and UC. This overlap of clinical traits has been mirrored by genetic studies where some genes were uniquely associated to either CD or UC and a similar percentage to both forms (Section 1.4.2).

The challenging diagnostic process and lines of evidence of biological overlap between CD and UC are increasingly suggesting the need of a new classification system that can better represent the wide range of IBD presentations. Further investigations of endoscopic data (the main evidence used to assign a diagnosis through the current classification criteria), reported a high discordance with histological and other clinical data [51, 8]. To date, several studies already investigated the application of machine learning algorithms (ML) in order to identify novel strata within IBD patients using clinical and immunological data ([198, 126], Chapter 5).

As a consequence of more accessible costs for routinely obtain NGS data, such valuable information can be utilised to improve IBD patient classification and stratification. Based on this hypothesis, several models discriminating CD from UC and IBD from controls have been developed in recent years. Whilst first approaches to classify IBD subtypes focused on immunochip arrays and reached a maximal area under the ROC curve (AUC) of 85% [197], more recent supervised machine learning models based on WES data reached marginally higher AUC (87%) in classifying CD patients [28]. Conversely, there is a dearth of unsupervised approaches for stratification of IBD patient by modelling NGS data.

The marginal improvement in discriminating UC from CD using high-throughput NGS data suggests a technical limitation concerning the integration of NGS data into machine learning models. In Chapter 4 we presented GenePy, our scoring system for improving modelling of NGS data in machine learning frameworks. By integrating more biological information on a gene basis compared to conventional approaches, we demonstrated GenePy superior ability in detecting subtle biological differences.

For these reasons, in this chapter supervised and unsupervised machine learning approaches are applied to WES data from our cohort of paediatric IBD patients. Following the transformation of sequencing data into per-gene per-patient GenePy scores, we investigate two classification scenarios. First, we assess the performance of our model in discriminating CD from UC individuals. Secondly, we construct a model for discriminating IBD patients from healthy controls. Genes selected by both models are then investigated for enrichment. In the final section, we apply unsupervised modelling through both PCA and hierarchical clustering on paediatric IBD patients.

## 6.3 Methods

### 6.3.1 Sample data

Whole exome sequencing (WES) data were derived from two sources. This first group comprised 285 patients diagnosed in childhood with IBD. This cohort (described in Chapter 3) includes unrelated, Caucasian patients ascertained and recruited through Southampton Children's Hospital who were diagnosed under the age of 18 years according to the modified Porto criteria [97]. Additional WES data from a cohort of 180 anonymised individuals diagnosed with an infectious disease but unselected for any form of autoimmune disease were also used to give a total cohort size of 465 individuals with WES data.

Genomic DNA was extracted from peripheral venous blood using the salting out method [124] and fragmented DNA subjected to adaptor ligation and exome library enrichment using the Agilent SureSelect All Exon capture kit versions 4, 5 and 6. Enriched libraries were sequenced on Illumina HiSeq systems. DNA concentration was estimated using the Qubit 2.0 Fluorometer and the 260:280 ratio calculated using a nanodrop spectrophometer.

WES data was processed with our custom pipeline described in Section 2.2 and then processed through GenePy (Chapter 4) in order to obtain per-patient gene

scores. Sixteen deleteriousness metrics were individually implemented in GenePy resulting in the same number of output matrices each containing approximately 14,000 gene scores per individual. Since GenePy scores are given by the sum of weighted variants within a given gene, large genes have higher chance of presenting mutations. Therefore, GenePy gene scores were normalised according to the size of targeted gene region and then multiplied by the median gene size (1461 base pairs). Due to the adoption of Caucasian allelic frequency in the calculation of GenePy scores, non-Caucasian samples were excluded from downstream analyses.

### 6.3.2 Gene selection

Inflammatory bowel disease, as described in Section 1.4.2, has been associated with a large number of genes involved in maintaining immune system homoeostasis and gut barrier integrity. These genes act through specific signalling pathways. According to the current literature [86, 192, 10] it is possible to identify fifteen main pathways (Table 6.1) involved to some extent in the IBD pathogenesis. Lists of all genes within these pathways were obtained either from the KEGG pathway [84] repository or through PathCards [19] as reported in table 6.1.

In order to include in our analyses most of the current biological knowledge about IBD and remove unnecessary background noise, we restricted both supervised and unsupervised approaches to the combined non-redundant list of 989 genes derived from the fifteen IBD associated pathways.

### 6.3.3 Supervised classification

Two supervised frameworks were investigated, the first aiming to classify Crohn's disease and ulcerative colitis patients and a second aiming to distinguish IBD patients from unaffected controls. Despite the two different classification tasks, the same supervised approach was applied to both frameworks and consisted of four main steps: (*i*) variance univariate feature selection; (*ii*) ANOVA feature selection; (*iii*) parameters grid search and; (*iv*) final model training and testing.

Table 6.1: **Pathways involved in IBD pathogenesis.**

| **Pathway** | **Genes** | **Repository** | **Search term** |
|---|---|---|---|
| Autophagy | 40 | KEGG | hsa04140 |
| B cell receptor signaling pathway | 73 | KEGG | hsa04662 |
| Cytokine-cytokine receptor interaction | 270 | KEGG | hsa04060 |
| GPCRs | 336 | PathCards | GPCRs |
| IBD | 65 | KEGG | hsa05321 |
| IFN-gamma pathway | 72 | PathCards | IFN-gamma |
| IL-10 pathway | 35 | PathCards | IL-10 |
| IL-9 signalling pathway | 12 | PathCards | IL-9 |
| Jak-STAT signaling pathway | 158 | KEGG | hsa04630 |
| NOD-like receptor signaling pathway | 170 | KEGG | hsa04621 |
| Intrinsic NOD2 pathway | 56 | KEGG | hsa04621 |
| T cell receptor signalling pathway | 105 | KEGG | hsa04660 |
| Th17 cell differentiation | 107 | KEGG | hsa04659 |
| Tight Junctions | 139 | PathCards | Tight junctions |
| TNF signalling pathway | 67 | PathCards | TNF signalling |
| **TOTAL (non-overlapping)** | 989 | | |

In order to introduce as much biological knowledge as possible and to remove uninformative data represented by invariant genes, a univariate feature selection based on variance was applied. With a threshold set to zero, genes that exhibit the same score across the entire dataset, in both cases and controls, were removed. This step filtered out genes without missense variants (all subject scores equal to zero) but also genes in which only common variants were observed (all patient scores are >0 but equal). Since the univariate feature selection is solely based on the variance of each feature and not on the classification target, it can be interpreted as a first unsupervised filtering approach.

There are multiple advantages in applying univariate feature selection prior to any supervised approach in which the number of features is much greater than the sample size. Removing invariant genes would filters out redundant data that are non-informative for the model which, especially when employing support vector machines (SVMs), would reduce classification performance (Section 2.4.1). A reduced number of genes would translate to a lower computational power and time required to train and test the classifiers. In frameworks requiring cross-validation this effect scales exponentially due to the multiple times a model gets fitted and tested.

Following the variance-based selection, a second feature selection step was implemented in the model to further refine the feature set. This second step implements a univariate feature selection based on a F statistic (ANOVA) testing the discriminatory power of each gene after providing the classification target. This approach provides a ranked list where the first gene is the most effective in separating the two target labels. Since there is not a recommended threshold for selection, we tested classifier performances applying an increasing 10 percent cut-off from 10% to 100%.

All supervised models described in this chapter implemented a support vector machine (SVM). In order to avoid overfitting and at the same time optimising models, a grid search was performed within a 10 fold cross-validation scheme for selecting the optimal SVM parameters. The cross-validated grid search tested two different kernels, a linear and a radial basis function (rbf) to cover both linear and non-linear decision functions. As explained in Section 2.3.1, the $C$ parameter controls the tolerance to error in the training process, here we tested three possible levels: 0.1, 1 and 10. Whilst $C$ is the only parameter to be tuned in linear SVM, rbf kernels also require tuning of the shape parameter $\gamma$ (kernel coefficient). Six possible kernel coefficients were tested: 0.001,0.01, 0.1, 1, 10 and 1/(number of features). The resulting optimal combination of kernel, tolerance and coefficient was then used in the final classifier. Due to imbalance between the number of patients with CD, UC and the controls classes, all SVM models were forced to correct weights accordingly (class_weight = " balanced").

Following the tuning step the model was fitted and tested using 10 fold cross-validation and performance metrics such as accuracy, F1 statistic and the area under the ROC curve (AUC) were corrected.

A total of 16 SVMs were trained and tested, each of which was implementing GenePy data based on one of the 16 different deleteriousness metrics

**Gene enrichment** The list of 989 gene was ranked according to the times each gene was recurrently selected across the 16 SVMs using different GenePy delete-

riousness metrics. The set of genes that were selected by all 16 GenePy deleteriousness scores were analysed for biological process enrichment using the Enrichr online tool [35]. In order to visualise any gene acting as potential hub of regulation, selected genes were analysed through STRING[183].

### 6.3.4 Unsupervised learning

Genomic data from paediatric IBD cases was modelled using unsupervised machine learning algorithms. This methodology was applied to all 16 different GenePy score matrices. The process was iterated for all 16 versions of GenePy integrating different deleteriousness metrics.

Supervised feature selection approaches (e.g ANOVA, $\chi^2$, entropy) were not applied since it would bias clustering towards the provided target labels. Therefore, once invariant genes were removed, data were directly modelled with unsupervised algorithms.

Principal component analysis and hierarchical clustering approaches were applied to IBD patient data to assess stratification into distinct patient groups that coul be clinically informative. Prior to PCA each feature was centred on the mean and scaled to unit variance. Whilst PCA does not require the definition of any parameter, HC was performed using average linkage method and euclidean distances. In order to establish the accuracy of the HC representation the cophenetic correlation coefficient was calculated.

## 6.4 Results

### 6.4.1 Supervised Classification of IBD subtypes

The first supervised approach aimed to classify CD and UC patients using a SVM classifier. In order to identify the optimal percentage of ranked genes to retain

after the ANOVA feature selection, we tested all bespoke thresholds for each of the 15 gene pathways while fixing the deleteriousness metric implemented in GenePy. This optimisation was performed on the total number of 989 genes previously identified and using CADD as deleteriousness metric. Table 6.2 shows accuracies, AUCs and F1 statistics for each tested percentage of ANOVA selected genes. We observed a curve in performance when the selection retains from 5 to 40 percent of the genes in the ANOVA ranked gene list with the optimum peaking at twenty percent. Table 6.2 reports statistics resulting from a 10 fold cross-validation approach. By including only 20% of the initial gene set, the SVM reaches its maximal performance in terms of accuracy, AUC and F1. Whilst there is a small difference between 10% and 20% in terms of accuracy and F1 metrics (one percent), the AUC at the 20% threshold is three percent better than the one calculated at 10%.

Table 6.2: **Performance of CD *vs.* UC classification for different ANOVA thresholds.** Performance were tested using the initial complete set of 989 genes (then variance-filtered down to 639) and CADD deleteriousness metric.

| **Anova threshold** | **Genes** | **Accuracy ($\sigma$)** | **AUC($\sigma$)** | **F1($\sigma$)** |
|---|---|---|---|---|
| **0.05** | 31 | 0.77 (±0.06) | 0.86 (±0.07) | 0.81 (±0.06) |
| **0.1** | 63 | 0.81 (±0.07) | 0.87 (±0.07) | 0.85 (±0.05) |
| **0.2** | **127** | **0.82 (±0.08)** | **0.89 (±0.07)** | **0.86 (±0.05)** |
| **0.3** | 191 | 0.79 (±0.04) | 0.84 (±0.07) | 0.84 (±0.06) |
| **0.4** | 255 | 0.77 (±0.08) | 0.83 (±0.05) | 0.82 (±0.04) |
| **0.5** | 319 | 0.73 (±0.06) | 0.78 (±0.07) | 0.76 (±0.08) |
| **0.6** | 383 | 0.72 (±0.06) | 0.78 (±0.05) | 0.78 (±0.05) |
| **0.7** | 447 | 0.68 (±0.06) | 0.70 (±0.08) | 0.75 (±0.06) |
| **0.8** | 511 | 0.66 (±0.06) | 0.59 (±0.08) | 0.76 (±0.04) |
| **0.9** | 575 | 0.64 (±0.01) | 0.58 (±0.11) | 0.78 (±0.01) |
| **1** | 639 | 0.64 (±0.01) | 0.34 (±0.09) | 0.78 (±0.01) |

After determining that 20% of the initial ranked gene list represented the most efficient threshold, we tested the efficiency of all genes within each of the fifteen pathways previously defined. In order to test such discriminatory power, the same SVM model, using the same ANOVA threshold and deleteriousness metric (CADD), was applied. Table 6.3 and Figure 6.1 shows the performance of each gene list each of which underwent the two feature selection steps. Here we observe that the maximal score is obtained when using the combined list of all 15 pathways resulting in a 89% AUC. We observe a strong positive correlation between SVM

performance and the number of genes used in the classification ($R^2 = 0.8336, p = 6 \cdot 10^{-5}$). Despite this effect, the NOD-like receptor signalling shows the second best AUC (73%) despite not being the second largest gene set. This indicates an enrichment for genes helpful in the classification of the CD/UC subtypes.

Overall, models based on larger gene sets perform better compared to those using pathways including few genes. The highest performance observed with the set of all genes can be imputed to a better representation and classification of the genotypic heterogeneity within our patient cohort. However the high classification performance of *NOD*-full pathway genes echoes the known strong biological role of the *NOD2* genes and its close molecular interactions.

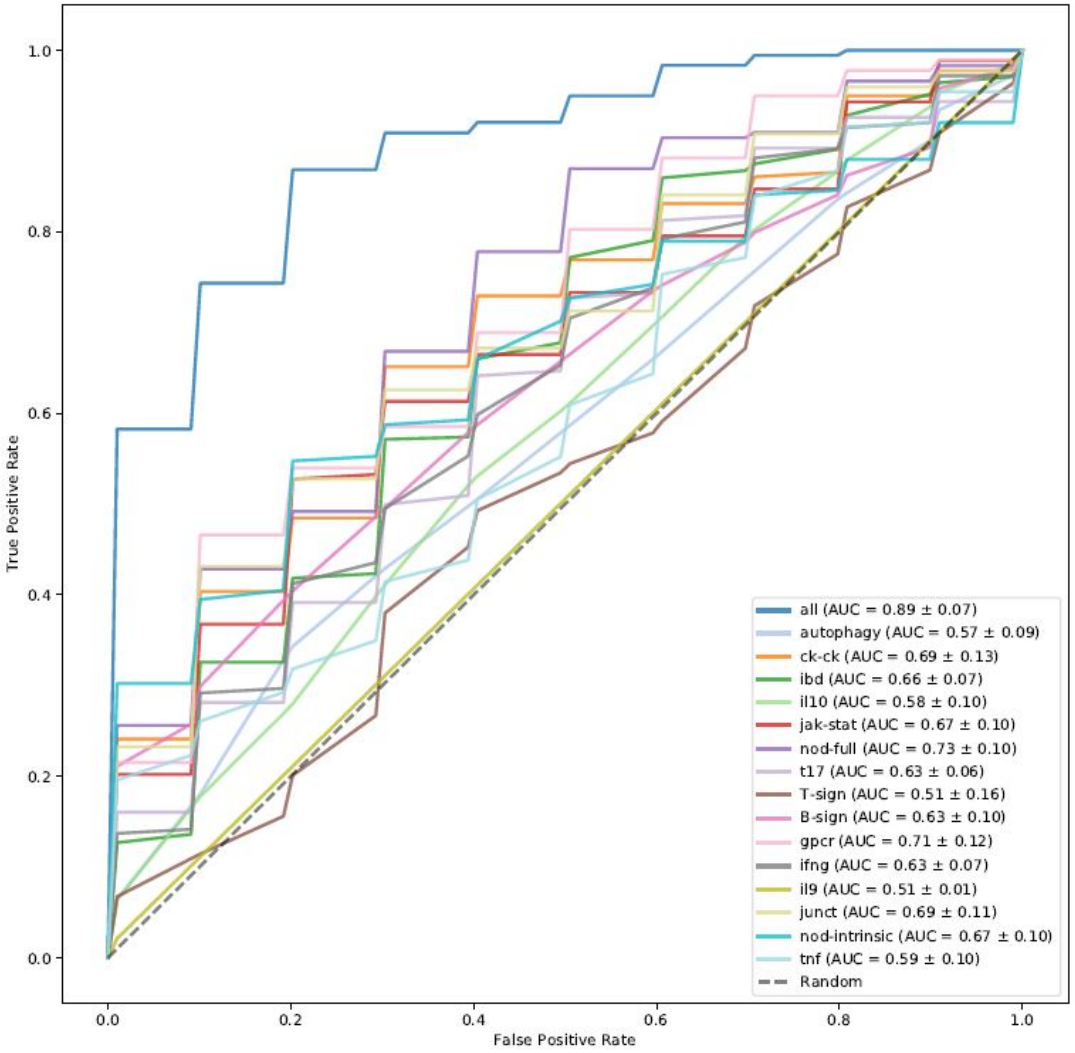


Figure 6.1: **CD *vs.* UC areas under the ROC curve for each IBD related pathway.** Each colour indicates a different pathway tested in a 10 fold cross validation framework following a 20% ANOVA selection. GenePy scores based on CADD metric were used to perform these test.

Table 6.3: **Supervised classification of CD vs UC using GenePy scores and CADD metric.**

| Pathway | Genes | GenePy available genes | Post variance selection | Post ANOVA selection | AUC | $\sigma$ |
|---|---|---|---|---|---|---|
| **Autophagy** | 40 | 17 | 17 | 6 | 0.59 | 0.10 |
| **B cell receptor signalling** | 73 | 47 | 45 | 13 | 0.64 | 0.10 |
| **Cytokine-cytokine receptor interaction** | 270 | 154 | 150 | 30 | 0.69 | 0.13 |
| **GPCRs** | 336 | 235 | 229 | 91 | 0.71 | 0.09 |
| **IBD** | 65 | 36 | 36 | 7 | 0.66 | 0.07 |
| **IFN-gamma pathway** | 72 | 46 | 46 | 9 | 0.63 | 0.07 |
| **IL-10 signalling** | 35 | 23 | 22 | 6 | 0.59 | 0.10 |
| **IL-9 signalling** | 12 | 5 | 5 | 3 | 0.51 | 0.06 |
| **Jak-STAT signalling** | 158 | 84 | 81 | 16 | 0.67 | 0.10 |
| **NOD-like receptor signalling** | 170 | 106 | 105 | 31 | 0.74 | 0.12 |
| **Intrinsic NOD2** | 56 | 31 | 31 | 21 | 0.68 | 0.09 |
| **T cell receptor signalling** | 105 | 64 | 63 | 12 | 0.63 | 0.09 |
| **Th17 cell differentiation** | 107 | 67 | 66 | 19 | 0.66 | 0.06 |
| **Tight Junctions** | 139 | 102 | 96 | 19 | 0.69 | 0.11 |
| **TNF signalling** | 67 | 39 | 38 | 15 | 0.60 | 0.11 |
| **All genes** | 989 | 659 | 639 | 127 | 0.89 | 0.07 |

Given the accurate predictions obtained implementing CADD as deleteriousness metric within the GenePy score calculation, we explored the impact of implementation of the 15 other deleteriousness metrics applying three different ANOVA thresholds (10%, 20%, 30%). Figure 6.2 graphically display the classification performance in terms of AUC across all deleteriousness metrics. Excepting the model using GenePy scores generated with M-CAP which consistently under performs, all other SMVs report highly similar AUCs (Table 6.4). Similar to previous observations, optimal classifications are obtained when selecting only the top 20% of genes following ANOVA univariate ranking. The best classifier was the SMV implementing DANN-based GenePy scores with an AUC of 0.91 using a 20% ANOVA threshold.

In order to provide a more complete assessment of model performance, we repeated this last test replacing the ANOVA univariate selection with a $\chi^2$ selection. Using the same threshold we observed lower AUCs (0.78 ($\pm$0.07) on average at 20% selection).

These results led to the application of a 20% ANOVA threshold and the use of all genes rather than a subset of genes within individual biological pathways in all

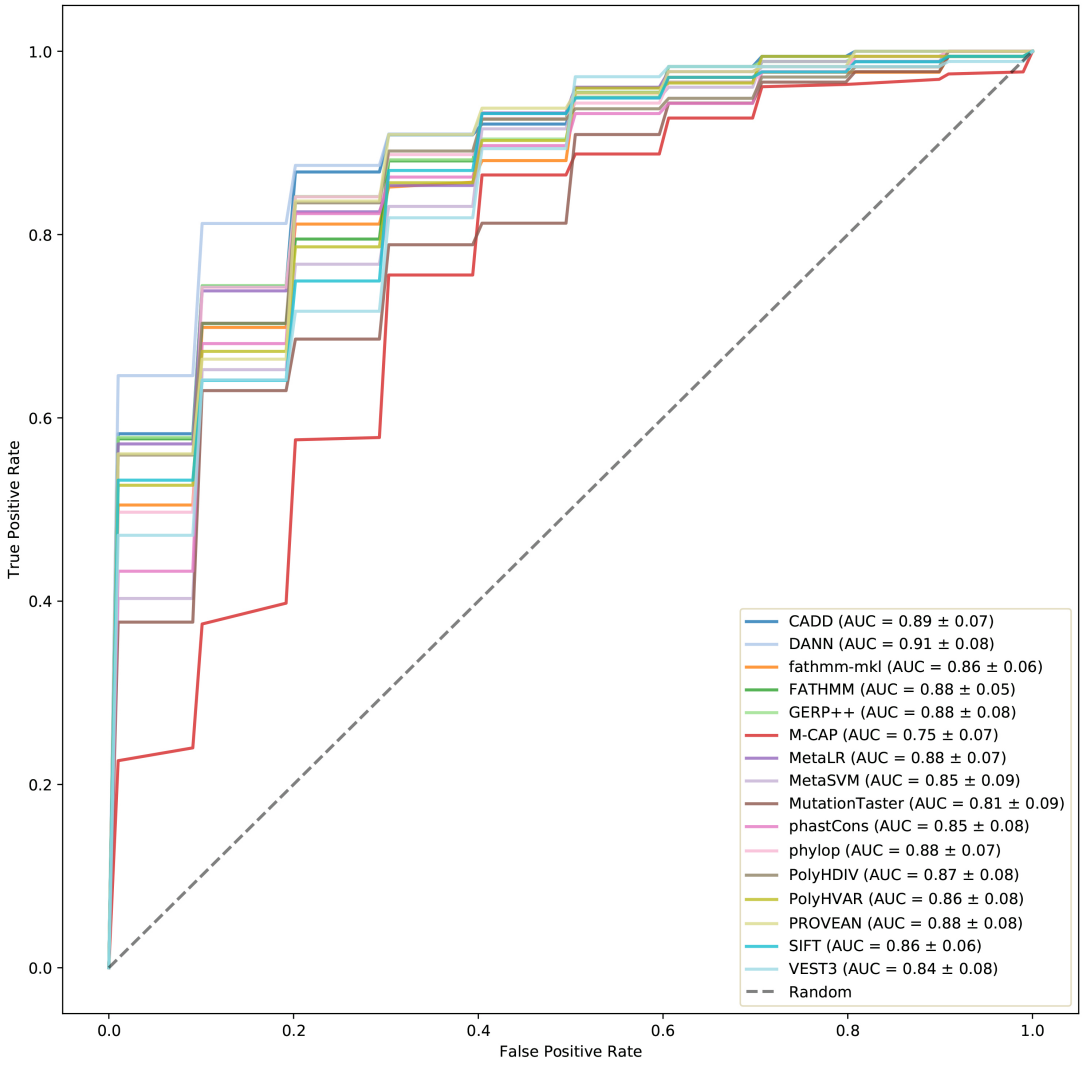


Figure 6.2: **CD *vs.* UC areas under the ROC curve for each deleteriousness metric.** Each colour indicates a different tested metric in a 10 fold cross validation framework following a 20% ANOVA selection over the complete set of genes (n=989).

subsequent analyses. Conversely, testing for all 16 versions of GenePy scores will be required to model different aspects of the same cohort.

### 6.4.2 Supervised Classification of IBD vs Control

The same classification framework was applied in order to distinguish IBD patients from healthy controls using solely genomic data transformed through GenePy. Using the same feature set of 898 genes involved in IBD related pathways and a 20% ANOVA selection threshold, sixteen SVM classifiers were trained and test

Table 6.4: **CD *vs.* UC classification performance for each deleteriousness metric.**

| **Metric** | **Post variance selection** | **Post ANOVA 10%** | **AUC 10%** | **Post ANOVA 20%** | **AUC 20%** | **Post ANOVA 30%** | **AUC 30%** |
|---|---|---|---|---|---|---|---|
| CADD | 639 | 63 | 0.87 (±0.07) | 127 | 0.89 (±0.07) | 191 | 0.84 (±0.07) |
| DANN | 639 | 63 | 0.88 (±0.04) | **127** | **0.91 (±0.08)** | 191 | 0.89 (±0.04) |
| fathmm-MKL | 639 | 63 | 0.87 (±0.08) | 127 | 0.86 (±0.06) | 191 | 0.82 (±0.08) |
| FATHMM | 619 | 61 | 0.86 (±0.06) | 123 | 0.88 (±0.05) | 185 | 0.86 (±0.05) |
| GERP++_RS | 645 | **64** | **0.89 (±0.05)** | 129 | 0.88 (±0.08) | **193** | **0.90 (±0.03)** |
| M-CAP | 573 | 57 | 0.73 (±0.13) | 114 | 0.75 (±0.07) | 171 | 0.80 (±0.11) |
| MetaLR | 631 | 63 | 0.85 (±0.09) | 126 | 0.88 (±0.07) | 189 | 0.86 (±0.05) |
| MetaSVM | 636 | 63 | 0.85 (±0.05) | 127 | 0.85 (±0.09) | 190 | 0.83 (±0.08) |
| MutationTaster | 595 | 59 | 0.81 (±0.06) | 119 | 0.81 (±0.09) | 178 | 0.86 (±0.06) |
| phastCons | 645 | 64 | 0.87 (±0.07) | 129 | 0.85 (±0.08) | 193 | 0.87 (±0.09) |
| phyloP | 645 | 64 | 0.88 (±0.05) | 129 | 0.88 (±0.07) | 193 | 0.86 (±0.05) |
| Polyphen2_HDIV | 669 | 66 | 0.88 (±0.04) | 133 | 0.87 (±0.08) | 200 | 0.84 (±0.08) |
| Polyphen2_HVAR | 672 | 67 | 0.83 (±0.07) | 134 | 0.86 (±0.08) | 201 | 0.87 (±0.05) |
| PROVEAN | 624 | 62 | 0.87 (±0.05) | 124 | 0.88 (±0.08) | 187 | 0.88 (±0.05) |
| SIFT | 661 | **66** | **0.89 (±0.05)** | 132 | 0.86 (±0.06) | 198 | 0.86 (±0.06) |
| VEST3 | 638 | 63 | 0.85 (±0.11) | 127 | 0.84 (±0.08) | 191 | 0.87 (±0.06) |

implementing different deleteriousness metrics.

Figure 6.3 shows the AUC results of all the tested models. Similarly to observations in the classification of IBD subtypes, the choice of deleteriousness metric selection had only a very small impact on classifier performance. Whilst DANN results are marginally optimal, M-CAP performs least well with almost 10% lower AUC compared to other metrics. Likely due to the fact this recent metric has most missing data for individual variants and subsequently less information for modelling.

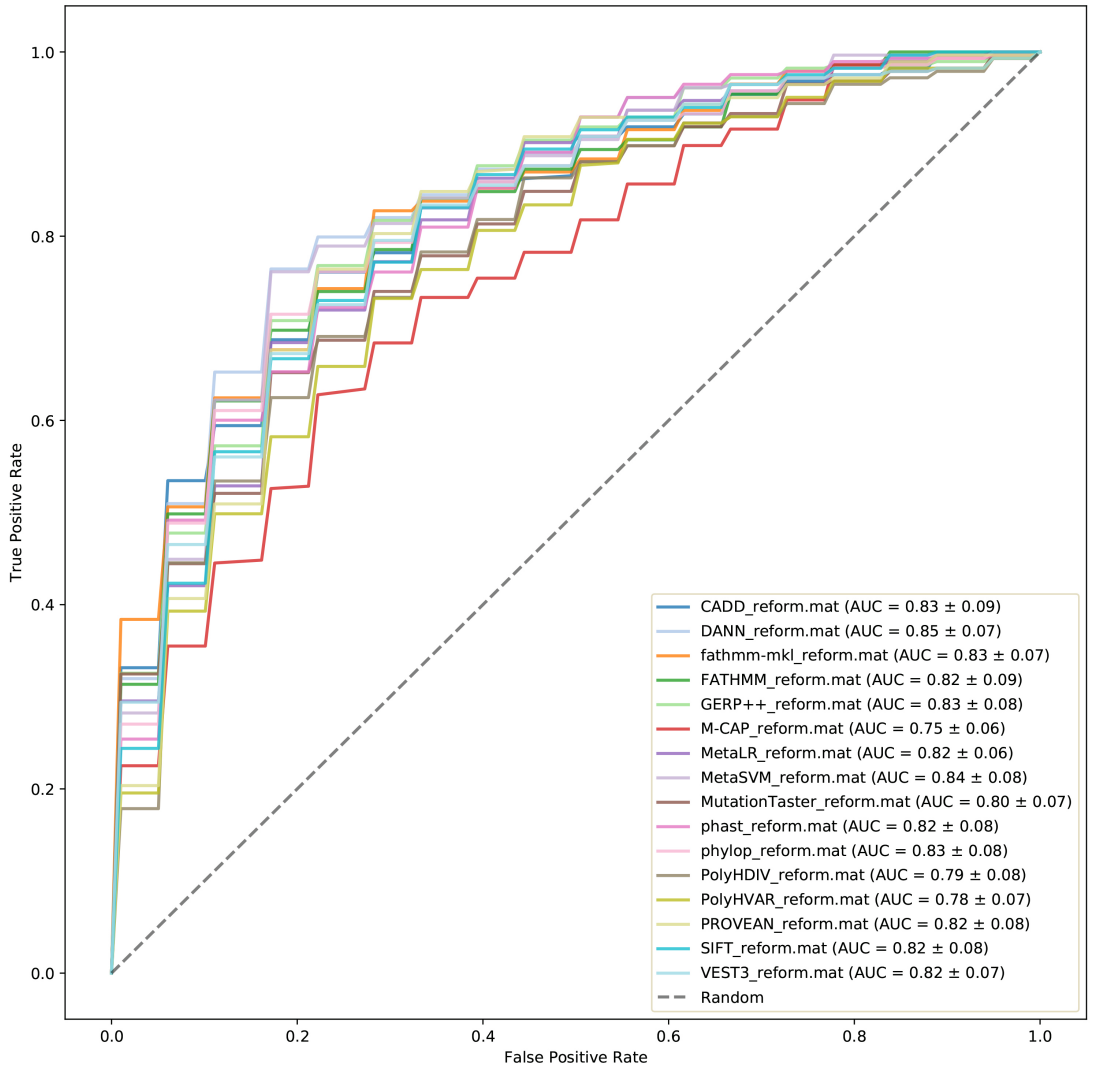


Figure 6.3: **IBD *vs.* control areas under the ROC curve for each deleteriousness metric.** Each colour indicates a different tested metric in a 10 fold cross validation framework following a 20% ANOVA selection over the complete set of genes (n=989).

### 6.4.3 Gene enrichment

The similar performance observed in both classification approaches might indicate the presence of a core gene set that provides the maximal discriminatory power when classifying IBD subtypes or IBD from controls. In order to identify such selection pattern, we extracted the list of genes employed by each SMV implementing different deleteriousness metrics. Subsequently, we ranked such lists depending of the times each gene was selected and analysed with Reactome the list of genes that were selected by all models (n=16).

This process was performed for for investigations of both CD *vs.* UC and IBD *vs.* controls and the consensus gene set analysed for enrichment through Reactome. In total, 40 genes were utilised by all SVMs in the classification of IBD subtypes and these were strongly enriched for *FGFR3* signalling cascade (FDR p = $6.48 \cdot 10^{-5}$). Figure 6.4 shows the protein-protein interaction network based on the 40 genes selected by all models. It is notable that *IL10*, *PLCG1*, *MTOR* and *PPP3CA* are the nodes with the highest degree (number of edges) and therefore important hubs for maintaining the network connectivity. While *IL10* and *MTOR* are genes known to be associated with the IBD phenotype [94, 69], there are no records concerning the role of *PLCG1* and *PPP3CA* in disease characterisation.

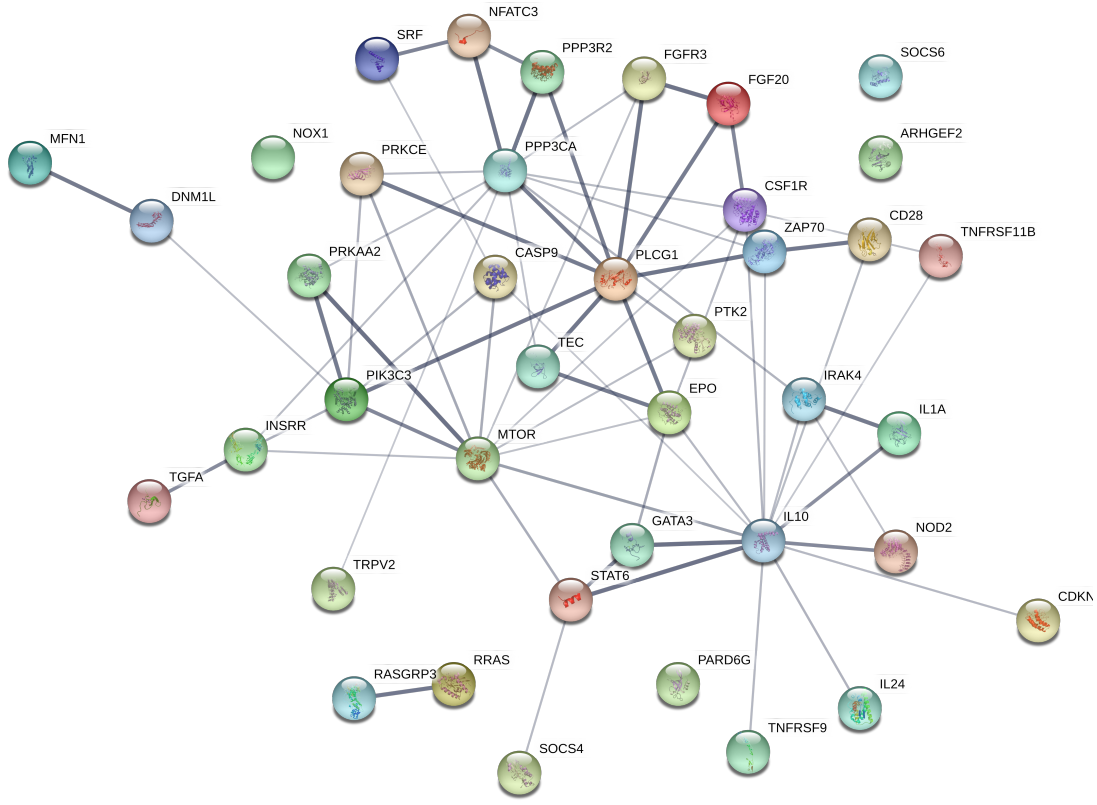


Figure 6.4: **Protein-protein interaction network based on the 40 genes selected for the CD *vs.* UC classification.** Each node represent a protein (gene product) and edges represent interaction between nodes. The thickness of the edges is directly proportional to the strength of the interaction.

The same analysis identified 28 genes selected throughout the 16 SVM models for the classification of IBD patient and controls. This gene list was significantly enriched for the interleukin family (*IL2*, *IL20*, *IL21*, *IL35*) signalling (FDR p = $5.55 \cdot 10^{-4}$). Figure 6.5 shows the interaction network obtained by analysing the

fundamental genes for distinguishing cases from controls. Although there is lower overall connectivity, *JAK2* and *FGF2* act as hubs of the network. *JAK2* is a known IBD-associated gene which mutation induces a gain of function triggering a pro-inflammatory reaction mediated by cytokines and macrophages [66]. Conversely, *FGF2* cooperates with interleukin 17 to repair the intestinal epithelial damage induced by chronic inflammation [177].

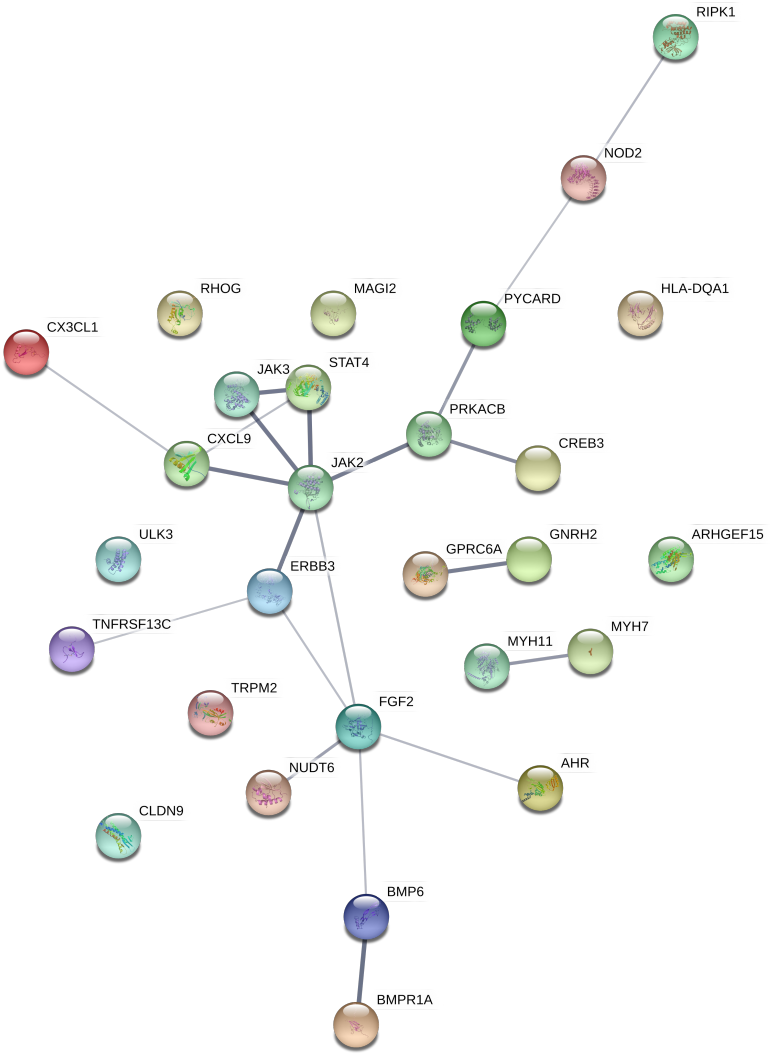


Figure 6.5: **Protein-protein interaction network based on the 28 genes selected for the IBD *vs.* controls classification.** Each node represent a protein (gene product) and edges represent interaction between nodes. The thickness of the edges is directly proportional to the strength of the interaction.

Both networks exhibit the distinctive scale-free network topology that characterise most of the known biological networks. Interestingly, *NOD2* is the only gene that is always selected when using any deleteriousness metric in both classification tasks.

This observation represents a further evidence of the central role that *NOD2* plays in the IBD pathogenesis and characterisation of specific subtypes.

### 6.4.4 Unsupervised Stratification of IBD patients using genomic data

Unsupervised stratification can lead to the discovery of novel strata which might better reflect molecular pathology and inform treatment choices. The true biology of IBD is likely to be more complex than is reflected by the currently applied subtype categories of UC and CD that are historically based uniquely on pathology data. We therefore fist modelled genomic data from IBD patients to observe unsupervised grouping of affected individuals.

A principal component analysis was performed on the merged and unlabelled set of all IBD cases (CD, UC and IBDU) using all 989 genes identified in IBD related pathways. Figure 6.6 shows the resulting PCAs for GenePy generated using 16 different deleteriousness metrics. None of the PCAs show a significant separation of cases when plotting the first two principal components. Both components on average are capable of explaining only $\sim 1\%$ of the original variance. Except for few outliers, almost all IBD patients generate a tight cluster, not showing any stratification. This lack of separation and the low variance explained indicates a poor performance of the PCA algorithm in stratifying such data rather the complete absence of strata.

HC was performed on all versions of GenePy implementing each of the 16 available deleteriousness metrics. According to the cophenetic correlation coefficient (CC), M-CAP based HC provides the closest representation of the original distances of the unmodelled data (CC = 0.88). The lowest CC was observed when implementing the VEST3 metric (CC=0.67) whilst on average the mean CC is 0.79. HCs based on Meta-LR and SIFT GenePy scores show the second best CC of 0.85. Using average linkage and euclidean distances, HC stratifies IBD patients into different number of clusters depending on the metric implemented in GenePy

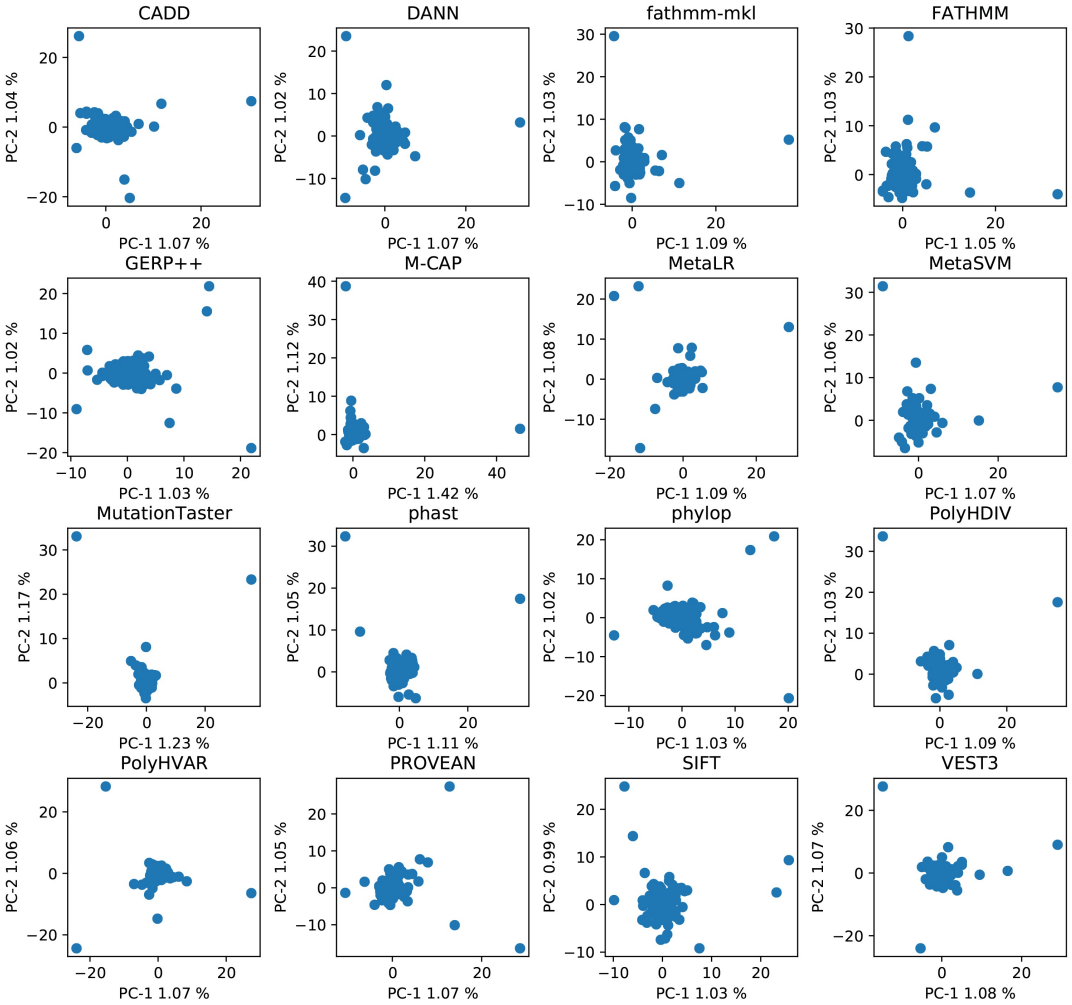


Figure 6.6: **Principal component analyses of IBD cases 16 different deleteriousness metrics.**

scores (Figure 6.7).

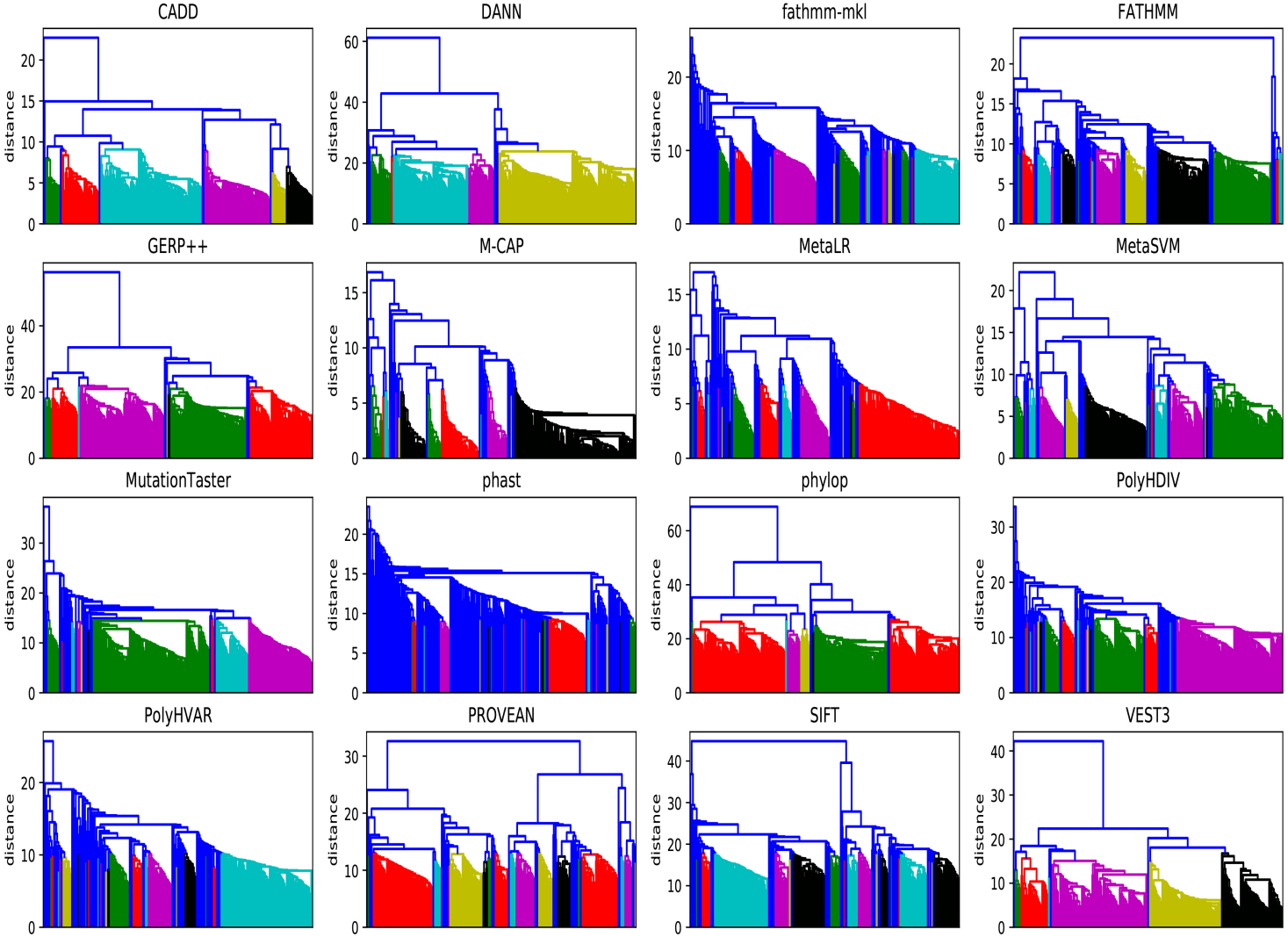


Figure 6.7: **Hierarchical clustering of IBD cases for all 16 deleteriousness metrics. Cophenetic correlation coefficient (CC) is shown in the top right corner of each subplot.**

These groups are distinct from the current classification of IBD subtypes into UC and CD. A highly similar structure can be observed in M-CAP and Meta-LR hierachical clustering models (first and second best performing model), whilst is absent in SIFT based HC (Figure 6.7). Clustering based on CADD, DANN, GERP++, phyloP and VEST3 show a recurrent pattern characterised by four equal in size clusters generating from a recent common ancestor.

The optimal model defined by CC implements M-CAP scores and identifies five clear groups (Figure 6.8). Group V is the largest group (n=136) accounting for 48% of the total IBD cohort while group I is the smallest including only 7% of the cohort (Table 6.5 and Figure 6.8). Five percent of the individuals (n=15) were not assigned to any specific cluster. Cluster I is generated by the second split of the dendrogram starting from the common ancestor. This event indicates a strong separation and greater distance of cluster I from remaining clusters (II to V). Clusters II to V occur five splits following the separation from cluster I.

In order to identify which genes were contributing to defining these novel groups, we performed a Mann-Whitney test between GenePy scores of the largest cluster V against all others. In total, GenePy scores from 54 genes were significantly different between cluster V versus clusters I to IV (Supplementary Table 7.3). Following the correction for multiple tests (Bonferroni correction for 898 genes), two genes significantly influenced cluster formation: *HLA-DQA1* ($p=1.50 \cdot 10^{-27}$) and *HLA-DRB5* ($p=7.72 \cdot 10^{-19}$). These two adaptive immune system genes has been repeatedly reported as involved in IBD pathogenesis and disease markers [119]. Examining the GenePy scores for these two *HLA* genes across the clusters, cluster I is characterised by high mutation of *HLA-DRB5* and moderate mutation of *HLA-DQA1* genes (Table 6.5). Clusters II and III exhibit the highest mean mutation of *HLA-DQA1* whilst cluster IV shows the minimal burden for both genes. Finally, cluster V is characterised by a modest mean mutation of the *HLA-DRB5* gene.

Clinical markers of disease severity (surgery, AZA and steroids) were regressed against cluster structure using a multiclass logistic regression and did not exhib-

ited statistical significant results. Borderline statistical significance was observed between patients that underwent surgery and those who did not in terms of mutational burden of *HLA-DQA1* ($p = 0.052$). Further investigation of the obtained clusters with regard to their clinical data shows an age of diagnosis of individuals largely equal across clusters with a mean age of 11.7 years old (Table 6.5). Clusters II and III exhibit the lowest number of individuals that required surgery (9.7% and 7.1% respectively), cluster V follows with approximately 14.7% of individuals.

Interestingly, the mutational burden of the *HLA-DQA1* is negatively correlated with surgery in that as the GenePy score for this gene decreases the number of patients requiring surgery increases. Clusters I and IV have a markedly higher percentage of individuals that underwent surgery (19%).

Azathioprine (AZA) treatment was equally prescribed to individuals across the five clusters with the exception of cluster I which exhibited a higher percentage of patients treated with such drug (81%). The same percentage (75%) of patients were treated with steroids within each cluster. Cluster I shows the highest mutational burden of *HLA-DRB5* as well as the highest percentage of patients treated with AZA.

Table 6.5: **Unsupervised IBD groups regression data.** Summary representation of demographic and genomic data for disease severity markers distinguishing the five novel groups. Group zero indicates all individuals not fitted in any cluster. Asterisks indicate elements for which the mean value is reported. Clinical features, such as surgery, treatment with azathioprine (AZA) or steroids, are reported as the number of patients that experienced one episode since diagnosis.

| **Groups** | **0** | **I** | **II** | **III** | **IV** | **V** | **Total** |
|---|---|---|---|---|---|---|---|
| **N** | 15 | 21 | 31 | 56 | 26 | 136 | 285 |
| **Age at diagnosis*** | 12.41 | 11.86 | 11.79 | 11.90 | 11.39 | 11.42 | 11.64 |
| **Surgery** | 0 (0) | **19.0% (4)** | **9.7% (3)** | **7.1% (4)** | **19.2% (5)** | 14.7% (20) | 12.6% (36) |
| **AZA** | 53.3% (8) | **80.9% (17)** | 67.7% (21) | 64.3% (36) | 69.2% (18) | 66.2% (90) | 66.7% (190) |
| **Steroids** | 80.0% (12) | 76.1% (16) | 71.0% (22) | 75.0% (42) | 69.2% (18) | 81.6% (111) | 77.5% (221) |
| **HLA-DQA1*** | 4.59 | 3.06 | 7.40 | 8.55 | 0.13 | 0.14 | 3.03 |
| **HLA-DRB5*** | 3.39 | 15.30 | 6.93 | 0.00 | 0.00 | 2.75 | 2.75 |

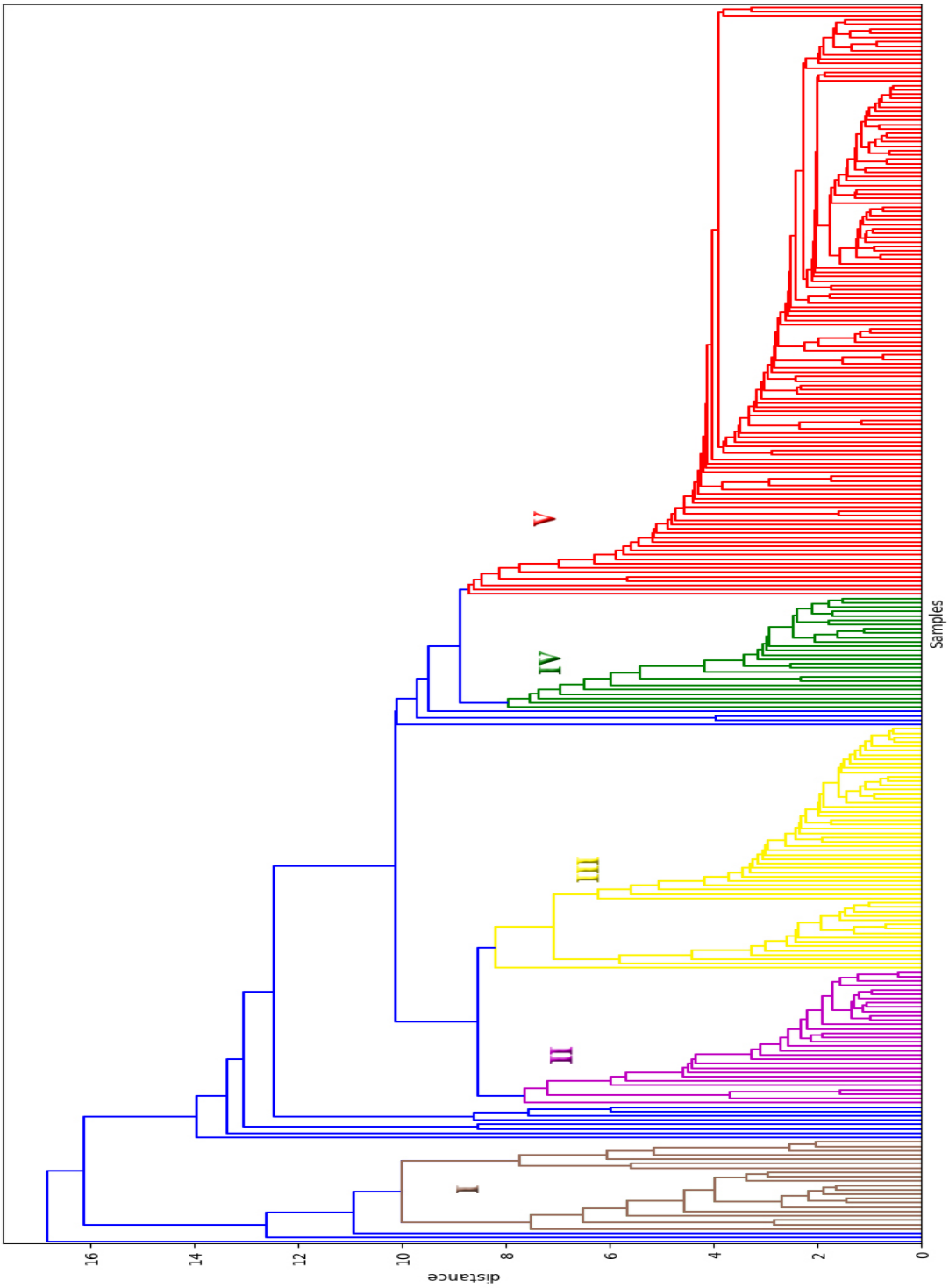


Figure 6.8: **Hierarchical clustering of only IBD cases using M-CAP deleteriousness metric.**

## 6.5 Discussion

Machine learning algorithms were historically developed to solve complex discriminatory problems and their application in clinical and biological contexts is not new (Section 1.3). Supervised machine learning models can be used to classify samples, usually represented by a collection of hundreds or thousands features, according to existing classification labellings. The main advantage of such approaches resides in the integration of large datasets that would not be possible to analyse with conventional statistics.

In recent years, several models have being proposed for classifying IBD patients into the two main clinical subtypes (CD and UC). The need for machine learning approaches is driven by the complex diagnostic process of IBD. Similar to other complex diseases, IBD presents as wide range of phenotypes with a large overlap of symptoms between subtypes. In addition to the unclear aetiology, the phenotypic similarity between IBD subtypes increases the challenge of assigning specific diagnoses. The diagnostic uncertainty is often increased in children in whom the label of IBD unspecified is more frequently assigned than in adults [42]. Lack of a clear diagnosis impacts clinical decision making. The genetic heterogeneity is demonstrated when patients frequently presenting with a positive family history of one subtype of IBD often have closely related individuals diagnosed with the other subtype. Whilst most of these methods focussed on modelling clinical data, only few supervised approaches were able to perform sufficiently well implementing unbiased genomic data. During the 2013 CAGI challenge, 14 research groups presented various approaches to distinguish a cohort of 51 Cronh's disease patients from 15 healthy individuals using solely whole exome sequencing data. As a result of an AUC-based performance evaluation, two models reached the highest score of 0.87[28].

We hypothesised improved classification performance would follow implementation of genomic data. The main purpose for developing the GenePy model was to fill a gap in systematic approaches for manipulating and transforming sequencing data in a format that would accommodate more complex analyses while incorporating

additional biological information.

Utilising upstream gene selection processes and support vector machines, we reported the highest AUC (91%) in the discrimination of CD from UC patients. The improved performance of our model compared to those so far reported is largely attributable to the biological higher information content entered through GenePy. Moreover, the accurate feature selection process, thorough pathway-oriented pruning and subsequent systematic univariate selection, removed confounding factors that could have reduced our model performance.

A common problem affecting supervised machine learning models is data overfitting, leading to almost perfect classifiers that in reality would perform less well on novel unseen data due to the lack of generalisation. In this chapter, we ensured that each step of our methodology was approached with a 10 fold cross-validation, protecting from overfitting and leading to more conservative results.

Despite the superior performance, our model did not reach the 95% AUC threshold required for a clinical application [155]. By investigating all the current ML models for the classification of CD and UC patients using genomic data, it is possible to observe a virtual upper bound (approximately 91% AUC) that every model struggles to overcome. The reason for this limitation might be attributable to two factors. First, the current CD/UC diagnosis is based on clinical observations (predominately endoscopic and histological) and imposes a historic pathology based classification that may not be fit for the purpose of accurately reflecting the true molecular aetiology of IBD. Secondly, due to the complexity of IBD, changes in diagnosis are not infrequent, particularly in children. Both these features reflect the need for revision and potential reformulation of disease classifiers. The need of a new revised classification based on deep clinical phenotyping and modern immuno/genomics traits has been already extensively discussed [126, 198] and might explain the inability of supervised model to classify accurately the totality of IBD patients. Our supervised model for the classification of IBD samples against controls reported the highest AUC of 85% but also fell short of the clinical threshold of 95%.

From the analysis of genes selected by each supervised classifier we observed two distinct pathways enriched in the classification of CD versus UC and IBD versus controls. Genes involved in the cytokine signalling pathway appear key elements in distinguishing paediatric IBD patients from healthy individuals. In addition to a degenerate cytokine response, changes in the *FGFR3* pathway appear to drive patients in the current CD and UC subtypes. Both cytokines and grow factors have been associated with IBD pathogenesis [83, 133]. The absence of cytokines enriched in the CD/UC distinction might indicate that such pathway is similarly affected in all IBD patients. Interestingly, the only gene selected by both classifiers is *NOD2*. Since *NOD2* is specifically associated with CD, it is expected to play a role in the CD versus UC discrimination. Since our IBD cohort is enriched with CD cases, selection of *NOD2* in the IBD versus controls classification is also not unexpected.

Unsupervised approaches to stratify IBD cases using genomic data demonstrated that whilst PCA was not effective in identifying strata, the best hierarchical clustering model identified five novel clusters. These new strata, driven uniquely by genomic data, do not follow the current classification into just two CD or UC subtypes and is a further evidence of the possible need for improved complex subtyping systems in IBD. Investigating the main clinical features used as marker of severity against novel clusters we observed interesting results indicating the presence of two clusters that have lower frequency of surgery events and one cluster with higher azathioprine treatment rates. Similar results could be of high value in a clinical framework aiding clinicians in choosing the appropriate intervention based on genetic evidence.

From a genetic perspective, we observed a borderline significance between the level of *HLA-DQA1* gene damage (indicated by GenePy scoring with M-CAP) and the need for surgery. The role of *HLA* genes in IBD are well established [188, 53, 119]. It has been estimated that approximately 10 to 33 percent of the total risk of developing CD is due to mutations in HLA genes [115]. Our data indicate the relationship between IBD and *HLA* might be specific to subsets of patients harbouring risk genes. Although these patients may manifest endoscopic disease

feature similar to other groups. Moreover, IBD patients with mutations in the *HLA-DQA1* gene tend to exhibit intolerance to AZA treatment which might lead to pancreatitis [199]. Whilst *HLA-DQA1* gene has been strongly associated with IBD, the role of *HLA-DRB5* is not fully understood[188]. Additional genomic data and refined clinical evidence will play an important role in elucidating the precise nature of this relationship. However, these findings are a blueprint for a more personalised approach to treating IBD patient.

The novel identified clusters require further investigation, possibly in independent cohorts, and (perhaps longitudinal) clinical data. A multi-class supervised ML model could be employed to better characterise the specific genomic and clinical attributes of each group.

Herein, both supervised and unsupervised models were based solely on genomic data, which represent only one of the many 'omics data available to describe the IBD phenotypes. As well as being a more intuitive than existing methods aiming to integrate NGS data through presence or absence of variants, the continuous nature of GenePy scores affords relatively straightforward merging of genomic data with metabolomic, transcriptomic and clinical data will be simpler. A novel multi-level approach may provide a more complete picture of IBD complexity, leading to more accurate IBD strata and classifiers.

Ultimately, this will aid clinicians in making early accurate diagnoses and rapid assignment of treatments specific to the underling molecular biology and not the superficial appearance at endoscopy.

# Chapter 7

# Conclusions and Future work

This thesis describes novel approaches to analyse next-generation sequencing data in order to untangle the complexity of IBD genetics by using a broad range of supervised and unsupervised machine learning methods. The preceding chapters depicted the potential of machine learning and mathematical models for more accurate diagnosis and patient stratification.

Machine learning can assist and accelerate the decision making process, reduce uncertainty and provide confidence for medical decisions. The application of unsupervised machine learning to clinical data demonstrated the complexity of classifying and stratify patients with inflammatory bowel disease as a consequence of the substantial overlap of disease symptoms. Overlapping characteristics of IBD subtypes are established in the current literature and represent the main source of uncertainty when assigning a specific subtype diagnosis. Through the integration of multiple clinical features machine learning algorithms were capable of partially solving such complexity by combining eight features between histological and endoscopic observations. The simultaneous modelling and interpretation of multiple traits is one of the key reasons for applying ML to health care problems. Assignment of any given patient with the correct IBD subtype is a crucial step that impacts specific treatment plans. Our data provided the highest accuracy in classifying CD and UC patients using solely histological and endoscopic evidence. Histopathology data is routinely collected for the majority of suspected

IBD patients making our model easy and cost effective to implement in other studies. Analysis of larger paediatric cohorts could increase model performance as machine learning methodologies have direct correlation between the size of learning dataset, prediction accuracy and generalisation. However, paediatric IBD has a substantial genetic component and so classification of patients using uniquely clinical features ignores informative genetic discriminants of disease. The optimal ML models for clinical application would incorporate contemporary genomic data where available.

We investigated the opportunity of modelling patients using their genomic data in order to increase the performance of machine learning algorithms to classify and stratify IBD patients. Genomic data represents rich molecular data that is objective and less influenced by observation bias in its generation. The integration of NGS data in non-canonical analytical approaches (e.g machine learning and network analyses) requires the transformation of substantially binary data (mutation/no-mutation) into a format capable of both maximising the information content and ease data interpretability. We hypothesised several strategies for modelling genomic data through machine learning models with the aim of magnifying the biological information carried by NGS data. Most of the currently available metrics and approaches focus either on single mutations or large cohorts of individuals making it impossible for clinicians to integrate genomic data at the patient level. GenePy is a novel per-patient score capable of modelling genomic data at a gene-level. While about two percent of IBD patients disease can be explained by individual very rare mutations in single genes (monogenic disease), most patient disease is consequent to the cumulative effect of multiple genetic alterations (SNVs, large indels and CNVs). The gene-level approach of GenePy can model this additive behaviour and, following correction for gene length, gene scores can be compared between genes or combined into higher level systems such as signalling pathways. This innovative approach to genomic data performed optimally on our cohort of PIBD patients for which we obtained both immunology and WES data. Through combining GenePy scores of genes closely interacting with *i*) receptor; *ii*) signal modulation and; *iii*) cytokine production we were able to

discriminate genes specifically inducing the general hypo-inflammatory response in a subset of treatment naïve patients.

Although our data demonstrated the value of GenePy in modelling NGS data, there is clear scope for additional improvements. GenePy performance is dependent on the quality of NGS data, the bioinformatic pipeline and the efficiency/completeness of annotation software (e.g. ANNOVAR) deleteriousness scores. Such annotation tools are constantly updated to provide better performance. For example, the VEST3 deleteriousness metric was originally developed to score insertion/deletion (indels) variants that alter protein sequence. The newer release, VEST4 (not yet implemented in dbsnfp database or ANNOVAR) was improved to score in-frame and frameshift indels. In the current version of GenePy, frameshift indels are arbitrarily set to the maximal deleteriousness value, however, with better annotated data, the score would better model the effect of such mutation leading to more accurate representation of the gene burden. Combined scoring systems, such as CADD that rely on multiple third-party scores would require a complete score rebuild to implement up to date features. Future versions of GenePy could include novel deleteriousness metrics capable of scoring (non)frameshift and splice site altering variants. This could be done through the implementation of refined metrics such as VEST4[32] or Gwava[156], a recent deleteriousness metric based on a random forest classifier trained to score coding and non-coding variants combining existing genomic and epigenomic metrics.

GenePy scoring of genes is currently limited to genomic regions captured by whole exon enrichment capture kits used in the IBD cohort. GenePy provided scores for approximately 14,000 genes out of the 21,000 currently reported in the RefSeq database. Whole exome sequencing is bound by capture kit design and a solution to capture missing regions would be through obtaining whole genome sequencing data. Large scale projects, such Genomics England which is collecting complete whole genome sequencing data from thousands of patients across England, will maximise the potential of GenePy as model for integrating and interpreting WGS data. Due to the flexibility of GenePy, it is also possible to investigate non-coding regions making the score suitable and ready for the analysis of whole genome se-

quencing data (WGS). Instead of focussing on pre-defined gene regions, GenePy could be used to assess deleteriousness across contiguous sliding windows to reflect the burden of pathogenic annotation across whole chromosomes. In order to efficiently estimate the deleteriousness of non-coding regions, GenePy scores could integrate deleteriousness metrics based either on conservation frequencies (e.g. phastCons, GERP++, phyloP) or through methods capable of modelling genome-wide features such as distance to transcription start sites, chromatin state, GC-content, etc... (e.g. DANN, Eigen, Gwava).

One of the greatest opportunities that might substantially increase the power of GenePy would be incorporation of information about allelic phase (Figure 7.1). The allelic phase, defines on which of the two parental alleles (maternal or paternal) a variant is observed. For example, given two heterozygous mutations (where one allele is mutated and one is not) without information about phase it is not possible to ascertain whether both mutations are on the paternal allele; both on the maternal allele or one on each. Since the majority of the proteins are coded by both alleles, having pathogenic mutations on both maternal and paternal alleles at the same time might result in the absence of any gene product. On the other hand, if both mutations are on the same allele, there would be still one unaffected allele capable of producing a normally functioning protein. Phase of genomic data is therefore biologically important to predict whether the organism is capable of producing any normal/non-pathogenic protein. Phased data requires either long paired-end approach sequencing or parental data. Some tools are currently available to impute phase from WES data but are reported to perform poorly [29].

Although there is a clear scope for further refinement, the implicit benefit of GenePy is already conferred through its ability to provide per-patient per-gene scores that can be used for machine learning purposes. Modelling GenePy scores from IBD patients through supervised and unsupervised machine learning methodologies is efficient and leads to new insights on the genetics of paediatric IBD. To date, the performance reached through GenePy and SVMs are the highest observed in the classification of IBD subtypes (CD/UC) using WES data. Such models were based on support vector machines, powerful yet simple algorithms that can be used

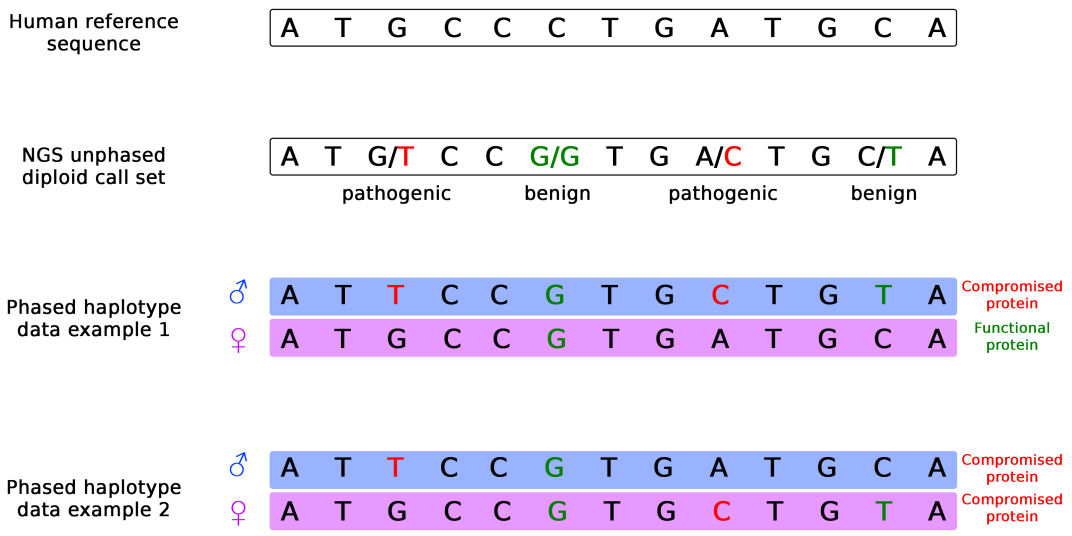


Figure 7.1: **Schematic representation of phased data.** Representation of phased sequencing data. NGS data obtained with short-reads cannot resolve on which chromosome a heterozygous variant was observed (paternal or maternal). With phased data, it is possible to observe the exact origin of a mutation. Given an unphased callset, the phase can result in multiple scenarios. In phased example 1 both pathogenic heterozygous variants are on the same paternal chromosome leaving the maternal copy unaltered. In phased example 2 deleterious mutations are distributed on both chromosome copies which may lead to the complete absence of the coded protein.

for either classification and regression purposes. However, more computationally intensive algorithms can be taken into consideration, such as Random Forest (RF) and Artificial Neural Network (ANN) approaches. These two algorithms are becoming increasingly popular due to their flexibility across different data types and their ability to model non-linear problems. Also, such models have less constraints on input data with respect to their distribution. Whilst Random Forest models are amongst the most interpretable machine learning approaches, ANNs act can as black-box making the relationship between features and prediction hard to interpret[190].

Unsupervised stratification of genomic data from patients with common complex disease can be challenging and harder to interpret but may lead to novel insights. In this work it was possible to link ML-defined genomic subgroups of paediatric IBD patients with different frequency of surgery or treatment with azathioprine indicating an important role of genetics in predicting disease course and optimal clinical management. In order to further improve the understanding of natural strata characterising paediatric IBD, more complex and advanced models can be

applied. Network based approaches could be employed to detect clusters of patients with specific genomic signatures. This has already proved effective in the analysis of patients diagnosed with asthma [23], a common complex autoimmune condition not dissimilar to IBD. This novel approach is based on topological data analysis (TDA) theories, particularly well suited for the analysis of high dimensional and noise data [50]. GenePy scores are applicable to define the backbone network generated through TDA and then, hypothetically identified clusters can be regressed against clinical, metabolomic, transcriptomic and microbiome data.

The framework presented in this thesis represents a blueprint applicable to any common complex disease where genetics plays a central role. This work presents new models for more accurate representation of IBD subtypes and the complexity of their genetic component. The current categorisation of IBD by CD and UC subtypes is more frequently showing its limitations and evidence of novel strata based solely on clinical observations have been already reported[7]. An efficient integration and modelling of genomic data, alongside other 'omics, will pave the way towards a personalised approach to treatment where clinical management will depend on medical history, genetic predisposition and objectively reported phenotypical traits (e.g. HPO terms).

The work presented in this thesis is and the GenePy algorithm in particular were developed using data from a cohort of paediatric IBD patients which represent a subset with a higher genetic component of the more broad IBD phenotype. Findings discussed following the supervised/unsupervised modelling of PIBD genomics data are therefore valid for this specific subset of paediatric patients and would require additional validation before generalising to the adult form of IBD. Moreover, the investigations herein reported were based on an homogeneous Caucasian group for which mutation frequencies may largely differ from the mean frequency of the admixed population. GenePy and subsequent approaches can be tailored to better model the available genomic data.

The collection of accurate longitudinal digital data from health care organisations will provide a solid base on which detailed genomic, proteomic, transcriptomic and

other types of biological "Big Data" will layer. At the same time, Artificial Intelligence approaches will have to evolve with a similar pace. Currently, many machine learning industries are investing in the analysis of health care and biological data with the promise of early detection and precise diagnosis.

We are still at the early stages of a healthcare revolution where the human expertise of clinicians is coupled to the extraordinary power of machine learning to solve the complexity behind many human conditions. Machine learning models will represent the main set of tools for enhancing diagnosis, drug discovery and clinical management.

# Supplementary Material

Table 7.1: **Monogenic IBD genes.**

| Genes | | |
|---|---|---|
| ADA | HPS4 | NCF2 |
| ADAM17 | HPS6 | NCF4 |
| AICDA | ICOS | PIK3R1 |
| BACH2 | IKBKG | PLCG2 |
| BTK | IL10 | RAG2 |
| CD40LG | IL10RA | RTEL1 |
| COL7A1 | IL10RB | SH2D1A |
| CYBA | IL21 | SKIV2L |
| CYBB | IL2RA | SLC37A4 |
| DCLRE1C | IL2RG | STAT1 |
| DKC1 | ITGB2 | STXBP2 |
| DOCK8 | LIG4 | TRIM22 |
| FERMT1 | LRBA | TTC37 |
| FOXP3 | MASP2 | TTC7A |
| G6PC3 | MEFV | WAS |
| GUCY2C | MVK | XIAP |
| HPS1 | NCF1 | ZAP70 |

Table 7.2: **All single nucleotide variants in the NOD2 gene used in GenePy validation.** Statistical significance was assessed through a Cochran-Armitage trend test using Plink v1.9 only for common variants (MAF >0.05). Significant associations smaller than $1x10^{-2}$ or smaller than $5x10^{-2}$ are highlighted by two (**) or one (*) asterisks respectively. p-values are not corrected for multiple testing.

| Chr | POS | Ref. allele | Alt. allele | MAF | Function† | Nucleotide change | Amino acid change | Controls vs IBD | Controls vs UC | Controls vs CD |
|---|---|---|---|---|---|---|---|---|---|---|
| chr16 | 50699512 | C | A | 1.18E-03 | NS | c.C17A | p.A6D | . | . | . |
| chr16 | 50699554 | C | T | 2.30E-03 | NS | c.C59T | p.S20L | . | . | . |
| chr16 | 50699710 | C | T | 1.15E-03 | NS | c.C215T | p.A72V | . | . | . |
| chr16 | 50699948 | C | G | 0.356 | SYN | c.C453G | p.S151S | 0.236 | 0.551 | 0.209 |
| chr16 | 50707880 | C | T | 3.66E-03 | NS | c.C485T | p.T162M | . | . | . |
| chr16 | 50710654 | T | G | 2.74E-03 | NS | c.T662G | p.L221R | . | . | . |
| chr16 | 50710713 | C | T | 0.316 | NS | c.C721T | p.P241S | 0.030* | 0.422 | 0.007** |
| chr16 | 50710777 | A | G | 4.70E-03 | NS | c.A785G | p.N262S | . | . | . |
| chr16 | 50710842 | C | T | 1.17E-03 | NS | c.C850T | p.R284W | . | . | . |
| chr16 | 50711028 | C | T | 1.17E-03 | NS | c.C1036T | p.R346C | . | . | . |
| chr16 | 50711101 | C | T | 1.17E-03 | NS | c.C1109T | p.P370L | . | . | . |
| chr16 | 50711203 | C | T | 4.68E-03 | NS | c.C1211T | p.S404L | . | . | . |
| chr16 | 50711204 | G | T | 1.17E-03 | SYN | c.G1212T | p.S404S | . | . | . |
| chr16 | 50711231 | C | T | 1.17E-03 | SYN | c.C1239T | p.T413T | . | . | . |
| chr16 | 50711288 | C | T | 0.316 | SYN | c.C1296T | p.R432R | 0.053 | 0.599 | 0.012* |
| chr16 | 50711492 | C | G | 1.17E-03 | SYN | c.C1500G | p.P500P | . | . | . |
| chr16 | 50711514 | C | T | 1.17E-03 | SYN | c.C1522T | p.L508L | . | . | . |
| chr16 | 50711600 | C | T | 1.17E-03 | SYN | c.C1608T | p.Y536Y | . | . | . |
| chr16 | 50711672 | T | G | 0.355 | SYN | c.T1680G | p.R560R | 0.300 | 0.799 | 0.169 |
| chr16 | 50711699 | G | A | 1.13E-03 | SYN | c.G1707A | p.T569T | . | . | . |
| chr16 | 50711744 | C | T | 0.018 | SYN | c.C1752T | p.A584A | . | . | . |
| chr16 | 50711811 | A | G | 1.13E-03 | NS | c.A1819G | p.R607G | . | . | . |
| chr16 | 50711867 | G | A | 1.13E-03 | SYN | c.G1875A | p.S625S | . | . | . |
| chr16 | 50712015 | C | T | 0.06 | NS | c.C2023T | p.R675W | 0.408 | . | 0.037* |
| chr16 | 50712018 | C | T | 9.05E-03 | NS | c.C2026T | p.R676C | . | . | . |
| chr16 | 50712034 | G | A | 1.13E-03 | NS | c.G2042A | p.R681H | . | . | . |
| chr16 | 50712049 | G | A | 4.53E-03 | NS | c.G2057A | p.R686H | . | . | . |
| chr16 | 50712058 | G | A | 1.13E-03 | NS | c.G2066A | p.R689H | . | . | . |
| chr16 | 50712085 | C | G | 1.13E-03 | NS | c.C2093G | p.A698G | . | . | . |
| chr16 | 50712141 | C | T | 1.13E-03 | NS | c.C2149T | p.R717W | . | . | . |
| chr16 | 50712175 | C | T | 4.53E-03 | NS | c.C2183T | p.A728V | . | . | . |
| chr16 | 50712243 | G | A | 1.13E-03 | NS | c.G2251A | p.E751K | . | . | . |
| chr16 | 50712288 | G | A | 4.53E-03 | NS | c.G2296A | p.V766M | . | . | . |
| chr16 | 50712317 | G | T | 2.26E-03 | SYN | c.G2325T | p.V775V | . | . | . |
| chr16 | 50716594 | G | A | 1.12E-03 | NS | c.G2389A | p.D797N | . | . | . |
| chr16 | 50716899 | A | G | 1.16E-03 | NS | c.A2474G | p.N825S | . | . | . |
| chr16 | 50722629 | G | C | 0.015 | NS | c.G2641C | p.G881R | . | . | . |
| chr16 | 50722660 | C | A | 1.13E-03 | NS | c.C2672A | p.A891D | . | . | . |
| chr16 | 50723365 | G | A | 0.085 | NS | c.G2782A | p.V928I | 0.079 | 0.159 | 0.123 |
| chr16 | 50723375 | A | G | 1.12E-03 | NS | c.A2792G | p.E931G | . | . | . |
| chr16 | 50725494 | A | G | 1.14E-03 | NS | c.A2807G | p.E936G | . | . | . |
| chr16 | 50729867 | G | GC | 0.039 | FSI | c.2936dupC | p.A979fs | . | . | . |

Table 7.3: **Genes driving the unsupervised clustering of IBD patients.** List of the genes that resulted significant in discriminating cluster V from clusters I to IV. Statistical significance was assessed using a Mann-Withney U test and p-values were corrected using Bonferroni correction for 898 tested genes (p-corr).

| Gene | U | p-value | p-corr |
|---|---|---|---|
| HLA-DQA1 | 3053.5 | $2.48 \cdot 10^{-30}$ | $1.50 \cdot 10^{-27}$ |
| HLA-DRB5 | 4757 | $1.28 \cdot 10^{-21}$ | $7.72 \cdot 10^{-19}$ |
| NOD1 | 9588 | 0.003 | 1.894 |
| MFN2 | 9522 | 0.004 | 2.594 |
| EPB41L2 | 9685 | 0.005 | 2.952 |
| MYH3 | 9685 | 0.005 | 2.952 |
| CSNK2A3 | 9656 | 0.005 | 3.235 |
| SOS2 | 9473 | 0.006 | 3.440 |
| CCL26 | 9724 | 0.009 | 5.546 |
| ARHGEF12 | 9676 | 0.011 | 6.636 |
| RAPGEF1 | 9682 | 0.012 | 7.183 |
| NLRP12 | 9792 | 0.016 | 9.572 |
| BMP1 | 9834 | 0.018 | 10.765 |
| MAVS | 9834 | 0.018 | 10.765 |
| TRPM7 | 9728.5 | 0.021 | 12.892 |
| ARHGEF11 | 9734 | 0.023 | 13.772 |
| SEMA4B | 9671 | 0.024 | 14.219 |
| PIK3R5 | 9674 | 0.024 | 14.659 |
| IFNA5 | 9860 | 0.028 | 16.668 |

| | | | |
|---|---|---|---|
| P2RX7 | 9860 | 0.028 | 16.668 |
| CTTN | 9860 | 0.028 | 16.668 |
| EPHA3 | 9860 | 0.028 | 16.668 |
| MAP3K1 | 9860 | 0.028 | 16.669 |
| ZAK | 9860 | 0.028 | 16.669 |
| MYL5 | 9908.5 | 0.035 | 20.983 |
| PIAS4 | 9908.5 | 0.035 | 20.984 |
| TANK | 9908.5 | 0.035 | 20.984 |
| ARHGEF4 | 9908.5 | 0.035 | 20.984 |
| BMP10 | 9908.5 | 0.035 | 20.984 |
| IL12RB2 | 9908.5 | 0.035 | 20.984 |
| IRF1 | 9908.5 | 0.035 | 20.984 |
| TGFBR2 | 9908.5 | 0.035 | 20.984 |
| NFATC4 | 9796 | 0.036 | 21.676 |
| CD28 | 9798.5 | 0.037 | 22.322 |
| NRG2 | 9800.5 | 0.038 | 22.852 |
| NFAT5 | 9801.5 | 0.038 | 23.120 |
| CTNNA2 | 9825.5 | 0.038 | 23.137 |
| IL20RA | 9827 | 0.039 | 23.577 |
| GNA13 | 9827.5 | 0.039 | 23.723 |
| TNFSF18 | 9928 | 0.049 | 29.478 |
| DNM1L | 9928 | 0.049 | 29.479 |
| FOS | 9928 | 0.049 | 29.479 |
| MYL2 | 9928 | 0.049 | 29.479 |
| BMPR1B | 9928 | 0.049 | 29.480 |
| CCR4 | 9928 | 0.049 | 29.480 |
| CD79A | 9928 | 0.049 | 29.480 |
| F11R | 9928 | 0.049 | 29.480 |
| GFAP | 9928 | 0.049 | 29.480 |
| IRAK4 | 9928 | 0.049 | 29.480 |
| MYH13 | 9928 | 0.049 | 29.480 |
| NUDT6 | 9928 | 0.049 | 29.480 |
| PIK3R1 | 9928 | 0.049 | 29.480 |
| PPP2R1B | 9928 | 0.049 | 29.480 |
| TRAF3 | 9928 | 0.049 | 29.480 |

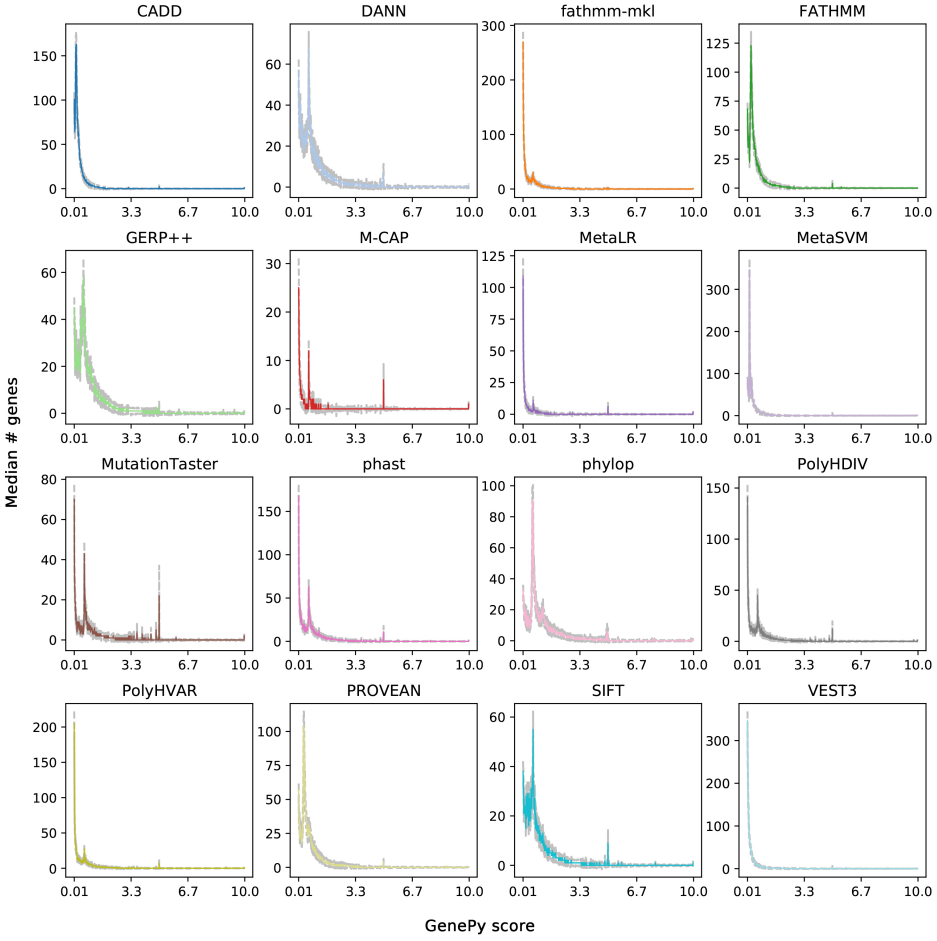


Figure 7.2: **Median whole gene GenePyuncorrected score profiles observed across the cohort of 508 patients with WES data depicted separately for each of the sixteen deleteriousness metrics.** For ease of comparison, x-axes are truncated at scores of 10. Bin size was set to 0.01 with the first bin shown 0.0.1-0.02. Grey dashed lines represent the standard deviation of each bin.

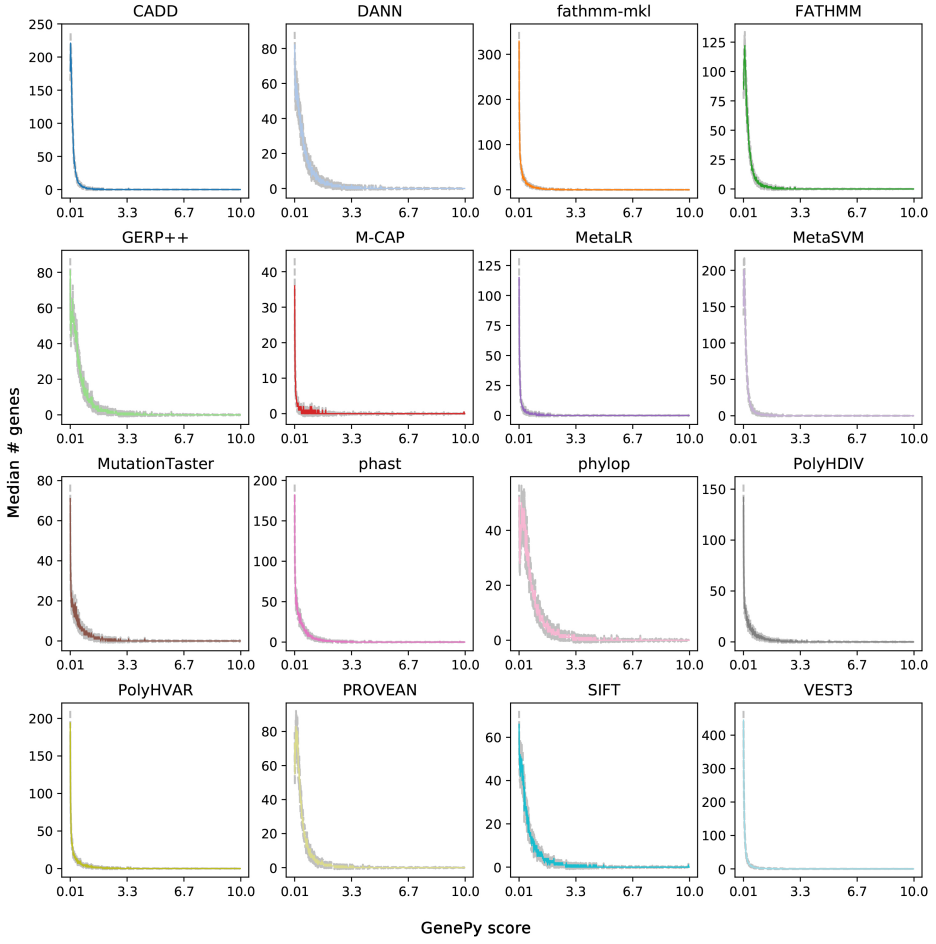


Figure 7.3: **Median whole gene GenePy$_{cgl}$ score profiles observed across the cohort of 508 patients with WES data depicted separately for each of the sixteen deleteriousness metrics.** For ease of comparison, x-axes are truncated at scores of 10. Bin size was set to 0.01 with the first bin shown 0.0.1-0.02. Grey dashed lines represent the standard deviation of each bin.

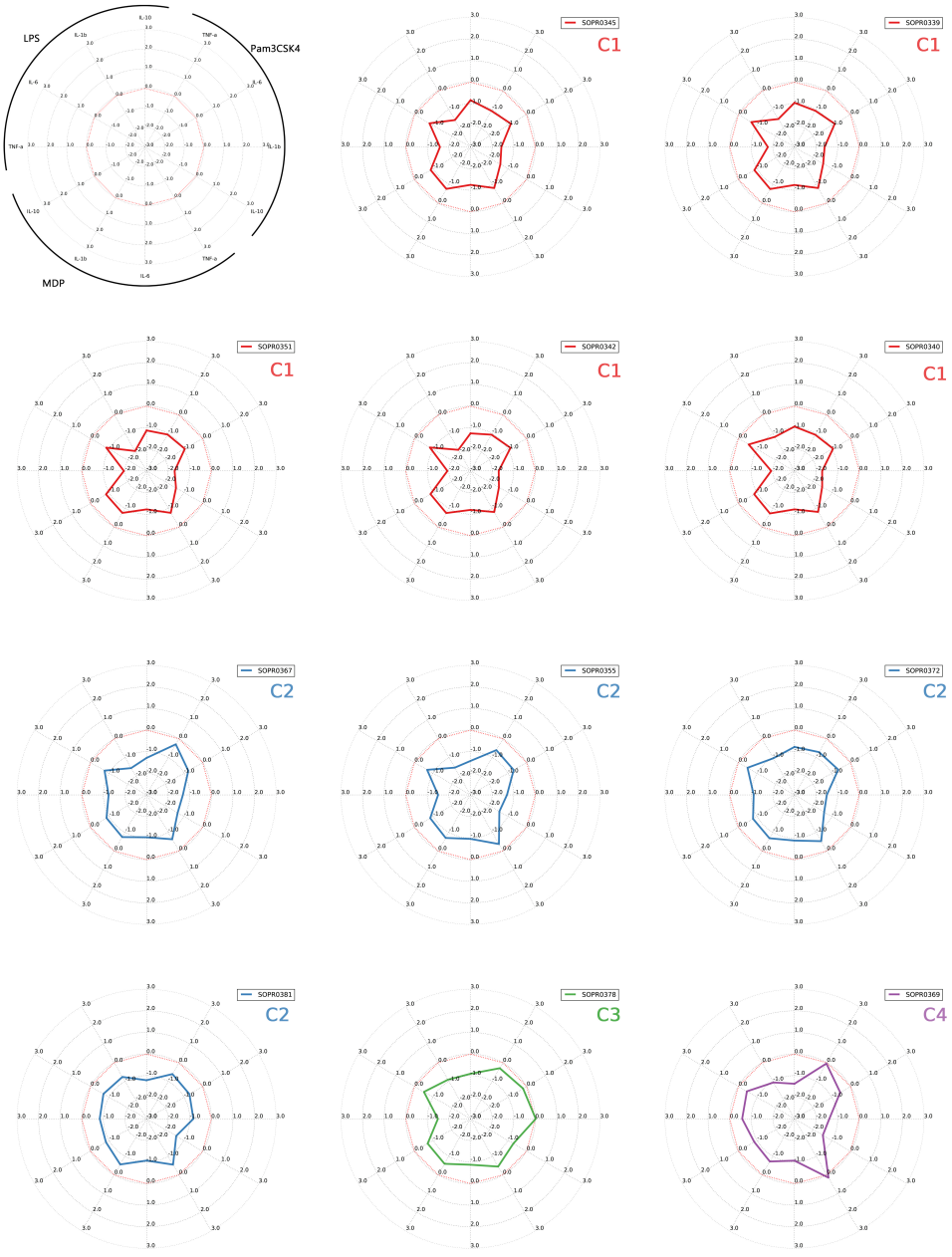


Figure 7.4: **Individual radar plots of cytokine responses per patient.** Each spoke of the radar represents a ligand-cytokine combination with the red dashed line indicating the mean response of the control cohort. Each plot is coloured according to the eight different immuno-types identified by hierarchical clustering.

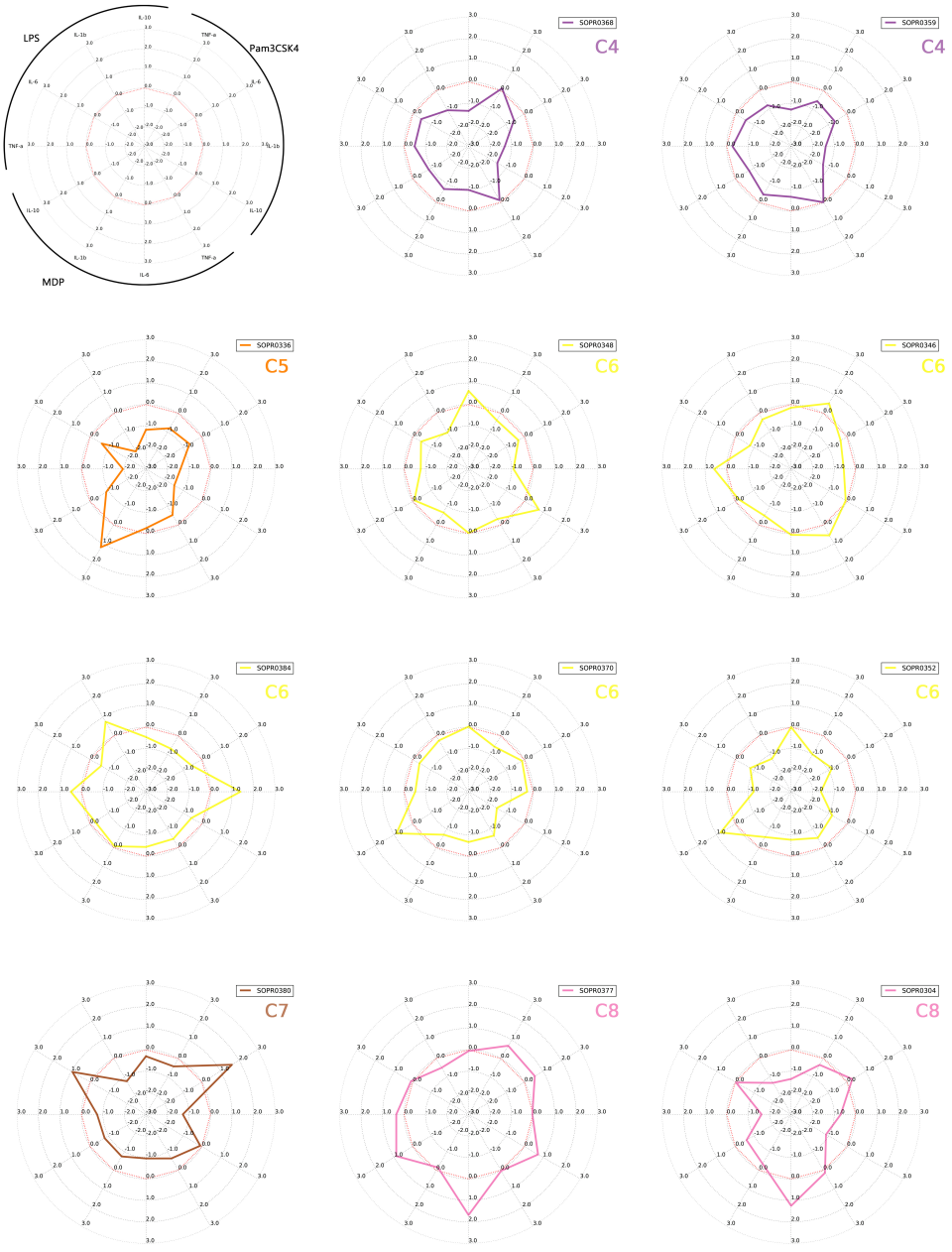


Figure 7.5: **Individual radar plots of cytokine responses per patient (2).** Each spoke of the radar represents a ligand-cytokine combination with the red dashed line indicating the mean response of the control cohort. Each plot is coloured according to the eight different immuno-types identified by hierarchical clustering.

# Bibliography


[1] Goncalo R Abecasis et al. "An integrated map of genetic variation from 1,092 human genomes." In: *Nature* 491.7422 (2012), pp. 56–65.

[2] Marit Ackermann et al. "Teamwork: Improved eQTL Mapping Using Combinations of Machine Learning Methods". In: *PLoS ONE* 7.7 (2012). Ed. by Avi Ma'ayan, e40916.

[3] Ivan A Adzhubei et al. "A method and server for predicting damaging missense mutations." In: *Nature methods* 7.4 (2010), pp. 248–9.

[4] Subramanian S Ajay et al. "Accurate and comprehensive sequencing of personal genomes." In: *Genome research* 21.9 (2011), pp. 1498–505.

[5] Carl A Anderson et al. "Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47". In: *Nature Genetics* 43.3 (2011), pp. 246–252.

[6] Gaia Andreoletti et al. "Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway". In: *Scientific Reports* 7 (2017), p. 46454.

[7] Ingrid Arijs and Isabelle Cleynen. "RISK stratification in paediatric Crohn's disease." In: *Lancet (London, England)* 389.10080 (2017), pp. 1672–1674.

[8] J. J. Ashton et al. "Rising incidence of paediatric inflammatory bowel disease (PIBD) in Wessex, Southern England". In: *Archives of Disease in Childhood* 99.7 (2014), pp. 659–664.

[9] James J. Ashton et al. "Endoscopic Versus Histological Disease Extent at Presentation of Paediatric Inflammatory Bowel Disease". In: *Journal of Pediatric Gastroenterology and Nutrition* 62.2 (2016), pp. 246–251.

[10] I. Atreya, R. Atreya, and M. F. Neurath. "NF-$\kappa$B in inflammatory bowel disease". In: *Journal of Internal Medicine* 263.6 (2008), pp. 591–596.

[11] Raja Atreya et al. "Antibodies against tumor necrosis factor (TNF) induce T-cell apoptosis in patients with inflammatory bowel diseases via TNF receptor 2 and intestinal CD14+ macrophages." In: *Gastroenterology* 141.6 (2011), pp. 2026–38.

[12] Amiirah Aujnarain, David R Mack, and Eric I Benchimol. "The role of the environment in the development of pediatric inflammatory bowel disease." In: *Current gastroenterology reports* 15.6 (2013), p. 326.

[13] Monya Baker. "Next-generation sequencing: adjusting to data overload". In: *Nature Methods* 7.7 (2010), pp. 495–499.

[14] S Bamford et al. "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website." In: *British journal of cancer* 91.2 (2004), pp. 355–8.

[15] Michael J Bamshad et al. "Exome sequencing as a tool for Mendelian disease gene discovery." In: *Nature reviews. Genetics* 12.11 (2011), pp. 745–55.

[16] Stavros Bashiardes et al. "Direct genomic selection." In: *Nature methods* 2.1 (2005), pp. 63–9.

[17] Daniel C Baumgart and William J Sandborn. "Inflammatory bowel disease: clinical aspects and established and evolving therapies". In: *The Lancet* 369.9573 (2007), pp. 1641–1657.

[18] Mélissa Beaudoin et al. "Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis." In: *PLoS genetics* 9.9 (2013), e1003723.

[19] Frida Belinky et al. "PathCards: multi-source consolidation of human biological pathways." In: *Database : the journal of biological databases and curation* 2015 (2015).

[20] Richard Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton Univ. Press, Princeton, N. J, 1961.

[21] Bradley E Bernstein et al. “An integrated encyclopedia of DNA elements in the human genome.” In: *Nature* 489.7414 (2012), pp. 57–74.

[22] Charlotte I de Bie et al. “Disease phenotype at diagnosis in pediatric Crohn’s disease: 5-year analyses of the EUROKIDS Registry.” In: *Inflammatory bowel diseases* 19.2 (2013), pp. 378–85.

[23] Jeannette Bigler et al. “A Severe Asthma Disease Signature from Gene Expression Profiling of Peripheral Blood from U-BIOPRED Cohorts”. In: *American Journal of Respiratory and Critical Care Medicine* 195.10 (2017), pp. 1311–1320.

[24] Eckart Bindewald and Bruce A Shapiro. “RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers.” In: *RNA (New York, N.Y.)* 12.3 (2006), pp. 342–52.

[25] Denise K Bonen and Judy H Cho. “The genetics of inflammatory bowel disease.” In: *Gastroenterology* 124.2 (2003), pp. 521–36.

[26] Joseph P Boyle et al. “Insights into the molecular basis of the NOD2 signalling pathway.” In: *Open biology* 4.12 (2014), pp. 12955–12958.

[27] Jesper Bertram Bramsen et al. “Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer”. In: *Cell Reports* 19.6 (2017), pp. 1268–1280.

[28] Steven E Brenner. *Critical Assessment of Genome Interpretation*. 2013.

[29] Sharon R Browning and Brian L Browning. “Haplotype phasing: existing methods and new developments.” In: *Nature reviews. Genetics* 12.10 (2011), pp. 703–14.

[30] Mariusz Butkiewicz and William S Bush. “In Silico Functional Annotation of Genomic Variation.” In: *Current protocols in human genetics* 88 (2016), Unit 6.15.

[31] Emidio Capriotti and Russ B Altman. “A new disease-specific machine learning approach for the prediction of cancer-causing missense variants.” In: *Genomics* 98.4 (2011), pp. 310–7.

[32] Hannah Carter et al. “Identifying Mendelian disease genes with the variant effect scoring tool.” En. In: *BMC genomics* 14 Suppl 3.3 (2013), S3.

[33] Robert Castelo and Roderic Guigó. “Splice site identification by idlBNs.” In: *Bioinformatics (Oxford, England)* 20 Suppl 1 (2004), pp. i69–76.

[34] Siow-Wee Chang et al. “Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods.” In: *BMC bioinformatics* 14.1 (2013), p. 170.

[35] Edward Y Chen et al. “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool”. In: *BMC Bioinformatics* 14.1 (2013), p. 128.

[36] Chandra Sekhar Reddy Chilamakuri et al. “Performance comparison of four exome capture systems for deep sequencing.” In: *BMC genomics* 15 (2014), p. 449.

[37] Yongwook Choi et al. “Predicting the functional effect of amino acid substitutions and indels.” In: *PloS one* 7.10 (2012), e46688.

[38] S. Chun and J. C. Fay. “Identification of deleterious mutations within three human genomes”. In: *Genome Research* 19.9 (2009), pp. 1553–1561.

[39] Kristian Cibulskis et al. “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.” In: *Nature biotechnology* 31.3 (2013), pp. 213–9.

[40] Peter J A Cock et al. “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.” In: *Nucleic acids research* 38.6 (2010), pp. 1767–71.

[41] Gregory M Cooper et al. “Distribution and intensity of constraint in mammalian genomic sequence.” In: *Genome research* 15.7 (2005), pp. 901–13.

[42] Giulia D’Arcangelo and Marina Aloi. “Inflammatory Bowel Disease-Unclassified in Children: Diagnosis and Pharmacological Management”. In: *Pediatric Drugs* 19.2 (2017), pp. 113–120.

[43] Eugene V Davydov et al. “Identifying a high fraction of the human genome to be under selective constraint using GERP++.” In: *PLoS computational biology* 6.12 (2010), e1001025.

[44] Mark A DePristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data.” In: *Nature genetics* 43.5 (2011), pp. 491–8.

[45] MatLab Documentation. *Matlab documentation*. 2012.

[46] C. Dong et al. “Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies”. In: *Human Molecular Genetics* 24.8 (2015), pp. 2125–2137.

[47] Jack J Dongarra et al. “TOP500 supercomputer sites”. In: *Supercomputer* 13 (1997), pp. 89–111.

[48] Alexander Edwards et al. “Corticosteroids and infliximab impair the performance of interferon-$\gamma$ release assays used for diagnosis of latent tuberculosis”. In: *Thorax* 72.10 (2017), pp. 946–949.

[49] Evan E. Eichler et al. “Missing heritability and strategies for finding the underlying causes of complex disease”. In: *Nature Reviews Genetics* 11.6 (2010), pp. 446–450.

[50] Charles Epstein, Gunnar Carlsson, and Herbert Edelsbrunner. “Topological data analysis”. In: *Inverse Problems* 27.12 (2011), p. 120201.

[51] Melissa A. Fernandes et al. “Addition of Histology to the Paris Classification of Pediatric Crohn Disease Alters Classification of Disease Location”. In: *Journal of Pediatric Gastroenterology and Nutrition* 62.2 (2016), pp. 242–245.

[52] Paul Flicek, MR Amode, and Daniel Barrell. “Ensembl 2012”. In: *Nucleic acids ...* (2012).

[53] D G Forcione et al. “An increased risk of Crohn’s disease in individuals who inherit the HLA class II DRB3*0301 allele”. In: *Proc Natl Acad Sci U S A* (1996).

[54] Luigi Franchi et al. “Function of Nod-like receptors in microbial recognition and host defense.” In: *Immunological reviews* 227.1 (2009), pp. 106–28.

[55] Manuel Garber et al. “Identifying novel constrained elements by exploiting biased substitution patterns.” In: *Bioinformatics (Oxford, England)* 25.12 (2009), pp. i54–62.

[56] C Gasche et al. “A simple classification of Crohn’s disease: report of the Working Party for the World Congresses of Gastroenterology, Vienna 1998.” In: *Inflammatory bowel diseases* 6.1 (2000), pp. 8–15.

[57] Richard A. Gibbs et al. “The International HapMap Project.” In: *Nature* 426.6968 (2003), pp. 789–96.

[58] Walter Gilbert. “Why genes in pieces?” In: *Nature* 271.5645 (1978), pp. 501–501.

[59] Christian Gilissen et al. “Disease gene identification strategies for exome sequencing”. In: *European Journal of Human Genetics* 20.5 (2012), pp. 490–497.

[60] Alan E Guttmacher and Francis S Collins. “Genomic medicine–a primer.” In: *The New England journal of medicine* 347.19 (2002), pp. 1512–1520.

[61] Isabelle Guyon et al. “Gene Selection for Cancer Classification using Support Vector Machines”. In: *Machine Learning* 46.1/3 (2002), pp. 389–422.

[62] R. W. Hamming. “Error Detecting and Error Correcting Codes”. In: *Bell System Technical Journal* 29.2 (1950), pp. 147–160.

[63] Ada Hamosh et al. “Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.” In: *Nucleic acids research* 33.Database issue (2005), pp. D514–7.

[64] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. “The Elements of Statistical Learning”. In: *Elements* 1 (2009), pp. 337–387. arXiv: 1010.3003.

[65] Matthew S Hayden and Sankar Ghosh. “Shared principles in NF-kappaB signaling.” In: *Cell* 132.3 (2008), pp. 344–62.

[66] Matija Hedl, Deborah D. Proctor, and Clara Abraham. "JAK2 Disease-Risk Variants Are Gain of Function and JAK Signaling Threshold Determines Innate Receptor-Induced Proinflammatory Cytokine Secretion in Macrophages". In: *The Journal of Immunology* 197.9 (2016), pp. 3695–3704.

[67] Paul Henderson et al. "Rising incidence of pediatric inflammatory bowel disease in Scotland". In: *Inflammatory Bowel Diseases* 18.6 (2012), pp. 999–1005.

[68] B Hope et al. "Rapid rise in incidence of Irish paediatric inflammatory bowel disease." In: *Archives of disease in childhood* 97.7 (2012), pp. 590–4.

[69] Shurong Hu et al. "mTOR Inhibition Attenuates Dextran Sulfate Sodium-Induced Colitis by Suppressing T Cell Proliferation and Balancing TH1/TH17/Treg Profile". In: *PLOS ONE* 11.4 (2016). Ed. by Hossam M Ashour, e0154564.

[70] Jean-Pierre Hugot et al. "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease". In: *Nature* 411.6837 (2001), pp. 599–603.

[71] Human Genome Sequencing ConsortiumInternational. "Finishing the euchromatic sequence of the human genome." In: *Nature* 431.7011 (2004), pp. 931–45.

[72] David J Hunter. "Gene-environment interactions in human diseases." In: *Nature reviews. Genetics* 6.4 (2005), pp. 287–98.

[73] N Inohara et al. "Human Nod1 confers responsiveness to bacterial lipopolysaccharides." In: *The Journal of biological chemistry* 276.4 (2001), pp. 2551–4.

[74] Yuval Itan et al. "The human gene damage index as a gene-level approach to prioritizing exome variants". In: *Proceedings of the National Academy of Sciences* 112.44 (2015), pp. 13615–13620.

[75] Karthik A Jagadeesh et al. "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity". In: *Nature Genetics* 48.12 (2016), pp. 1581–1586.

[76] G James et al. *An introduction to statistical learning: with applications in R*. Vol. XIV. 2013, p. 426.

[77] Saumya Shekhar Jamuar and Ene-Choo Tan. "Clinical application of next-generation sequencing for Mendelian diseases". In: *Human Genomics* 9.1 (2015), p. 10.

[78] Jyh Shing Roger Jang. "ANFIS: adaptive-network-based fuzzy inference system". In: *IEEE Transactions on Systems, Man and Cybernetics* 23 (1993), pp. 665–685.

[79] Luke Jostins et al. "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease." In: *Nature* 491.7422 (2012), pp. 119–24.

[80] Goo Jun et al. "Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data". In: *The American Journal of Human Genetics* 91.5 (2012), pp. 839–848.

[81] Latha Kadalayil et al. "Exome sequence read depth methods for identifying copy number changes." In: *Briefings in bioinformatics* (2014), bbu027–.

[82] Nobuhiko Kamada et al. "Unique CD14 intestinal macrophages contribute to the pathogenesis of Crohn disease via IL-23/IFN-gamma axis." In: *The Journal of clinical investigation* 118.6 (2008), pp. 2269–80.

[83] Shigeo Kanazawa et al. "VEGF, basic-FGF, and TGF-beta in Crohn's disease and ulcerative colitis: a novel mechanism of chronic intestinal inflammation". In: *The American Journal of Gastroenterology* 96.3 (2001), pp. 822–828.

[84] Minoru Kanehisa et al. "KEGG as a reference resource for gene and protein annotation." In: *Nucleic acids research* 44.D1 (2016), pp. D457–62.

[85] Gilaad G Kaplan. "The global burden of IBD: from 2015 to 2025." In: *Nature reviews. Gastroenterology & hepatology* 12.12 (2015), pp. 720–727.

[86] Bernard Khor, Agnès Gardet, and Ramnik J Xavier. "Genetics and pathogenesis of inflammatory bowel disease." In: *Nature* 474.7351 (2011), pp. 307–17.

[87] Min-su Kim et al. “RDDpred: a condition-specific RNA-editing prediction model from RNA-seq data”. In: *BMC Genomics* 17.S1 (2016), p. 5.

[88] Martin Kircher et al. “A general framework for estimating the relative pathogenicity of human genetic variants.” In: *Nature genetics* 46.3 (2014), pp. 310–5.

[89] P Klein. “Prediction of protein structural class by discriminant analysis.” In: *Biochimica et biophysica acta* 874.2 (1986), pp. 205–15.

[90] Daniel Kotlarz et al. “Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy.” In: *Gastroenterology* 143.2 (2012), pp. 347–55.

[91] Hugo Y K Lam et al. “Performance comparison of whole-genome sequencing platforms”. In: *Nature Biotechnology* 30.1 (2011), pp. 78–82.

[92] E S Lander and N J Schork. “Genetic dissection of complex traits.” In: *Science (New York, N.Y.)* 265.5181 (1994), pp. 2037–48.

[93] Katrina M de Lange et al. “Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease.” In: *Nature genetics* 49.2 (2017), pp. 256–261.

[94] Michael W. Leach et al. “The Role of IL-10 in Inflammatory Bowel Disease: ”Of Mice and Men””. In: *Toxicologic Pathology* 27.1 (1999), pp. 123–133.

[95] Seunggeun Lee et al. “Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.” In: *American journal of human genetics* 91.2 (2012), pp. 224–37.

[96] Monkol Lek et al. “Analysis of protein-coding genetic variation in 60,706 humans”. In: *Nature* 536.7616 (2016), pp. 285–291.

[97] Arie Levine et al. “ESPGHAN revised porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents.” In: *Journal of pediatric gastroenterology and nutrition* 58.6 (2014), pp. 795–806.

[98] Arie Levine et al. “Pediatric modification of the Montreal classification for inflammatory bowel disease”. In: *Inflammatory Bowel Diseases* 17.6 (2011), pp. 1314–1321.

[99] Arie Levine et al. “Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification.” In: *Inflammatory bowel diseases* 17.6 (2011), pp. 1314–21.

[100] Arie Levine et al. “The ESPGHAN Revised Porto Criteria for the Diagnosis of Inflammatory Bowel Disease in Children and Adolescents”. In: *Journal of Pediatric Gastroenterology and Nutrition* (2013), p. 1.

[101] Bingshan Li and Suzanne M. Leal. “Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data”. In: *The American Journal of Human Genetics* 83.3 (2008), pp. 311–321.

[102] Heng Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: (2013), p. 3. arXiv: 1303.3997.

[103] Heng Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: *Genomics* (2013). arXiv: 1303.3997.

[104] Heng Li and Nils Homer. “A survey of sequence alignment algorithms for next-generation sequencing.” In: *Briefings in bioinformatics* 11.5 (2010), pp. 473–83.

[105] Jiao Li et al. “A survey of current trends in computational drug repositioning.” In: *Briefings in bioinformatics* 17.1 (2016), pp. 2–12.

[106] Jinchen Li et al. “VarCards: an integrated genetic and clinical database for coding variants in the human genome.” In: *Nucleic acids research* 46.D1 (2018), pp. D1039–D1048.

[107] Jun Z Li et al. “Worldwide human relationships inferred from genome-wide patterns of variation.” In: *Science (New York, N.Y.)* 319.5866 (2008), pp. 1100–4.

[108] Yun R Li et al. “Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases.” In: *Nature medicine* 21.9 (2015), pp. 1018–27.

[109] Angélica Nakagawa Lima et al. “Use of machine learning approaches for novel drug discovery”. In: *Expert Opinion on Drug Discovery* 11.3 (2016), pp. 225–239.

[110] Jimmy Z Liu et al. “Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.” In: *Nature genetics* 47.9 (2015), pp. 979–86.

[111] Kirk E. Lohmueller et al. “Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease”. In: *Nature Genetics* 33.2 (2003), pp. 177–182.

[112] Mark Lutz. *Learning Python*. Vol. 78. 1. O’Reilly, 2007, p. 700. arXiv: `1011.1669v3`.

[113] Laurens Van Der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[114] Michael Maes, Frank N.M. Twisk, and Cort Johnson. “Myalgic Encephalomyelitis (ME), Chronic Fatigue Syndrome (CFS), and Chronic Fatigue (CF) are distinguished accurately: Results of supervised learning techniques applied on clinical and inflammatory data”. In: *Psychiatry Research* 200.2-3 (2012), pp. 754–760.

[115] Batool Mutar Mahdi. “Role of HLA typing on Crohn’s disease pathogenesis.” In: *Annals of medicine and surgery (2012)* 4.3 (2015), pp. 248–53.

[116] Khalid Mahmood et al. “Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics.” In: *Human genomics* 11.1 (2017), p. 10.

[117] Lira Mamanova et al. “Target-enrichment strategies for next-generation sequencing.” In: *Nature methods* 7.2 (2010), pp. 111–8.

[118] Catherine Mathé et al. “Current methods of gene prediction, their strengths and weaknesses.” In: *Nucleic acids research* 30.19 (2002), pp. 4103–17.

[119] Julien Matricon, Nicolas Barnich, and Denis Ardid. “Immunopathogenesis of inflammatory bowel disease.” In: *Self/nonself* 1.4 (2010), pp. 299–309.

[120] John Mccarthy et al. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”. In: *AI Magazine* 27 (2006), pp. 12–14.

[121] Dermot P B McGovern et al. “Genome-wide association identifies multiple ulcerative colitis susceptibility loci”. In: *Nature Genetics* 42.4 (2010), pp. 332–337.

[122] Aaron McKenna et al. “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.” In: *Genome research* 20.9 (2010), pp. 1297–303.

[123] Gil Y Melmed and Stephan R Targan. “Future biologic targets for IBD: potentials and pitfalls.” In: *Nature reviews. Gastroenterology & hepatology* 7.2 (2010), pp. 110–7.

[124] S A Miller, D D Dykes, and H F Polesky. “A simple salting out procedure for extracting DNA from human nucleated cells.” In: *Nucleic acids research* 16.3 (1988), p. 1215.

[125] Matthew Mort et al. “MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing.” In: *Genome biology* 15.1 (2014), R19.

[126] E Mossotto et al. “Classification of Paediatric Inflammatory Bowel Disease using Machine Learning”. In: *Scientific Reports* 7.1 (2017), p. 2427.

[127] Bjorn Moum et al. “Change in the extent of colonoscopic and histological involvement in ulcerative colitis over time”. In: *The American Journal of Gastroenterology* 94.6 (1999), pp. 1564–1569.

[128] Allan M. Mowat and William W. Agace. “Regional specialization within the intestinal immune system”. In: *Nature Reviews Immunology* 14.10 (2014), pp. 667–685.

[129] Tim W. Nattkemper et al. “Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods”. In: *Artificial Intelligence in Medicine* 34.2 (2005), pp. 129–139.

[130] Nature. “E pluribus unum”. In: *Nature Methods* 7.5 (2010), pp. 331–331.

[131] Benjamin M. Neale et al. “Testing for an Unusual Distribution of Rare Variants”. In: *PLoS Genetics* 7.3 (2011). Ed. by Suzanne M. Leal, e1001322.

[132] S B Needleman and C D Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” In: *Journal of molecular biology* 48.3 (1970), pp. 443–53.

[133] Markus F. Neurath. “Cytokines in inflammatory bowel disease”. In: *Nature Reviews Immunology* 14.5 (2014), pp. 329–342.

[134] AY Ng and MI Jordan. “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes”. In: *Advances in neural information processing* 14 (2002), pp. 841–848.

[135] NHS. *NHS standard contract for colorectal: complex inflammatory bowel disease (adult)*. Tech. rep. 2013.

[136] John Novembre et al. “Genes mirror geography within Europe”. In: *Nature* 456.7218 (2008), pp. 98–101.

[137] Andrea Oeckinghaus, Matthew S Hayden, and Sankar Ghosh. “Crosstalk in NF-$\kappa$B signaling pathways”. In: *Nature Immunology* 12.8 (2011), pp. 695–708.

[138] Y Ogura et al. “A frameshift mutation in NOD2 associated with susceptibility to Crohn’s disease.” In: *Nature* 411.6837 (2001), pp. 603–6.

[139] David T Okou and Subra Kugathasan. “Role of genetics in pediatric inflammatory bowel disease.” In: *Inflammatory bowel diseases* 20.10 (2014), pp. 1878–84.

[140] Travis E Oliphant. “SciPy: Open source scientific tools for Python”. In: *Computing in Science and Engineering* 9 (2007), pp. 10–20.

[141] Thomas Oommen et al. “An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing”. In: *Mathematical Geosciences* 40.4 (2008), pp. 409–424.

[142] Alexander Panda et al. “Statistical approaches for analyzing immunologic data of repeated observations: a practical guide.” In: *Journal of immunological methods* 398-399 (2013), pp. 19–26.

[143] Brent S Pedersen and Aaron R Quinlan. “Who’s Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy.” In: *American journal of human genetics* 100.3 (2017), pp. 406–413.

[144] Fabian Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2012), pp. 2825–2830.

[145] Reuben J. Pengelly et al. “Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation”. In: *Briefings in Bioinformatics* (2017).

[146] Britt-Sabina Petersen et al. “Opportunities and challenges of whole-genome and -exome sequencing.” In: *BMC genetics* 18.1 (2017), p. 14.

[147] *Picard Tools - By Broad Institute*.

[148] Scott Plevy et al. “Combined serological, genetic, and inflammatory markers differentiate non-IBD, Crohn’s disease, and ulcerative colitis patients.” In: *Inflammatory bowel diseases* 19.6 (2013), pp. 1139–48.

[149] Daniel K. Podolsky. “Inflammatory Bowel Disease”. In: *New England Journal of Medicine* 325.13 (1991), pp. 928–937.

[150] Katherine S Pollard et al. “Detection of nonneutral substitution rates on mammalian phylogenies.” In: *Genome research* 20.1 (2010), pp. 110–21.

[151] Daniel Quang, Yifei Chen, and Xiaohui Xie. “DANN: a deep learning approach for annotating the pathogenicity of genetic variants”. In: *Bioinformatics* 31.5 (2015), pp. 761–763.

[152] A. R. Quinlan and I. M. Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. In: *Bioinformatics* 26.6 (2010), pp. 841–842.

[153] R&D Systems. *NOD-like Receptor Signaling Interactive Pathway: R&D Systems*.

[154] Boris Reva, Yevgeniy Antipin, and Chris Sander. "Predicting the functional impact of protein mutations: application to cancer genomics." In: *Nucleic acids research* 39.17 (2011), e118.

[155] Marnie E. Rice and Grant T. Harris. "Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r." In: *Law and Human Behavior* 29.5 (2005), pp. 615–620.

[156] Graham R S Ritchie et al. "Functional annotation of noncoding sequence variants". In: *Nature Methods* 11.3 (2014), pp. 294–296.

[157] Manuel A Rivas et al. "Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease." In: *Nature genetics* 43.11 (2011), pp. 1066–73.

[158] Laurie A Robak et al. "Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease". In: *Brain* 140.12 (2017), pp. 3191–3203.

[159] Sara Ruiz-Pinto et al. "Exome array analysis identifies GPR35 as a novel susceptibility gene for anthracycline-induced cardiotoxicity in childhood cancer". In: *Pharmacogenetics and Genomics* 27.12 (2017), pp. 445–453.

[160] S Salzberg. "Locating protein coding regions in human DNA using a decision tree algorithm." In: *Journal of computational biology : a journal of computational molecular cell biology* 2.3 (1995), pp. 473–85.

[161] F Sanger, S Nicklen, and A R Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74 (1977), pp. 5463–5467.

[162] E A Sankey et al. "Early mucosal changes in Crohn's disease." In: *Gut* 34.3 (1993), pp. 375–81.

[163] J Satsangi et al. "The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications." In: *Gut* 55.6 (2006), pp. 749–53.

[164] Nicholas J. Schork. "Personalized medicine: Time for one-person trials". In: *Nature* 520.7549 (2015), pp. 609–611.

[165] Jana Marie Schwarz et al. “MutationTaster evaluates disease-causing potential of sequence alterations.” In: *Nature methods* 7.8 (2010), pp. 575–6.

[166] Jana Marie Schwarz et al. “MutationTaster2: mutation prediction for the deep-sequencing age”. In: *Nature Methods* 11.4 (2014), pp. 361–362.

[167] Tony Shen et al. “The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders”. In: *Genetics Research* 97 (2015), e15.

[168] David Q Shih and Stephan R Targan. “Immunopathogenesis of inflammatory bowel disease.” In: *World journal of gastroenterology* 14.3 (2008), pp. 390–400.

[169] Hashem A. Shihab et al. “An integrative approach to predicting the functional effects of non-coding and coding sequence variation”. In: *Bioinformatics* 31.10 (2015), pp. 1536–1543.

[170] Hashem A Shihab et al. “Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models.” In: *Human mutation* 34.1 (2013), pp. 57–65.

[171] Margaret A Shipp et al. “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.” In: *Nature medicine* 8.1 (2002), pp. 68–74.

[172] Adam Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.” In: *Genome research* 15.8 (2005), pp. 1034–50.

[173] Ngak-Leng Sim et al. “SIFT web server: predicting effects of amino acid substitutions on proteins.” In: *Nucleic acids research* 40.Web Server issue (2012), W452–7.

[174] T.F. Smith and M.S. Waterman. “Identification of common molecular subsequences”. In: *Journal of Molecular Biology* 147.1 (1981), pp. 195–197.

[175] R.R. Sokal and C.D. Michener. “A statistical method for evaluating systematic relationships”. In: *The University of Kansas Science Bulletin* 38 (1958), pp. 1409–1437.

[176] Ray L. Somorjai, B. Dolenko, and R. Baumgartner. “Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions”. In: *Bioinformatics* 19.12 (2003), pp. 1484–1491.

[177] Xinyang Song et al. “Growth Factor FGF2 Cooperates with Interleukin-17 to Repair Intestinal Epithelial Damage”. In: *Immunity* 43.3 (2015), pp. 488–501.

[178] Lincoln D Stein. “The case for cloud computing in genome informatics.” In: *Genome biology* 11.5 (2010), p. 207.

[179] Peter D Stenson et al. “The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.” In: *Human genetics* 133.1 (2014), pp. 1–9.

[180] Stephen M. Stigler. “Gauss and the Invention of Least Squares”. EN. In: *The Annals of Statistics* 9.3 (1981), pp. 465–474.

[181] Warren Strober, Ivan J. Fuss, and Richard S. Blumberg. “The immunology of mucosal models of inflammation”. In: *Annual Review of Immunology* 20.1 (2002), pp. 495–549.

[182] Samuel P Strom. “Current practices and guidelines for clinical next-generation sequencing oncology testing.” In: *Cancer biology & medicine* 13.1 (2016), pp. 3–11.

[183] Damian Szklarczyk et al. “The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D362–D368.

[184] Shinichi Takahashi et al. “De novo and rare mutations in the HSPA1L heat shock gene associated with inflammatory bowel disease”. In: *Genome Medicine* 9.1 (2017), p. 8.

[185] Lun Tan et al. “FBN1 mutations largely contribute to sporadic non-syndromic aortic dissection”. In: *Human Molecular Genetics* 26.24 (2017), pp. 4814–4822.

[186] Haiming Tang and Paul D Thomas. “Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation.” In: *Genetics* 203.2 (2016), pp. 635–47.

[187] Robert Tibshirani. *Regression Shrinkage and Selection Via the Lasso.*

[188] E A Trachtenberg et al. “HLA class II haplotype associations with inflammatory bowel disease in Jewish (Ashkenazi) and non-Jewish caucasian populations.” In: *Human immunology* 61.3 (2000), pp. 326–33.

[189] Daniel Trujillano et al. “Clinical exome sequencing: results from 2819 samples reflecting 1000 families”. In: *European Journal of Human Genetics* 25.2 (2017), pp. 176–182.

[190] J V Tu. “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.” In: *Journal of clinical epidemiology* 49.11 (1996), pp. 1225–31.

[191] Dan Turner. “Microscopic Assessment in Inflammatory Bowel Disease”. In: *Journal of Pediatric Gastroenterology and Nutrition* 62.2 (2016), p. 191.

[192] Holm H Uhlig. “Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease”. In: *Gut* 62.12 (2013), pp. 1795–1805.

[193] Rosanna Upstill-Goddard et al. “Support Vector Machine classifier for estrogen receptor positive and negative early-onset breast cancer.” In: *PloS one* 8.7 (2013), e68606.

[194] Johan Van Limbergen et al. “Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease.” In: *Gastroenterology* 135.4 (2008), pp. 1114–22.

[195] Heming Wang et al. “Variants in angiopoietin-2 ( ANGPT2 ) contribute to variation in nocturnal oxyhaemoglobin saturation level”. In: *Human Molecular Genetics* 25.23 (2016), ddw324.

[196] K. Wang, M. Li, and H. Hakonarson. “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data”. In: *Nucleic Acids Research* 38.16 (2010), e164–e164.

[197] Zhi Wei et al. “Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease”. In: *American Journal of Human Genetics* 92 (2013), pp. 1008–1012.

[198] Matthew Weiser et al. “Molecular classification of Crohn’s disease reveals two clinically relevant subtypes.” In: *Gut* (2016).

[199] A. Wilson et al. “HLA-DQA1-HLA-DRB1 polymorphism is a major predictor of azathioprine-induced pancreatitis in patients with inflammatory bowel disease”. In: *Alimentary Pharmacology & Therapeutics* 47.5 (2018), pp. 615–620.

[200] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. 2nd. Morgan Kaufmann Publishers Inc, 2005.

[201] Sunny H Wong et al. “Effect of immunosuppressive therapy on interferon $\gamma$ release assay for latent tuberculosis screening in patients with autoimmune diseases: a systematic review and meta-analysis”. In: *Thorax* 71.1 (2016), pp. 64–72.

[202] Elizabeth A Worthey et al. “Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease.” In: *Genetics in medicine : official journal of the American College of Medical Genetics* 13.3 (2011), pp. 255–62.

[203] Hui Y Xiong et al. “RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease.” In: *Science (New York, N.Y.)* 347.6218 (2015), p. 1254806.

[204] Gene Yeo and Christopher B Burge. “Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.” en. In: *Journal of computational biology : a journal of computational molecular cell biology* 11.2-3 (2004), pp. 377–94.

[205] Daria V Zhernakova et al. “DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts.” In: *PLoS genetics* 9.6 (2013), e1003594.

SCIENTIFIC REPORTS

OPEN

# Classification of Paediatric Inflammatory Bowel Disease using Machine Learning

uary 2017
il 2017
25 May 2017

E. Mossotto[1,2], J. J. Ashton[1,3], T. Coelho[1,3], R. M. Beattie[3], B. D. MacArthur[2] & S. Ennis[1]

**Paediatric inflammatory bowel disease (PIBD), comprising Crohn's disease (CD), ulcerative colitis (UC) and inflammatory bowel disease unclassified (IBDU) is a complex and multifactorial condition with increasing incidence. An accurate diagnosis of PIBD is necessary for a prompt and effective treatment. This study utilises machine learning (ML) to classify disease using endoscopic and histological data for 287 children diagnosed with PIBD. Data were used to develop, train, test and validate a ML model to classify disease subtype. Unsupervised models revealed overlap of CD/UC with broad clustering but no clear subtype delineation, whereas hierarchical clustering identified four novel subgroups characterised by differing colonic involvement. Three supervised ML models were developed utilising endoscopic data only, histological only and combined endoscopic/histological data yielding classification accuracy of 71.0%, 76.9% and 82.7% respectively. The optimal combined model was tested on a statistically independent cohort of 48 PIBD patients from the same clinic, accurately classifying 83.3% of patients. This study employs mathematical modelling of endoscopic and histological data to aid diagnostic accuracy. While unsupervised modelling categorises patients into four subgroups, supervised approaches confirm the need of both endoscopic and histological evidence for an accurate diagnosis. Overall, this paper provides a blueprint for ML use with clinical data.**

Paediatric inflammatory bowel disease (PIBD), comprising Crohn's disease (CD), ulcerative colitis (UC) and inflammatory bowel disease unclassified (IBDU) are a group of autoimmune inflammatory conditions affecting children, the incidence of which is increasing[1,2]. The major feature of inflammatory bowel disease is chronic inflammation of the gastrointestinal (GI) tract. Symptoms of PIBD include diarrhoea, abdominal pain, blood in the stool and weight loss[3]. Although both Crohn's disease and ulcerative colitis are considered to fall within the same disease group, there are often differences in disease location within the bowel, observable through endoscopic and histological assessment. Endoscopic investigation of disease is macroscopic and typically determines initial treatment and provisional diagnosis, however the endoscopic assessment of the gastrointestinal system is not always sufficient for diagnosis and histological (microscopic) examination of biopsies from the upper and lower GI tracts is vital to determine disease extent and confirm diagnosis. Typically, Crohn's disease is characterised by a non-continuous inflammation of the entire gastrointestinal system, while the inflammation pattern of ulcerative colitis is continuous and restricted to the colon and rectum. There is a well-established discordance between endoscopic (macroscopic) and histological (microscopic) disease extent[4–6]. Mucosal healing (histological) is frequently cited as a 'true' measure of remission. Despite this, the major clinical classification tool for PIBD, the Paris classification, is based exclusively on endoscopic and radiological disease extent[7–9]. Previous data has indicated histological disease extent to be significantly greater than endoscopic disease extent, at both diagnosis and follow-up[4,5]. This raises the possibility of a modification of classification to account for histological disease as an additional measure of disease extent. However, the current endoscopic Paris classification remains a validated tool to guide treatment[6,10].

Diagnosis of PIBD is challenging, the aetiology is not fully understood and deciding on management and prognostication is complex. The accuracy of diagnosis in PIBD is key to prompt and effective treatment[11]. The treatment for PIBD is highly dependent on disease location and disease extent, as well as accurately classifying as CD, UC and IBDU. Surgical intervention may be necessary for pancolitis in UC but would not provide a cure for

[1]Human Genetics and Genomic Medicine, University of Southampton, Southampton, UK. [2]Institute for Life Sciences, University of Southampton, Southampton, UK. [3]Department of Paediatric Gastroenterology, Southampton Children's Hospital, Southampton, UK. E. Mossotto and J. J. Ashton contributed equally to this work. Correspondence and requests for materials should be addressed to S.E. (email: s.ennis@southampton.ac.uk)

pancolitis in CD. Additional decisions about escalation of therapy, including use of monoclonals, rely on precise understanding of an individual patient's disease. The use of these therapies is not without drawbacks and accurate diagnosis is vital to achieve remission without putting the patient at risk of harm.

Uncertainty in the classification or the severity/extent of disease can lead to delays or inappropriate treatment[12]. Tools to assist clinicians in making a more accurate diagnosis are attractive and may assist in the better categorisation of disease into a number of specific phenotypes with implications for how best to treat. Plevy *et al.* previously developed a multi-component machine learning model (including serological and genetic markers) in adult IBD to assist with diagnosis achieving good CD/UC discrimination[13]. However, these markers are expensive, time consuming to generate and not routinely available in most hospitals; to date there are no mathematical models based solely on simple clinical data such disease location to assist with diagnosis and classification.

Machine learning is a contemporary branch of statistics particularly well suited for analysis of complex data. Machine learning algorithms aim to find patterns within data and use them to make predictions and classifications or infer new knowledge[14]. These methods are broadly grouped in two categories: (1) unsupervised machine learning algorithms do not need *a priori* knowledge of classes, instead they aim to infer classes on the basis of presenting features; (2) supervised algorithms are better suited to solve classification problems where the class of each sample/patient is known *a priori* – these samples are then used to train a model to classify subsequent samples of *unknown* class. This study utilises unsupervised models to examine the evidence for clearly distinguishable strata identifiable through endoscopic and histopathological data and examines the properties of any inferred groups. The study then applies a supervised support vector machine (SVM) and patient samples with established diagnoses of either CD or UC to construct a classification model. The resultant model is tested for accuracy and implemented on an unseen validation cohort. Such methodology has been used successfully in medicine and biology for cancer subtype classification, novel drug discovery and genomics[15–19]. Here we use paediatric patient endoscopic and histological data to assess the utility of such approaches for the diagnosis and management of this complex disease.

## Materials and Methods

Patients were recruited from the Genetics of Paediatric Inflammatory Bowel Disease study at Southampton Children's Hospital. Data were collected from prospectively entered electronic clinical records using a standardised proforma[5]. Fully anonymised patient data were obtained from endoscopy and histology at initial diagnosis, all patients were diagnosed in line with Porto criteria[20]. Disease type was confirmed by two investigators (RMB, JJA). The dataset comprised manually collected data from 287 patients, 178 with Crohn's disease, 80 with ulcerative colitis and 29 with inflammatory bowel disease unclassified (Supplementary dataset 1). The ratio of CD to UC is typical of paediatric onset disease[2].

Informed consent was obtained for all participants. The study has full ethical approval from Southampton & South West Hampshire Research Ethics Committee (09/H0504/125). All methods were performed in accordance with the relevant guidelines and regulations.

Ten gastrointestinal (GI) locations were investigated for the presence of macroscopic and microscopic abnormalities: mouth, oesophagus, stomach, duodenum, ileum, ascending colon, transverse colon, descending colon, rectum and perianal. Clinical observations were converted into numerical variables $[-1, 0, +1]$ depending on tissue abnormalities. At each location, abnormal tissues observations were coded as $+1$ and normal were coded as $-1$. Null values (0) were assigned for missing data such as in the case of restriction at endoscopy. Mouth and perianal locations are not typically biopsied for histology, therefore these feature were excluded in the unsupervised approach and automatically excluded in the supervised approach.

**Unsupervised machine learning.** In order to observe whether clinical features can induce the formation of two clusters representing CD and UC, data were modelled using principal component analysis (PCA) and multidimensional scaling (MDS) algorithms as unsupervised machine learning approaches. In unsupervised machine learning the diagnosis of CD, UC or IBDU is hidden from the model, leaving the algorithm to impose the most relevant strata. Both PCA and MDS are dimensionality reduction algorithms that convert a high dimensional space (here each dimension corresponds to a measured traits), to a lower dimensional space (usually 2D or 3D). The main difference between PCA and MDS is the search space of those two algorithms. While PCA investigates linear feature associations, MDS can also uncover non-linear associations. However, if the associations between the features are essentially linear then multidimensional scaling will provide a similar representation to that of PCA.

To better visualise the relationship between patients and traits, hierarchical clustering with Hamming distance[21] and average linkage[22] was performed.

Groups identified by hierarchical clustering were assessed with respect to: age of onset and C-reactive protein levels at diagnosis, using ANOVA; disease subtype, gender, family history and personal history of autoimmune disease using $\chi^2$. Statistical analyses were performed applying Python SciPy package[23].

**Supervised machine learning.** In order to discriminate CD and UC patients, a model was assembled utilising different techniques of supervised machine learning. We applied a supervised machine learning model where the diagnosis of CD and UC was seen by the model.

In order to isolate the key histological and endoscopic features that determined diagnostic subgrouping, we tested a range of classification strategies including ensemble learners (Boosted and Bagged Trees), linear discriminant analysis and support vector machines (SVMs) with a variety of different kernels[14,24].

Data were split in order to construct and then validate the model, 210 patients ($n_{CD} = 143$; $n_{UC} = 67$) patients were included in the model construction step. Forty-eight patients ($n_{CD} = 35$; $n_{UC} = 13$) were set aside to validate

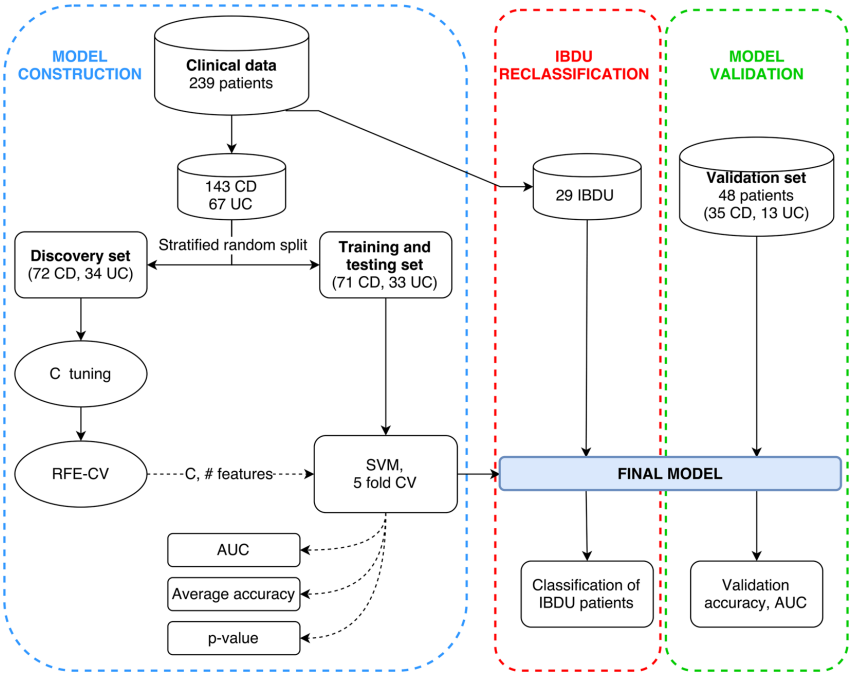


**Figure 1.** Model and data processing. Schematic representation of the model construction (blue section), validation (green section) and IBDU reclassification (red section) phases. Solid arrows represent data stream while dashed arrows represent parameters or metrics stream. The discovery set was used to identify the optimal penalty parameter (C) and number of features using the recursive feature elimination with cross validation algorithm (RFE-CV). These two elements were then passed to the training and testing set which was then modelled using a support vector machine (SVM). Three metrics were collected: area under the ROC curve (AUC); accuracy over the 5 folds and; a permutation-generated p-value.

the model on unseen data. Data from IBDU patients (n = 29) were used only for a final reclassification. Figure 1 is a schematic representation of the model and shows the usage of the different subsets.

To create a model which is applicable to unseen data, the 210 CD and UC samples were randomly split in two subsets preserving the original disease subtype ratio. The first data subset was used for searching the best parameters for the CD *versus* UC classification (discovery set). The second data subset was used for training and testing the model according to the parameters determined during the discovery phase. After assessing the performance of the final model, data from IBDU patients were passed to the model in order to classify them as either CD or UC.

Construction of optimal model utilised a linear support vector machine, allowing for regression of weights for each feature and assessment of the relative importance of each variable. Additionally, linear SVMs require estimation of a single penalty parameter (C) that allows for misclassification within the training set. In an attempt to improve model performance when optimizing the classifier we allowed the search space for C values to range from $1 \times 10^{-3}$ to $1 \times 10^{2}$. Large values of C are less prone to misclassify data points, but perform suboptimally when classifying outliers in unseen data. Small C values generate models that are more robust to outliers by allowing more misclassified data points at the expense of the training accuracy.

Machine learning approaches are weakened by the inclusion of features that are not relevant to the classification problem (confounding factors or 'noise') and reduce model performance. In order to minimise noise from non-informative features, we applied a recursive feature elimination algorithm combined with a 5-fold cross validation scheme (RFE-CV) selecting pertinent features as described by Guyon *et al.*[25]. Including a 5-fold cross validation avoids overfitting the model to the discovery set by selecting parameters and features that are specific to this set but do not generalize well, and therefore perform poorly on the test subset. The selection of the best feature subset and optimal C were chosen to maximise the classification accuracy over the discovery set.

Following the identification of the optimal C and set of features, we trained a new support vector machine and tested its efficiency (Fig. 1). With a 5-fold cross-validation scheme the algorithm repeatedly fitted and tested data from the training/testing set, providing the average accuracy in the CD *vs*. UC classification. The area under the receiver operating characteristic curve (AUC) was used to assess model efficiency. Statistical significance of the observed accuracy was determined through permutation testing of 1,000,000 randomly generated models in

which sample labels were shuffled. The p-value was then determined by calculating the frequency at which the observed accuracy was replicated by the random models. Finally, the overall performance of the model was verified by classifying unlabelled data from the validation dataset of 48 patients.

Once the model had been fully trained and validated, it was used to classify IBDU patients and posterior probabilities for membership to both the UC and CD classes were obtained. These probabilities depend on the distance between an observation and the decision function that SVM uses in order to discriminate between the two groups. The uncertainty in the classification of an individual increases as its profile is closer to the decision boundary (which is defined by the SVM decision function).

Data manipulation and modelling was performed using Matlab[24] (R2016b), Python[26] (2.7) and the Scikit-Learn[27] (0.17.1) package.

## Results

Endoscopic and histological data were collected for 287 patients; 178 patients with Crohn's disease, 80 with ulcerative colitis and 29 patients with inflammatory bowel disease unclassified. Machine learning was applied to 239 patients (CD = 143, UC = 97, IBDU = 29). Females account for 37% (107) of the individuals in the dataset. Average age of onset was 11.5 years (range 1.6 to 17.6 years). Twenty-six (9%) of patients were diagnosed below 6 years of age (very-early onset IBD). The remaining 48 patients (CD = 35, UC = 13, average age of onset 13.2 years) were used to validate the model.

**Unsupervised clustering shows the overlap of CD and UC phenotypes.** Endoscopic and histological data underwent principal component analysis with the first three components being representative of 52.2% of the total variance of data. According to both PCA and multidimensional scaling, there was no clear separation of Crohn's disease and ulcerative colitis (Fig. 2A,B).

Despite the lack of distinct clusters, CD and UC individuals are differently distributed across the 3D space with regions predominantly populated by one or the other class. As anticipated, IBDU patients were distributed uniformly throughout the CD and UC data. The same clustering pattern was observed with MDS (Fig. 2B) strongly suggesting linear relationships between the measured features. The lack of clear clusters confirms the complexity in distinguishing CD and UC phenotypes from microscopic and macroscopic observations.

**Hierarchical clustering identifies four PIBD subtypes.** In accordance with PCA and MDS analyses, hierarchical clustering did not stratify patients according to CD, UC and IBDU diagnosis (Fig. 2C). However, it did reveal the presence of distinct subgroups of patients, corresponding to complex patterns of abnormalities. As expected, most of the macroscopic and microscopic dysregulations were observed in the colorectal region. Considering only the colorectal region, it is possible to observe four distinct groups (Fig. 2C,i–iv). In the first group (i) patients exhibit tissue abnormalities identified by both endoscopy and histology. The second group (ii) shows colorectal abnormalities only after a microscopic investigation. Patients belonging to the third group (iii) present with inflammation of the rectum and the descending colon. Finally, the fourth group (iv) does not show any disruption of the colorectal region. Some patients are not placed within any of these four groups since they do not show any clear colorectal pattern. These patients have higher numbers of disease locations with null values (reflecting restriction at endoscopy).

The ileum exhibited an inconsistent pattern of disruption, acting as interface between mostly-abnormal and mostly-normal regions (left hand side *vs*. right hand side of Fig. 2C). Additionally, endoscopic or histological abnormalities in the upper GI tract are less frequent compared to lower GI tract abnormalities, this is equally applicable to all patients, regardless of their diagnosis (of CD or UC).

The four groups were analysed for any difference in their composition of patients with: a diagnosis of CD or UC; gender; positive or negative family history and clinical diagnosis of any other personal autoimmune disease. There was no significant difference between the groups with regard to any of these variables with the exception of diagnosis. Group iii (inflammation of the rectum and the descending colon) was significantly enriched for patients with ulcerative colitis patients ($p = 0.046$) and group iv (no colorectal involvement) was significantly enriched for patients with Crohn's disease ($p = 0.007$). Groups i and ii were not significantly enriched either for CD or UC indicating presence of both disease types.

Regression analysis of the four groups identified a significant ($p = 0.003$) increase in CRP for patients in group iii compared to the other groups (Fig. 2D). There was no significant difference in age of diagnosis across groups.

**A combined model distinguishes Crohn's disease from ulcerative colitis with the greatest accuracy.** Model selection was based by testing a range of different algorithms and kernels. Table 1 reports classification accuracies obtained fitting and testing models on the whole dataset excluding IBDU patients and the validation cohort. Reported accuracies are only informative in terms of comparing different models and were not validated on external dataset. Linear discriminant and linear support vector machine outperformed other tested algorithms. Linear models performed better than Tree-based model and non-linear SVMs. Although 0.5% less accurate compared to a linear discriminant model, linear SVM represented the best choice in terms of adaptability and interpretation. Linear discriminant models assume data have the same covariance and a normal distribution, while SVMs does not have such requirements and is better suited for discriminative tasks[28]. Therefore, an SVM[14] with a modified linear kernel was used as core classifier in our model.

In order to elucidate which observations are needed for optimal disease classification of patients, three supervised models were generated implementing endoscopic features, histological features and both endoscopic and histological features.

The combined model outperforms the other two models achieving the highest accuracy; the model correctly assigns the diagnosis of CD or UC to a patient in 82.7% of cases (Table 2). All metrics that assess model

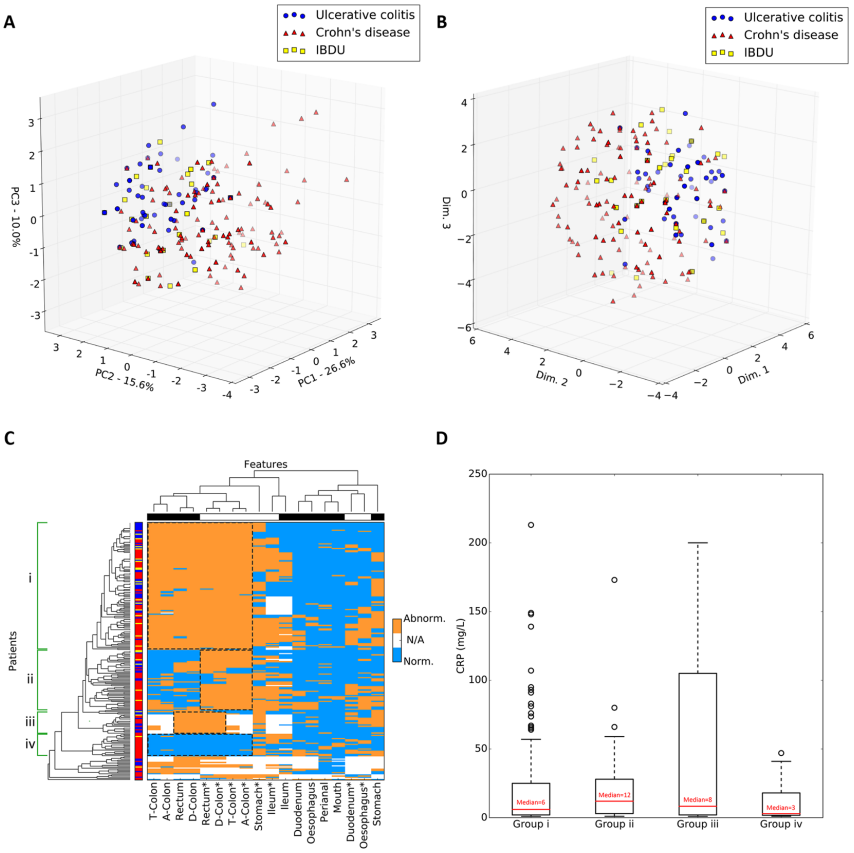


**Figure 2.** Dimensionality reduction approaches and hierarchical clustering of PIBD data. (**A**,**B**) Principal component analysis (**A**) and multidimensional scaling (**B**) of clinical data from 239 PIBD patients. The first three PCA components account for 52.2% of the total variance. Important note – UC/CD/IBDU diagnoses were used only to retrospectively colour data points and were not included in actual modelling. (**C**) Heatmap of endoscopic and histological tissue abnormalities in PIBD patients. Abnormal manifestations are shown in orange, normal in light blue and missing data in white. Asterisks indicate histology features. Ascending colon, transverse colon and descending colon labels were shortened to A-Colon, T-Colon and D-Colon respectively. Left hand side bar shows the referred diagnosis: CD in red, UC in blue, IBDU in yellow. Again, UC/CD/IBDU diagnoses were not used to model data but only to retrospectively colour each element. The top bar shows the type of investigation: histology in white, endoscopy in black. Identified colorectal groups are shown by dashed boxes and labelled from one (i) to four (iv). (**D**) Box and whisker plot depicting C-reactive protein (CRP) levels recorded at diagnosis across the four identified groups. Each box represents data from the first (bottom edge) and the third (top edge) quartile. Red bars and numbers are the median CRP level. Dashed whiskers show the lowest and highest CRP within each group. Black circles are outlier data points.

performance agree in the superior efficiency when using combined endoscopy and histology data. The combined model shows the highest accuracy, precision and F1-score; recall is close to that observed in the histological model. The endoscopy model performs well in terms of precision but is poorer in recall. Conversely, the histological model has the lowest precision but highest recall. This indicates that using endoscopy data the model is highly precise in identifying most of individuals from both classes (CD and UC). However, the endoscopy model is prone to produce more false negatives (recall) compared to the histology model. Both the accuracy and the F1 score, which combines precision and recall metrics, indicate that histology model is superior to the endoscopy model although having a lower precision. Moreover, the combined model selects all the features selected by the endoscopy and histology models plus two additional histological features (oesophagus and ascending colon). As expected, the ileum location appears to be consistently informative for the discrimination of CD and UC patients in every model, and in the histological model is sufficient to diagnose CD or UC in 76.9% of cases. Features with

| Method | Accuracy |
|---|---|
| Simple Tree (4 splits) | 78.1% |
| Medium Tree (20 splits) | 75.2% |
| Complex Tree (100 splits) | 76.7% |
| Linear discriminant | 81.0% |
| Linear SVM | 80.5% |
| Quadratic SVM | 78.1% |
| Cubic SVM | 73.8% |
| Boosted Trees | 74.8% |
| Bagged Trees | 77.6% |

**Table 1.** Preliminary assessment of linear and non-linear models. Linear support vector machine (SVM) was the selected model.

| Input | Accuracy % (AUC) | Precision | Recall | F1-score | (#) Features |
|---|---|---|---|---|---|
| Endoscopy | 71.0% (0.78) | 0.89 | 0.68 | 0.75 | (5) Duodenum, Ileum, D-Colon, Rectum, Perianal |
| Histology | 76.9% (0. 82) | 0.81 | 0.86 | 0.83 | (1) Ileum |
| Combined (E + H) | 82.7% (0.87) | 0.91 | 0.83 | 0.87 | (8) Duodenum, Ileum, D-Colon, Rectum, Perianal, Oesophagus*, Ileum*, A-Colon* |

**Table 2.** Performance of the three optimised supervised models, asterisks indicate histological features. All metrics represent the average over the 5-folds of the cross validation.

similar observations in both CD and UC patients are not informative for the classification while locations with a more variable manifestation of tissue damage were typically selected in the RFE-CV selection.

The greatest area under the curve (AUC) was observed in the combined model (0.87) followed by the histology (0.82) model and then the endoscopic model (0.78) (Fig. 3A). The endoscopic, the histological and the combined models showed a statistical significance of $p = 3 \times 10^{-3}$, $p = 5 \times 10^{-6}$ and $p = 1 \times 10^{-6}$ respectively (Fig. 3B).

For each training fold of the combined model, the observed accuracies (in decimals) were 0.86, 0.67, 0.95, 0.85 and 0.80 respectively. Overall, the mean accuracy was 0.83, the median 0.85, the standard deviation 0.09 and the standard error 0.05. Over the 1,000,000 permutations, none of the randomised models achieved an accuracy equal or greater than the observed (p-value $= 1 \times 10^{-6}$). These metrics indicate good overall performance and no overfitting of the model.

**Assessment of the combined model in an additional cohort.** In order to further validate the combined histological and endoscopic model we applied it to classify 48 anonymised PIBD patients (validation set, Fig. 1). These data had not been used in the optimisation or training of the model. The model was accurate in classifying this additional cohort, correctly assigned the diagnosis of CD or UC in 83.3% of cases (Table 3). The performance metrics calculated on the validation set confirm the previous results in terms of accuracy and recall. However, precision, and consequently the F1-score, are lower when compared to the performance calculated over the test set. F1-score of the validation set is still higher than the histology and endoscopy only models.

Since the validation set never took part in any phase of the model generation, and since the model was already trained and tested avoiding overfitting, the accuracy over the validation set did not required any additional shuffling.

**IBDU patients can be categorised by the combined model.** The combined model was used to attempt to classify the 29 IBDU patients by assigning them to either a CD or UC subtype and computing the posterior probability of belonging to each class (Fig. 3C). It should be noted that the model was not trained to classify IBDU therefore patterns restricted to this class were not learnt by the algorithm. Instead the model aims to identify patterns learnt from UC and CD data in these previously unseen IBDU cases.

When applied to the 29 IBDU patients, 17 patients were assigned as Crohn's disease and 12 as ulcerative colitis. In 17 of these patients the IBD subtype classification was estimated with a probability greater than 80% (Fig. 3D). Exploring the distribution of the posterior probabilities (Fig. 3D), patients are not equally distributed across the entire probability range. The sigmoidal distribution reflects higher certainty of the model predication where patients present with a pattern learnt during the construction step but prediction accuracy declines rapidly for patients exhibiting previously unseen patterns.

## Discussion

In this study we have mathematically modelled endoscopic and histological data to aid with classification of IBD diagnosis in paediatric patients. The resulting model demonstrates high accuracy in discriminating CD and UC patients and also provides an effective visualization of the complex overlap of these two disease subtypes.

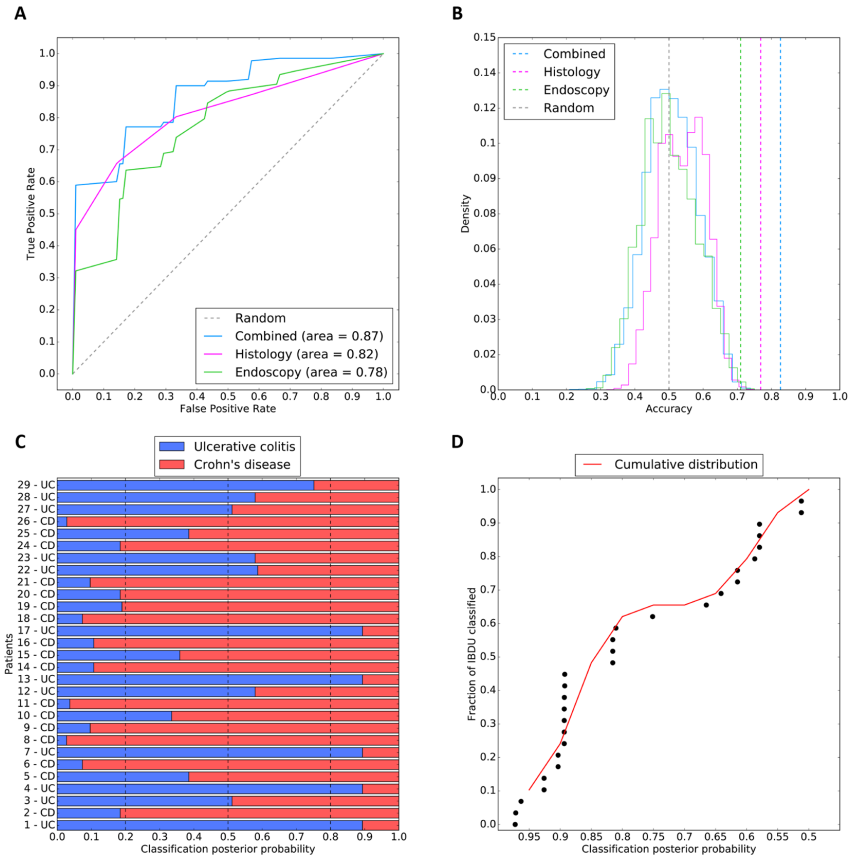


**Figure 3.** Supervised classification performance and metrics. (**A**) Receiver operating characteristic of the combined (light blue), histology (purple) and endoscopy (green) models. The grey dashed line represents the expected performance of a random model. (**B**) Permutation tests of models: dashed lines represent the observed accuracy of the combined (light blue), histology (purple) and endoscopy (green) models. The endoscopic, histological and combined models have a p-value of $p = 3 \times 10^{-3}$, $p = 5 \times 10^{-6}$ and $p = 1 \times 10^{-6}$ respectively. The grey dashed line represents the average expected performance of random model. Solid coloured lines show the distribution of random permutations for each model. (**C**) Classification of IBDU patients with the combined model in Crohn's disease (red) or ulcerative colitis (blue) subtypes. The classification posterior probability indicates the confidence of the model in assigning UC or CD labels. (**D**) Cumulative confidence in IBDU reclassification represented as cumulative density function (red line) of posterior probabilities for 29 IBDU patients. Each dot represents an IBDU patient.

| Validation set | Accuracy % | Precision | Recall | F1-score | Support |
|---|---|---|---|---|---|
| UC | — | 0.65 | 0.85 | 0.73 | 13 |
| CD | — | 0.94 | 0.83 | 0.88 | 35 |
| Average/Total | 83.3% | 0.86 | 0.83 | 0.84 | 48 |

**Table 3.** Performance of the trained combined model over the validation set.

Interpretation of the unsupervised models confirms uncertainty in discriminating CD and UC subtypes with overlapping and undefined clusters based only on disease location. We observed a limited separation of Crohn's disease and ulcerative colitis patients, with UC presenting less variance than CD cases.

Based on the endoscopic and histological disease location the unsupervised models did not classify disease into distinct CD/UC subtypes, instead four distinct groups of patients were characterised by different colorectal

involvement. The hierarchical clustering was not able to fit some individuals in those previously described groups. There are clear challenges in diagnostic categorisation based solely on disease location, however this model points to further subcategorization of disease, with significant overlap between UC and CD in groups i and ii. Whilst group iv is almost exclusively CD all colonic involvement has some overlap between disease types suggesting sub-classification of disease may be useful in distinguishing subtypes of CD or UC, potentially with impacts on management decisions. This theory has been raised previously through mathematical modelling of complex IBD data including serological and genetic markers[13,29]. Regression analysis of CRP level at diagnosis with groups i-iv indicates a statistically significant increase in CRP in group iii, whilst the reason behind this are uncertain there is a need to identify patients with increased systemic inflammation in order to optimise treatment. Here we provide potential evidence of the need for further subcategorization of disease based on solely on clinical parameters used in standard practice.

It is well established that ileal inflammation is key to diagnosis of Crohn's disease. Here we found that ileal inflammation (endoscopic or histological) is the only feature selected as important in all the models we constructed, providing evidence that ileal disease is the single most important factor for disease classification. Additionally, whilst colonic inflammation is important in paediatric UC, we find that it is also frequently present in CD with significant overlap between the 2 diseases.

There is significant interest in application of machine learning to clinical problems to aid with diagnosis, disease classification and personalising treatment. Nevertheless, the main focus of machine learning should not be to replace the human decision-making but to provide help in uncertain situations. There will always be an innate limitation of mathematical models to replicate the human intuition built with experience. However, some examples of machine learning applied to clinical data have been proved successful in situations to such as providing risk scoring systems[30], imaging interpretation[31], new patient stratification models[32] and diagnostic tools[33].

Our machine learning models have been utilised for solving a classification problem (CD vs UC) and additionally to observe data structure and complexity with a view to improvement of current classification. Through the application of machine learning to these data we confirmed the higher accuracy of histological over endoscopic data if used in isolation. We also demonstrated that both investigations are needed for an optimal classification, although the current Paris classification only accounts for endoscopic disease location.

Recently there has been interest in discrepancies between endoscopic and histological disease extent, with some calls to review the Paris classification of paediatric IBD to incorporate an additional histological score[4–6]. This model provides further evidence to suggest that there are significant differences between endoscopic and histological disease extent, with notable differences seen in Fig. 2C. Additionally the classification accuracy of the model of endoscopic disease alone is less than a combined model, further raising the need to discuss a modification to the Paris classification.

The potential clinical utility of machine learning models such as the one we have developed are significant, by placing these basic data into the model a clinician will get a disease probability score at this, the model is open to incorporating additional data coming from independent clinics, leading to increasing accuracy over time.

IBDU presents an ongoing challenge to clinicians. There is broad guidance on treatment but increasingly there is uncertainty with diagnosis and reclassification of disease at a later stage[20]. The model described here has been developed in an attempt to classify Crohn's disease and Ulcerative Colitis at diagnosis, and not to reclassify IBDU based on disease location. Despite this, IBDU patients appear throughout the PCA/MDS plots and do not cluster, indicating a heterogeneous disease phenotype. We applied the model to 29 patients diagnosed with IBDU at initial endoscopy, 17 of these patients were assigned a probability of greater than 80% to either CD or UC based on their disease location. Posterior probabilities obtained from the classification of IBDU patients as either CD or UC, resulted in either high ($p > 0.85$, $n = 14$) or low ($p < 0.65$, $n = 10$) values, with few ($n = 5$) exceptions. This distribution suggests the presence of at least two subgroups within IBDU patients. The first, where the model assigns the CD/UC label with high confidence, might represent a subset of patients with a clinical presentation similar to those already observed and learnt in CD and UC cases. The second subgroup, labelled with low confidence, might instead reflect a distinct clinical presentation that does not fit in the current classification criteria. Support from ML modelling may be particularly attractive for IBDU cases.

The strengths of this study lie in the robust nature of data collection. Patients recruited to this study were diagnosed by 4 different clinicians from Southampton Children's Hospital, therefore the pattern discovered by the model is not that of a single gastroenterologist. The supervised model combines different machine learning elements, but its relative simplicity makes it quick and easily interpretable. The feature selection step (RFE-CV) implicated the most informative GI locations for diagnosing IBD subtypes.

Through this model we report a diagnostic accuracy of 82.7% with an area under the ROC curve of 0.87, although for clinical application this would need to be increased to exceed 0.95. This may be possible with the addition of more patients or more data (e.g. blood data, granulomata). Comparing the metrics of the trained model with the performance over the validation set we conclude that: (1) the combined model performs better than individual histology or endoscopy models; (2) that both endoscopic and histological evidences are needed for an optimal classification of PIBD and (3) performance over the validation set is similar to that observed over the test set, confirming the absence of overfitting and good generalisation. Moreover, performance metrics seen in the validation set, suggest that classification of UC patients is much more complex than for CD patients, reflecting the uncertainty observed in clinics. In total, 94% of Crohn's disease patients were successfully labelled as CD while only 65% of UCs were correctly labelled. In conclusion, the missing 17% percent in accuracy can be mostly attributed to a lower discriminability of patients affected by UC. Additionally, this work can be seen as a blueprint for improvement of IBD categorisation in the future, through modelling of additional data, such as variants from whole-exome sequencing, transcriptome profiles and microbiome signatures it may be possible to gain further, clinically relevant, disease groups[34]. In the future this may aid with treatment selection, prognostication and ongoing management.

This study employs a mathematical model of histological and endoscopic data within IBD; it provides a model with high diagnostic accuracy on unseen data (83.3%). We present 4 novel subgroups of disease identified by unsupervised machine learning based on colonic disease.

The purpose of this study was two-fold, to better understand disease aetiology, heterogeneity and classification and to understand the potential for machine learning to assist with disease classification. Through further work machine learning can aid clinicians to accurately subtype disease and personalise treatment. Additionally this may help with classification of IBDU. Whilst existing methods for diagnosis appear robust, the opportunity to improve and personalise therapy for patients through new and more accurate subtyping of disease is exciting and increasingly tangible.

## References


1. Henderson, P. *et al.* Rising incidence of pediatric inflammatory bowel disease in Scotland. *Inflamm. Bowel Dis.* **18**, 999–1005, doi:10.1002/ibd.21797 (2012).
2. Ashton, J. J. *et al.* Rising incidence of paediatric inflammatory bowel disease (PIBD) in Wessex, Southern England. *Arch. Dis. Child.* **99**, 659–664, doi:10.1136/archdischild-2013-305419 (2014).
3. Podolsky, D. K. Inflammatory Bowel Disease. *N. E. J. Med.* **325**, 928–937, doi:10.1056/NEJM199109263251306 (1991).
4. Fernandes, M. A., Verstraete, S. G., Garnett, E. A. & Heyman, M. B. Addition of Histology to the Paris Classification of Pediatric Crohn Disease Alters Classification of Disease Location. *J. Pediatr. Gastroenterol. Nutr.* **62**, 242–245, doi:10.1097/MPG.0000000000000967 (2016).
5. Ashton, J. J. *et al.* Endoscopic Versus Histological Disease Extent at Presentation of Paediatric Inflammatory Bowel Disease. *J. Pediatr. Gastroenterol. Nutr.* **62**, 246–251, doi:10.1097/MPG.0000000000001032 (2016).
6. Turner, D. Mcroscopic Assessment in Inflammatory Bowel Disease. *J. Pediatr. Gastroenterol. Nutr.* **62**, 191–2, doi:10.1097/MPG.0000000000001049 (2016).
7. Sankey, E. A. *et al.* Early mucosal changes in Crohn's disease. *Gut* **34**, 375–81, doi:10.1136/gut.34.3.375 (1993).
8. Moum, B., Ekbom, A., Vatn, M. H. & Elgjo, K. Change in the extent of colonoscopic and histological involvement in ulcerative colitis over time. *Am. J. Gastroenterol.* **94**, 1564–1569, doi:10.1111/j.1572-0241.1999.01145.x (1999).
9. Levine, A. *et al.* Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm. Bowel Dis.* **17**, 1314–21, doi:10.1002/ibd.21493 (2011).
10. de Bie, C. I. *et al.* Disease phenotype at diagnosis in pediatric Crohn's disease: 5-year analyses of the EUROKIDS Registry. *Inflamm. Bowel Dis.* **19**, 378–85, doi:10.1002/ibd.23008 (2013).
11. Levine, A. *et al.* The ESPGHAN Revised Porto Criteria for the Diagnosis of Inflammatory Bowel Disease in Children and Adolescents. *J. Pediatr. Gastroenterol. Nutr.* **1**, doi:10.1097/MPG.0000000000000239 (2013).
12. Levine, A. *et al.* Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm. Bowel Dis.* **17**, 1314–21, doi:10.1002/ibd.21493 (2011).
13. Plevy, S. *et al.* Combined serological, genetic, and inflammatory markers differentiate non-IBD, Crohn's disease, and ulcerative colitis patients. *Inflamm. Bowel Dis.* **19**, 1139–48, doi:10.1097/MIB.0b013e318280b19e (2013).
14. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning. *Elements* **1**, 337–387, doi:10.1007/978-0-387-84858-7 (2009).
15. Upstill-Goddard, R. *et al.* Support Vector Machine classifier for estrogen receptor positive and negative early-onset breast cancer. *PLoS One* **8**, e68606, doi:10.1371/journal.pone.0068606 (2013).
16. Capriotti, E. & Altman, R. B. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**, 310–7, doi:10.1016/j.ygeno.2011.06.010 (2011).
17. Li, J. *et al.* A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **17**, 2–12, doi:10.1093/bib/bbv020 (2016).
18. Lima, A. N. *et al.* Use of machine learning approaches for novel drug discovery. *Expert Opin. Drug Discov* **11**, 225–239, doi:10.1517/17460441.2016.1146250 (2016).
19. Mathé, C., Sagot, M.-F., Schiex, T. & Rouzé, P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* **30**, 4103–17, doi:10.1093/nar/gkf543 (2002).
20. Levine, A. *et al.* The ESPGHAN Revised Porto Criteria for the Diagnosis of Inflammatory Bowel Disease in Children and Adolescents. *J. Pediatr. Gastroenterol. Nutr.* 1, doi:10.1097/MPG.0000000000000239 (2013).
21. Hamming, R. W. Error Detecting and Error Correcting Codes. *Bell Syst. Tech. J* **29**, 147–160, doi:10.1002/bltj.1950.29.issue-2 (1950).
22. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull* **38**, 1409–1437 (1958).
23. Oliphant, T. E. SciPy: Open source scientific tools for Python. *Comput. Sci. Eng.* **9**, 10–20 (2007).
24. Documentation, M. Matlab documentation. *Matlab* R2012b, doi:10.1201/9781420034950 (2012).
25. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn* **46**, 389–422, doi:10.1023/A:1012487302797 (2002).
26. Lutz, M. Learning Python. *Icarus* **78** (O'Reilly, 2007).
27. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2012).
28. Ng, A. & Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Adv. neural Inf. Process.* **14**, 841–848 (2002).
29. Wei, Z. *et al.* Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **92**, 1008–1012, doi:10.1016/j.ajhg.2013.05.002 (2013).
30. Kannel, W. B., Doyle, J. T., McNamara, P. M., Quickenton, P. & Gordon, T. Precursors of sudden coronary death. Factors related to the incidence of sudden death. *Circulation* **51** (1975).
31. Criminisi, A. Machine learning for medical images analysis. *Medical Image Analysis* **33**, 91–93, doi:10.1016/j.media.2016.06.002 (2016).
32. Woodruff, P. G. *et al.* T-helper Type 2–driven Inflammation Defines Major Subphenotypes of Asthma. *Am. J. Respir. Crit. Care Med.* **180**, 388–395, doi:10.1164/rccm.200903-0392OC (2009).
33. Hu, X. *et al.* Artificial neural networks and prostate cancer–tools for diagnosis and management. *Nat. Rev. Urol* **10**, 174–82, doi:10.1038/nrurol.2013.9 (2013).
34. Weiser, M. *et al.* Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut*, doi:10.1136/gutjnl-2016-312518 (2016).


## Acknowledgements


The authors would like to thank Rachel Haggarty for assistance with management of the genetics of PIBD study database. We also would like to thank: the Hilary Marsden IfLS Scholarship; the University of Southampton NIHR academic clinical fellowship and; the Crohn's in Childhood Research Association.

## Author Contributions

S.E., B.D.M. and R.M.B. conceived the study design. J.J.A., T.C. and R.M.B. collected the data. E.M. analysed the data. E.M. and J.J.A. wrote the manuscript. All authors contributed to the final revision and have approved the manuscript for submission.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-02606-2

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.