UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND PHYSICAL SCIENCE

Next Generation Computational Modelling

Genomic Informatics Group

Mathematical tools for analysis of genome function, linkage disequilibrium structure and disease gene prediction

by

Norma Alejandra Vergara Lope Gracia

ORCID ID 0000-0001-6262-2233

A thesis submitted for the degree of Doctor of Philosophy

Supervisors: Prof. Andrew Collins and Dr. Reuben Pengelly

Examiners: Dr. Jane Gibson and Prof. Denis Shields

September, 2021

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCE
Next Generation Computational Modelling

Doctor of Philosophy

MATHEMATICAL TOOLS FOR ANALYSIS OF GENOME FUNCTION, LINKAGE DISEQUILIBRIUM STRUCTURE AND DISEASE GENE PREDICTION

by Norma Alejandra Vergara Lope Gracia

Next-generation sequencing (NGS) help to identify disease-causing genes underlying any given monogenic or complex disease. Concurrently, mathematical tools and statistical methods, including 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.

This thesis has been divided into three main sections. Firstly, the Linkage disequilibrium (LD) patterns were observed to understand the combined impact of recombination, natural selection, genetic drift and mutation. LD is the non-random association of alleles at different loci in a given population. To this end, LD patterns were constructed using 454 whole-genome sequences (WGS) from the Wellderly study based on the Malécot Morton model (exponential distributions with restricted parameters). Therefore, the extent of the LD was computed for genic, intergenic, exon and intron regions. The main result demonstrated that significant differences between exonic, intronic and intergenic components demonstrate that fine-scale LD structure provides important insights into genome function, which cannot be revealed by LD analysis of much lower resolution array-based genotyping and conventional linkage maps.

Secondly, machine learning methodologies were applied to classify genes into four groups: essential genes, Mendelian genes, genes associated with complex disorders, and non-essential–non-disease genes. To this end, the dataset was extracted from published studies of biological and functional properties of the genes. Hence, different supervised machine learning (ML) models were studied to select the most important features relevant for classifying genes. Simultaneously, Bayesian inference in a Gaussian graphical model (BGGM) was carried out to investigate recognising the significant features to enclose genes. Once the relevant features had been selected, a proposed unsupervised ML approach was developed to cluster genes into those four groups. The combined analysis of genomic data for gradient boosting and random forest models showed that more than 50% of the variance was explained and the results from BGMM showed that the connectivity between these gene metrics was 40%. The proposed unsupervised model showed an improvement for classifying genes into Mendelian group. However, results suggested that some genes involved in developing Mendelian disorders overlap with complex disorders.

Thirdly, a polygenic risk score (PRS) was developed to quantify the cumulative effect of low-penetrance genetic variants on breast cancer (BC), following the hypothesis that the polygenic component has an important impact on BC patients, as do BRCAs variants. Genome data from POSH and WTCCC were used to generate the PRS. This score was computed based on the surprisal theory. As a result, relative genome information per individual (RGI) was estimated to understand how unusual a genome is related to the reference genome. Thus, a person with a higher RGI has a more unusual genome. Likewise, a lower RGI corresponds to having more common alleles, and therefore a less surprising genome. The PRS for women who carry BRCA1/2 mutations or intermediate-risk/common variants demonstrated the hypothesis that the BC cases contain a strong inherited polygenic component. Furthermore, the polygenic component carriers tend to have more significant changes in allele frequencies compared to *BRCA1* and *BRCA2* variants.

This thesis presents methodological contributions to predictive models based on machine learning techniques and mathematical programming, together with relevant insights into disease mechanisms and potential treatment options.

List of Abbreviations

AF allele frequency

AIC Akaike information criterion

BGGM Bayesian inference in Gaussian graphical models

BC Breast Cancer

BIC Bayesian information criterion

cM centimorgan

CM Complex-Mendelian

CNM Complex non-Mendelian

DDR DNA damage response

DNA deoxyribonucleic acid

DNE gene constraint *de novo* excess

EIL expected information per locus

EM expectation-maximization

END Essential non-disease

EOBC early-onset breast cancer

FUSIL full spectrum of intolerance to loss of function

GWAS genome-wide association studies

GDI gene damage index

GGM Gaussian graphical models

GHIS genome-wide haploinsufficiency

GIMS gene-level integrated metric of negative selection

GMM Gaussian mixture model

GMM-non Gaussian mixture clustering with outlier removal

GTB Gradient tree boosting

GWAS genome-wide association studies

HRC haplotype reference consortium

HD human disease

HI Haploinsufficiency

HRI Hill-Robertson Interference

HWE Hardy-Weinberg equilibrium

kb kilobase

KNN K–nearest neighbour

KL Kullback–Leibler

LD Linkage disequilibrium

LDU linkage disequilibrium unit

LoF loss of function

Mb Megabases

MAF minor allele frequency

MDS multi-dimensional scaling

MNC Mendelian non-complex

ML machine learning

MLE maximum likelihood estimation

MSE mean squared error

NBS UK National Blood Services

NET gene position in networks

NGS next-generation sequencing

NDNE Non-disease non-essential

OMIM Online Mendelian Inheritance in Man

PBWT Burrows-Wheeler transform

PCA principal components analysis

pLI loss intolerance probability

POSH Prospective Study of Outcomes in Sporadic versus Hereditary breast cancer

pre-mRNA precursor messenger RNA

PRS polygenic risk score

PUDI positive-unlabelled learning algorithm

QC Quality control

RF Random forest

REC recessive

RLI MakeLowercaserelative local information

RNA ribonucleic acid

mRNA messenger RNA

ncRNA non-coding RNA

rRNA ribosomal ribonucleic acid

RGI relative genome information

RVIS residual variation intolerance score

SIS substitution intolerance score

SNP single nucleotide polymorphism

SVM supervised vector machine

t-SNE t-distributed stochastic neighbour embedding

WES whole-exome sequencing

WGS whole-genome sequencing

WTCCC Wellcome Trust Case Control Consortium

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Norma Alejandra Vergara Lope Gracia , declare that the thesis entitled *Mathematical tools for analysis of genome function, linkage disequilibrium structure and disease gene prediction* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- 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;
- Where I have consulted the published work of others, this is always clearly attributed;
- 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;
- I have acknowledged all main sources of help;
- 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;
- Parts of this work have been published as detailed overleaf:

Signed: Norma Alejandra Vergara Lope Gracia

Date: September - 2021

List of Publications

1. **Vergara-Lope A**, Ennis S, Vorechovsky I, Pengelly RJ, Collins A. Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding RNA and intergenic chromosome regions. Eur J Hum Genet. 2019;27(9):1436-1444. DOI: 10.1038/s41431-019-0419-0

2. **Vergara-Lope A**, Jabalameli MR, Horscroft C, Ennis S, Collins A, Pengelly RJ. Linkage disequilibrium maps for European and African populations constructed from whole genome sequence data. Sci Data. 2019;6(1):208. DOI: 10.1038/s41597-019-0227-y

3. Jabalameli MR, Horscroft C, **Vergara-Lope A**, Pengelly RJ, Collins A. Gene-dense autosomal chromosomes show evidence for increased selection. Heredity (Edinb). 2019;123(6):774-783. DOI: 10.1038/s41437-019-0272-5

4. Pengelly RJ, **Vergara-Lope A**, Alyousfi D, Jabalameli MR, Collins A. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. Brief Bioinform. 2019;20(1):267-273. DOI: 10.1093/bib/bbx110

Acknowledgements

First and foremost, I am greatly indebted to my main supervisor, Professor Andrew Collins, who has provided me with constant support, insightful discussions, and the freedom to pursue interesting research problems throughout my five-year journey at the University of Southampton. In addition to being a researcher, Andrew is also the kindest and most supportive individual with whom I have had the pleasure of working. He has a great sense of research taste, which also shapes mine to work on important and elegant questions. I would like to sincerely thank Andrew for his intuitive insights in understanding complex problems, for sharing his broad knowledge with me in almost every aspect of genetics and mathematics, and for his unwavering faith in me that pushes me to become a better person. I hope to become a researcher, an advisor and a person like him in the future.

I also would like to thank my supervisor Dr. Reuben Pengelly, Dr. William Tapper, Prof. Niranjan Mahesan and Prof. Ben MacArthur. They all taught me, directly and indirectly, many valuable things about research. The project, as it was, would not have been possible without their input in the different fields.

This research would not have been possible without the financial support of the Consejo Nacional de Ciencia y Tecnología (the National Council of Science and Technology) (abbreviated CONACYT). The University of Southampton also provided financial support through the Engineering and Physical Sciences Research Council (EPSRC). Thank you to Margaret Schroeder for her help with proofreading this thesis. No changes in intellectual content were made as a result of her advice.

Many thanks to all the members of both research groups I was luckily part of. They were available to help me or discuss my doubts about methodologies and results whenever possible since the day I arrived. Good friendships were established and essential to keeping me encouraged in this journey. Big thanks to Gabriella Galata, Lucy Upton, Gabriele Boschetto, Alvaro Perez, Guillermo Romero, Alexandry Augustin and Enrico Mossotto.

A special thanks to all my friends who helped make my time at UoS a wonderful and unforgettable period of my life: especially Reyna Peñaililllo, Gerardo Espíndola and Eduardo Pérez, because since the day one were part of this journey. I also express my deep sense of gratitude to my friends Carina Valenzuela, Yann Gelister and Domenico Balsamo for all the trips, words, random chats, fun times and laughs we shared while I was doing this Ph.D. Thanks also to my friends Félix Catalán and Yurani Moreno. They all helped me pursue several non-academic trajectories of my life, and I cannot thank them enough for their unconditional friendship and support, which has definitely been priceless.

I would also like to give a shout out to my amazing bunch of Mexican friends who contribute to making it a vibrant and fun adventure to pursue graduate studies: Martha Moreno, Cristina Pérez, Brenda Carrasco, Gerardo Escaroz, Liv Lafontaine, Itzel Corrales, Diana Oropeza, Grisell Orozco, Edith Oropeza, Karina Barrios, Rodrigo Aranda, Rocío Espinoza, Dulce Cano, Ulises Andraca, and Enrique Minor despite being far away they were always there.

Last, but most importantly, I would like to thank my mom Norma Gloria, my siblings, Don Allhan, Meyhanni, Jackeline, my sister in law, Sandra Yolanda, my beautiful niece, Zuri and my father, Alejandro, for all of your unconditional love and support during my Ph.D. journey. This thesis would not have been possible without your encouragement along the way. You have always been and will be an essential part of my life.

(Spanish translation of last paragraph) Por último, pero más importante, quisiera agradecer a mi mamá Norma Gloria, a mi hermano Don Allhan, a mi cuñada Sandra y a mi hermosa sobrina Zuri, por todo su amor y apoyo incondicional durente mis estudios de doctorado en el extranjero. Esta tesis no hubiera sido posible sin su aliento y ayuda durante todo este camino. Siempre han sido y serán parte esencial de mi vida.

To my mother and brother, for their encouragement, over all aspects, during these years.

Chapter 1

Introduction

1.1 The human genome

Genes make up the core building blocks of life and their genomic and functional properties, along with various interactions, are linked to health or disease conditions. The human genome is a diploid genome encompassing a complete sequence of repeated nitrogenous bases (nucleotides): adenine (A), guanine (G), thymine (T) and cytosine (C) with a haploid size of ∼3 Gbp [1]. This deoxyribonucleic acid (DNA) sequence comprises 22 autosomal homologous pairs (1–22) and two allosomes (X & Y) (Figure 1.1) [2]. Chromosomes range from 250 million bases (250 Megabases (Mb)) for chromosome 1, to 50 Mb for chromosome 21. The total length of the genome is approximately 3,000 Mb.

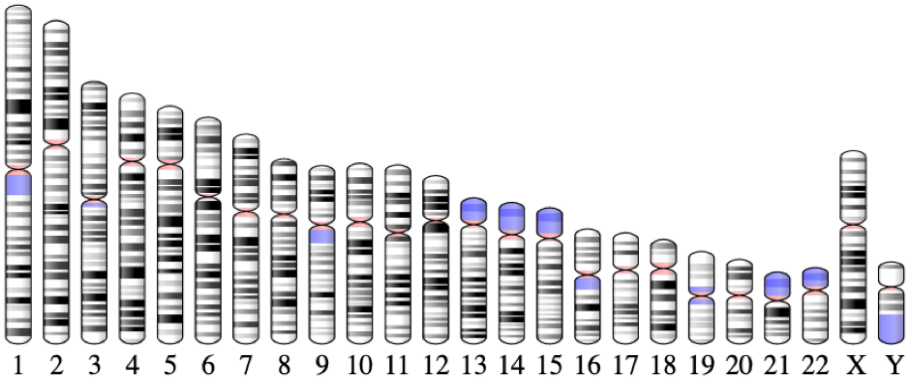


Figure 1.1: Human chromosome idiogram. Human chromosome is shown following Giemsa staining. Black and grey indicate Giemsa positive, pink indicates centromeric regions, light blue indicates heterochromatin and dark blue indicates AT-rich regions of chromosomes. From the National Center for Biotechnology Information, U.S. National Library of Medicine (https://www.ncbi.nlm.nih.gov/genome/tools/gdp/)

The human genome can be organised by the function of a specific genomic region. The DNA regions that code for proteins are defined as genes and there are approximately 20,000 protein-coding and 12,000 non-coding genes [3]. These genes have a divided structure in segments of coding sequence called exons which are separated by non-coding sequences termed introns (Figure 1.2). The transcription process is the first stage in gene expression into proteins. This process begins

when in the cell nucleus a ribosome is exported to the cytoplasm to translate the information in messenger RNA (mRNA) into a protein. As part of the mRNA processing pathway, precursor messenger RNA (pre-mRNA) transcript is transformed into a mature mRNA removing the introns by an alternative splicing mechanism [2]. Once introns are removed, protein synthesis results and the ribosomes translate the messenger RNA sequence into an amino acid sequence. Transcribed DNA is estimated to comprise 1–2% of the human genome [1], while the remaining ∼98% is formed of repeat elements and regions of DNA that are non-coding [4].

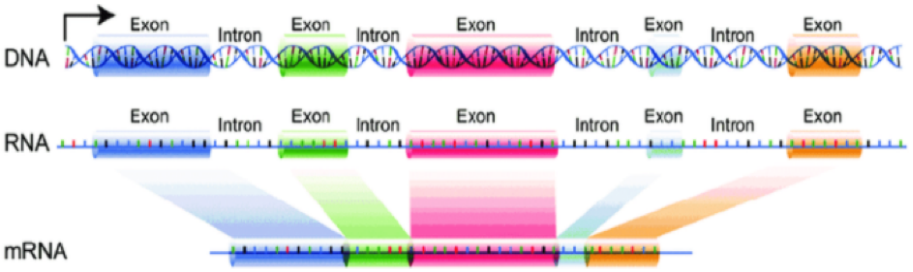


Figure 1.2: Gene structure. Exon and intron regions for eukaryotic DNA are represented by the coloured barrels. In the transcription of eukaryotic DNA into messenger RNA, introns are omitted (splicing) and only exons are translated into proteins. Taken from Saberkari *et al.* (2013) [5]

1.2 Next-Generation Sequencing and Genotyping

Next-generation sequencing (NGS) or high throughput sequencing of DNA molecules makes it possible to analyse whole-genome sequencing (WGS), whole-exome sequencing (WES) or sequencing of a particular genome region [6] within a reasonable timeframe and cost [7]. NGS has the capacity for a far higher genotyping density than even the highest density array. NGS-reads display direct information of base pairs sequenced from a DNA fragment to elucidate genes containing disease-causing variations underlying Mendelian/monogenic or complex diseases [8].

Several studies found that NGS technologies have a high rate for identifying causal genes across diverse disorders with varying modes of inheritance, including dominant, recessive and de novo arising mutations [9]. These methods yield the entire genome (through WGS) or parts of it (through WES or targeted panels) to be sequenced faster, at greater depth and with greater sensitivity, and this decrease in costs has made the clinical application of WES and WGS more feasible. WES is a genomic technique for sequencing all protein-coding regions of genes, while WGS covers both introns and exons of the genome without using techniques to isolate specific regions of DNA [10]. Given this, several studies have examined the benefits derived from the application of WES and WGS in the clinic. These studies have mainly focused under Mendelian conditions more broadly, consistently reporting diagnostic yields between 25 and 30% [11]. In a recent study analysing the application experience WES clinic indicated a 25% rate of putative molecular diagnoses in large cohorts with various Mendelian diseases [12]. However, cohort studies have shown that WGS provides a higher diagnostic yield than WES, with a 34% diagnosis rate in Mendelian disease increasing to 57% [13].

In addition to the inclusion of non-exonic regions of the genome, WGS provides complete coverage of the exome, providing greater sensitivity for detection of variants [14].

Genotyping arrays remain a viable option for SNP detection due to the computational challenges posed by the large amounts of sequence data produced at WGS [15]. Genotyping matrices have been used primarily in large-scale genome-wide association studies (GWAS). In a GWAS, a large number of markers, generally SNPs, are genotyped using these high-density genotyping matrices, followed by SNP tests for association [16]. According to the GWAS catalogue (December 2017 version), 2,724 unique GWAS studies have been completed and published. This is particularly useful for assessing known markers associated with cancer in the human genome, enabling researchers to find DNA copy-number alterations in human cancers, as discussed in Chapter 6. Since 2007, GWAS have identified roughly 100 common genetic susceptibility loci for breast cancer risk [17]. Mullighan *et al.* illustrated the power of SNP genotyping arrays to identify significant genetic abnormalities in children acute with lymphoblastic leukaemia [18].

1.3 Genetic diseases in humans

There is a lot to gain from understanding the relationship between an individual's genotype and their phenotypes. Identifying the genetic component becomes extremely valuable to treat or cure a disease or understand more about a particular trait. A genetic disorder is a health complication originated by one or more abnormalities on the genome, such as a mutation in a single gene (monogenic), a small number of genes (oligogenic), multiple genes (polygenic) or by a chromosomal abnormality. There are 6,711 well known genetic disorders and new genetic disorders are continually being reported in the medical literature [19]. The prevalence of people that are affected by a known single-gene condition is approximately 1 in 50 people, while 1 in 263 people have been identified with a chromosomal disorder [20]. Online Mendelian Inheritance in Man (OMIM) reported that the molecular basis is not known for 50% of confirmed Mendelian phenotypes. OMIM also stated that 1,769 phenotypes related to Mendelian basis, but they have not been fully established yet or may overlap with other characterised phenotypes. Studies of monogenic diseases contribute to the knowledge of polygenic forms of human disease [21]. Although polygenic or complex disorders are more common than single-gene disorders, these disorders are difficult to study and treat because different factors have not been identified. Moreover, complex diseases do not have Mendelian inheritance patterns, making it complicated to define of person's risk of inheriting these disorders or transmitting them to their descendants [22].

1.3.1 Mendelian disorders

Mendelian (monogenic) diseases are typically rare due to a mutation at a single genetic locus with a clear inheritance pattern in pedigrees. Mendelian disorders are also recognised by their different inheritance patterns, such as autosomal dominant, co-dominant or recessive and X-linked (sex-linked). Recessive diseases occur because of inheriting two mutated genes or in the allele. Dominant

disorders involve damage to only one gene copy. X-linked conditions are linked to defective genes in the X chromosome. The X-linked alleles might also be dominant or recessive. These disorders are not affected by exogenous factors. However, these factors can affect the phenotype [23].

At least 3,933 genes which underlie 5,646 Mendelian/monogenic diseases have been identified; however, there are thousands of Mendelian disorders that are yet to be uncovered in OMIM [19]. Since 2010, next-generation sequencing (NGS) has accelerated the rate estimation for monogenic disorders diagnosis by only about 25% to 50%. The remaining significant fraction of monogenic disorders is unfounded, or several genes may yield to be involved in complex phenotypes or require additional environmental triggers. A further possibility for a low rate of the known Mendelian diseases may fail in identifying genetic variation such as copy number variation, epistatic or epigenetic mechanisms [24]. For instance, Taylor *et al.* [13] stated that in 217 cases from WGS with a broad range of disorders, only 35% of these cases were molecularly resolved as Mendelian diseases. In 2015, Jamuar *et al.* [25] found that NGS has been employed in clinics, with a reported diagnostic yield of approximately 25% in Mendelian diseases. Boycott *et al.* (2013) stated that roughly 25% of reported mutations in known disease-causing genes were associated with a phenotype that, in retrospect, matched the clinical presentation of the patient being investigated. Although a large number of underlying genetic disorders, mostly Mendelian/monogenic, have been resolved at the molecular level by NGS, most of these genetic disease markers have no clear functional roles in disease aetiology [26]. Hence, many genes–disease relationships remain poorly understood in terms of identifying causal alleles for Mendelian disorders, most common diseases and complex diseases [27, 28].

1.3.2 Complex disorders

Complex diseases are caused by the interaction of multiple genetic, environmental and lifestyle factors. These factors confer a small risk individually of population risk [29]. To describe complex diseases, it is essential to understand Mendel's two main principles of inheritance, segregation and independent assortment of genes. These principles determine how inherited traits, comprising these underlying diseases, are passed from generation to generation. Following this, the factors that influence complex genetics involve reduced penetrance, variable expressivity, phenotypic characterisation, gene–gene interactions and gene–environment interactions [22, 30].

Genome-wide genotyping with high-throughput approaches for analysing complex diseases has enabled the identification of >2,600 associated common risk alleles that have positive associations with >350 distinct complex traits [31, 32]. However, the majority of associated alleles, the identities of causal genes and variants and their function remain uncertain due to multiple interactions for their manifestation, with genetic variants that predispose an individual to the condition, intergenic signals/tag SNP or LD markers. Therefore, it requires different approaches to recognise Mendelian or complex disease genes, as discussed in Chapter 5.

1.4 Genome dynamics

1.4.1 Population structure

Population genetic structure refers to the total genetic diversity and distribution within and among a population set. The shape patterns of genetic variability across the human genome are driven by the combined effects of recombination, mutation, genetic drift, evolutionary history and natural selection. Some of these variants are highly common in a population with a high alternative allele frequency (AF), while other variants are found only in a single population among a broader collection of people [33, 34]. These variants are not evenly distributed across populations, as a variant could be rare in one population and common in another [35]. Therefore, it is essential to consider these variations among populations to minimise population differences that may limit some studies.

1.4.2 Linkage disequilibrium

In the human genome, alleles at two loci on the same chromosome often show stochastic dependence. Linkage disequilibrium (LD) patterns are non-random pairwise allelic associations over many generations that indicate that two genes are physically linked [36]. LD patterns are measured as the difference between the observed frequency of a particular combination of alleles at two loci and the rate expected for random associations. These random associations assume that given enough evolutionary time, the event of recombination will affect the distribution of allele frequencies. Thus the prevalence of a specific allele at a given locus will be independent of alleles at other linked loci. Understanding this distribution of alleles frequencies and the recombination frequencies between markers during crossover of homologous chromosomes, as reflected in a genetic linkage map, is critical for genetic mapping. Therefore, the location of a disease-causing allele can be found by coinheritance with a marker allele in a pedigree where transmitted recombined haplotypes might target region to search for the disease-causing gene [37] (Figure 1.3).

LD patterns have been demonstrated that have an impact on population background in the form of bottlenecks, genetic drift in small populations and admixture [38]. LD throughout the genome reflects the population's history and the pattern of geographic subdivision [39]. In contrast, LD in each genomic region reflects the genetic recombination and elucidates patterns of the history of natural selection, gene conversion, mutation and blocks of LD that might arise through genetic hitchhiking [40, 41]. LD patterns help identify the genomic regions in which hitchhiking has taken place by showing correlations between single nucleotide polymorphisms (SNPs). Thus the regions which promote low recombination can be identified [42, 43, 44]. For example, Jeffreys *et al.*[44] argued the recombination rates across the genome in the narrow regions (< 2 Kilobase (Kb)), also known as hotspots, are much higher than would be expected by the average genome recombination rate. A degenerate 13 base-pair motif was identified (CCNCCNTNNC-CNC) that is overrepresented

in recombination hotspots compared to recombination cold regions [43], and this is the predicted binding site for the *PRDM9* which is involved in recombination double-strand breaks [45, 46, 47].

The extent and strength of LD patterns enable genes of unknown sequence that influence susceptibility to gene diseases to be allocated. In previous studies, it has been demonstrated that high recombination rates and mutation frequency reduce the extent of LD, resulting in increased diversity of haplotypes. Likewise, low recombination rates in certain regions of the genome increase LD [48]. Thus, the relationship between LD and recombination is directly correlated between regions of weak LD and the presence of recombination hotspots [49].

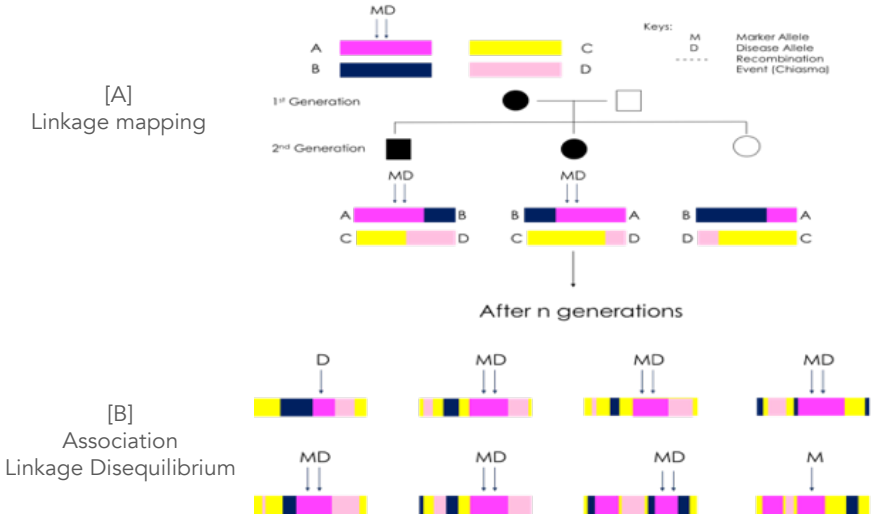


Figure 1.3: [A] Linkage mapping: the location of a disease-causing allele (D) can be marked by co-inheritance with a marker allele (M) in a pedigree. Linkage mapping tracks and limits the linked region of interest by meiotic recombination help reduce the target region for searching genes associated with disease susceptibility. [B] Association mapping: population association between D in a founder haplotype is encoded with M over many generations. Disease genes can be mapped by estimating the association in LD between alleles M and D. Figure adapted from Fig. 1 and Fig. 3 from Collins A., (2007) [37].

Previously, LD analysis has enabled the development of cost-effective genome-wide association studies and the consequent mapping of numerous common disease genes through development of arrays of 'tag' SNPs. However, Pengelly *et al.* [50] demonstrated that LD maps of SNP genotype data from arrays of tag SNPs do not fully recover the LD structure. Thus, high-resolution maps of the LD constructed by WGS data may enable the patterns of LD to be understood at a much higher resolution. Specifically, in human data, it has been shown that the set of genes with 'strong' LD are enriched for 'core' biological functions such as phosphorylation, cell division, cellular transport and metabolic processes. By contrast, genes with weak LD are enriched for functions relating to sensory perception or some immune functions and similar strategies for which high haplotype diversity is likely to be adaptive [27, 51]. Genes that harbour significant disease gene variation have been shown

to have intermediate levels of LD between these extremes. The much higher resolution LD maps derived from large WGS samples enable much more comprehensive analysis of LD structure, and genome function [52, 53].

As a result, some studies have suggested that fine-scale genetic maps of humans provide an opportunity to determine how recombination rates are influenced by genomic context. For instance, McVean *et al.* [54] stated that recombination rates are lower in genic regions than in non-coding regions. However, these frequencies could not be explained by recombination rates and selection alone [55]. Similarly, Kong *et al.* [56] found that for females and males, the recombination rates tend to be lower at genic regions, especially in bins containing exons, and higher for those containing only introns. Berger *et al.* [57] observed a 13.6% increase in LD in humans in genic regions compared to non-genic.

1.4.3 Recombination

Recombination is a genetic mechanism that 'mixes' or 'reshuffles' the genetic material of different individuals from generation to generation; it takes place during the reproductive cycle of sexually reproducing organisms. Genetic recombination involves a set of genetic exchanges between homologous chromosomes (derived from each of the individuals' parents) during the process of meiosis. Meiotic recombination allows sites that are subject to purifying selection to segregate independently and form new combinations of alleles (Figure 1.4) [58]. This process confers a significant evolutionary advantage through the breakdown of associations between alleles at linked loci generated by genetic drift, thus reducing the accumulation of deleterious variants [59]. Due to the shortening of haplotypes, regions of LD breakdown align with recombination areas (hotspots) might determine and map the locations of disease-associated variation in the genome [50].

Understanding the recombination mechanism is crucial for interpreting the patterns of genome evolution and identifying variation in the recombination rate across the genome [60]. In the absence of recombination, the accumulation of deleterious variants arising by mutation cannot be eliminated because the original haplotype cannot be regenerated over many generations, a process called Muller's ratchet [61]. Previous studies have analysed this problem in regions with high recombination rates for variations that are likely harmful, and they have found that the variants are reduced in these regions because purifying selection removes them [62]. However, some studies have suggested that recombination may be directly mutagenic, leading to sequence structural changes because of non-allelic homologous recombination [60]. It has been found that the gene mutation rate is correlated with local GC content; however, recombination rate and recombination hotspots have a negligible effect on the frequency of mutation [63].

1.4.4 Selection

Selection can be divided into three main categories: 1) Positive selection or a hard selective sweep occurs when advantageous genetic variants rapidly increase in frequency, thereby promoting the

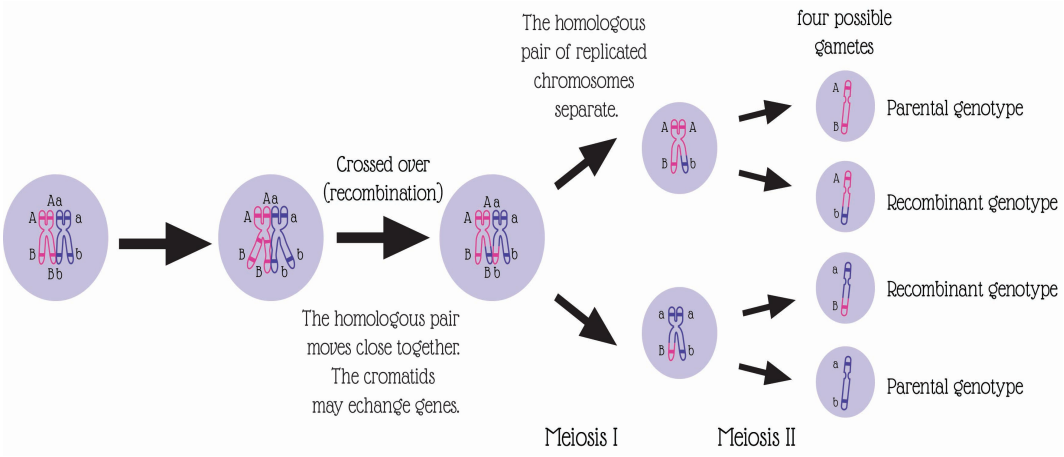


Figure 1.4: Crossing over or recombination, is the exchange of chromosome segments between nonsister chromatids in meiosis. Meiosis occurs in two stages, called meiosis I and II. Meiosis I divides homologues from each other. Meiosis II breaks up sister chromatids from each other. Figure adapted from Pearson Education Inc., (2012) [64].

spread of beneficial alleles in the population. 2) Soft selective sweep is when a neutral mutation becomes a weakly beneficial variation because of environmental changes. 3) Negative (purifying) selection removes deleterious variants from a population [60]. Selective sweeps thus work in opposite directions on negatively associated variants, resulting in Hill-Robertson Interference (HRI) [61].

HRI describes the effects of linkage in reducing the effectiveness of selection by segregating sites (Figure 1.5) [65]. The HRI effect shows that the efficacy of selection is reduced where there is diminished recombination [60]. The selective sweep could also mean that disease variation is the probable result of random mutation and penetrant monogenic variants. As a consequence, these variants are maintained at low frequencies by purifying selection [52]. Linkage disequilibrium between alleles at selected loci, caused by the stochastic nature of mutation and sampling in finite populations, 'interferes' with the action of selection at any one locus [59].

1.4.5 Mutation

Genes that have effective mutation rates might generate disease mutations and hence be associated with a disease [66]. Mutations arise from errors in DNA replication or spontaneous DNA alterations. Mutation rates depend upon many factors; sequence context, replication timing, transcription, expression level and recombination rate [67]. Mutations can be small or large scale insertions or deletions which may be missense, nonsense or frameshift mutations [68] (see Figure 1.6).

Mutation rates may also depend on the mutation rate per site according to gene length [69]. For instance, Eyre-Walker *et al.* [66] found that the size of the gene has an impact related to the disease. However, the mutation rate has historically been a complex genetic parameter to measure accurately due to the variability of sequenced genomes [70]. The sequence context may explain patterns of mutation rates. The substitution intolerance scores for genes improve the resolution

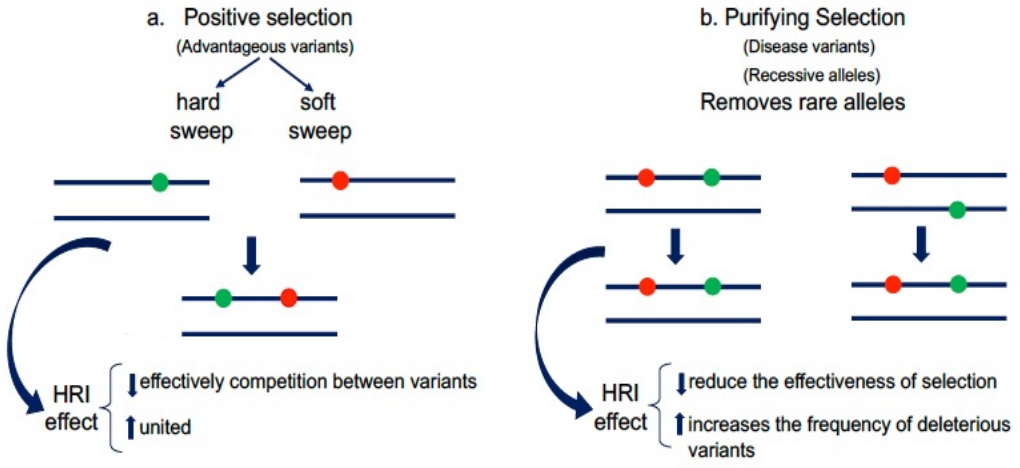


Figure 1.5: Associations between selected alleles may reduce the efficacy of selection. HRI decreases molecular adaptation and alters patterns of polymorphism in low recombination regions. The interference effect increases the frequency distribution of segregating sites to resemble that expected from more weakly selected mutations and generates specific linkage disequilibrium patterns. HRI has a significant impact on positive selection, reducing competition between beneficial alleles (fixing or losing one of the beneficial mutations), whereas, for variants subject to purifying selection, it diminishes the chance of eliminating potentially deleterious variations.

of substitution models and identify new mutation-promoting motifs. Aggarwala and Voight argued that the statistical substitution models used explain >81% of the variability in substitution probabilities across all substitution classes, covering 84% of all mutational events [71].

1.4.6 Gene essentiality

Essential genes are highly conserved and mostly encode proteins that drive basic cellular functions such as transcription, translation, DNA replication, cell division cycle control and fundamental metabolism [72]. Some mutations of essential genes could drastically alter phenotypes, almost without exception being lethal or deleterious. Mutations within 'non-essential' genes can be further defined as mutations that do not affect the phenotype in a particular environment [73].

Measuring the degree of gene essentiality of the proteins may quantify the tolerance for non-synonymous variants. However, estimating gene essentiality is challenging due to the protein structure complexity and the high density of interactions within a group of proteins. Recently, many approaches have been developed to study gene essentiality genome-wide, such as protein interaction networks integrated with gene expression or histone marks [72]. This information can be *a priori* knowledge of essential genes, which may promote the identification of disease-causing genes among multiple candidates.

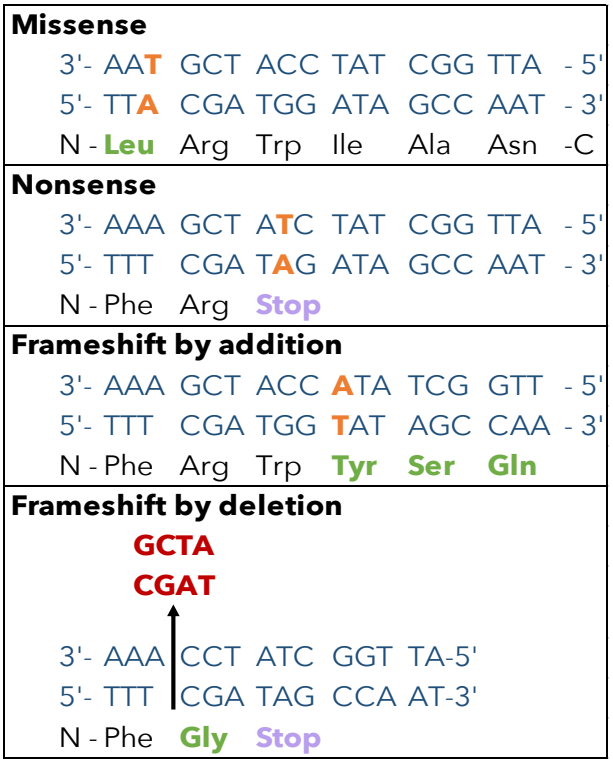


Figure 1.6: Substitution mutations include the change in a single base pair. Small deletions affect the function of only one gene. Missense mutations drive a change in a single amino acid in the protein. A nonsense mutation breaks a nucleotide base to a stop codon, resulting in premature translation termination to produce a truncated protein. Frameshift mutations include insertions or deletions of nucleotides, making a change in the reading frame. Figure adapted from Lodish *et al.* (2016) [68]).

1.4.7 Hypothetical relationship between gene essentiality, linkage disequilibrium, recombination and selection

The hypothetical relationships between gene essentiality, recombination and selection can provide insights into disease and non-disease genes (see Figure 1.7). The hypothetical model proposed by Pengelly *et al.* [53], shows the relationships between gene essentiality, recombination and selection. ∼72% (12,062/16,736) of the genes included in this study had a reduced LD, reflecting high recombination rates and potentially indicating that these genes may be associated with low essentiality and weakly affected by selection. The properties of these genes might include greater tolerance of mutation and include, for example, genes involved in sensory perception, such as genes encoding olfactory receptors. The high recombination rate regenerates the less harmful haplotypes, although residual variation is low and is unlikely to be associated with the disease.

The second group, which includes ∼9% (1,509/16,736) of the genes, has high essentiality, and its genes are associated with reduced recombination rates and strong selection. For these genes, any damaging variation is correlated with lethality. The measure of essentiality has a limited haplotype diversity and strong LD. This suggests that the genes are involved in the metabolism of DNA and RNA, damaging the cell's DNA cycle. The third group comprises ∼19% (3,165/16,736) of the genes, which contain or are affected by disease variation. These genes and their LD patterns are in

an intermediate position. They are genes with low-intensity recombination and selection, enabling identification of some deleterious variation associated with disease, which might involve these genes in common diseases.

Finally, the paper also highlights that genes are affected by recombination and selection. Thus, the impact of HRI reduces the efficiency of selection and Muller's ratchet permits the accumulation of deleterious variants.

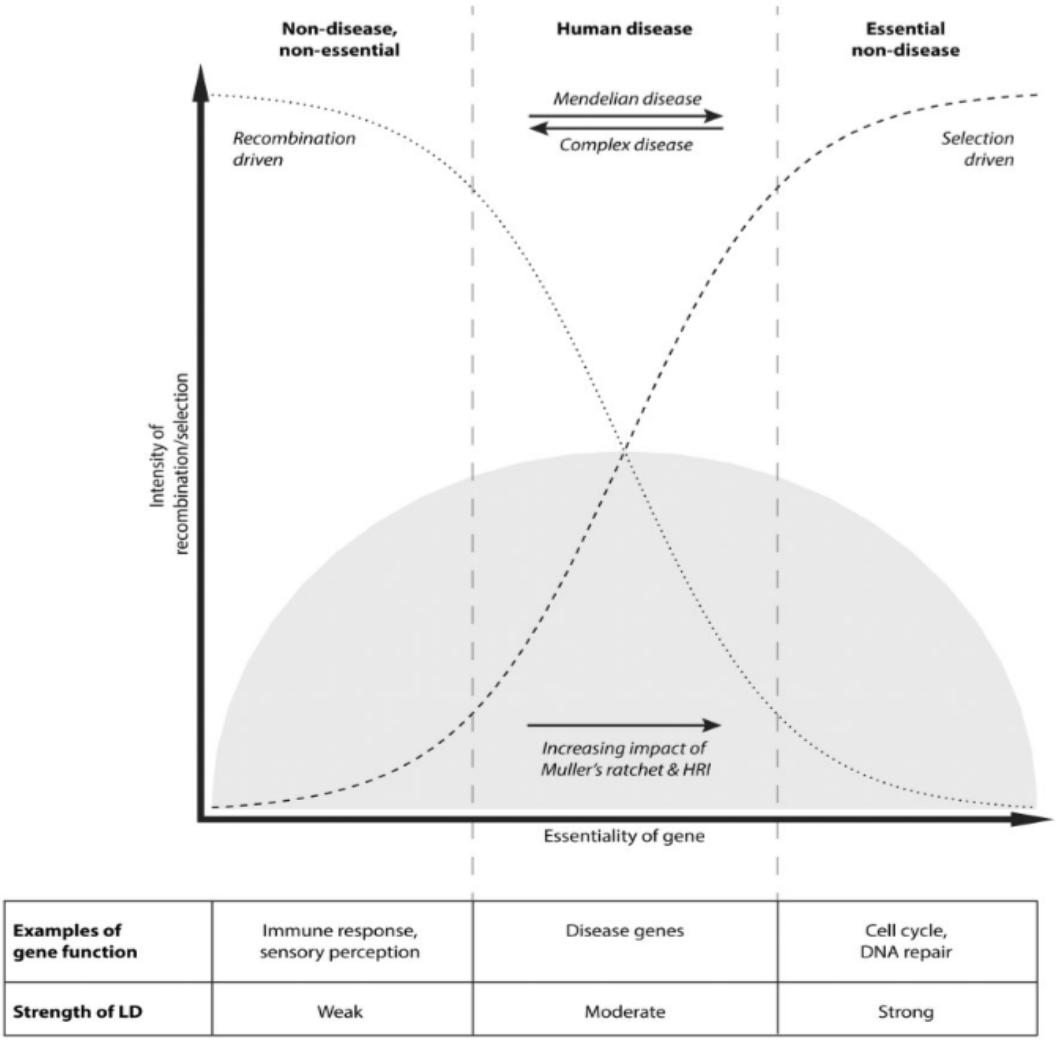


Figure 1.7: Hypothetical relationships between gene essentiality, recombination and selection. The dotted line denotes recombination, and the dashed line denotes selection. The shaded area indicates that deleterious variants are removed through recombination for 'non-disease, non-essential' gene groups and intense selection for 'essential non-disease' gene groups. Muller's ratchet describes the random accumulation of slightly deleterious mutations in finite populations with limited amounts of recombination. HRI means that linkage between sites under selection will reduce the overall effectiveness of selection in finite populations. Reprinted by permission from Pengelly *et al.*, ©(2017) [53].

1.5 Gene-specific metrics of disease genes

Establishing a causal link between gene function and human phenotypes may identify Mendelian and complex phenotypes. In order to do this, existing gene-specific predictors related to evolutionary and functional properties of the genes can potentially improve recognition of genes likely to be disease-related.

1.5.1 Machine learning algorithms

The role of machine learning (ML) algorithms or statistical learning is to find an empirical solution (target function) from the observed data. ML attempts to understand and predict mechanisms handling an automatic or semi-automatic process in which the algorithm corrects its outputs. Firstly, it assumes a target function to be learned (unknown) based on examples generated by the target. Therefore, the learning algorithm applies these examples to look for a hypothesis that approximates the target. This learning process follows the algorithm's rules, continuously updating as it receives more information to adapt the different inputs and then improve performance (see Figure 1.8) [74].

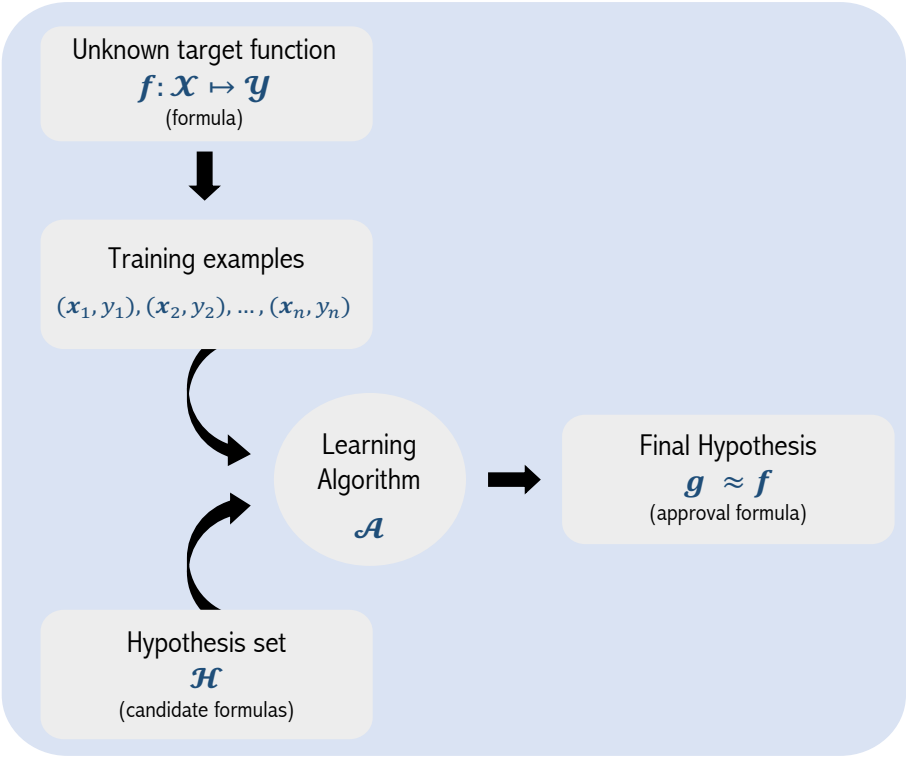


Figure 1.8: The figure illustrates the components of the learning problem. $\mathbf{x}$ represents the inputs, $f: \mathcal{X} \to \mathcal{Y}$ describes the unknown target function, where $\mathcal{X}$ is the input space, $\mathcal{Y}$ is the outputs space. There is a data set $\mathcal{D}$ of inputs-outputs examples $(\mathbf{x_1}, y_1), ..., (\mathbf{x_n}, y_n) where y_n = f(\mathbf{x_n})$ for $n = 1, ..., N$. Finally, there is the learning algorithm that uses the data set $\mathcal{D}$ to pick a formula $g: \mathcal{X} \to \mathcal{Y}$ that approximates f. The algorithm chooses g from a set of candidates formulas under consideration, called the hypothesis set $\mathcal{H}$. The decision will be good only to the extent that g faithfully replicates f. To achieve that, the algorithm chooses g that best matches f on the training examples. Figure adapted from Abu-Mostafa *et al.* (2012) [74].

As the basic premise of ML from data is using a set of observations to uncover an underlying process, there is a vast premise and challenging to fit into a single framework. As a result, various learning paradigms have arisen to deal with different situations and different assumptions. The most significant variations on the different types of ML have to do with the nature of the data set [74].

1.5.2 Supervised machine learning application in genetics

Supervised ML algorithms typically build an algorithm that uses a dataset of candidate features as input and then predicts a specific outcome. Their learning process is based on a comparison of the predicted labels to the provided ones. This implies that the input data includes parallel information about the correct output of each element. Supervised learning includes different algorithms to solve prediction and inference problems such as linear regression, support vector machines, decision trees or artificial neural networks.

Supervised ML methods have been applied to a wide variety of problems in genomics and genetics. For example, ML techniques can aid gene prioritisation, that is, recognising which genes are more likely to be associated or to interact to share a functional relationship [75]. The assumption behind applying ML techniques is that causal genes are similar to those already known to be associated with a disease. As a result, most strategies based on guilt-by-association need a set of seed genes to train a model. They then use the trained models to rank a set of candidate disease genes for the biological process, phenotype or disease under investigation [76].

Adie *et al.* [77] used a decision tree algorithm to identify disease-related genes based on the assumption that these genes underly human hereditary diseases, which share certain distinctive and sequence-based features. The authors selected evolutionary features such as conservation, coding sequence length, and closeness of paralogs in the human genome. Similarly, Xu J. *et al.* [78] proposed a classifier capable of identifying genes more likely to be involved in hereditary disease based on the topological patterns of genetic products in protein-protein interaction networks by using the K–nearest neighbour (KNN) algorithm. Radivojac *et al.* [79] proposed the PhenoPred algorithm for candidate genes using the human protein-protein interaction network, protein sequence and protein functional information at the molecular level. PhenoPred was built through a supervised framework using two layers of supervised vector machine (SVM). Yang *et al.* [80] designed a positive-unlabelled learning algorithm (PUDI) for disease gene identification. This algorithm combines biological process, molecular function, cellular component, protein domain and protein-protein interaction data and divides the negative likelihood set into four groups based on their likelihoods to be positives on gene affinity networks. Subsequently, a multi-level weighted SVM used these four sets for classifying disease genes.

1.5.3 Unsupervised machine learning in genetics

Unsupervised ML approaches are harder to interpret because there is no predefined outcome to be predicted. The task here is to derive an algorithm able to explore data patterns and to discover structure. Unsupervised ML algorithms try to represent and reduce the complexity of data by observing relationships between samples and features. This process is known as clustering, and dimensionality reduction [81, 82].

Various unsupervised ML algorithms have been explored to split genes up into different groups by their genomic and functional similarity. For such reasons, several different approaches to clustering have been proposed in the literature. For example, Oyelade *et al.* [83] considered functional information related to gene expression. This information reveals underlying structures and identifies patterns by using different clustering techniques such as hierarchical methods, hybridised k-means, model-based methods, soft clustering, etcetera. Lopez *et al.* [84] proposed an unsupervised machine learning method for recognising patient clusters based on their genomic similarity. Specifically, their method classified genetically distinct subtypes of patients within genomic datasets. Karmakar *et al.* [85] adopted a scalable version of the tight clustering method for significant gene expression data sets or other large data sets. It applies a truncation of a hierarchical clustering tree to overcome the local minimum problem in k-means clustering. They modified a tight clustering algorithm based on decreasing order of tightness by resampling the sample at each iteration. However, it is still challenging for unsupervised clustering to separate the genes when those genes may present noise or irrelevant genes (outliers) [86]. For example, traditional clustering algorithms such as hierarchical and k-means clustering are known to perform inefficiently for datasets with a small number of outliers [83].

Because unsupervised ML algorithms involve several parameters, often operate in high dimensional spaces, and cope with noisy, incomplete and sampled data, their performance can vary substantially for different applications and types of data. Thus, previous approaches for comparing the performance of clustering algorithms can be divided according to the nature of used datasets. As an example, a comparative analysis using a real-world dataset is presented by Souto *et al.* [87]. The authors conducted a comparative analysis considering five clustering methods: k-means, multivariate Gaussian mixture, hierarchical clustering, spectral and nearest neighbour. The authors used different proximities measures in the experiments, such as Pearson and Spearman correlation coefficient, cosine similarity and the euclidean distance. The algorithms were evaluated using the adjusted rand index for performance evaluation in the context of 35 gene expression data from either Affymetrix or cDNA chip platforms. The multivariate Gaussian mixture method provided the best performance in recovering the actual number of clusters of the datasets. K-means approach displayed similar results. In this same analysis, the hierarchical method led to limited performance, while the spectral method showed to be particularly sensitive to the proximity measure employed [87].

1.5.4 Bayesian estimation for Gaussian graphical models in genetics

Bayesian inference in Gaussian graphical models (BGGM) allows for learning conditional independence structures that are encoded by partial correlations among random variables (features) based on the Wishart prior distribution. This approach has been widely applied in genomics and proteomics to infer various types of networks, including co-expression, gene regulatory, and protein interaction networks [88].

Verzilli *et al.* [89] implemented a Bayesian graphical model to estimate candidate genes from GWAS. The authors included prior knowledge of the spatial dependencies to constrain marker nodes due to linkage disequilibrium. The proposed model was used in the simulation studies and analysis of the synthetic *CYP2D6* data. Zakery *et al.* [75] categorised genes by combining genotype and phenotype data sources using Bayesian matrix factorisation. This Bayesian data fusion method provides results on a variety of diseases such as those impacting the nervous system, metabolic diseases, and congenital malformations [75].

In addition to the above, Williams *et al.* [90] introduced a Bayesian method to estimate sparse matrices to estimate a protein-signalling network, in which conditional relationships are determined with a projection of predictive selection. The authors employed Kullback-Leibler (KL) divergence and cross-validation for neighbourhood selection to construct the network. Similarly, Yang *et al.* [80] built a classification model to tackle the feature selection problem. The authors proposed a computational method based on sparse Bayesian learning to produce a classifier and select highly correlated predictive features simultaneously.

1.6 Breast cancer

Breast Cancer (BC) is developed in breast tissues, most commonly from the inner lining of milk ducts (small tubes that carry the milk), leading to progressive aggregation of genetic and epigenetic changes in breast cancer cells. BC is a complex polygenic disorder caused by the interaction of genetic risk factors and environmental factors [91]. The genetic component is accompanied by other risk factors attributed to moderate to high-penetrance variants defining the BC phenotype and identifying causative or associated elements [92]. The two most important breast cancer susceptibility genes are *BRCA1* and *BRCA2*. These genes are expressed at different phases in the DNA damage response (DDR) and DNA repair. *BRCA1* is a multifaceted DDR protein that functions in both DNA damage sensing and DNA repair effectors, whereas *BRCA2* is a mediator of the homologous recombination repair. Inactivation of both leads to carcinogenesis [93, 94]. However, rare genetic variants account for only up to 5% of BC, indicating that a polygenic component is involved in disease liability. Much of this missing heritability may be either very rare highly penetrant genes not currently known or much larger numbers of rare genetic variants with small effect sizes. The cumulative effect of these genetic variants may be associated with increased relative risk [95, 96].

1.6.1 Polygenic risk score for breast cancer prediction

Breast cancer is a highly heritable disorder and is known to have a substantial common polygenetic component [97]. This component can be directly estimated through a polygenic risk score (PRS) which measures common genetic susceptibility to the disease in individuals regardless of their affected status [98]. A PRS also combines relevant SNPs and therefore predicts the risk of breast cancer. Recent studies have found that the effect of PRS on absolute risks of the breast can be higher in women in the highest quartile of polygenic distribution, with at least two-fold increased risk for BC compared to those with PRS in the lowest quartile [99, 100]. In the study conducted by Sawyer *et al.* 2012, the polygenic information subdivided the group of women with uninformative genetic testing results for monogenic causes for their ongoing breast cancer risk and the risk of collateral disease for previously affected women [101]. Data for multiple common susceptibility alleles for breast cancer may identify women at different levels of breast cancer risk. The authors developed a PRS that stratifies breast cancer risk in women with and without a family history of breast cancer. Based on these results, the level of risk discrimination can be used to inform targeted screening, and prevention strategies [100].

1.7 Thesis outline, aims and contribution

The overall purpose of the work herein is to develop and apply sophisticated mathematical, statistical and ML modelling techniques to large-scale genetic data. This introduction has aimed to provide a brief background on the genetic definition. The thesis also describes the mathematical, statistical and ML methodologies used to understand better the genetic background that will help the reader to contextualise what will be discussed in the following chapters. Since most of the results will be focused on analysing different data types, it is necessary to describe state-of-art from a research perspective in each chapter. Following the methods chapter, more details are provided on the bioinformatics tools, mathematical models and machine learning algorithms that will be extensively applied and cited throughout the result chapter. The method section is essential to facilitate an understanding of the sophisticated mechanism that would blur the reader's focus if it were covered individually in each result chapter. The work conducted in this thesis was mainly supervised by Prof Andrew Collins and Dr Reuben Pengelly. The following section presents a summary of the specific intentions of each research chapter.

Research hypothesis

Mathematical and statistical approaches can be used to analyse genome function, linkage disequilibrium structure and improve disease gene prediction for patient benefit.

Hypothesis 1

Mapping LD at high resolution may reveal the highly variable intensity of LD structure in exonic, intronic and intergenic regions, providing a novel understanding of the impact of recombination and selection on genome structure and function.

Aim 1 –Evaluate highly variable intensity of linkage disequilibrium in exonic, intronic and intergenic regions reflecting recombination and selection on fine scales

The first aim was focused on the construction of LD patterns from WGS to understand the historical impact of recombination, natural selection, genetic drift and mutation. It was demonstrated that accurate determination of the extent of LD at high resolution provides increased insights into the different regions of the genome, as discussed in detail in Chapter 3. My contribution consisted of carrying out every step described in the chapter, including curation of the research database, data quality control, data processing, LD map construction, model application and manuscript preparation.

This chapter was predominantly my own work, with significant input from Dr Reuben Pengelly and Prof. Andrew Collins contributing to research perspective and outputs interpretation.

Hypothesis 2

Gene-level metrics related to evolutionary and functional properties may improve recognition of genes likely to be Mendelian disease-related using supervised ML classifier and BGGM approaches.

Aim 2 – Estimate gene-level score of genome function to predict disease genes through supervised machine learning

Chapter 4 focuses on applying supervised ML algorithms to identify which determinants are acting on human disease-causing genes. Thus, different gene-level metrics related to the gene's evolutionary and functional properties were required to identify genes involved in Mendelian disease. My contribution was to develop supervised ML models, analyse the public data, perform statistical tests, and interpret results.

This work was predominantly my own, with significant contributions from Prof. Andrew Collins, Prof. Niranjan Mahesan and Dr Reuben Pengelly.

Hypothesis 3

The multi-objective clustering technique proposed using the unsupervised ML method can be used to predict and categorise known and novel Mendelian disease genes separate from complex diseases and non-disease genes with improved performance over state-of-the-art methods.

Aim 3 – Construct a robust prediction to identify disease genes using unsupervised machine learning

Chapter 5 covers applying a proposed unsupervised machine learning methodology using evolutionary and functional properties (discussed earlier in Chapter 4) as markers to classify and stratify

genes according to their degree of essentiality. I was responsible for the models' development and their application.

This work was predominantly my own, with significant contributions from Prof. Andrew Collins and Dr Reuben Pengelly.

Hypothesis 4

A PRS may quantify the cumulative effect of low-penetrance alleles on BC risk to examine whether non-BRCA breast cancer patients have a significant risk of developing BC.

Aim 4 – Estimate a polygenic risk score to quantify cumulative effect of low-penetrance alleles on breast cancer subtypes

The primary aim was to develop a polygenic risk score based on surprisal theory to measure the risk of BC within specific subtypes of BRCA mutation and non-BRCA breast cancer patients, and additionally to demonstrate that early-onset breast cancer (EOBC) has a strongly polygenic component. My contribution consisted of developing the PRS and its application. For the genomic data I was in charge of curating the research database, data quality control, data processing, the application of statistical analysis, and the interpretation.

I mainly carried out this work with significant input from Prof. Andrew Collins, Prof. Ben MacArthur, Dr William Tapper and Dr Reuben Pengelly.

Chapter 7 summarises the thesis findings and discusses future work.

Chapter 2

Methods and data analysis tools

The exponential growth in biological data has increased since the emergence of high-throughput technologies such as next-generation sequencing (NGS). This method has revolutionised biological research by providing a more comprehensive understanding of biological systems under study and the mechanisms that underlie disease development [102]. NGS's large amount of data requires applying advanced bioinformatics techniques and mathematical tools to obtain new insights about linkage disequilibrium and disease genes to improve diagnoses and design personalised treatments. This chapter provides an overview of the bioinformatics tools, mathematical and statistical models used in this research.

2.1 Malécot-Morton

Linkage disequilibrium (LD) research has underpinned the past decade of medical genetics research. LD patterns show the dominant process of recombination, mutation, selection, genetic drift and its cumulative effects of multiple historical bottlenecks. Maps expressed in linkage disequilibrium units (LDUs) are constructed using the Malécot-Morton model, which predicts the background levels of LD resulting from evolutionary history. LDU maps are analogous to linkage maps and discriminate blocks of conserved LD with additive distances. Newton Morton developed this methodology of LD drawing on the population genetics and geographical isolation work of Gustave Malécot to apply his models to loci in the genome. The model is based upon isolation by distance to separate populations by geographic distance. Formally, LD decreases exponentially with physical distance d in kilobases, and the Malécot-Morton model represents the association ρ between any pair of SNPs as

$$\rho = (1 - L)Me^{\sum d_i \epsilon_i} + L, \tag{2.1}$$

where the asymptote level L is the residual association at large distance, which acts as the correction factor for spurious association. M is the initial value of the LD before decay begins; that is, association at zero distance, with values of 1 consistent with monophyletic inheritance and less than one otherwise. ϵ is the exponential decline of LD with physical distance d in kilobases between

single nucleotide polymorphisms (SNPs). The product $\epsilon_i \cdot d_i$ at the i^{th} interval is equivalent to estimating the recombination product θ and time t in generations where recombination has taken place.

The parameters L, M and ϵ are not known, but can be estimated iteratively by using composite likelihood assumed from the observed pairwise comparison as:

$$\hat{\rho} = \frac{D}{Q(1-R)}, \tag{2.2}$$

where Q, R are the minor allele frequencies and the covariance D is obtained from the 2×2 rearrangement χ^2 contingency table. D is always a positive difference between a haplotype frequency and its equilibrium value as the product of allele frequencies [37, 103, 104, 105, 106, 41].

LD map generation was performed using the LDMAP program in C. The software LDMAP iteratively fits the Malecot-Morton model for values of $\hat{\rho}$ between multiple markers to identify the values of L, M, and ϵ provide the closest fit for the observed data. The software ultimately produces a map in LDU, equal to ϵd, such that one LDU corresponds to the (highly variable) physical distance over which LD declines to background levels.

2.2 Machine learning algorithms

Currently, statistical learning or machine learning (ML) algorithms are used for regression, classification, clustering or dimensionality of large datasets with high dimensions. ML algorithms aim to optimise the performance of a particular task by using examples based on experience. Usually, ML can be divided into two broad groups: supervised and unsupervised learning algorithms.

2.3 Supervised machine learning

Supervised machine learning is based on the same principles as probability distribution fitting. In supervised learning, the machine is given a sequence of desired outputs $y_1, y_2, \ldots$, attempting to find the unknown function linking known inputs to unknown outputs. The result for unknown domains is estimated by extrapolating patterns found in the labelled training data [107].

Supervised ML approaches select or engineer relevant features to predict the output accurately. This is accompanied by the selection of a supervised ML algorithm that will fit the desired output value. Applying the ML algorithm requires first generating, finding, and cleaning the data to ensure consistency and accuracy. Once the data pre-processing and model selection stages are completed, the model is iteratively improved to reduce prediction error using an optimisation technique. This entails adjusting hyperparameters that control the training process, structure, and characteristics of the model. Some authors consider model selection and hyperparameter tuning as parts of the same process. In other words, both the selection of the model and the hyperparameters are part of

the model selection. The validation dataset is separated from the test and training sets to optimise these hyperparameters [108, 107].

The two supervised ML approaches selected for classifying disease genes from non-disease genes are detailed in the next section. Further information regarding the supervised ML used in this research is given in Appendix A.

2.3.1 Random forest

Random forest (RF) is a supervised ensemble learning algorithm that constructs different models of several decision trees on the training set to improve the prediction performance. RF creates classification trees, and the bootstrap technique is applied to train each tree and therefore select the best solution by employing voting. This ensemble method reduces over-fitting by averaging the result.

Formally, RF is a combination of a collection of tree-structured classifiers $(h(\mathbf{x}, \theta_k)|k = 1, \dots)$ where $\mathbf{x}$ is an input vector, and θ_k are independent and identically distributed random vectors. Each tree shapes a unit vote for the most popular class at input $\mathbf{x}$. The algorithm draws a bootstrap sample $\mathbf{Z}^*$ of size N from the training data. Then a random forest tree T_k is grown to the bootstrapped data by recursively repeating the selection of m features at random from the p features for each terminal node of the tree. Thus, the best variable/split-point is chosen from among m possibilities. Next, the node is separated into two daughter nodes. The last steps are replicated until the minimum node size n_{min} is reached. The classification at a new point x is then estimated by

$$\hat{C}_{rf}^{k}(x) = \text{majority vote}\{\hat{C}_k(x)\}_1^K, \tag{2.3}$$

where $\hat{C}$ is the class prediction of the $k_t h$ random forest tree [109].

2.3.2 Gradient tree boosting

Gradient tree boosting (GTB) is a numerical optimisation problem where the objective is to minimise the loss of function by adding weak learners using a gradient descent–procedure. The idea of this methodology is to convert multiple weak learners into strong learners.

GTB is a generalization of boosting to arbitrary differentiable loss functions. GTB builds an additive model in a forward stage-wise fashion, which is written as

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \tag{2.4}$$

where $h_m(x)$ are weak learners. Thus, this algorithm uses decision trees of fixed size as weak learners. In addition, the model allows for the optimization of arbitrary differentiable loss functions. At each stage n class trees are fit on the negative gradient of the multinomial deviance loss function. The

l_2 regularization parameter is based on the loss function. GTB considers additive models of the following form:

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x), \tag{2.5}$$

where the newly added tree h_m tries to minimize the loss L, given the previous ensemble $F_m - 1$:

$$h_m(x) = \arg\min \sum_{i=1}^{n} L(y_i, F_m - 1(x_i) + h(x_i)) \tag{2.6}$$

The initial model F_0 is problem-specific; for least-squares regression, the mean of the target values is usually chosen [110].

2.3.3 Cross-validation

Each clustering algorithm is based on a set of parameters that require to be tuned for viable performance. A prevalent problem in ML is defining a proper procedure for setting the parameter values [111]. In general, an optimisation procedure can be applied to find these hyperparameters framework that provides the best performance of a given algorithm, like genetic algorithms [112]. Nevertheless, there are two significant problems with such an approach. First, fitting the parameters to a given data set can lead to overfitting [113]. Second, parameter optimisation can be unfeasible depending on the time complexity of many algorithms, combined with their large number of parameters [111, 114]. Therefore, different applications are required for evaluating and comparing the performance of clustering algorithms under default and optimisation situations.

In the following aspects of overfitting, the model could lead to high training error due to the complexity of the data or the features not adequately describing the output. An overfitted model might interpret a part of the noise in the training data as relevant information, resulting in a high variance in the model, thus failing to predict new data reliably. Similarly, an underfitted model may fail to fit the training dataset or generalise to a new data set, producing a high bias in the model. In order to avoid these two causes of poor performance, it is necessary to monitor the process during the training and validation to obtain an unbiased evaluation of a final model fit on the training dataset [113].

Cross-validation is a statistical method used to assess and compare learning algorithms by randomly dividing the dataset into two parts: one used to train or teach the model and the other utilised to validate the model's performance. The training and validation sets are crossed in successive rounds to validate each data point. Firstly, the training set is built as a sample of data to fit the model. Therefore the validation set takes place as a data sample to perform an unbiased evaluation for the fit model on the training dataset while tuning model hyperparameters. Then, the evaluation becomes more biased as a trade on the validation dataset is incorporated into the model configuration [115].

The simple form of cross-validation is k-fold cross-validation, while other forms are special cases, including repeated rounds of k-fold cross-validation. In k-fold cross-validation, the data is first randomly divided into k equal groups or folds. Consequently, k iterations of training and validation are performed, leaving out one fold for testing and training the model on the remaining $k-1$ folds. The process is repeated for every fold, returning an mean squared error (MSE) relative to that particular observation, and then the MSE is estimated as the average MSE over the k iterations by

$$CV_(x) = \frac{1}{k}\sum_{i=1}^{k} MSE_i. \tag{2.7}$$

The trade-off between bias and variance is then associated with the choice of k, and it has been empirically demonstrated that 5-to-10 fold cross-validation is the optimal choice for both variance and bias [115, 116]. Lastly, the final model must be assessed on previously unseen sample data, denoted a test set, to estimate its generalisation/extrapolation and performance.

2.3.4 Resampling methods

A dataset is class imbalanced if the categories are not represented approximately equally. Class imbalance is one of the problems that may arise if the data source provides unequal classes; examples of one class in a training dataset outnumber examples of the other class.

The class imbalance makes identification of the minority class by a learner challenging because this imbalance problem introduces a bias in the majority class. A biased learning process could classify all instances as the majority class and produce a false high-accuracy metric. Resampling strategies are methods for dealing with imbalanced domains. The resampling method draws repeated samples from the training set on each sample in order to balance them before training the classifiers. There are different types of resampling methods, such as oversampling and undersampling techniques. In the random undersampling method, all of the training data points from the minority class are used. Instances from the majority class are removed randomly from the training set until the desired balance is achieved [117].

2.4 Unsupervised learning

Unsupervised learning algorithms are models used when the observations are unlabelled, and the analysis aims to find patterns between features and samples. In unsupervised learning, the machine receives inputs $x_1, x_2, \ldots$, but no supervised target output is available. However, it is possible to develop a formal framework of the data, which assumes that the data points are independently and identically drawn from some distribution $P(x)$ [118]. These algorithms are also applied as methods for data visualisation or pre-processing before applying other supervised techniques.

Unsupervised learning models are utilised for two main tasks; clustering and dimensionality reduction. Clustering is a technique that groups unlabelled data based on similarities or differences to find hidden structures or patterns in the data. Clustering algorithms can be classified into a few types; exclusive, hierarchical, and probabilistic. Dimensional reduction is a technique that preserves the structure of the original dataset as much as possible while also decreasing the number of data inputs to a manageable size. This technique is commonly used in the preprocessing data stage. The most popular algorithms for dimensionality reduction are principal component analysis and t-distributed stochastic neighbour embedding [118].

2.4.1 Clustering

2.4.1.1 k-means clustering

K-means clustering is a clustering method in which data points are assigned into K groups, specifying that a data point exists only in one cluster. K denotes the number of clusters based on the distance from each group's centroid. The closest data points to a given centroid are grouped under the same class. K-means is one of the most popular clustering methods for analysing gene expression data [119].

2.4.1.2 Hierarchical clustering

Hierarchical clustering is an unsupervised clustering algorithm that can be categorised in two ways: agglomerative or divisive. Agglomerative clustering is considered a "bottoms-up approach." Its data points are initially isolated as separate groupings and then merged iteratively based on similarity until one cluster has been achieved. These clusterings are usually visualised using a dendrogram or a tree diagram that displays the splitting or merging of data points at each iteration [119].

2.4.2 Probabilistic clustering

A probabilistic model is an unsupervised technique that the data points are clustered based on the likelihood of belonging to a particular distribution. The Gaussian Mixture Model (GMM) is one of the most commonly used probabilistic clustering methods.

2.4.2.1 Gaussian Mixture Models

Gaussian mixture model (GMM) are classified as mixture models that are made up of an unspecified number of probability distribution functions. GMMs are primarily leveraged to determine which Gaussian, or normal, probability distribution a given data point belongs to. If the mean or variance are known, it can determine which distribution a given data point belongs to. However, in GMM, these variables are unknown, so we assume that a latent or hidden variable exists to cluster data

points appropriately. Expectation-Maximization (EM) algorithm is commonly used to estimate the assignment probabilities for a given data point to a particular data cluster.

Formally, GMM is a parametric distribution assembled from weighted multivariate Gaussian distributions. The density of each data point is a weighted sum of K component Gaussian densities which can be written as:

$$p(\mathbf{y}|\theta) = \sum_{k=1}^{K} \pi_k p(\mathbf{y}|\theta_k), \tag{2.8}$$

where the K components of the mixture are Gaussian distributions with differing means and covariances $\theta_k = (\mu_k, \Sigma_k)$ where π_k is the mixing proportion for component k, and they satisfy the constraint that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0, \forall k$. The complete GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities.

Since the data point assignment is not known, this describes a form of unsupervised learning. GMMs measure continuous features, and their parameters are estimated from training data using the iterative expectation-maximisation (EM) algorithm.

The expectation-maximisation algorithm for Gaussian mixture models starts with an initialisation step, which assigns reasonable values to the model parameters based on the data. Then the model iterates over the expectation (E) and maximisation (M) steps until the parameter estimates converge; i.e., for all parameters, θ_{kt} at iteration t satisfies $|\theta_{kt} - \theta_{kt-1}| \leq \varepsilon$ where ε is the tolerance error [120, 121].

K-means groups data points using distance from the cluster centroid; data objects are divided into non-overlapping groups. K-means using a pre-specified number of clusters based on expert knowledge. GMM generates density-based to assign data points to clusters. Each cluster is described by a separate Gaussian distribution [122]. In contrast, hierarchical clustering builds a hierarchy of clusters without having a fixed number of groups.

2.4.3 Dimensionality reduction

2.4.3.1 Principal component analysis

Principal Component Analysis (PCA) is an unsupervised linear dimensionality reduction and data visualisation technique for high dimensional data. The fundamental idea of this technique is to reduce the dimensionality of highly correlated data by transforming the original vector set to a new set known as principal components (PCs). This can be achieved by computing the eigenvectors corresponding to the largest eigenvalues of the data's covariance matrix and returning the data's projection on these eigenvectors, namely, the principal components (PCs). These PCs are given by an orthonormal linear transformation from a set of random features and therefore are ordered, with the first component retaining the most variation from the original variables. Application of this technique includes feature extractions, stock market predictions, and gene data analysis.

Formally, the uncorrelated PCs are represented by the linear projection of the original random vector of a p-dimensional variable $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ with covariance matrix Σ. Thus, the linear combinations of the columns of matrix $\mathbf{x}$ with maximum variance are given by

$$z_{(1)} = \alpha_1' x = \alpha_{11} x_1 + \alpha_{12} x_2 + \cdots + \alpha_{1p} x_p = \sum_{j=1}^{p} \alpha_1 j x_j, \tag{2.9}$$

where α_1 is a vector of p constants $\alpha_1, \alpha_2, \ldots, \alpha_p$ and $'$ denotes the transpose. The linear function $\alpha_2' x$ is uncorrelated with $\alpha_1' x$, having maximum variance, and so on at the k^{th} stage. Thus, α_1 is the eigenvector corresponding to the largest eigenvalue of Σ, and $\mathrm{var}(\alpha_1' x) = \alpha_1' \varsigma \alpha_1 = \lambda_1$ is the largest eigenvalue. In general, z can be denoted by the vector whose k^{th} element is z_k, the k^{th} PC, $k = 1, 2, \ldots, p$. Then

$$z = \mathbf{A}' x, \tag{2.10}$$

where $\mathbf{A}$ is the orthogonal matrix. Thus, the PCs are defined by an orthonormal linear transformation of $\mathbf{x}$ and the p PCs are orthogonal and ordered with respect to their variances (eigenvalues of the covariance matrix Σ). Hence, the eigenvectors are orthonormal; they satisfy the unit length $\alpha_k' \alpha_k = 1$, $k = 1, 2, \ldots, p$, and are called PC loading vectors. Since PCA is scale-dependent, the variables are often mean-centered and scaled to unit variance before PCA is carried out. With standardised variables, the correlation matrix is instead used to derive the eigenvector–eigenvalue pairs. The first principal component is their normalised linear combination that has maximum variance. Then the second principal component follows the same property, accounting for as much of the remaining variance as possible with the constraint that the second component is orthogonal to the first component; their projections will be uncorrelated. Then all subsequent principal components must be uncorrelated to those already computed, and the explained variance will decrease after an empirically derived number of components [123, 124].

2.4.3.2 t-distributed stochastic neighbour embedding

The method of t-distributed stochastic neighbour embedding (t-SNE) is a nonlinear dimensionality reduction algorithm and data visualisation technique. It embeds the points from a higher dimension to a lower dimension trying to preserve the neighbourhood of that point. This technique is commonly applied in music analysis, cancer research, bioinformatics, and biomedical signal processing.

t-SNE computes low-dimensional coordinates of high-dimensional data, converting the high-dimensional Euclidean distance between data points into conditional probability by:

$$p_{i|j} = \frac{\exp\left(\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(\|x_i - x_k\|^2 / 2\sigma_i^2\right)}, \tag{2.11}$$

where x_i and x_j are the data point in the Cartesian plane, the probability density distribution assumes a Gaussian distribution around each data point in the high-dimensional space and models

the target distribution of pairwise similarities (the joint probability) in the lower-dimensional space using the Cauchy distribution (Student t distribution with 1 degree of freedom) to calculate the joint probability between y_i and y_j as:

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|y_i - y_k\|^2)^{-1}}, \tag{2.12}$$

If the map points y_i and y_j correctly estimate the similarity between the high dimensional data points x_i and x_j, the joint probability q_{ij} should be close to p_{ij}. Then, Kullback–Leibler (KL) divergence between the conditional probabilities is iteratively minimised via gradient descent to evaluate the projection from the high-dimensional structure and joint probability distribution P to the low-dimensional representation Q as [125, 126, 127]:

$$\mathrm{KL}(P \parallel Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2.13}$$

PCA embedding preserves the global structure; however, the local information is lost due to the noise and the nonlinear structure. Conversely, t-sne preserves the local data structure by minimising the Kullback–Leibler divergence between the two distributions concerning the locations of the points in the map.

2.5 Bayesian inference in Gaussian graphical models

The application of Gaussian graphical models (GGM) comprises two main parts: the first is qualitative, given by the graphs, which represent the structure of dependency among the studied variables, and the second is quantitative, which refers to the conditional or joint variables of the same distribution. GGM estimates the conditional independence among a set of random variables using an undirected graph. In this graph model, nodes represent the random variables, and directed edges represent stochastic dependencies among the variables. A set of conditional distributions is assumed to follow a multivariate normal distribution [128, 129].

Bayesian inference in Gaussian graphical models (BGGM) [130] uses the G-Wishart distribution as *a priori*. This is a generalisation of the Wishart distribution, which is the conjugate prior for the precision matrix whose elements associated with edges, not in the underlying graph, are constrained to be equal to zero [131, 128]. Then the Bayesian inference estimates the sparse matrices[1] in which conditional relationships are determined with predictive projection [90]. Consequently, assuming the posterior joint density is normally distributed (asymptotic normality and consistency [132]), it allows for constructing credible intervals and computing posterior probabilities.

[1] A sparse matrix is a matrix that contains mostly zero values.

Formally, BGGM characterises the undirected[2] and conditional dependence structure of a set of random variables. Formally stated, the undirected graph is $G = (N, E)$, consisting of a node set $N = 1, \ldots, p$ and an edge set $E \subset N \times N$. Let $X = (X_1, \ldots, X_p)$ be a $n \times p$ matrix, where each X is a n-dimensional vector that is indexed by the graph nodes. The edge set E contains pairs (i, j), where $(i, j) \in E$[3]. Assume that the X_p follow a Wishart prior distribution $W(k, \epsilon \mathbf{I}_p)$ with k degrees of freedom and identity matrix $\mathbf{I}_p$. This prior distribution provides an analytic solution for determining E, and allows for conveniently drawing posterior samples. Thus, the Wishart distribution is the conjugate prior of the inverse covariance-matrix Σ_{-1}, where the $p \times p$ positive definite covariance matrix Σ, which is the posterior distribution, also has a G-Wishart distribution:

$$\Theta|\mathbf{X} \sim W(k + n, (\mathbf{S} + \epsilon \mathbf{I}_p)^{-1}), \tag{2.14}$$

where $\mathbf{S}$ is the sum of squares matrix $Y'Y$ and ϵ is a constant [128, 133, 134].

2.6 Surprisal theory

Surprisal analysis identifies the probability of different events of a physical system when these events may have a rich internal structure. The procedure assumes that a reference system is in a steady-state, to estimate a probability measure on the space. These bits of information are considered to be sufficient to characterise deviations of the distribution from the system due to the conditions imposed on the balance system. Then the procedure also uses the estimated probability to assess how unusual any other system is concerning the reference system that minimises the expected number of bits required to specify a system drawn at random from the reference system [96, 135].

Surprisal analysis is based on the principle of maximal entropy. Entropy can be defined as a measure of disorder in a physical system or uncertainty (lack of information) of a random variable; i.e., it quantifies the amount of information required on average to describe the random variable. Entropy decreases when the probability of the system of being in a particular state is much larger than the probabilities of being in any other state; otherwise, the entropy is minimal [135, 136].

Formally, let X be a discrete random variable with alphabet Ω and probability mass function $p(x) = Pr\{X = x\}$, $x \in \Omega$. Then $p(x)$ and $p(y)$ refer to two different random variables with different probability mass functions. The entropy $H(X)$ of a discrete random variable is defined by:

$$H(X) = -\sum_{x \in \Omega} p(x) \log p(x). \tag{2.15}$$

The logarithm is base two, and the entropy is expressed in bits. The entropy will then be measured in bits. For example, the entropy of a fair coin toss is 1-bit [136].

[2] An undirected graph encodes a factorisation of the joint probability distribution in terms of clique potentials.

[3] There is an edge between two nodes X_i and X_j if and only if X_i and X_j are conditionally dependent given all the other variables $\{X_k, k \neq i, j\}$.

2.7 Computational tools and resources

Iridis 4

Most of the analysis performed in this research was conducted using the IRIDIS High-Performance Computing Facility. Iridis 4 is the fourth-generation computational facility at the University of Southampton. The cluster is made up of:

- 750 computing nodes with dual 2.6 GHz Intel Sandybridge processors
- each compute node has 16 CPUs per node with 64 GB of memory
- 4 high-memory nodes with two 32 cores and 256 GB of RAM
- 24 Intel Xeon Phi accelerators
- 3 login nodes with 16 cores and 125 GB of memory
- 12,320 processor cores providing 250 TFlops peak
- 1.04 PB of raw storage with Parallel File System
- InfiniBand network for interprocess communication
- 12 × GPU nodes (2.6 GHz Intel Sandybridge 16-core nodes with ∼62 GB usable memory and 2 K20 GPU cards each)

Iridis 4 does not have a graphical user interface (GUI), and each operation or software must be executed using the bash command line. The Iridis 4 command line is based on the operating system Red Hat Enterprise Linux.

C

C is a procedural programming language. Dennis Ritchie initially developed it as a system programming language to write an operating system. C language includes low-level memory access to make it suitable for system programmings like operating systems or compiler development [137].

Python

Python is an interpreted high-level, general-purpose programming language with dynamic semantics. It is a multi-paradigm programming language, as it supports object orientation in combination with imperative, functional and procedural programming. The Python interpreter and standard libraries are available in source or binary form. The versions used for all the analyses performed on

this research were versions 3.5 and 3.7.3 [138]. The custom scripts ran on Iridis were in .py format, and some analyses were run on a personal computer in Jupiter notebook 6.2.0 format.

Packages

`SciPy` is a library of programs for mathematics, science and engineering [139]. It includes fundamental tools such as `NumPy` for array (matrix) calculations and `Matplotlib` for graphs and 2-3D plotting.

`Scikit-learn` is the main package for statistical learning in Python [140]. It includes algorithms for supervised and unsupervised machine learning with tools for choosing and validating parameters and models.

`Pandas` is a library that allows to you perform data manipulation and analysis in Python [141]. This library facilitates data manipulation and operations for numerical tables and time series. Pandas is built on top of NumPy.

R

R is a language and environment for statistical computing and graphics. GNU project is similar to the S language and environment, developed at Bell Laboratories by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification or clustering) and graphical techniques and is highly extensible. R can be extended via packages through the CRAN family of Internet sites covering multiple modern statistics. The version used in this dissertation was and R 3.6.3 (`http://www.r-project.org/`) [142].

Plink

PLINK is a free, open-source whole-genome association analysis toolset designed to perform a range of fundamental large-scale analyses in a computationally efficient manner. The focus of PLINK is purely on analysis of genotype/phenotype data such as study design and planning, generating genotypes, or copy number variation calls from raw data. Through integration with gPLINK and Haploview, there is some support for subsequent graphing, annotation and storage of results. PLINK was developed by Shaun Purcell at the Center for Human Genetic Research, Massachusetts General Hospital, and the Broad Institute of Harvard and MIT, with the support of others. The versions used in this dissertation were 1.07, and 1.9 [143].

Perl

Practical Extraction and Reporting Language (Perl) is a general-purpose, high level interpreted and dynamic programming language. Larry Wall developed it in 1987. The Perl is used widely for text

processing, like extracting the required information from a specified text file and converting the text file into a different form. Perl supports both procedural and Object-Oriented programming. The versions used in this dissertation were Perl [144].

2.7.1 Data and code availability

All data, code and results obtained in this study are available for download at the GitHub repository `https://git.soton.ac.uk/nvlg1e15/thesis_navlg`.

Chapter 3

Highly variable intensity of linkage disequilibrium in exonic, intronic and intergenic regions reflecting recombination and selection on fine scales

3.1 Introduction

Patterns of Linkage disequilibrium (LD) reflect the combined impacts of recombination, natural selection, genetic drift and mutation. The analysis of high-resolution of LD structure provides essential insights into human evolutionary history, the nature of recombination, and disease gene mapping [105, 145]. In addition, fine-scale LD analysis has improved the understanding of population structure and migration, biological mechanisms such as the nature of recombination hotspots, mutation and selection and the identification of sequence determinants that promote recombination [146, 43].

The majority of LD analyses have focused on the mapping of numerous common disease genes by developing arrays of 'tag' single nucleotide polymorphisms (SNPs). However, the cost-effectiveness and high sequence quality of whole-genome sequencing (WGS) have made it feasible to investigate the properties of genomes at high resolution. For example, Pengelly *et al.* [50] proved that LD maps from WGS produce up to 2.8-fold more regions of intense LD breakdown (which align with recombination hotspots) compared to array-based tag genotypes, which miss valuable information.

Developing LD maps at the gene and sub-genre levels may characterise the recombination rate, which may contribute to predict the degree of LD and target marker densities for genomic selection. Furthermore, improving the prioritisation of candidates genes and variants depends on

understanding the interplay between recombination and selection processes at genic and subgenic level [51, 53].

The construction of LD maps is based on the Malécot-Morton model [103, 104, 49]. This equation combines pairwise association data between SNPs to quantify the decline of LD across SNP intervals. LD map distances are additive and are analogous to the linkage map centimorgan (cM) scale, but expressed in linkage disequilibrium units (LDUs) where one LDU is the physical distance along the chromosome over which LD declines to 'background' levels (generated by genetic drift). Plots of the LDU scale compared to the physical kilobase (kb) maps show 'steps' where LD breaks down over narrow sequence intervals (often aligning with recombination hotspots) and 'plateaus' where LD is strong in regions which align with 'blocks' of low haplotype diversity [147].

Earlier construction of genome-wide LD maps using this approach for array-based data [106, 49], considering HapMap phase II [148], indicated that the CEU genome has 57,819 LDUs. The genome sequence spans ∼3,100,000 kb,[1] showing a average genome extent of LD (kb/LDU) of ∼54 kb. Nevertheless, this approach has lower resolution than WGS [50]. Therefore, improving the map resolution from WGS yields the analysis of the LD structure at sub-genre levels such as gene exonic and intronic sequences.

Previous analyses have estimated the intensity of LD along and within genes. These estimates found that recombination rates tend to be higher in intergenic regions close to genes compared with genic areas, implying that LD is likely to be stronger within genes [54]. Eberle *et al.* [55] observed that the extent of LD mostly increases in genic regions as compared to intergenic regions. This finding could not be explained purely by differences in the recombination rates, suggesting increased selection in genic regions. Their analyses also indicated that fewer than ∼3% (600) genes contribute to the observed excess LD in genic regions.

Recently, the recombination map at a fine-scale – 10 kb of resolution – has been analysed by estimating 10 kb bins classified as genic, intergenic, or at gene boundaries [56]. The authors found that the recombination rates are reduced in genic versus intergenic regions along with some sex-specific differences. Moreover, lower recombination rates were detected in bins having only exons. Additionally, intergenic regions close to genes have a lower recombination rate to the $5'$ ends of genes than the $3'$ ends.

Berger *et al.* [57] observed approximately 13.6% more LD in genic regions of the genome than in non-genic regions in their study based on array-based genotyping (684,990 SNPs). However, their results could have a bias in the average extent of LD because they do not correct for chromosome size dependence; therefore, these results may indicate higher recombination rates of smaller chromosomes [49].

In the present study, this chapter aims to estimate LD maps of the autosomal genome at a fine-scale based on WGS data from individuals in the Wellderly study [149]. To this end, LD patterns were computed at high-resolution sub-gene levels such as exons, introns and non-coding ribonucleic

[1] http://www.ensembl.org/Homo_sapiens/Location/Genome

acids (RNAs). LDU maps will be created using LDMAP in-house software. These estimates use information from the reference human genome build hg19/GRCh37 from the UCSC Genome Browser [150]. The UCSC Genome Browser files information specifying the locations of these features. The underlying objective is to augment the much-increased resolution of LD structure, which enables analysis of the LD structure on a very fine scale at the level of individual gene exons, contributing novel understanding of the impact of recombination and selection on genome structure and function.

3.2 Methods

3.2.1 Sample used

The data analysed consist of ∼60 million SNPs genotyped for chromosomes 1–22. Genomic data from the Wellderly study has been made freely available to the scientific community by Erikson *et al.* [149] from the Scripps Wellderly Genome Resource.[2] The WGS genotype data are from 597 unrelated individuals. The Wellderly sample is characterised by individuals aged 80 to 105 without chronic diseases and who are not taking any long-term medications. The Wellderly cohort has been collected over the course of eight years.

The demographic characteristics of this cohort are shown in Table 3.1. The majority of the healthy ageing cohort were enrolled between 80 and 85 with similar overall age distributions for females and males. The Wellderly individuals contain a low but significantly elevated rate of male smokers, and more than half of the total population do regular exercise.

The data was downloaded in tab-separated values (TSV) format. The file was converted using a custom script of Complete Genomics 'Small-Variant-Table' multi-individual files to VCF.

Table 3.1: Wellderly cohort data

Characteristic		**Wellderly Cohort**
Average age	years	84.2 (±9.3)
Gender	male	39.3% (±1.3%)
	female	60.7% (±1.3%)
Average weight (kg)	male	76.2 (±21.3)
	female	59.9 (±20.4)
Ever smoked	male	61% (±2.6%)
	female	42% (±2.6%)
Exercise	-	66.8% (±2.5%)

The 95% confidence interval is in parentheses. Taken from Erikson *et al.*, (2016).

[2] https://www.scripps.org/news_items/4757-scripps-wellderly-genome-resource-now-available-to-researchers.

3.2.2 Single nucleotide polymorphism processing

The Wellderly genomes were filtered to retain only 454 individuals of self-reported European ethnicity who were unrelated. Quality control (QC) metrics were applied to the Wellderly data to identify poorly genotyped SNPs. QC at this stage led to removing 7,238,157 SNPs from the analysis owing to poor genotyping quality (see Table 3.2). In order to avoid differences in deoxyribonucleic acid (DNA) quality and bias towards one genotype per sample, further QC was carried out as part of this study to remove an additional ∼15 million SNPs, which were genotyped in fewer than 95% of the samples, with no individuals having more than 5% of SNPs missing.

Following Pengelly *et al.* [50] and Purcell *et al.* [143], SNPs with 5% missing genotyping and variants that deviated significantly from Hardy-Weinberg equilibrium (HWE) (χ^2, P <0.001) were excluded. Since rare SNPs are uninformative for LD, previous studies have demonstrated that including these variants with low minor allele frequency (MAF) can bias LD estimates [151, 152]. The impact of excluding SNPs with MAF frequency cut-offs of <0.05 and also <0.01 were evaluated on chromosome 22 as an example Figure 3.4. The close similarity between <0.01 and <0.05 MAF cut-off were observed, but using a <0.01 MAF cut-off produces a 3.4% longer map and retains many more SNPs (103,367, compared to 70,579 SNPs retained using the MAF <0.05 cut-off). Therefore, all SNPs with a MAF of <0.01 were used for all subsequent work, which may help better resolve LD structure in genomic regions with higher recombination rates. Outcomes of the HWE and MAF screening are given in the results per chromosome. The heterozygosity rates were computed to identify sex, taking into account the frequency threshold of less than 20% for females. However, the X chromosome was not included in the subsequent analysis due to a more significant proportion of SNPs being lost on this chromosome; almost half of SNPs in both female and male failed >5% missing genotyping. Therefore, only autosomal chromosomes were used in all subsequent analyses.

Table 3.2: Quality filtering

Variant Filtration	**SNPs**
Raw data	59,818,579
SNPs with minor allele threshold (<0.01)	37,376,474
SNPs with missing genotype data (>0.05)	15,170,966
SNPs failing Hardy-Weinberg exact test (<0.001)	32,982
Variants after filtration	7,238,157

Genotype data were filtered out using PLINK version 1.07 [143] and VCFtools version 0.1.15 [153]. Samples were processed on IRIDIS 4, the University Southampton cluster, and required 16 processors nodes per chromosome.

3.2.3 Linkage Disequilibrium map construction

Prior to LD map construction, quality control data had been determined with sufficient numbers of individuals and SNPs. The next step was to compute LDU maps on autosomal chromosomes 1–22 by the LDMAP programme available at `http://soton.ac.uk/genomicinformatics/research/ld.page/`. LD maps based on the Malécot-Morton model combine pairwise association data between SNPs to quantify the variable rate of decline of LD with distance across SNP intervals. LD patterns are represented in the form of a metric map with additive LDUs distances. LDUs are analogous to cM,[3] reflecting accumulated recombination over generations. The LDU scale is comparable to the cM [103, 104].

Maps are expressed by LDU ratio length to the linkage map in Morgans showing the impact of the dominant process of recombination and its duration, genetic drift, selection, mutation, gene conversion and the partly cumulative effects due to demographic events (the effective bottleneck time) between populations [106]. Thus, the ratio of the LDUs estimates the number of generations over which recombination has driven the decay of LD along with the impact of other factors such as selection and systematic errors in estimating interface in the linkage map [37].

One LDU represents the distance in kb between loci on the chromosome over which LD declines to background levels [103, 105]. These units are constructed by estimating the decline of association between SNP markers. This LDU distance between SNPs represents the product of a small frequency of recombination θ and the effective number of generations t over which recombination has accumulated after one or more population bottlenecks.

The LDMAP algorithm implements a sequential program built from the Malécot equation. It analyses pairwise measures of association through a Newton-Raphson iterative process. The Malécot equation estimates LD map distances from SNP data in LDUs whilst an iterative process examines the convergence of composite $-2\ln$ (likelihood) for combining pairwise SNP data at every interval. This model is defined as:

$$\rho = (1 - L)Me^{\sum d_i \epsilon_i} + L, \tag{3.1}$$

where ρ is the probability of association between SNPs, the asymptote L accounts for the bias introduced by residual association at large distance, and M is the maximal association at zero distance. Here, a value of $M \sim 1$ means monophyletic inheritance (allele arising a single haplotype) and $M < 1$ means polyphyletic inheritance. d_i represents the physical distance (kb) between SNP_i and $\text{SNP}_i + 1$,b and ϵ_i at the i^{th} interval denotes the exponential decline of the LD in kb.

The LDU distances are estimated for every interval between adjacent SNPs. The LDU map distance is then a product of ϵ_i and d_i at the i^{th} interval. Thus the cumulative LDU for the whole map $\sum_{i=1}^{n} \epsilon_i d_i$ is the sum of SNP intervals. The Malécot equation combines the multiple informative

[3] A centimorgan is a unit of genetic distance. Two loci are one cM apart if there is a 1% chance of recombination between two markers in a given meiotic event.

SNP pairwise data to estimate the association ρ. The predicted association between SNPs and $\widehat{\rho}$ is fitted using composite likelihood:

$$lk = e^{-\Lambda/2}, \tag{3.2}$$

where

$$\Lambda = \sum K_\rho(\widehat{\rho} - \rho)^2. \tag{3.3}$$

In this way, each SNP interval uses a sliding window that weights association data from all SNP pairs in the region that includes the interval of interest. These measures are combined by weighting them according to their information (K_{rho}) so that interval 1–2 is given the biggest weight and 1–5 the smallest weight.The corresponding LDU distance for the interval is $\sum_i \epsilon_i d_i = 1$, where d_i is the physical distance in the i^{th} interval and ϵ_i is a Malécot parameter whose LDU distances are additive to form a map contour. However, LDU maps may include a small proportion of holes; that is, intervals with indeterminate values of ϵ_i. For those intervals, the maximum value assigned by default is 3 LDU [41].

To construct the LDU maps with the LDMAP programme, it was necessary to create an intermediate file. PLINK [143] was used to transform the pre-filtered VCF files into flat files (MAP/PED format). These files were converted into LDMAP input format (TPED format) for each chromosome. Subsequently, the new TPED files were split into overlapping subfiles to allow parallel processing. Having produced the intermediate data, the Malécot model was used to estimate values of ϵ_i for the whole genome.

At this stage, to enhance computational feasibility, optimise time and parallel processing, and minimise information loss at the ends of each segment, the genotype data was split into ∼25,000 SNPs segments, adding a 200-marker overlap at the ends of each segment. These segments had 25 markers trimmed off at the end of the LD maps segments and were then joined to complete the entire chromosome for assembled annotation. Map segments were submitted and constructed as individual jobs on the computer cluster. The simultaneous submission of all segments accomplishes the parallel processing. Plots of the LDU scale compare to the physical maps (kb) represent the weak LD as a 'step' interval (often aligning with recombination hotspots) and the strong LD as a 'plateaus' region (often aligning with 'blocks' of low haplotype diversity) [49].

3.2.4 SNP annotation and characterisation of gene-specific maps

The extent, or intensity, of LD computed as kb/LDU for all genes were obtained. The boundaries of the gene, exon, intron, intergenic and non-coding RNA regions were identified using the UCSC Genome Browser files available at https://genome.ucsc.edu/ based on the reference human genome build GRCh37 (UCSC name: hg19) released by the Genome Consortium in 2012 [150]. Since chromosome coordinates and context change depending on the completeness of sequences, it is essential to select the appropriate release consistently in the case of sample comparison. Obtaining the boundaries of these features into analogous locations on the LDU map, two algorithms were developed for: 1) Interpolating LDU locations and 2) Merging gene names.

The first algorithm computes the location (beginning and ending) of gene transcripts both on physical maps in kb characterised by a physical map of the reference human genome build on hg19/GRCh37 and LDU maps from the Malécot equation. Then, custom Python scripts were used for linear interpolation to convert the sequence positions of the boundary genes into corresponding locations on the LDU map. The left and right boundaries from the genomic intervals (starting and ending locations of the genes) along the physical map in kb were located for the whole genome. Once both positions were found, the LDU position was re-calculated employing a linear interpolation to have the real and accurate positions of the gene in the LDU maps. The resultant database contains the sequence positions of the boundaries of these genes in LDU and kb map locations. Although the LDU maps are not linear, the use of linear interpolation for analysis of high-resolution maps is justified over short distances.

The second algorithm computes a function to match the physical map in kb in the genomic interval based on the reference human genome build hg19/GRCh37 from UCSC Genome Browser, thus providing the boundaries of the new LDU location and the approved gene names for autosomal genes. Another custom Python script was developed to match the approved gene names using the NCBI Refseq gene definitions [154]. After matching to the approved names, the complete set of interval locations produced a map comprising 18,268 autosomal genes in the kb, LDU map and the extent of the LDU/kb gene. The canonical transcript was used for each gene.

3.2.5 LDU map analysis

Genic and intergenic regions were obtained applying the methodology of Berger *et al.* [57]. This method indicates that all genes that overlap with other genes are merged into smaller genic regions. Intergenic regions were taken as any areas flanked without overlapping by genic regions (see Figure 3.1). A custom Python script was run for all overlapping intervals merged in genic and intergenic regions. The total number of resulting intervals was 16,742. For calculations of these intervals, the script applies a binary search algorithm based on $O(\log(n))$. The idea is that, in a sorted array of intervals, if interval $[i]$ does not overlap with interval $[i-1]$, then interval $[i+1]$ cannot overlap within interval $[i-1]$ because the starting time of interval $[i+1]$ must be greater than or equal to interval $[i]$.

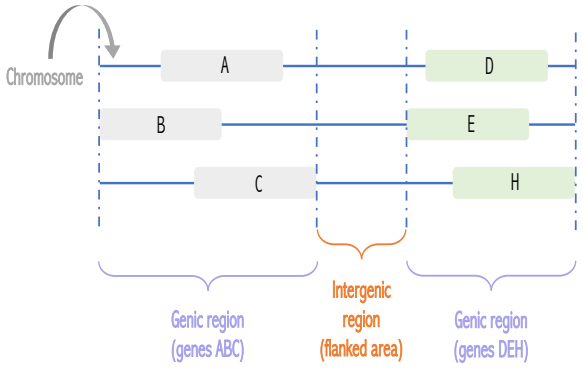


Figure 3.1: The general process of the integration sites between genic and intergenic regions. All genes which overlap with other genes were merged into smaller number of genic regions. Intergenic regions were taken as any areas situated next to but not overlapped by genic regions.

The LD maps should be estimated under the assumption that the number of generations over which recombination has accumulated is constant between chromosomal arms and their deciles. Moreover, the exclusion of heterochromatic regions from acrocentric chromosome p-arms was not included in the construction of LDU maps. To avoid bias due to the different properties of intergenic regions, all centromeric intervals between the last gene on chromosome p-arms and the first gene on chromosome q-arms of non-acrocentric chromosomes were omitted for subsequent analyses (see Figure 3.2).

Exon and intron boundaries were identified as genic and intergenic regions. Intronic regions were defined as the difference between exon base pairs. Data for the exons and introns were then partitioned into those transcribed on either positive or negative strands, genes in the $5'$ to $3'$ direction. The exon/intron analysis excluded the small number of exons, and introns involved in producing transcribed products on both forward and reverse strands.

Following Kong *et al.*, [56], the bins closest to the $5'$ end of autosomal chromosomes were excluded. The recombination into these bins is less reliable because introns have more extensive LD than exons. Non-coding RNA (ncRNA) data were also clustered if overlapping but were not distinguished from other genomic features (such as our definition of intergenic regions). For all output data, we removed the isoform repetition regions of DNA, taking the longest one. The extent of LD across genes was analysed independently in exonic and intronic regions due to variable gene size. The analysis consisted of dividing all genes into five bins oriented from $5'$ to $3'$, with equally sized bins for each given exon and intron within a gene. The location of the mid-point in the sequence of each exon and/or intron was used to sum the LDU and kb length of that exon or intron into the respective bin, that is LDU ($\sum_{n=1}^{18,268} LDU_{bin}$) and kb ($\sum_{n=1}^{18,268} kb_{bin}$), and then the extent of LD ($^{k}b/LDU$) was added into each bin (see Figure 3.3). To examine the impact of highly variable size LD extent profiles was constructed using the set of $18,268$ genes divided into two groups of $9,134$ genes, each corresponding to small genes of size < 23.5 kb and large genes of size > 23.5 kb.

Finally, the ratio kb/LDU was used to quantify the extent of LD in kilobases for any genomic region.

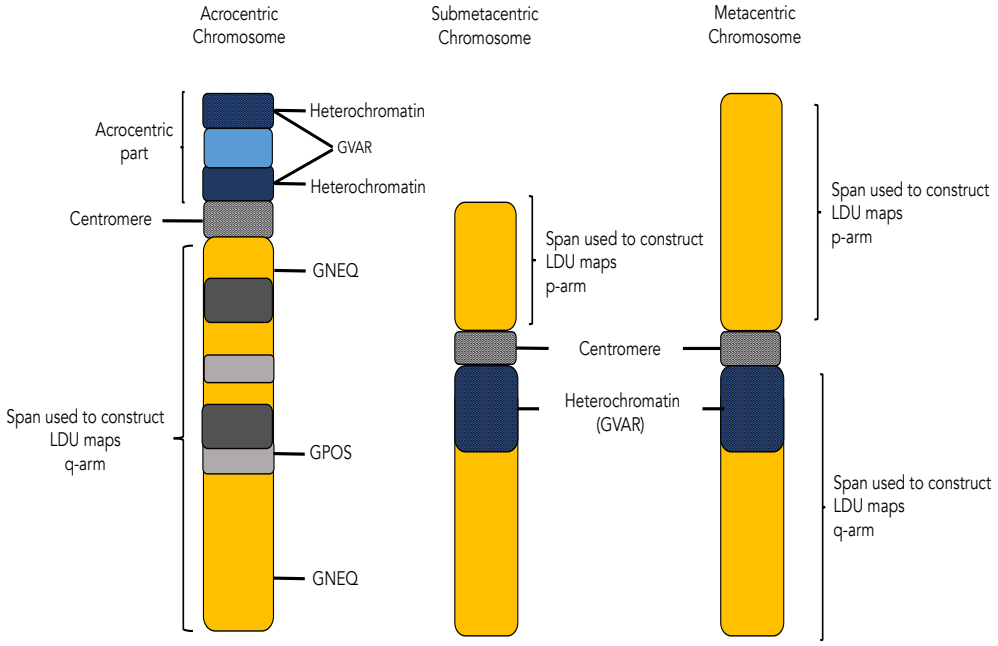


Figure 3.2: Chromosome Types. Acrocentric: centromere severely off-set from centre with shorter p-arm (chromosomes 13–15, 21, 22). Submetacentric: centromere off-centre, leading to shorter p-arm compared to q-arm (chromosomes 2, 4–12, 17, 18). Metacentric: centromere is in the middle, meaning p and q arms are of comparable length (chromosomes 1, 3, 16, 19, 20). Black and grey indicate Giemsa positive (GNEQ/GPOS). GPOS are classes containing progressively lighter staining G-positive bands, while GNEQ class consists of the nonstaining G-negative light bands. GVAR in blue determinates heterochromatin part. The yellow parts of the chromosomes (shown in the figure) were considered in constructing the LDU maps.

3.2.6 Variation in extent of LD for different gene groups

To examine the extent of LD and the relationship to gene essentiality and disease, the extent of LD at gene level was matched to one of the five gene groups defined by Spataro *et al.* [52]. Around fifteen thousand genes were matched, considering names and locations in both datasets. The gene groups are defined as:

- Essential non-disease (END) genes, 1572 putatively essential genes defined as orthologues of essential mouse genes detected by knock-out experiments and not involved in any human disease.
- Non-disease non-essential (NDNE) genes, 13,135 genes not known to be involved in any human disorder and not known to be essential.
- Complex non-Mendelian (CNM), 2388 genes uniquely associated with complex diseases.
- Complex-Mendelian (CM), 203 genes associated with both complex and Mendelian disorders.
- Mendelian non-complex (MNC), 684 genes uniquely causing Mendelian disease traits.

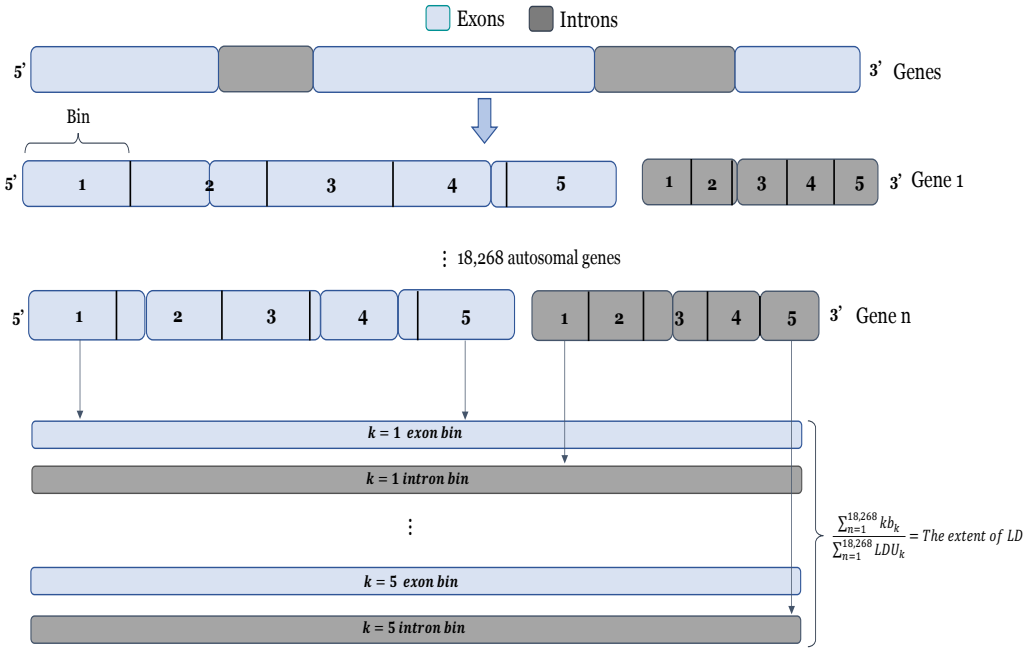


Figure 3.3: LD intensity across all genes at exonic and intronic levels. The exonic and intronic regions were analysed separately. All genes were divided into five bins oriented from 5′ to 3′, with equally sized bins for each gene. The location of the mid-point in the sequence for each exon and/or intron was used to add the LDU. Therefore, kb length of those exons/introns were located into the respective bin, and then the extent of LD was calculated per bin.

3.2.7 Data and code

Quality control analyses were performed using Plink v1.90p 64-bit [143]. The LDU maps construction and analysis were conducted in Python version 3.7.3 (`https://www.python.org/`), `awk` and R version 3.2.2 (`https://www.r-project.org/`) using custom-written scripts.

3.3 Results

3.3.1 SNP genotyping

SNP genotypes were obtained from WGS data from the Scripps Wellderly Genome Resource comprising 454 unrelated individuals of European ethnic origin from the Wellderly study [149]. Of the ∼60 million autosomal SNPs genotyped, ∼7.2 million (12.1%) remained after filtering, resulting in more than 20 million pairwise comparisons. Of those SNPs excluded, ∼32 thousand SNPs were excluded for not being in HWE and a further ∼37 million were excluded with MAF<0.10. Lastly, ∼15 million SNPs with missing genotypes above 5% were removed. The number of SNPs per autosome remaining after exclusions ranged from 103,367 to 610,013 and were closely related to chromosome length, as shown in Table 3.3.

Table 3.3: Quality filtering of SNPs

Chromosome	**Data**[1]	**MAF**[2]	**Geno**[3]	**HW**[4]	**Final Data**[5]
1	4,805,958	3,061,059	1,178,848	2,990	563,061
2	5,183,336	3,288,369	1,283,087	1,867	610,013
3	4,391,564	2,745,126	1,131,916	1,080	513,442
4	4,573,833	2,720,842	1,340,110	3,764	509,117
5	4,019,742	2,488,929	1,065,794	1,192	463,827
6	3,849,523	2,357,732	1,010,552	5,047	476,192
7	3,531,207	2,158,659	952,271	1,504	418,773
8	3,265,054	2,068,305	791,981	913	403,855
9	2,488,441	1,558,292	614,896	749	314,504
10	2,781,879	1,766,058	643,533	1,694	370,594
11	2,900,610	1,794,783	749,409	1,008	355,410
12	2,933,769	1,827,370	756,805	736	348,858
13	2,291,013	1,375,766	650,705	542	264,000
14	1,974,646	1,231,611	499,922	1,213	241,900
15	1,659,110	1,079,433	369,162	1,377	209,138
16	1,711,170	1,109,285	367,433	3,313	231,139
17	1,664,107	1,093,372	372,093	1,433	197,209
18	1,659,986	1,044,868	405,232	258	209,628
19	1,376,672	850,610	359,134	1,391	165,537
20	1,219,655	804,629	246,453	285	168,288
21	817,191	488,733	227,798	355	100,305
22	720,113	462,643	153,832	271	103,367
Total	59,818,579	37,376,474	15,170,966	32,982	7,238,157

[1] Raw data.
[2] Variants removed due to minor allele threshold (<0.01).
[3] Variants removed due to missing genotype data (>0.05).
[4] Variants removed due to Hardy-Weinberg exact test (>0.001).
[5] Number of valid variants which pass the filters and Quality Control.

3.3.2 Whole chromosome LDU maps and comparison with linkage maps

The extent of LD in the maps was determined for genes and intergenic regions and also for exons and introns within genes using the UCSC Genome Browser files (hg19/GRCh37). LD maps were constructed based on the Malécot mode explained in Chapter 2 for autosomal chromosomes 1–22. Each SNP has an LD location, and the distance between adjacent SNPs in the LD map was constrained to a maximum of 3 LDU. Following Tapper *et al.* [49], the intervals called holes were removed. Nine intervals holes were detected between adjacent markers where the upper limit on ϵ (set at $\epsilon_i d_i = 3$) was requested because LD is unlikely for $\epsilon_i d_i > 3$ and of unreliable for $\epsilon_i d_i > 2$. This implies that the LD map can be used to determine the local density of SNPs and that increased density within these holes is required to refine LDU map length.

Subsequent, the boundaries of these features were calculated to convert the sequence positions of these boundaries into corresponding locations on the LDU map. As a result, LDU locations were matched with approved gene names for 18,268 autosomal genes. The final LD map of the autosomes has ∼63,428 LDUs. The completed LDU maps of chromosomes 1–22 contain 7,162,973 SNPs spanning 2,791,110 kb, indicating a density of one SNP for every ∼400 base pairs (see Table 3.4).

Descriptive analysis of the LD maps showed that chromosome 21 has the smallest LDU length (∼1,110 LDUs) and 99,550 SNPs representing 1.39% of the DNA total. The largest LDU length is for chromosome 2 (557,873 LDUs) with ∼5.9 million SNPs, 8.3% of the DNA.

LDU/cM ratios were computed, and it was observed the effective number of generations over LD has declined consistently with the extent of LD to date. Following Zhang *et al.* [106], the effective population bottleneck time was estimated based on 63,428 LDUs for an autosomal euchromatic genome spanning 34.36 Morgans = 1,846 generations or 46,150 years since an effective bottleneck, assuming 25 years per generation. The 'swept radius', [4] for European population is 54.16 kb [(3,435.71 × 1,000)/63,427.68 = number of kb per LDU, where 1 LDU is the average extent of LD, see Table 3.4)].

Since rare SNPs are uninformative for LD, we evaluated the impact of excluding SNPs with alternative MAF of >0.01 and >0.05 on chromosome 22 as an example. The LD maps were very similar but using a >0.01 MAF cut-off produced a 3.4% longer map, as shown in Figure 3.4. The inclusion of many more SNPs in the analysis by excluding rarer SNPs (using a MAF >0.01 cut-off retains 103,367 SNPs, compared to 70,579 SNPs retained using the MAF >0.05 cut-off). Consequently, it was decided that MAF >0.01 would be applied to all SNPs for consecutive analyses.

There is a high correspondence between the linkage map in Morgans and LDU maps for all autosomal chromosomes (see Figure 3.5). The linear regression analysis showed that linkage maps explain ∼99% of variation in the LDU map for all chromosomes. The regression coefficient shows that the size of each chromosome map tends to be $1,784$ times longer compared to the smallest chromosome in centimorgans. $R^2 = 0.985$ for all chromosomes also shows highly high correspondence, showing

[4] the swept radius estimates the distance in kb at which LD declines to $e^1 \sim 0.37$ of its initial value [155].

that recombination dominates patterns of LD. These values are calculated as the average of the results for each chromosome, and the values are consistent across these chromosomes. A ratio of LDU weighted by centimorgans also can give the effective bottleneck time in generations.

Table 3.4: Characteristics of whole chromosome maps

Chromosome	Chromosome start location (kb)	Chromosome end location (kb)	Chromosome (kb) coverage*	Chromosome LDU length	Number of SNPs	Chromosome length (cM)**	LDU/cM
1	69.51	249,222.53	249,153.02	5,078.92	557,873	270.27	18.79
2	11.94	239,856.97	239,845.03	4,736.82	593,868	257.48	18.4
3	60.20	197,880.78	197,820.58	4,138.09	509,066	218.17	18.97
4	13.26	191,033.02	191,019.76	3,936.59	504,243	202.8	19.41
5	13.33	180,716.00	180,702.67	3,785.07	459,987	205.69	18.4
6	148.00	170,919.74	170,771.73	3,604.75	472,261	189.6	19.01
7	21.95	159,127.02	159,105.07	3,460.39	415,335	179.34	19.3
8	161.47	146,296.84	146,135.37	3,101.18	400,025	158.94	19.51
9	62.10	141,102.87	141,040.77	2,953.02	311,320	157.73	18.72
10	92.19	135,506.38	135,414.19	3,140.77	367,619	176.01	17.84
11	189.67	134,945.77	134,756.10	2,943.56	351,378	152.45	19.31
12	83.15	133,838.99	133,755.84	2,990.16	345,765	171.09	17.48
13	19,168.01	115,108.80	95,940.79	2,309.50	261,818	128.6	17.96
14	19,050.28	107,288.38	88,238.10	2,158.42	239,704	118.49	18.22
15	20,010.01	102,486.12	82,476.10	2,151.48	207,177	128.76	16.71
16	83.89	90,180.71	90,096.83	2,562.56	229,203	128.86	19.89
17	0.83	81,153.78	81,152.95	2,287.03	195,607	135.04	16.94
18	11.28	78,015.56	78,004.28	2,079.02	208,014	120.59	17.24
19	94.62	59,097.93	59,003.31	1,869.27	163,978	109.73	17.04
20	61.1	62,964.27	62,903.17	1,846.33	166,816	98.35	18.77
21	9,495.96	48,100.71	38,604.75	1,110.60	99,550	61.86	17.95
22	16,054.80	51,223.99	35,169.19	1,184.17	102,366	65.86	17.98
Totals/Chromosome mean	-	-	2,791,109.60	63,427.68	7,162,973	3,435.71	18.36

*Table includes all heterochromatic and centromeric regions except acrocentric p-arms, which were not sequenced.
**Kong *et al.*, (2010) [56].

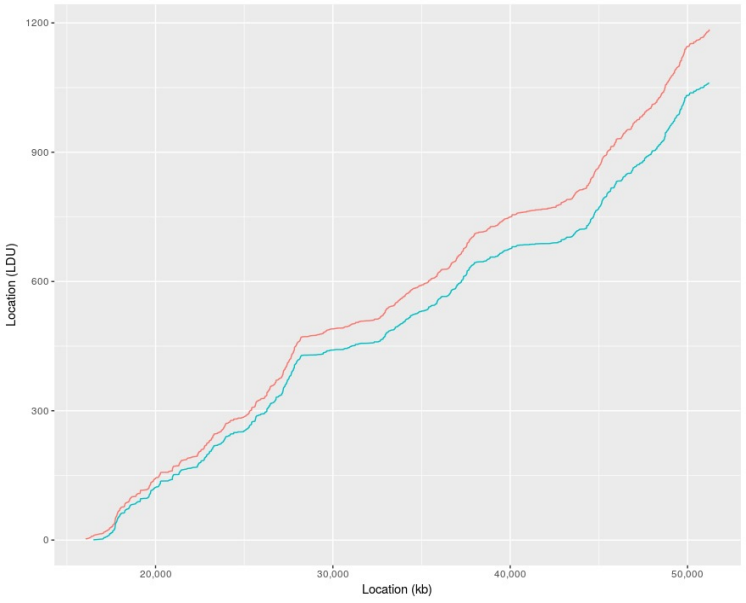


Figure 3.4: LDU maps for chromosome 22 at different MAF cut-offs. Excluding SNPs with MAF < 0.01 (red) includes many more markers but produces only a modest increase in map length compared to the cut-off at MAF < 0.05.

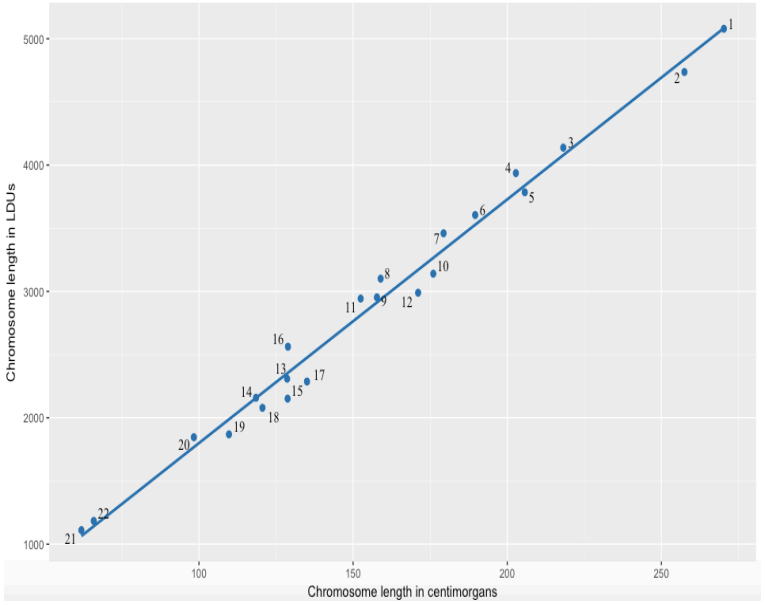


Figure 3.5: The genetic length in cM of the autosomal chromosomes (numbered), compared to LD length (LDUs), showing the close relationship between current and historical patterns of recombination.

3.3.3 Extent of LD in kb (kb/LDU) for genome regions

The total number of genic regions was 16,742, compared to 16,720 intergenic regions (see Appendix B Table B.1). Analyses were carried out for each genome region separately. The results showed (Table 3.5) that SNPs in genic region (introns and exons combined) make up ∼40% of the sequence, while this number is higher in intergenic regions ∼55%. The genetically inactive satellite DNA sequences regions enclose ∼4.3%. Moreover, ∼2.35% corresponds to the coding part of the genome, whilst intronic regions constitute 37.23% of the sequence (see Appendix B Table B.2 for additional details of whole chromosomes). The proportional distribution across the genome regions are shown in Table 3.6. Comparable LDU lengths are ∼38, ∼61 and ∼0.32% the greatly reduced LDU lengths in centromeric regions reflecting deeply suppressed recombination and therefore particularly strong LD. These extreme differences could be because the centromere is smaller in genic than intergenic regions (see Appendix B Table B.3 for more detail across chromosomes).

Table 3.5: Physical size of genome regions (kb)

Chromosomes	**Total**	**% coverage**
Whole chromosomes	2,791,109.60	-
Genic regions	1,121,267.26	40.17
Gene exons	65,524.61	2.35
Gene introns	1,039,236.88	37.23
Non-coding RNAs	334,132.38	11.97
Intergenic regions	1,541,437.90	55.23
Centromeric heterochromatin	120,967.95	4.33

*All calculations include centromeric regions.

Table 3.6: LDU size of genome regions

Chromosomes	**Total**	**% coverage**
Whole chromosomes	63,427.68	-
Genic regions	24,297.76	38.31
Gene exons	1,413.55	2.23
Gene introns	22,537.79	35.53
Non-coding RNAs	6,899.19	10.88
Intergenic regions	38,725.67	61.05
Centromeric heterochromatin	203.77	0.32

*All calculations include centromeric regions.

The extent of LD in kb (kb/LDU) showed similar averages between genome regions. However, the extent of LD in kb (kb/LDU) on centromeric regions (∼858.29 kb) is remarkably different from

the whole chromosome average (∼42 kb), as shown in Table 3.7; for example kb has a range of 139.52 kb in chromosome 19 to 3,773.83 kb in chromosome 1 (see Appendix B, Table B.4). A slight difference in the average extent of LD of ∼44.5 kb in genic regions was observed compared to ∼37.8 kb for intergenic regions. LD being about ∼16% more in genic regions compared to intergenic regions presumably reflects relatively reduced recombination and/or increased selection across genic regions (see Table 3.7).

Table 3.7: Extent of LD in kb (kb/LDU) for genome regions

Chromosomes	**Mean**	**Standard deviation**
*Whole chromosomes	42.03	6.40
Genic regions	44.54	7.23
Gene exons	46.39	8.20
Gene introns	44.49	7.42
Non-coding RNAs	46.52	7.61
**Intergenic regions	37.78	6.81
Centromeric regions	858.29	-
Gene exons + Non-coding RNAs	46.34	7.29

*Includes centromeric regions.
** Excludes centromeric regions.

The discrepancies in LDU lengths between exonic and intronic regions were analysed for individual gene exons and introns for all 18,268 genes, ignoring overlaps. For the extent of LD in exon and intron regions, a non-significant difference of ∼4% with $p = 0.078$ (paired t-test, 21 degrees of freedom; see Table 3.8) was observed. The results in Table 3.8 show that the extent of LD is longer in the exonic and intronic regions than in the intergenic regions, and the difference is highly significant ($p<0.001$). Non-coding RNA (ncRNA) regions comprise ∼11% of the DNA sequence, and the extent of LD across these regions is not significantly different from exons. LD is ∼4.5% more extensive than in intronic regions and ∼21% more extensive than in intergenic regions.

Table 3.8: Comparisions of extent of LD in kb

Variable1 (V1)	Variable 2 (V2)	Chromosome mean V1 (kb)	Chromosome mean V2 (kb)	Diff (kb)/ % difference	P (paired t-test, df=21)
Genic regions	Intergenic regions	44.54	37.78	6.76 / 16.40	<0.001
Exons	Introns	46.39	44.49	1.89 / 4.20	0.078
Exons	Non-coding RNAs	46.39	46.52	0.13 / 0.30	0.469
Exons	Intergenic regions	46.39	37.78	8.61 / 20.50	<0.001
Introns	Non-coding RNAs	44.49	46.52	2.02 / 4.50	0.005
Introns	Intergenic regions	44.49	37.78	6.71 / 16.30	<0.001
Non-coding RNAs	Intergenic regions	46.52	37.78	8.74 / 20.70	<0.001

*Pairwise comparison of length of LD in different genome regions in kb.
**P-value for differences in length of LD across all autosomal chromosomes.

The strong relationship between the extent of LD and chromosome recombination rate (see Figure 3.6) exceeds effects due to selection and mutation. This relationship shows elevated recombination rates across the smaller chromosomes (e.g., cM/Megabases (Mb)) displaying a markedly reduced extent of LD for these chromosomes. Intergenic regions show a significantly shorter extent of LD, intronic regions occupying an intermediate position. The observed patterns indicate elevated selection and/or reduced recombination in functionally sensitive genome regions. The similar patterns of strong LD observed for exonic and non-coding RNA (ncRNA) regions indicate increased positive selection, which might align with evidence for the functional significance of ncRNA regions.

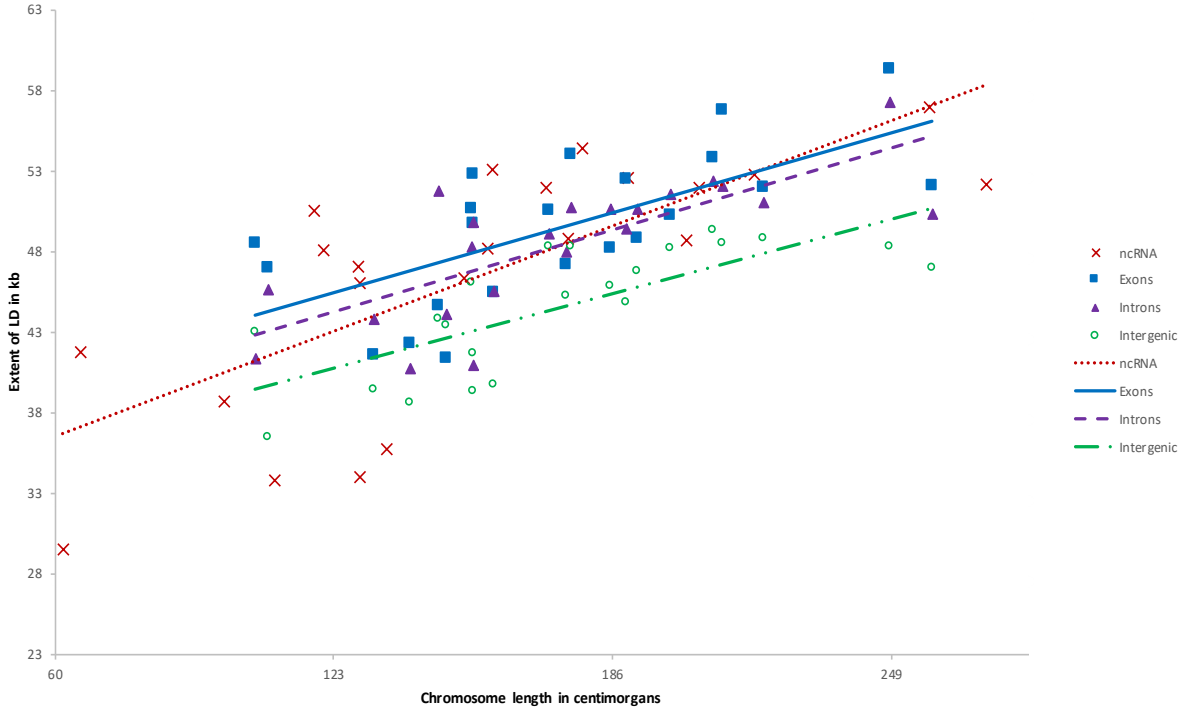


Figure 3.6: There is a strong linear relationship between chromosome genetic length and the extent of LD because smaller chromosomes have a higher rate of recombination per unit of physical length. Intergenic regions show a significantly reduced extent of LD, intronic regions occupying an intermediate position between exonic and ncRNA regions show the most extensive LD.The observed patterns indicate elevated selection and/or reduced recombination in functionally sensitive genome regions.

3.3.4 Variable extent of LD across gene endings from 5′ to 3′

Figure 3.7 summarises the relationships between the extent of LD in the exonic regions compared to intronic regions. For all bins except bin one, which is closest to the 5′ end of genes, introns have more extensive LD than exons. These results suggest that bins containing the first intron have higher recombination rates than the last intron or the first and last exons. For example, exons within the more central regions reflect high LD extending to ∼52 kb for bin 2 with a decrease in extent of LD towards the 3′ end. Introns show more uniform LD distribution across the gene, with a slight decrease in extent towards the 3′ end.

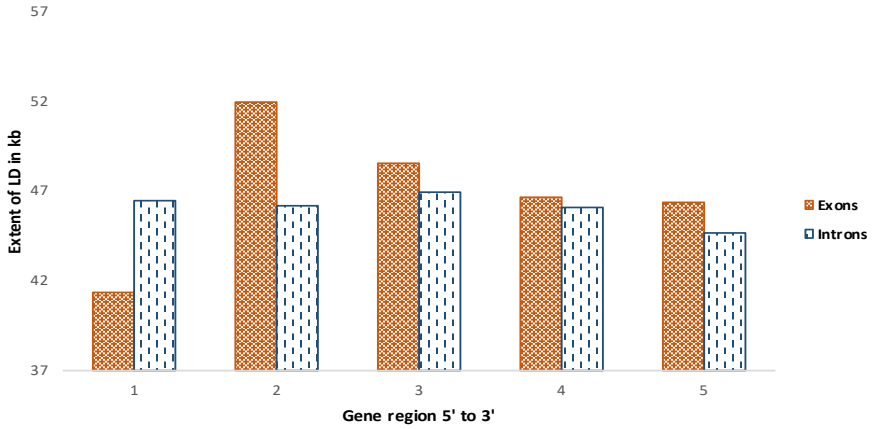


Figure 3.7: The profile of LD across all genes combined with LDU and kb data allocated to one of five (equally sized within a gene) positional bins. Exonic data shown in blue, intronic in red. The extent of LD is most variable for exonic regions: LD is less extensive, suggesting relatively increased recombination and/or reduced selection towards the 5′ end of genes.

Considering genes stratified into small and large size groups (see supplementary Figure 3.8), there was no significant difference in the extent of LD between exons and introns of small genes but increased evidence for a difference in some bins for larger genes. LD also extends further generally for large genes in both exonic and intronic regions compared with small genes. While larger genes may be subject to elevated selective pressure since they have a higher density of exons corresponding to multiple linked sites further studies are required to interpret this difference [156] fully.

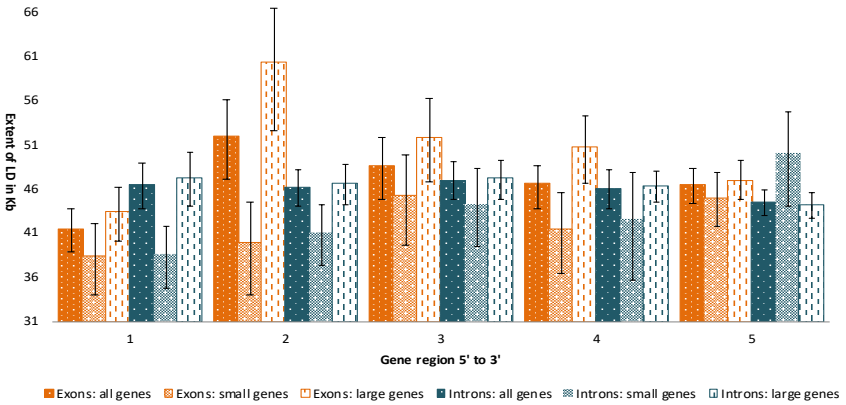


Figure 3.8: The profile showing extent of LD after allocating exon and intron mid-points to one of five regions 5' to 3' shows relatively elevated extent of LD fo exons, particularly in the second region. The sub-set of larger genes accounts for most of the difference.

3.3.5 Variable extent of LD across gene groups

END genes show significantly more extensive LD (53.9 kb) compared with genes implicated in complex phenotypes (CNM and CM groups, 40.9 and 35.2, respectively). However, the slight increase in extent of LD relative to MNC is not significant. Increased selective pressure in both END and MNC gene groups, relative to genes involved in complex phenotypes where variants have reduced phenotypic effect, might account for this difference. The large group of genes classed as NDNE also show extensive LD, although the extent to which some genes in this group are miss-classified because relationships with disease phenotypes and essentiality are not yet known is unclear (see Figure 3.9).

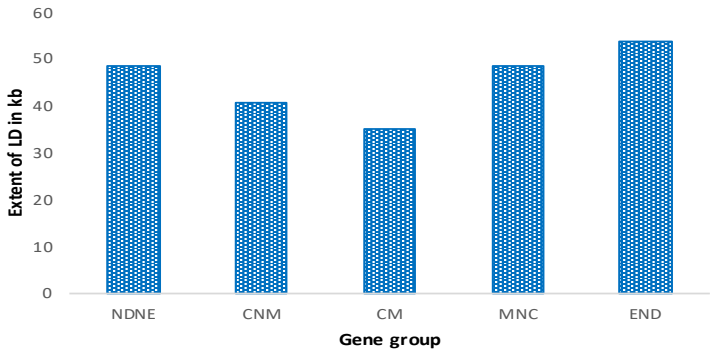


Figure 3.9: Groups of genes classified according to essentiality and relationship to disease phenotypes show wide variability in the extent of LD. Genes classed as essential (END) and Mendelian disease genes (MNC) show an elevated extent of LD compared with genes with variation related to complex disease phenotypes (CNM, CM). This might reflect elevated selective pressure within genes assigned to the Mendelian and essential groups. The large group of genes not known to be associated with disease and not known to be essential (NDNE) also show extensive LD, but this group is likely to include some miss-classified genes not currently identified as essential or disease-related. The 95% confidence intervals are shown.

3.4 Discussion

LD high-resolution maps from WGS were constructed in the present study showing estimates comparable to those from SNP array-based maps. The results of map length and the effective bottleneck are consistent with previous estimations. For example, Tapper *et al.* [49] estimated an effective bottleneck of 43,325 years which is comparable to ∼46,000 years in this study from the European population. This finding implies shorter LDU maps and more extensive LD.

The extent of LD over centromeres, as previously noted but not directly quantified, suggests that LD extends one megabase on average in these regions. These high-resolution patterns prove that the differences in fine-scale LD structure are detectable down to at least exon level, despite exons encompassing only 2% of the autosomal genome span of exons is ∼2% with an average exon size of just ∼300 base pairs. This resolution pattern is similar to the estimated pattern of Kong *et al.* [56]. They argue that a resolution down to 10 kb is effective for observing recombination patterns across a linkage map. Another related study to our results reported herein observed that the average of the extent of LD up to ∼42 kb can be detectable across autosomal chromosomes based on incompletely saturated maps from 'tag' SNP array data [49, 148].

Since the size of the chromosomes is a strong determinant of genome-wide recombination rates, substantial differences in the maps are presented here, indicating a significantly increased recombination rate of smaller chromosomes. In contrast, Berger *et al.* [57] did not show less extensive LD on the smaller chromosomes compare with larger chromosomes. For example, the LD extends an average ∼30 kb on chromosome 22 compared to ∼50 kb on chromosome 1. Similarly, Kong *et al.* [157] computed the recombination rate of chromosome 22 to be ∼2.1 cM/Mb compared to ∼1.1 for chromosome 1, and confirmed the close alignment between the extent of LD and recombination rate. However, it is worth noting that, among other differences between the two studies, Berger et al. [57] used the r2 metric to define pairwise LD.

The LD extent in genic regions is ∼16% greater than in intergenic regions. This result has a similar magnitude to that observed for the extent of LD shown by Berger *et al.* [57]. The authors found that genic regions are 13.6% higher than the intergenic region. Moreover, they have explained that genic regions are more conserved than non-genic and the higher haplotype diversity in non-genic regions may indicate that recombination has less effect on biological cycles or pathways. Thus, the results suggest that the stronger association observed between genic and intergenic regions might be a higher recombination rate in intergenic regions.

Haplotypes arising through recombination will have a more neutral impact on fitness and therefore reduced selection. Thus, regions with suppressed recombination are enriched to contain highly conserved genes with essential cellular functions showing an excess of mutations [62]. Similarly, Gibson *et al.* [51] found, in the opposite direction, high LDU/kb genes for which high levels of recombination might increase haplotype diversity in line with functions related to the immune system and sensory perception (where high levels of haplotype variation might be adaptive).

Nevertheless, regions or genes with low recombination rates have also been shown to have an excess of damaging mutations, including higher proportions of rare (MAF <0.01) and non-synonymous variants. This has a medical impact because recombination coldspot regions are more likely to hold rare variants. For example, Hussin *et al.* [62] have demonstrated that between ∼40 and ∼400-fold enrichment in disease-causing mutations have a higher recombination coldspot, which are more efficiently eliminated by natural selection.

Similarly, McVean *et al.* [54] reported that regions with weak LD and higher recombination rates tend to be impacted by selection, leading to a reduction in the accumulation of damaging variants. The binding *PRDM9* process seeks to fix the double-stranded breakthrough homologous recombination. This process breaks away from essential genomic functional elements that may have a protective role against potential mutagenic effects [46]. It has been shown that if the LD is more extensive in exonic regions, then these regions reflect selection against recombination within exons, which may imply that the recombination is mutagenic [158].

Empirical studies suggest that recombination rates influence positive or negative selection efficiency in moderate or high recombination regions. Still, there is no strong empirical evidence for the association between low recombination rates, which may permit the accumulation of damaging variation, and high recombination, which increases the rate of mutation and the interaction within the selection.

The analysis of LD profiles of genes enables an understanding of how recombination and selection are related, contributing to the identification of genes that are more likely to contain disease-related variation. Sun *et al.* [159] presented evidence to show how gene function is related to LD intensity, suggesting that these structures are related to the evolutionary history of genes based on their analysis of the LD structure of 'high' and 'low' LD genes.

Pengelly *et al.* [53] has demonstrated that elucidating LD structure at gene-specific level consists of studying the interaction between genomic properties with their functional elements through the identification of essential genes by observing intolerance to loss-of-function variants. Thus, genes that tolerate variation may carry a high number of variants with weak LD, whereas genes that are intolerant to variation will show a relative depletion, higher positive selection, reduced recombination, and strong LD [160].

The patterns of LD on exon-intron variation were consistent with the findings of Zhu *et al.* [161], which revealed that there is an ordinal reduction of length and divergence in both exons and introns from a $5'$ to $3'$ position. The authors calculated the correlation between exon and intron ordinal positions of 13 eukaryotic genomes and provided three main results: introns and exons being more extended or more divergent may reflect a time-orderly evolution; there are common factors that are likely to shape either exons or introns or both; and larger first and last exons may be splicing-required. Heterogeneity in patterns of LD intensity across genes provides another dimension to these analyses and are worth further investigation.

Several studies have suggested that mutation and recombination have created massive haplotype diversity in many species, including humans. These results imply that human-specific characteristics evolved primarily due to positive selection in non-coding regions involved in the regulation of genes. LD is 4.5% and 20.7% more extensive in ncRNA regions than in intronic and intergenic sequences respectively. It has been demonstrated that non-coding regions under selection are pervasively transcribed in critical regulatory mechanisms.

The ncRNA is implicated in chromosome conformation, regulation of enzymatic activity, coordination of cell state, differentiation, development, and disease, whereas selection acting on protein-coding regions is associated with immunity olfaction [162]. Furthermore, patterns of LD along ncRNA species should be researched further as it might indicate the relative functional importance of the sub-types.

In conclusion, all the results have shown here demonstrate that the LD patterns may provide distinguishable flags for understanding genome function at the sub-genic level, with conclusive differences between LD patterns within fine-scale levels such as exons. Further analysis of the LD structure at fine-scale in more sequence samples might contribute significant information to discovering candidate disease-causing genes and their biological implications.

Chapter 4

Gene-level score evaluation of genome function to predict disease genes through supervised machine learning

4.1 Introduction

Discovering the basis of genetic diseases has provided unparalleled opportunities to gain insights into the mechanisms linking genotype and phenotype to develop new diagnostics and therapeutics and transform healthcare delivery. However, the genetic basis for the majority of cases remains unsolved. Understanding the aetiology of Mendelian and complex disorders is far from complete and they are often poorly understood due to phenotypic diversity and complexity of gene–disease relationships [163, 52]. For example, the average diagnosis rate for monogenic diseases still remains low, ranging between 25 and 50% [164]. As of September 2020, 4,318 genes underlying 6,723 Mendelian conditions have been discovered, but the genes underlying ∼40% (i.e., 2,405) of all the known Mendelian phenotypes are still unknown, and many more Mendelian diseases have not yet been recognised [19].

Genes implicated in disease may differ in their genomic characteristics depending on whether they are associated with Mendelian or complex diseases. For example, genes that display allelic heterogeneity at their loci and low prevalence are more often implicated in Mendelian diseases [165]. In contrast, most complex disorders have a higher prevalence, low penetrance and are often involved in gene–environment interactions. Causal genes linked with both Mendelian and complex disorders tend to be more functionally important and expression levels in protein interactions are higher than the genes related exclusively to complex diseases. Moreover, genes that are implicated in both types of disorder have more significant effects on the likelihood of developing disease in comparison to those genes that are only correlated with the same complex disorders but not with Mendelian conditions [52].

In tackling this problem, it is necessary to understand disease aetiology and how genetic variation contributes to the phenotypes by incorporating gene-level information metrics. In recent years, ranking metrics have been developed to predict the functional effects of potential variant pathogenicity, based on the assumption that the genes that cause a particular disease have similar functions or in similar biological pathways to each other [166, 167]. These metrics or scores are frequently constructed from measures of genic intolerance of mutations using samples from the general population [168].

Several studies have integrated gene-level metrics related to evolutionary and functional properties to help recognise disease-causing variation. For example, Cacheiro *et al.* [169] proposed an approach for disease gene discovery across the full spectrum of intolerance to loss of function (FUSIL) score. They demonstrated that genes have different degrees of essentiality. Similarly, Spataro *et al.* [52] identified five main groups measuring the degree of essentiality. The first group is made up of a larger set of genes not currently known to be involved, neither essential nor associated with the disease; Non-disease non-essential (NDNE). The second group comprises known Mendelian non-complex (MNC) genes, the third and fourth ones contain Complex non-Mendelian (CNM) and Complex-Mendelian (CM) genes, respectively. The final group consists of Essential non-disease (END) genes that are not associated with NDNE genes or human disease (HD) genes. The systematic review of Alyousfi *et al.* [167] considered the literature related to gene-specific metrics and their applicability to improving the filtering of genome sequence data in order to identify disease genes.

Thus, this work aims to improve the recognition of genes likely to contain disease-related variation based on identifying gene-level metrics acting on human disease-causing genes. To this end, a supervised machine learning (ML) classifier and Bayesian inference in Gaussian graphical models (BGGM) were implemented to discern which of these metrics may help to identify genes that are potentially involved in Mendelian diseases. The gene-specific metrics related to evolutionary and functional properties were selected following the review of Alyousfi *et al.* [167]. These metrics described (i) essentiality and conservation (see Table 4.1), (ii) haploinsufficiency, (iii) genes under selection, (iv) recombination, and (v) genomic and biological characteristics. Moreover, for predicting gene diseases, the gene groups proposed by Spataro *et al.* [52] were used to categorise genes into NDNE, CNM, CM, MNC and END.

4.1.1 Properties of essential and conserved genes

Essential and conserved genes encode proteins that are required for the foundation of life. Genes at this level are enriched in fundamental biological processes, such as ribosomal ribonucleic acid (rRNA) processing, translational initiation, messenger RNA (mRNA) splicing and deoxyribonucleic acid (DNA) replication [170]. However, when **essential genes** include **loss of function (LoF)**[1] variation, this compromises organism viability (lethal to the cell) or results in a profound loss of fitness [171, 172]. Essential genes are more evolutionarily conserved, and therefore are subject to a

[1]Loss of function variants in genes are defined as those which impair or eliminate the function of the encoded protein.

more intense purifying (negative) selection than non-essential genes [173, 174]. Thus, the 'degree' of essentiality can be defined as a spectrum from essential to non-essential based upon the degree of tolerance to LoF mutation [175, 53]. Measuring patterns of intolerance to LoF variation in human genes might aid in recognition of the more essential genes, which might be associated with a wide range of diseases [167, 176].

Current gene-level scores aim to predict disease-causing variation by considering genic intolerance such as the loss intolerance probability (pLI) score obtained from the Genome Aggregation Database [177] dataset. This measure describes the probability of a gene being LoF intolerant, taking into account the frequency of synonymous variants. Another example of a ranking score is the residual variation intolerance score (RVIS), which categorises genes by the probability of carrying more, or less, functional genetic variation than expected compared to common functional variation [173, 166, 167]. Another score is the gene constraint *de novo* excess (DNE), which estimates the expected rate of *de novo* mutation excess per gene and gene set by calibrating a model of *de novo* LoF mutation [178].

A different method for measuring essentiality, given LoF variation, is the substitution intolerance score (SIS). Genes that show high SIS scores are functionally constrained and are thus more likely to contain pathogenic variation. In contrast, genes with low scores are highly tolerant of functional changes in the protein [71]. These metrics measure whether a given gene is likely to be a candidate for a specific genetic disease.

4.1.2 Properties of haploinsufficient genes

Haploinsufficiency (HI) describes a dominant phenotype caused by a heterozygous loss-of-function mutation. HI genes have highly conserved sequences and greater gene expression during early development, which leads to a sharp increase in tissue specificity [179]. Recent research suggests that haploinsufficiency is more often found for essential than non-essential genes. HI is one of the principal characteristics of Mendelian disease genes with a dominant inheritance pattern; therefore, several methodologies have been generated for predicting haploinsufficiency. These incorporate distinct biological properties by recognising LoF-tolerant genes, protein–protein interactions and dominant and recessive diseases [180].

Khurana *et al.* [181] devised the gene position in networks (NET) score, which quantitatively estimates the global perturbation caused by deleterious mutations in each gene. The NET score classifies functionally essential and LoF-tolerant genes to explain gene importance. Similarly, Steinberg *et al.* [182] proposed an unbiased HI score, the genome-wide haploinsufficiency (GHIS) score, which replaces biological networks with co-expression networks. Huang *et al.* [183] provided another important metric for predicting haploinsufficiency, which discriminates pathogenic and benign deletions among healthy individuals. The HI score gives the probability of potential candidate genes for causing dominant traits. McArthur *et al.* [184] calculated the recessive (REC) score, developed to distinguish genes with LoF tolerance from genes associated with recessive diseases.

4.1.3 Properties of genes under selection

The role of selection on patterns of genetic variation is termed positive selection if variants are advantageous, so that beneficial alleles are selected for and thereby increase their prevalence in the population. Negative selection, on the other hand, acts to remove deleterious alleles and thus reduce their frequency [52, 53]. However, deleterious alleles may not be entirely removed due to a balance between selection and mutation rates. This balance can be altered by alleles showing dominant/recessive properties and genetic drift. Models of the interactions between these forces have produced several tests for natural selection [185]. There are, therefore, some approaches to recognising genes that are more strongly impacted by negative and/or positive selection. These measurements might provide information on which genes are more likely to have variation that could have damaging consequences.

One approach is termed the gene-level integrated metric of negative selection (GIMS) score, which shows the selection intensity probability distribution in quantiles across the entire genome. Genes in the lowest quantile are scored under negative selection. GIMS combines multiple comparative genomics, functional genomics, and population genetic metrics to estimate the enrichment of negative selection for each gene [186].

Another score is the gene damage index (GDI), which predicts the mutational damage profile accumulated by each protein-coding human gene in both monogenic disease patients and in the general population. The GDI measures the combined influences of drift and selection. Genes with high GDI tend to be under less intense purifying selective pressure. A low GDI score is associated with highly conserved genes, reflecting essentiality [187].

4.1.4 Properties of genetic recombination

Homologous recombination is a fundamental biological process that might help to identify potentially disease-associated variants along the whole genome. Recombination is the exchange of genetic information between paired homologous chromosomes through the process of meiosis, which breaks up chromosome segments, increasing haplotype diversity [53, 42]. In the absence of recombination, deleterious mutations which accumulate on haplotypes cannot be eliminated by recombination – a process termed Muller's ratchet [62]. Additionally, homologous recombination is often associated with DNA repair and replication, which may also control the transcriptional context of essential genes [188]. Genes with high essentiality tend to have low recombination rates, and so purifying selection may be intense because, with increased essentiality, any deleterious variation is correlated with lethality [189].

Methods that recover the recombination structure using population data include the LDhat program, approximating the high-resolution linkage map. This method estimates recombination rates in the presence of hotspots, which might provide a basis for discriminating between genomic regions subject to selective sweeps and those with increased or reduced recombination [190].

4.1.5 Pattern of linkage disequilibrium

An alternative measure that recovers substantial information on local recombination rates and identifies recombination hotspots involves the construction of Linkage disequilibrium (LD) profiles using single nucleotide polymorphism (SNP) data. LD describes the nonindependence of alleles; this pattern is highly determined by recombination over many generations, although it is also impacted by selection and mutation. Recombination and mutation tend to increase the diversity of haplotypes, and therefore act to reduce LD locally; in contrast, selection tends to increase LD [191, 192]. The pattern of LD is an outcome of these processes and may have a close relationship with the disease genome [27].

Vergara-Lope *et al.* [192] examined the extent of LD and the relationship to gene essentiality and disease following the Spataro *et al.* [52] gene groups. They found that the extent of LD is more elevated in genes that are classified as essential and associated with Mendelian disorders compared to genes linked with complex traits [192]. Similarly, Gibson *et al.* [51] and Collins *et al.* [27] establish that genes with strong LD are enriched for essential functions (e.g., phosphorylation, cell division, cellular transport and metabolic processes), and genes with weak LD are enriched for functions related to sensory perception and some immune functions.

4.1.6 Functional genomic properties

Each disease-associated gene has a unique set of functional genomic properties. These properties can be described by measuring gene products such as coding transcripts, non-coding RNAs, gene expression levels, replication timing, levels of meiotic recombination and chromatin structure. These measures may potentially quantify the biological process and improve the filtering of genes associated with a phenotype. For example, Hsu *et al.* [180] found that autosomal recessive disease-associated genes tend to have more non-coding RNA isoforms and nonsynonymous variants [193]. The authors also observed that these recessive disease-associated genes had more complicated regulatory processes and tolerance of genetic variation. They have also demonstrated that early-onset disease-associated genes may also have *de novo* mutations with merely neutral effect [67]. Additionally, GC content is one of the significant promoter elements that control open chromatin status and support paused transcription [194]. Paused transcription can explain the regulation in gene expression in all domains of life, so understanding it is highly important [194, 195]. Hence, linking functional properties under essential human genes, non-disease genes, and genetic disorders could provide valuable information for gene annotation.

Table 4.1: Gene-specific metrics

Features	Definition	Method	Magnitude	Properties	Literature
pLI	pLI is the probability of being loss-of-function (LoF) intolerant	Recessive, where observed is 50% of expected (heterozygous LoFs are tolerated). Haploinsufficiency, where observed is <10% of expected (heterozygous LoFs are not tolerated). Posterior probabilities from a Poisson mixture model	I	Essential and conserved genes	Samocha *et al.* (2014) [178], Lek *et al.* (2016) [177]
RVIS	Residual variation intolerance score ranks genes from common missense and LoF variants versus the total number of protein-coding variants regardless of their frequency in the genetic population	A gene with a positive score has more common functional variation. A gene with a negative score has less and is referred to as 'intolerant'. Linear regression	I	Essential and conserved genes	Petrovski *et al.* (2013) [173]
DNE	Gene constraint *de novo* excess score recognises constrained genes using a neutral mutation model as a baseline	Gene-specific probabilities for different types of mutation: synonymous, missense, nonsense, essential split site and frameshift. Z-score	I	Essential and conserved genes	Samocha *et al.* (2014) [178], Hsu *et al.* (2016) [180]
SIS	Substitution intolerance score measures essentiality based on sample ascertainment, population history, selection and local context features that influence the rate of mutation	Probability that a nucleotide substitution occurs at a genomic site varies. Posterior substitution probabilities	I	Essential and conserved genes	Aggarwala *et al.* (2016) [71]
HI	Deletion-based haploinsufficiency score examines copy number variation in biological properties (genomic, evolutionary, functional and network) among many healthy individuals	Gene-based probability based on biological properties. Linear discriminant analysis	I	Haploinsufficient genes	Haung *et al.* (2010) [183]
NET	Gene position in network score calculates gene centrality and indispensability in various protein–protein interactions (PPI) and regulatory networks to assess the gene's importance	Gene-based probability to dissect the gene's importance. Logistic regression model	I	Haploinsufficient genes	Khurana *et al.* (2013) [181], Hsu *et al.* (2016) [180]

Table 4.1 : Gene-specific metrics description (continued)

Features	Definition	Method	Magnitude	Properties	Literature
GHIS	Genome-wide haploinsufficiency score eliminates study bias for the predictions	Gene-based probability of prioritising gene disruptions resulting from any genetic variant. Support vector machine	I	Haploinsufficient genes	Steinberg *et al.* (2015) [182]
REC	Recessive score explains conservation and adjacency to recessive disease genes in a protein interaction network to categorise genes into recessive disease and LoF-tolerant classes	Gene-based probability of LoF-tolerant and recessive disease classes. Linear discriminant analysis	D	Haploinsufficient genes	MacArthur *et al.* (2012) [184], Hsu *et al.* (2016) [180]
GIMS	Gene-level integrated metric of negative selection measures the strength of negative selection. GIMS integrates GERP++ scores as comparative genomic metric	GIMS expresses quantile across all genes	D	Genes under selection	Sampson *et al.* (2013) [186]
GDI	Gene damage index filters out the non-synonym mutational (false-positive) variants in genes that are susceptible to damaging variation in the general population	Probability of relationship to the mean of false-positive variants in damage-susceptible genes	D	Genes under selection	Itan *et al.* (2015) [186], Hsu *et al.* (2016) [180]
LDhat (residuals)	LDhat analyses patterns of linkage disequilibrium within the framework of the coalescent theory of the genealogical history of a sample of genes	Probability of the residuals adjusted for gene size using linear regression. Coalescent model. LDhat length adjusted by linear regression	I	Genetic recombination	Auton and McVean (2013) [190]
LDU (residuals)	Linkage disequilibrium units (LDU) quantify meiotic recombination over a few generations reflecting accumulated recombination over many generations	Probability of residuals adjusted for gene size using linear regression. LDMAP under Malécot–Morton model defines LD map distances in LDUs, analogous to the centimorgan scale of linkage maps. LDU length adjusted by linear regression	D	Linkage Disequilibrium	Vergara-Lope *et al.* (2019) [192], Lonjou *et al.* (2003) [197]
Non-synonym	Number of non-synonymous variants per gene	Frequency	-	Genomic features	Liu *et al.* (2013) [193], Hsu *et al.* (2016) [180]

Table 4.1 : Gene-specific metrics description (continued)

Features	Definition	Method	Magnitude	Properties	Literature
Synonym	Number of synonymous variants per gene	Frequency	-	Genomic features	Liu *et al.* (2013) [193], Hsu *et al.* (2016) [180]
Gene expression	Average expression level per gene	Global expression derived from RNA-Seq data and summed across the 91 cell lines in the Cancer Cell Line Encyclopaedia	-	Genomic features	Lawrence *et al.* (2013) [67], Hsu *et al.* (2016) [180]
DNA replication time	DNA replication time per gene	Expressed on a scale of 100 (early) to 1500 (late)	-	Genomic features	Lawrence *et al.* (2013) [67], Hsu *et al.* (2016) [180]
GC-content	Number of GC content per gene	Reads measured from the exome samples	-	Genomic features	Lawrence *et al.* (2013) [67], Hsu *et al.* (2016) [180]
Hi-C	Hi-C statistic measures the chromatin status	Measured from HiC experiment. This metric indicates which chromosomal compartment the gene is in (negative values = closed compartment 'B', positive values = open compartment 'A')	-	Genomic features	Lawrence *et al.* (2013) [67], Hsu *et al.* (2016) [180]
Gene in kb	Length of the gene	Total number of nucleotide base pairs of DNA	-	Genomic features	Vergara-Lope *et al.* (2019) [192]

Note. Adapted from 'Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data' by Alyousfi D, Baralle D, and Collins A. (2020) Briefings in Bioinformatics [196]. The magnitude column explains the direction of the score value. A 'D' value represents essentiality decrement, while an 'I' value corresponds to the essentiality increment.

4.1.7 Algorithms to select relevant gene-specific properties

Supervised ML techniques were carried out, along with BGGM, to identify genes that are likely to contain Mendelian variation. As described in Section 2.3, supervised ML algorithms aim to predict and find patterns within data and use them to make predictions and classifications or infer new knowledge. Thus, different supervised models are investigated to select relevant gene-specific properties. The BGGM methodology described in Section 2.5 is implemented to encode probabilistic relationships between gene properties, building connectivity between these features. Following this framework, the nodes represent the features, and the undirected edges describe the correlation among these nodes. The resulting network provides a comprehensive insight to explain the properties of genes.

All the approaches applied in this study focus on integrating diverse biological data sources and preserving the relevant features. To this end, scores that predict the potential pathogenicity of individual DNA at gene level were selected from the literature. Additionally, maps of linkage disequilibrium and recombination rates from the Wellderly population genetic data (Vergara-Lope *et al.*, [192]) were added to the models. The selected features were compared between these methodologies and were considered for the final machine learning classification model. The final approaches can provide biological knowledge and insights for disease genes to investigate further how they are related to the disease phenotypes.

4.2 Methods

4.2.1 Collection of gene-level and functional-related gene metrics

The datasets for this study were extracted from published gene prioritisation scores to assess each gene's functional genomic properties. These gene-level scores were selected to include known gene-specific properties and data on evolutionary and functional genomic properties. The following scores were downloaded from either dbNSFP [193], which provides a collection of scores, or the corresponding studies. Hence, for measuring essentiality and conservation, the following gene-specific scores were considered; pLI [177], RVIS [173], DNE [178], and SIS [71]. For assessing haploinsufficient genes the following were included; HI [183], NET [181], GHIS [182], and REC [184] scores. For quantifying genes under selection, GIMS [186] and GDI [187] metrics were included.

A total of six biological and genomic features were collected from the Ensembl database through the BioMart system. These features include gene length in kilobases [192], and GC content (GC) [198]. In addition, functional-related gene features were included: global expression (Expr) derived from RNA-Seq data and summed across ten different cell lines, DNA replication time (Reptime), the Hi-C statistic (HIc), and the measurement of chromatin status [67].

As described above, LDhat recombination maps [190] serve as a guide to the recombination rate per gene, while LD maps in LD units show the LD intensity per gene. These maps were constructed

using 454 whole-genome sequences from the Wellderly study [149, 192]. Linear regression for LDhat and extent of LD adjusting by physical gene lengths due to the strong correlation between LDhat/LDU and physical gene lengths. Therefore, the residual terms were used for both scores in the analysis as:

$$LDU_{res} = LDU_i - \hat{LDU}, \tag{4.1}$$

$$LDhat_{res} = LDhat_i - \hat{LDhat}, \tag{4.2}$$

where LDU_i and $LDhat_i$ are the observed values, and $\hat{LDU}$ and $\hat{LDhat}$ are the predicted values.

Once the gene-specific scores were integrated for all genes, an exploratory analysis was implemented to prepare the data for applying ML techniques and BGGM. First, a simple summary of central tendency and dispersion measurements are given to describe the data.

4.2.2 Genes groups

Following Spataro *et al.* [52], classified five discrete groups of genes based on roles in protein networks, rates of protein evolution and tests of neutrality. The five gene groups are ordered according to the degree of gene essentiality: non-disease and non-essential (NDNE), complex non-Mendelian (CNM), complex-Mendelian (CM), Mendelian non-complex (MNC) and essential non-disease (END).

Non-disease and non-essential (NDNE) comprises genes that are neither essential nor associated with a disease. These genes are related to being under the weakest purifying selection levels and have the least functional relevance.

The complex non-Mendelian (CNM), complex-Mendelian (CM), Mendelian non-complex (MNC) are named as well as Human Diseases (HD) genes. Genes at these ends are functionally relevant less than essential than NDNE group. Moreover, these genes are under stronger and longer-lasting purifying selection.

The last group, essential non-disease (END) includes genes that are defined as genes responsible for core biological functions in the organism and so are required for cell survival. These genes have no association with human disease.

Therefore, following the Spataro *et al.* [52] structure, here genes were classified into five groups which may represent different degrees of essentiality according to the model proposed by Pengelly *et al.* [53] (as described in Chapter 2). Because the small proportion of genes in the CM and CNM groups were joined into a single group named CNM, the gene groups considered were MNC, CNM, END and NDNE.

4.2.3 Supervised learning algorithms to select the relevant features

The supervised ML models were applied by training the pre-set learning algorithms to map the relationships between genes into the four gene groups based on 19 gene-specific metrics (see Table 4.1). The first step was collecting the relevant information of the genes to construct the data. The second step was data preparation using different methodologies to handle missing data, transforming skewed distributions and removing collinear features. Third, different ML algorithms were applied using a nested resampling strategy and adjustable parameters for feature distributions. The methods used in each of these steps are described in detail below.

To state the results more formally, some notations are introduced in this section. The classifier is mapped by the function $f\colon X \to Y$ based on the four gene groups within the selected X as biological features from the training set D where $Y = \{1, 2, \ldots, k | k \in \mathbb{N}\}$. Y represents a discrete set consisting of k classes, and $X \subseteq \mathbb{N}^D$ is the data domain such that X describes the attributes of gene i. The dataset of size n is a tuple $D = (X \times Y)^n$ consisting of n labelled examples $\{x^i, y^i\}_{i=1}^n$, where genes are assorted independently. The data set is queried by specifying a classifier $f\colon X \to Y$ and observing its accuracy $acc(f)$ on the training data D, which is simply the fraction of points that are correctly labelled $f_k(x_i) = y_i$. The accuracy of f is denoted by $acc(f) = Pr\{f(x) = y\}$ over the genes, attributes from which the joint probability distribution $P_{(X,Y)}$ of (x, y) is drawn. Proceeding in k iterations, the analyst specified a function in each round and observed its accuracy on the data set. The analysis is built as a sequence of adaptively chosen functions $f_1, \ldots, f_k$. Given a set of genes already causally linked to specific gene groups, the classifier is sought for any gene that predicts whether it would have belonged to any other gene group. This function is defined as follows:

$$f_k(x) = \begin{cases} x & \text{if } x \in \{1, \ldots, k\}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

Before the gene classification was carried out, the analysis was divided into three main steps. In the first step, the data was prepared, setting out a complete and detailed specification of the features. The second step identified the best possible learning algorithm based on prediction accuracy and model interpretability. In the final step, the hyperparameter tuning stage was implemented from the selected model.

Figure 4.1 shows the workflow applied in supervised learning. Along with this process, the machine learning algorithms to be used are selected to fit the desired target quantity. Most of the work consisted of generating, finding, and cleaning the data to ensure consistent and accurate data. It was also necessary to decide on the genes' functional properties, i.e., the inputs for the model that would be suitable for the chosen algorithm. The model was trained by optimising its performance, measured through cost function metrics. This entailed adjusting the hyperparameters that control the model's training process, structure, and properties. The data were split into various sets. A validation dataset separate from the test and training sets was used for optimisation of the hyperparameters.

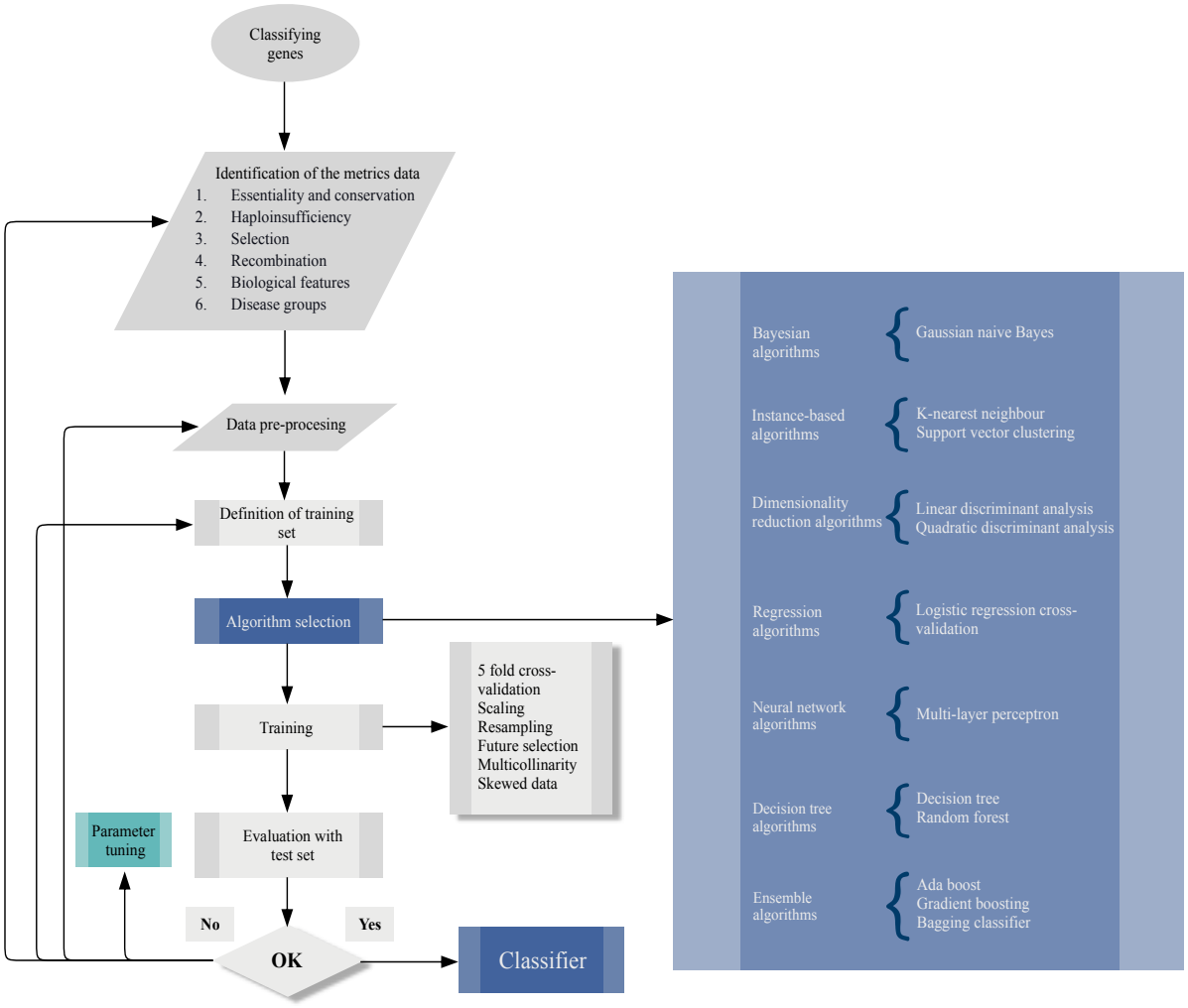


Figure 4.1: **Machine learning pipeline**. Generic pipeline of the methodologies used to classify genes into gene groups. The first step involves collecting the metrics data to describe functional genomic properties. The second step follows the data pre-processing; this process cleans the data, removes redundancy, and carries out feature scaling and data transformation. Once the data has been pre-processed, the next step is to divide the data into training and test datasets. Then the different machine learning models considered are evaluated using five-fold cross-validation to select the best-performing model. In each fold, the resampling technique is implemented. The final step in the pipeline is tuning the best-performing models.

4.2.4 Bayesian inference in Gaussian graphical models to select the relevant features

The BGGM is implemented to encode probabilistic relationships between gene properties to establish that the features proposed for this study are relevant for classifying the genes into the gene groups [52]. Following this framework, the nodes represent the features and the undirected edges represent the correlations among these nodes. The results describe the interactions between genes, reflecting gene properties based on functional properties.

4.2.5 Data processing step

The data was partitioned into a training set and a test portion was used to teach a model that covers highly complete patterns (fitting the model) from the training data and evaluate the model's accuracy (goodness of prediction) using the same features but different instances from the test data. Following Liu and Cocea [199], the data were randomly partitioned using 70% of the data for training and 30% for the test datasets.

4.2.5.1 Imputation

During the data preparation step, an imputation model was specified to handle missing data for each incomplete feature, conditioned on the observed data. It is important to note that the inference for parameters of a simpler data model can be biased and might decrease the accuracy of the classifier. The imputation model fits a statistical model to the observed data per feature as a function of the other features and uses it to estimate values for the missing data. At each step, one feature column was designated as output y and the other feature column treated as input X. The regressor from the model was fit to (X, y) for known y. In this way, the regressor was used to predict the missing value of y. This was done for each feature in an iterated round-robin fashion [200]. Each imputed feature was then analysed individually using standard statistical procedures (Student t).

The Bayesian ridge regression was used for imputing missing values. This is a probabilistic technique that introduces uninformative priors over the hyperparameters. This method is initialised using random sampling and runs univariate imputations sequentially until convergence is reached. Each iteration is a Gibbs sampler that draws from the distribution that is conditional on the imputed values. To obtain a fully probabilistic model, the output is assumed to be Gaussian distributed around X_ω:

$$p(y|X, \omega, \alpha) = \mathcal{N}(y|X_\omega, \alpha), \tag{4.4}$$

where the prior for the coefficient ω is given by a spherical Gaussian

$$p(y|X, \lambda) = \mathcal{N}(\omega|0, \lambda_{-1}\mathbf{I}_p). \tag{4.5}$$

The priors over α and λ are chosen to be gamma distributions, the conjugate prior for the precision of the Gaussian distribution. Then the parameters are jointly estimated by maximising the log marginal likelihood over the coefficients ω with precision λ^{-1} [201].

4.2.5.2 Data transformation

Data transformation strategies were applied both for training and for testing data. First, all distributions of the features were analysed for handling skewness and addressing scaling issues. A standard log transformation was used to reduce skewness and spread low values by expressing the values as orders of magnitude. In addition, this transformation might reduce the high degree of

variation among attributes within genes. Secondly, a scaling method was carried out to standardise the range of each independent numerical feature. By doing so, changes in different features become comparable and reduce the variation in magnitude and scope across features. The standard z score of X data was used to scale the features. It is calculated by removing the mean and scaling to unit variance:

$$z = \frac{(x - \mu)}{\sigma}, \tag{4.6}$$

where μ is the mean and σ is the standard deviation of the training and test data [202].

4.2.5.3 Multicollinearity analysis

During the data pre-processing step, a multicollinearity analysis was performed to reveal the linear dependence between the features [202]. The variance inflation factor (VIF) is used to quantify the severity of multicollinearity. The VIF is based on ordinary least squares regression analysis as:

$$VIF_j = \frac{1}{1 - R_j^2}, \tag{4.7}$$

where R_j^2 is the value obtained by regressing the j^{th} feature on the remaining predictors. Gene metrics with $VIF_j \geq 7$ [202, 203] were excluded from further estimation of ML algorithms in both cases, training and test data; this procedure is applied in the same way for the BGMM inference.

4.2.6 Model selection

Despite the broad range of ML approaches available today, there is no unique ML method that generally outperforms all others [204]. Different ML approaches were evaluated to recognise the most suitable model for determining the essential features (see Table 4.2). The final model was estimated using resampling methods and k-fold cross-validation from which mean scores were calculated and compared directly [205]. The best model was selected based on accuracy and minimum error (see Appendix A for further details of the methodologies).

Table 4.2: Supervised machine learning methodologies

Algorithms	Classifier
Bayesian	Gaussian naïve Bayes
Instance-based	K-nearest neighbour
	Support vector clustering
Dimensionality reduction	Linear discriminant analysis
	Quadratic discriminant analysis
Regression	Logistic regression cross-validation
Neural network	Multi-layer perceptron
Ensemble	Random forest
	Gradient boosting
	Bagging classifier

4.2.6.1 Cross-validation

The best classifier that fits the data was selected comparing the different models perform. However, training a model on the same train data means that the model will eventually learn well for only that data and fail on new data; this is called overfitting[2]. Thus, cross-validation strategy was implemented to avoid overfitting.

Cross-Validation in ML is a technique used to train and evaluate the model on a portion of the database before re-portioning the dataset and evaluating it on the new portions. Instead of splitting the dataset into train and test sets, the dataset is divided into multiple parts. Then, a different division to train is used and test our model. This ensures that the model is training and testing on new data at every iteration. The training data will be used by the model to learn. The model will use the testing dataset to predict unseen data called 'model's performance'. Then, a different portion is chosen to test on and use the other parts for training. The model performance is re-evaluated with the results obtained from the new portioned dataset to improve the results. This step is repeated multiple times until the model has been trained and evaluated on the entire dataset.

Stratified K-fold cross-validation was used here to mitigate the overfitting problem during the validation process. This method is useful when there are minority classes present in the data. The dataset was split into k parts one section is for testing and the rest for training. Then, another section will be chosen for testing and the remaining section will be for training. This will continue k number of times until all sections have been used as a testing set once. The final performance measure will be the average of the output measures of the k iterations. Thus, the gene-specific data was divided into five equal random non-overlapping folds. This was repeated five times such that each of the five sets was used as the test set exactly once [206].

4.2.6.2 Resampling

A balanced ratio of gene groups was required to address the multiclass problem. The undersampling technique was used to balance the skewed class distribution of the gene groups and improve the classification performance. A simple random undersampling method was applied for the majority class (NDNE) in the gene groups in the training data. This approach involves randomly selecting examples for the NDNE group and keeping all the observations from the minority class (MNC) from the training data until a balanced distribution is reached. Hence, the ratio is expressed as $\alpha = \frac{N_{MNC}}{N_{NDNE}}$ where N_{MNC} is the number of samples in the END gene group, and N_{NDNE} is the number of samples in the NDNE group after resampling. During classifier design, the NDNE group was undersampled in each training fold of the five cross-validations [207]. The method from the `imbalanced-learn` Python library was used, resulting in a subset of 25% per group. The five-fold cross-validation of classifiers built on undersampled datasets was repeated five times.

[2]When the model trains fit perfectly on the training data and generalise to it, but fails to perform on new, unseen data. It captures every single variation in training data and cannot perform on data with the same variations.

4.2.6.3 Tuning the model with a hyperparameter grid

Once the supervised ML model was chosen based on the parameters directly estimated from the data parameters in the training phase, the next step was to tune its hyperparameters to control the complexity of the training algorithm. The hyperparameters are the parameters predefined before the training process. During this tuning process, Bayesian optimisation cross-validation was applied to find the optimal hyperparameters combination. Bayesian optimisation proceeds by assuming the unknown function was sampled from a Gaussian process and keeps a posterior distribution for this function as the learning algorithm's results with different hyperparameters. That is, Bayesian optimisation works along with the probability distribution for each parameter on the sample [208, 209]. Therefore, a set of candidate tuning parameters was specified and then evaluated. The folds number and random seed were set to generate reproducible models. The next step was building a grid search using Bayesian optimisation with the selected model and fitting the entire training set. During this hyperparameter optimisation, the five-fold cross-validation process was also included [210].

4.2.6.4 Evaluation metrics

The performance of the classification algorithms was evaluated using several statistical measures: precision, recall and F_1-score. Precision is a measure of relevance, and recall is the fraction of retrieved relevant features over the total number of relevant features, so high values for both measures indicate better performance. The false-positive rate (FPR) is the fraction of wrong true predictions, and the true-positive rate (TPR) or sensitivity is the fraction of correct classifications (for a classifier). The TPR and FPR range between 0 and 1, and high TPR and low FPR reflect good performance [211]. Thus, the F_1-score supplements them by reducing the two measures into a single number by expressing the harmonic mean between precision and recall. The formulas are

$$Precision = \frac{TP}{TP + FP}, \tag{4.8}$$

$$Recall = \frac{TP}{TP + FN}, \tag{4.9}$$

$$F_1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{4.10}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false positives.

4.2.7 Bayesian inference for Gaussian graphical models

A BGGM was implemented to capture the conditional dependencies between the features for the functional properties of the genes, thus selecting the most important features by sparsifying the set of

edges. The 19 features represented the nodes, and the edges correspond to statistical dependencies between those features. Hence, the first step was preparing the data in the same way as the supervised ML application. In the next step, a Gaussian graphical model was constructed based on the Wishart prior distribution, which is directly conjugate to the precision matrix. These posterior probabilities determined the graphic structure, enabling conditional dependent and independent connections adjacent to the features to be assessed. Finally, the structured learning determined the topology of the probabilistic graphical model and its accuracy was estimated by the likelihood that the model explained the observed data [212, 90].

4.2.8 Data and code

The ML models were constructed and analysed in Python version 3.7.3 (`https://www.python.org/`), `awk` and R version 3.2.2 (`https://www.r-project.org/`) using custom-written scripts.

4.3 Results

4.3.1 Descriptive data analysis

The data formats for gene properties were obtained by merging the nine published datasets from either dbNSFP [193] or their corresponding studies. The combined data comprises the Spataro *et al.* [52] groups of the different subtypes of genes where n represents the number of genes, and p represents the number of biological features. The complete genetic matrix includes 19 features (see Table 4.1) along 14,708 genes, where all these genes were considered to be merged to all the datasets.

Figure 4.2 shows the gene distribution among the four groups defined by Spataro *et al.* [52]. The NDNE group has 10,421 genes not known to be involved in any human disorder and not known to be essential genes. The second group are those genes uniquely associated with MNC; it comprises 822. The third group contains 2,076 genes uniquely causing CNM. The END group includes 1,389 genes defined as orthologues of essential mouse genes detected by knock-out experiments and not involved in any human disease. Of the 14,708 genes, 70.85% were in the NDNE group, 14.11% were in MNC, 9.44% were in CNM, and 5.59% in END. The pie chart (Figure 4.2) shows that most genes tend to be in the NDNE group and that relatively few genes are in the END group. This is consistent with a previous study, which demonstrated that ∼10% of the ∼20,000 genes in human cells is essential for cell survival [170].

4.3.2 Statistical summary: Measures of central tendency

The means of the gene-specific metrics for each gene group are presented in Table 4.3. The features considered an approximation to gene essentiality show higher mutation intolerance on average than

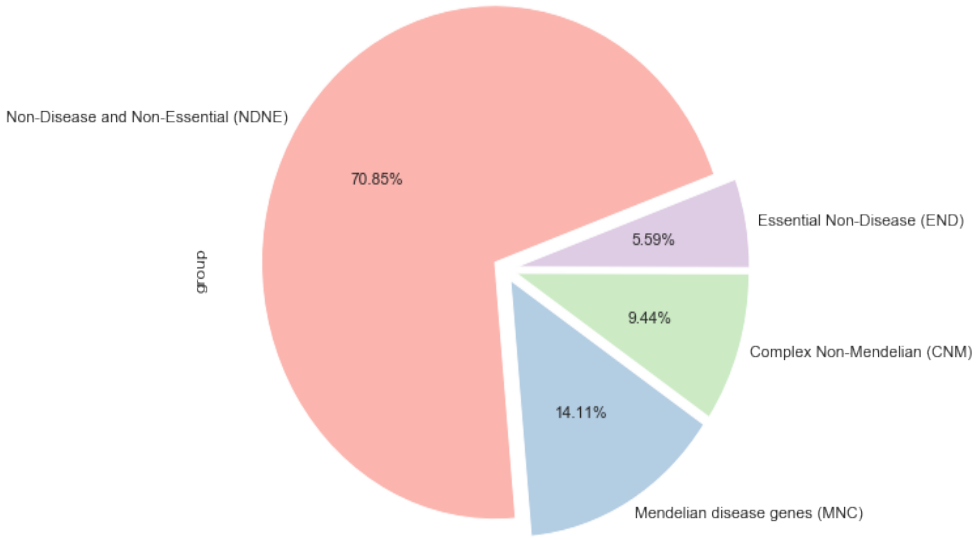


Figure 4.2: **Gene distribution by gene groups**. Gene number and percentage of genes belonging to NDNE, CNM, MNC, END groups based on Spataro *et al.* study [52].

the other gene groups (pLI m= 0.562, RVIS m= −0.428, DNE m= 1.733 and SIS m= 0.557 means). These results are consistent with essential genes that tend to have low recombination (LDhat m= −21.915), intense selection (GIMS m= 0.326, GDI m= 3.666) and strong LD (LD m= −0.323). In addition, at the END gene group, the degree of intolerance of genes to loss-of-function variation is higher than in the rest of the groups (HI m= 0.474, NET m= 0.743). Gene groups that contain disease variation occupy an intermediate place in mutation intolerance, recombination and selection and retain damaging variation associated with Mendelian disease (REC m= 0.361). Genes in the NDNE group tend to have low essentiality (pLI m= 0.252, RVIS m= 0.054, DNE m= 0.727 and SIS m= −0.048), high recombination and high haplotype diversity (LDhat m= −3.847 and LD m= −0.323) and haploinsufficiency (HI m= 0.264). They also show evidence for weak selection and weak LD (close to 0). Thus, genes located in this group might be more tolerant of mutation [53].

The correlation patterns are presented in Figure 4.3. Spearman correlation analysis suggests that most of the features have a modest correlation with each other. The highest correlations[3] in either direction (positive or negative) were between LD and LDhat ($\rho = 0.86$) as expected because both metrics model SNP association to estimate or approximate the population scaled recombination rate. In addition, synonymous and nonsynonymous rates per gene are highly correlated ($\rho = 0.78$); these results rely on common selective constraint due to mutations that might be related to selection for translational accuracy [213]. Meanwhile, the overall expression level, Hi-C status and GC content have positive correlations with each other. This result can be explained because chromatin structure contributes to controlling gene expression (Hi-C vs Expr $\rho = 0.58$) and replication (Hi-C vs Reptime $\rho = -0.61$) [214]. Moreover, GC content is a prominent for maintenance of open

[3]Spearman coefficient $|\rho| > 0.5$

Table 4.3: Distribution of the gene-specific mean among gene groups

Features	NDNE	MNC	CNM	END
pLI	0.252	0.245	0.379	0.562
RVIS	0.054	−0.220	−0.138	−0.428
DNE	0.727	0.914	1.006	1.733
SIS	−0.048	0.113	0.139	0.557
HI	0.264	0.363	0.350	0.474
NET	0.514	0.709	0.597	0.743
REC	0.156	0.361	0.218	0.261
GIMS	0.508	0.433	0.439	0.326
GDI	4.380	5.321	4.564	3.666
Ldhat (residuals)	−3.847	8.181	19.607	1.915
LDU (residuals)	−0.069	0.105	0.192	−0.323
Non-synonym	24.310	36.001	29.273	27.048
Synonym	13.350	20.934	17.407	19.048
Gene expression	6.716	6.319	5.752	7.122
DNA replication time	442.134	439.560	487.942	399.765
GC content	46.335	46.745	44.880	46.434
Hi-C	19.902	20.224	17.560	23.289
Gene size in kilo bases	53.543	60.701	123.959	74.193

chromatin structures (Hi-C vs GC $\rho = 0.61$) [195]. The pLI score has a high correlation with SIS and DNE scores (pLI vs SIS $\rho = -0.56$; pLI vs DNE $\rho = -0.58$); this relationship may imply that those scores share LoF measures such as gene essentiality and conservation. There is a negative correlation between RVIS and SIS scores (RVIS vs SIS $\rho = -0.70$); RVIS reports negative scores while SIS increases in quantifying the action of selective pressure. This result might reflect purifying selection on functional substitutions [173, 71]. Another high correlation is between RVIS and GIMS scores (RVIS vs GIMS $\rho = -0.70$), elucidating genes under negative selection [186]. SIS and DNE scores show a high correlation with each other. This correlation may reflect functionally essential genes and LoF tolerant genes [71, 181].

A descriptive statistical analysis was performed to describe gene–specific metrics and population genetics measures. First, the non-parametric Kruskal–Wallis (KW) test was employed to detect an association between gene-specific metrics and the four gene groups using $p \leq 0.05$ as a threshold criterion (see Table 4.3). All the gene-specific metrics are statistically significantly different between gene groups. Hartigan's dip test was used to detect multinomial distributions among all the features. This test reported possible multimodal distributions for HI, pLI, NET, Hi-C, DNA replication time, nonsynonymous, and synonymous. Understanding multimodal distributions may help better maximise classification accuracy, provide an exact Type I error rate, and allow for assessing cross-generalizability (Wu *et al.* [215]). Tukey's test suggested that DNE, REC, SIS, LDU, LDhat, RVIS, GDI, gene expression, Hi-C, DNA replication time, nonsynonymous, synonymous, and gene size have large values that may be considered as outliers.

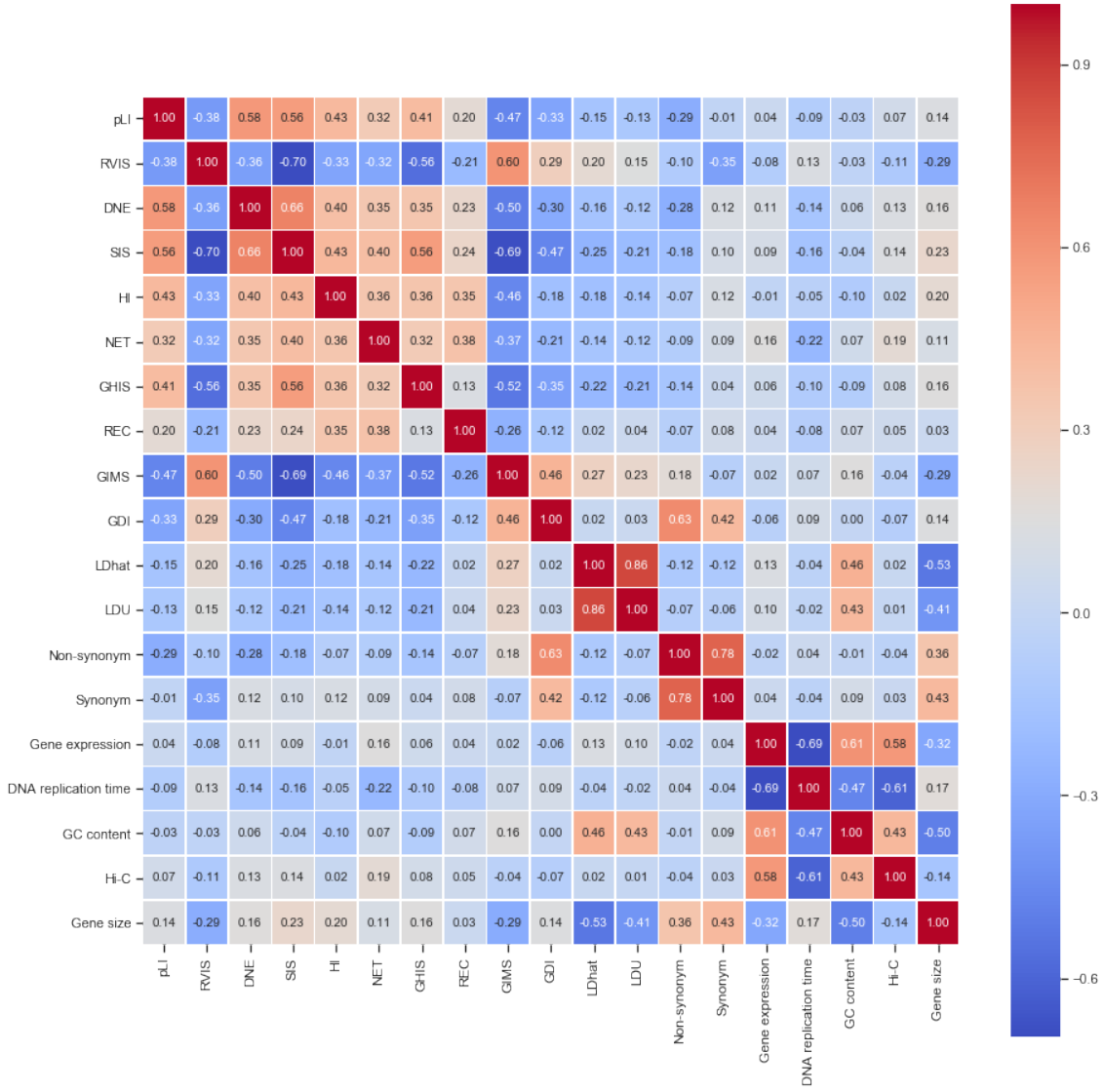


Figure 4.3: **Heatmap**. Sample correlation matrix based on the 14,708 genes using gene-level metrics. The matrix entries are Pearson correlation coefficients. Red indicates positive correlation and blue negative correlation. Darker colours indicate strong correlation and lighter colours indicate low dependency between features. The coloured bar at right shows the scale of the degree of correlation.

4.3.3 Data visualization

Figure 4.4 shows feature distribution by gene groups. The overall difference in pLI, NET, SIS, and HI scores between gene groups have the highest probability for increasing essentiality. In contrast, GDI, GIMS, LD, REC, and RVIS have a smaller degree with increasing essentiality. It also appears that GDI, LDU, LDhat, REC, SIS, gene size, nonsynonymous, and synonymous are heavily right-skewed. The features pLI, HI, NET, DNA replication time, and GC content follow a multimodal distribution.

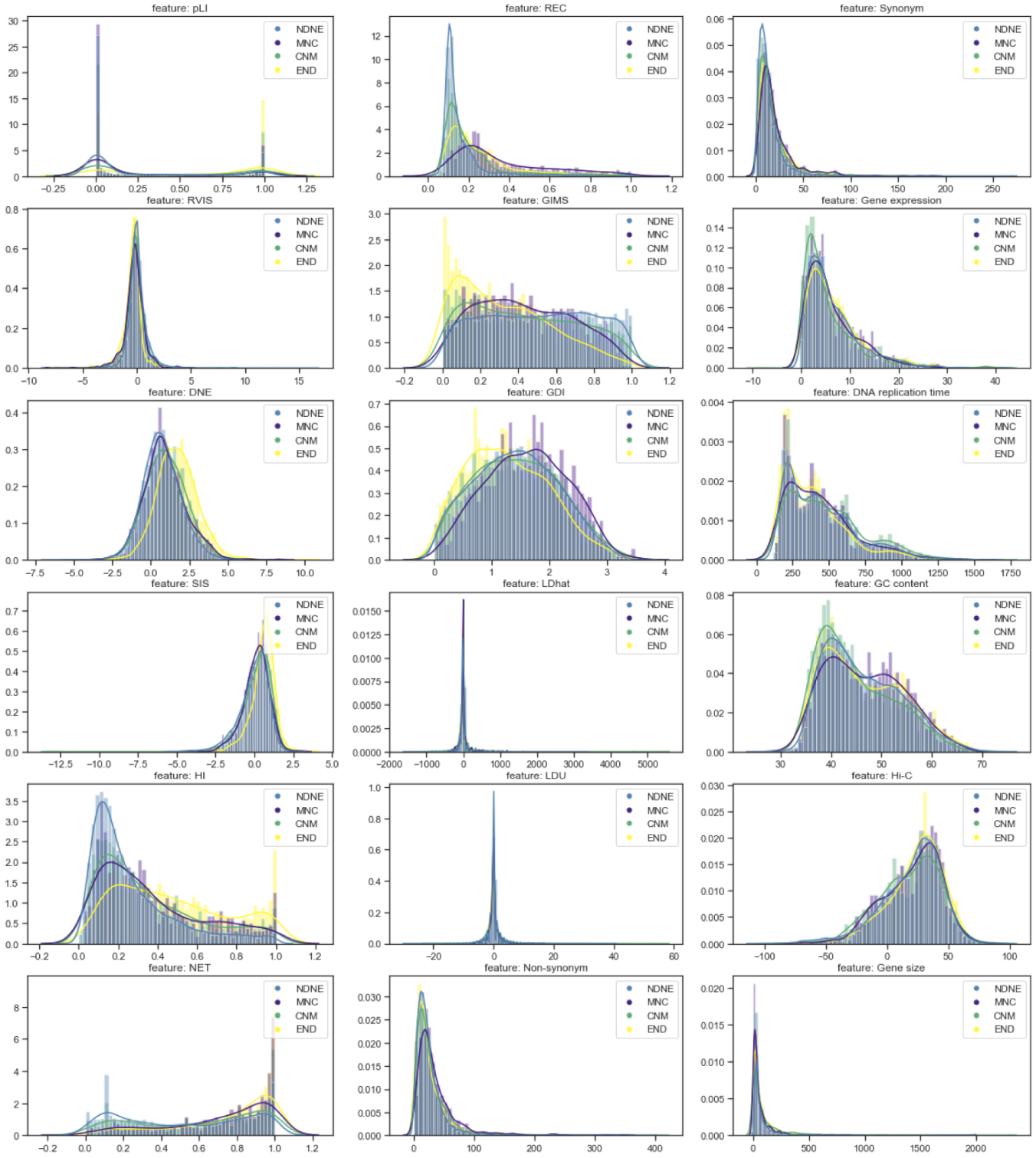


Figure 4.4: **Feature density distributions and correlation among features by gene group**. Illustration of modelling approach and prediction of the number of gene groups for single genes using the information on 18 functional genomic features of genes. The curve shows the density plot of a smooth version of the histogram. The y-axis indicates density, and the histogram is normalised on the same y-scale as the density plot.

Table 4.4: Statistical descriptive features by gene groups

Feature	NDNE	MNC	CNM	END	Kruskall-Wallis test
Number of genes	10421	822	2076	1389	
pLI	0.008 [0.000,0.483]	0.002 [0.000,0.399]	0.100 [0.000,0.901]	0.723 [0.024,0.992]	Significant
RVIS	−0.009 [−0.380,0.415]	−0.181 [−0.622,0.260]	−0.139 [−0.579,0.267]	−0.296 [−0.737,0.017]	Significant
DNE	0.704 [−0.109,1.581]	0.773 [0.045,1.721]	0.989 [0.085,1.966]	1.716 [0.852,2.590]	Significant
SIS	0.109 [−0.596,0.635]	0.197 [−0.385,0.679]	0.303 [−0.381,0.793]	0.677 [0.216,1.049]	Significant
HI	0.198 [0.113,0.390]	0.280 [0.137,0.558]	0.265 [0.133,0.535]	0.424 [0.220,0.728]	Significant
NET	0.533 [0.170,0.852]	0.801 [0.539,0.948]	0.682 [0.274,0.903]	0.846 [0.586,0.965]	Significant
REC	0.117 [0.101,0.166]	0.306 [0.194,0.541]	0.136 [0.108,0.239]	0.183 [0.128,0.315]	Significant
GIMS	0.513 [0.262,0.752]	0.410 [0.226,0.637]	0.408 [0.186,0.674]	0.278 [0.112,0.501]	Significant
GDI	3.143 [1.390,6.103]	4.088 [1.987,7.633]	3.156 [1.344,6.335]	2.387 [1.054,5.099]	Significant
Ldhat (residuals)	−1.197 [−25.512,8.837]	−1.730 [−27.846,10.457]	−4.325 [−47.276,9.665]	−10.837 [−51.647,5.309]	Significant
LDU (residuals)	0.021 [−0.460,0.211]	0.026 [−0.492,0.394]	−0.019 [−0.820,0.355]	−0.109 [−0.958,0.182]	Significant
Non−synonym	19.000 [11.000,31.000]	25.000 [15.000,44.000]	21.000 [11.000,37.000]	18.000 [10.000,34.000]	Significant
Synonym	10.000 [6.000,17.000]	14.500 [9.000,25.000]	12.000 [7.000,21.000]	14.000 [8.000,23.000]	Significant
Gene expression	4.955 [2.607,9.019]	4.656 [2.465,8.758]	3.756 [1.865,7.818]	5.449 [2.793,9.437]	Significant
DNA replication time	393 [236,581]	391 [249,564]	436 [269,627]	359 [219,506]	Significant
GC content	44.980 [40.150,51.860]	46.080 [40.360,52.370]	43.160 [38.990,50.050]	44.930 [39.700,52.695]	Significant
Hi−C	25 [6,37]	25 [5,38]	22 [1,37]	27 [11,39]	Significant
Gene size in kilo bases	22.646 [8.671,54.978]	30.386 [13.197,72.907]	45.237 [14.563,130.114]	36.934 [13.830,87.543]	Significant

Mean and 95% CI for all gene-level metrics. Kruskal−Wallis test was applied, p−value < 0.001.

4.3.4 Machine learning results

4.3.4.1 Training and testing dataset results

The gene-specific metrics data were divided into two parts; training data with 10,295 genes (70% of the whole data) and testing data with 4,413 genes (30% of the data; see Table 4.5). The distribution among the gene groups is highly unbalanced, as the minority class accounts for as little as ∼6% of the training data in the MNC group, followed by ∼10% and 14% for END and CNM, respectively. Therefore, there is a genuine lack of data in the END, MND and CNM groups due to the low frequency with which events occur or genes known to be associated with essentiality/disease variation are present [167]. In this study, the MNC and END are the classes of interest that can identify the underlying causes for disease genes, and NDNE potentially contains undiscovered disease genes. Therefore, the sampling method was applied in the supervised machine learning model for handling class imbalance during the pre-processing data step.

Table 4.5: Number of genes in the training and test sets by group

Gene group	Training set		Test set	
	Number	Percentage	Number	Percentage
NDNE	7252	70.4%	3169	71.8%
MNC	608	5.9%	214	4.8%
CNM	1438	14.0%	638	14.5%
END	997	9.7%	392	8.9%
Total	10295	100.0%	4413	100.0%

4.3.4.2 Gene dataset imputation

Of the 10,295 genes in the training data, 6,159 (∼60%) genes had data available for all features in the analysis model. HI and REC scores had the highest proportion of missing values, ∼24%, while DNE and SIS had ∼12% missing data. The rest of the features have less than 5% of the data missing. For the test data, 2,622 (∼60%) of 4,413 genes had complete data for all the features, showing the same proportion of missing values as the training data (see Table 4.6). Therefore, the imputation method was applied to reduce bias during the pre-processing data step for training and test data.

Table 4.6: Number of missing genes in the training and test data

Feature	Training data		Test data	
	Number	Proportion	Number	Proportion
pLI	0	0.0%	0	0.0%
RVIS	0	0.0%	0	0.0%
DNE	1,218	11.8%	544	12.3%
SIS	1,079	10.5%	506	11.5%
HI	2,478	24.1%	1,069	24.2%
NET	249	2.4%	125	2.8%
REC	2,478	24.1%	1,069	24.2%
GDI	5	0.0%	1	0.0%
GIMS	98	1.0%	42	1.0%
LDhat (residuals)	0	0.0%	0	0.0%
LDU (residuals)	0	0.0%	0	0.0%
Non-synonym	0	0.0%	0	0.0%
Synonym	0	0.0%	0	0.0%
Gene expression	253	2.5%	105	2.4%
DNA replication time	253	2.5%	105	2.4%
GC content	14	0.1%	5	0.1%
Hi-C	253	2.5%	105	2.4%
Gene size in kilo bases	0	0.0%	0	0.0%

Iterative imputation using Bayesian ridge regression was used for handling missing values. This process estimated each feature with missing values as a function of the other features. Thus, the Bayesian ridge approach completed the missing values by iteratively maximising the marginal log-likelihood of the observations. The imputed data for DNE, SIS, HI, NET REC, GIMS, GDI, gene expression, DNA replication time GC-content and Hi-C features are drawn from the joint posterior distribution of the missing data under a Bayesian ridge model. In order to appropriately analyse the performance of the imputation, Table 4.7 provides a summary of the two-sample t-test that was used. This test compared the empirical distributions of the observed and imputed data and flagged features as a potentially significant difference if they had a p-value below 0.05. DNE, REC and HI results suggest that the accuracy and robustness of imputation were affected because of an increment in the missing rate, showing statistically significant differences between the mean of the observed and imputed data. The distribution of the remaining features indicated no significant difference between the two averages. Although t-tests provide a suitable way to check a large number of imputed features, the results can be challenging to interpret because the magnitude of

the p-values depends on both the sample size and the proportion of missing values in the incomplete features. It is also important to note that discrepancies between observed and imputed data are not necessarily problematic since this gene data is missing at random; therefore, these differences can be expected to arise.

Table 4.7: Two sample t-test for difference between means of features in observed and imputed data

Feature	Percentage		Standard error		Original - Imputed		Z score	Statistical test	Level of significance*
	Original	Imputed	Original	Imputed	Diff. %	Diff. std. error			
DNE	0.91	0.87	0.0141	0.0128	−0.04	0.02	−1.99	0.05	Significant
SIS	0.03	0.05	0.0101	0.0093	0.02	0.01	1.24	0.22	Non-significant
HI	0.32	0.30	0.0029	0.0023	−0.02	0.00	−5.67	0.00	Significant
NET	0.56	0.56	0.0034	0.0033	0.00	0.00	−0.64	0.52	Non-significant
REC	0.20	0.19	0.0019	0.0015	−0.01	0.00	−3.48	0.00	Significant
GIMS	0.48	0.48	0.0028	0.0028	0.00	0.00	0.00	1.00	Non-significant
GDI	4.40	4.39	0.0418	0.0415	−0.02	0.06	−0.30	0.76	Non-significant
Gene expression	6.65	6.65	0.0585	0.0574	0.00	0.08	−0.01	0.99	Non-significant
DNA replication time	441.8	442.2	2.3950	2.3484	0.35	3.35	0.10	0.92	Non-significant
GC content	46.2	46.2	0.0747	0.0747	0.00	0.11	−0.01	0.99	Non-significant
Hi-C	20.3	20.2	0.2457	0.2405	−0.03	0.34	−0.08	0.94	Non-significant

*Two-tailed hypothesis test with a significance level of 0.05.

4.3.4.3 Skewed data

The log transformation was used to address skewed data; the log-transformed data follows a normal or near-normal distribution. Shown in the top panel in Figure 4.5 is the histogram of two distributions x_i, gene size and GDI, while the bottom panel is the histogram of y_i (the log-transformed version of x_i) based on a sample size of n=10,295. While both distributions of x_i are left-skewed, the log-transformed data y_i are not skewed.

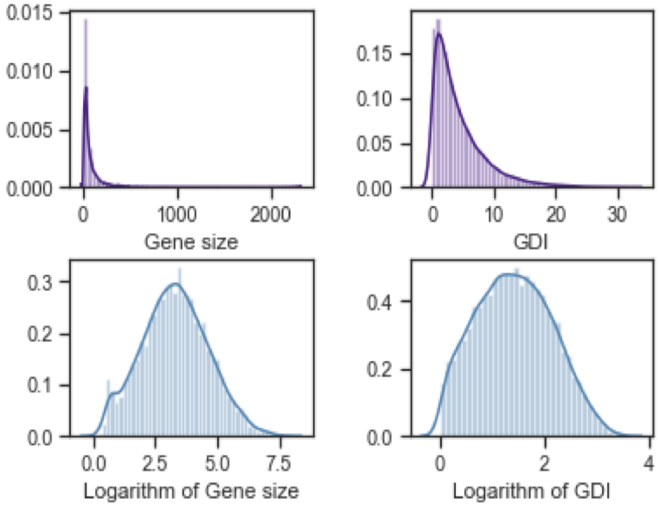


Figure 4.5: **Skewed data transformation**. Comparison of the KDE-plots of both gene size and GDI features vectors before (top row) and after (bottom row) transformation.

4.3.4.4 Feature scaling

Given the significant differences in the scales, magnitudes, and features, all these predictors were normalised before the machine learning model was developed for the training and test data. This feature scaling improved the model performance by speeding up gradient descent and avoiding many extra iterations that would be required when one or more features takes on much larger values than the rest. Moreover, in this study, most machine learning algorithms use Euclidian distance, which might be a problem because the results vary significantly between different ranges. For example, in the left column of Figure 4.6 the nonsynonymous, synonymous, LDU and LDhat residuals have high magnitudes, which will weigh a lot more in the distance calculations than features with low magnitudes. In contrast, in the right columns of Figure 4.6 all features have been normalised and are on the same scale (relative to one another).

4.3.4.5 Multicollinearity

The linear dependence or multicollinearity between feature means was tested to evaluate the proposed features' quality and measure instability and redundancy between them. These problems were addressed by measuring the variance inflation factor (VIF). The values with VIF above seven were removed and were considered highly collinear. The logarithm of GDI, Synonymous and gene size was not considered further for inclusion in the machine learning models (see Table 4.8).

Table 4.8: Multicollinearity among the features

Feature	**Variance Inflation factor**			
pLI	2.669	2.636	2.626	2.625
RVIS	2.927	2.896	2.348	2.342
DNE	4.078	4.041	3.200	3.102
SIS	4.144	4.101	3.867	3.758
HI	4.002	3.922	3.894	3.791
NET	4.994	4.631	4.602	4.439
REC	1.220	1.197	1.194	1.185
GIMS	6.388	6.145	6.111	3.835
GDI log	8.657	7.454	7.443	-
LDhat (residuals)	5.357	5.355	5.353	5.353
LDU (residuals)	5.468	5.467	5.435	5.435
Non-synonymous	6.992	6.979	1.420	1.128
Synonymous	7.545	7.542	-	-
Gene expression	1.836	1.813	1.810	1.807
DNA replication time	1.985	1.982	1.980	1.978
GC content	2.153	1.798	1.751	1.736
Hi-C	1.766	1.757	1.757	1.756
Gene size log in kilo bases	9.813	-	-	-

Note. Variables with VIF > 7 were removed from further analysis.

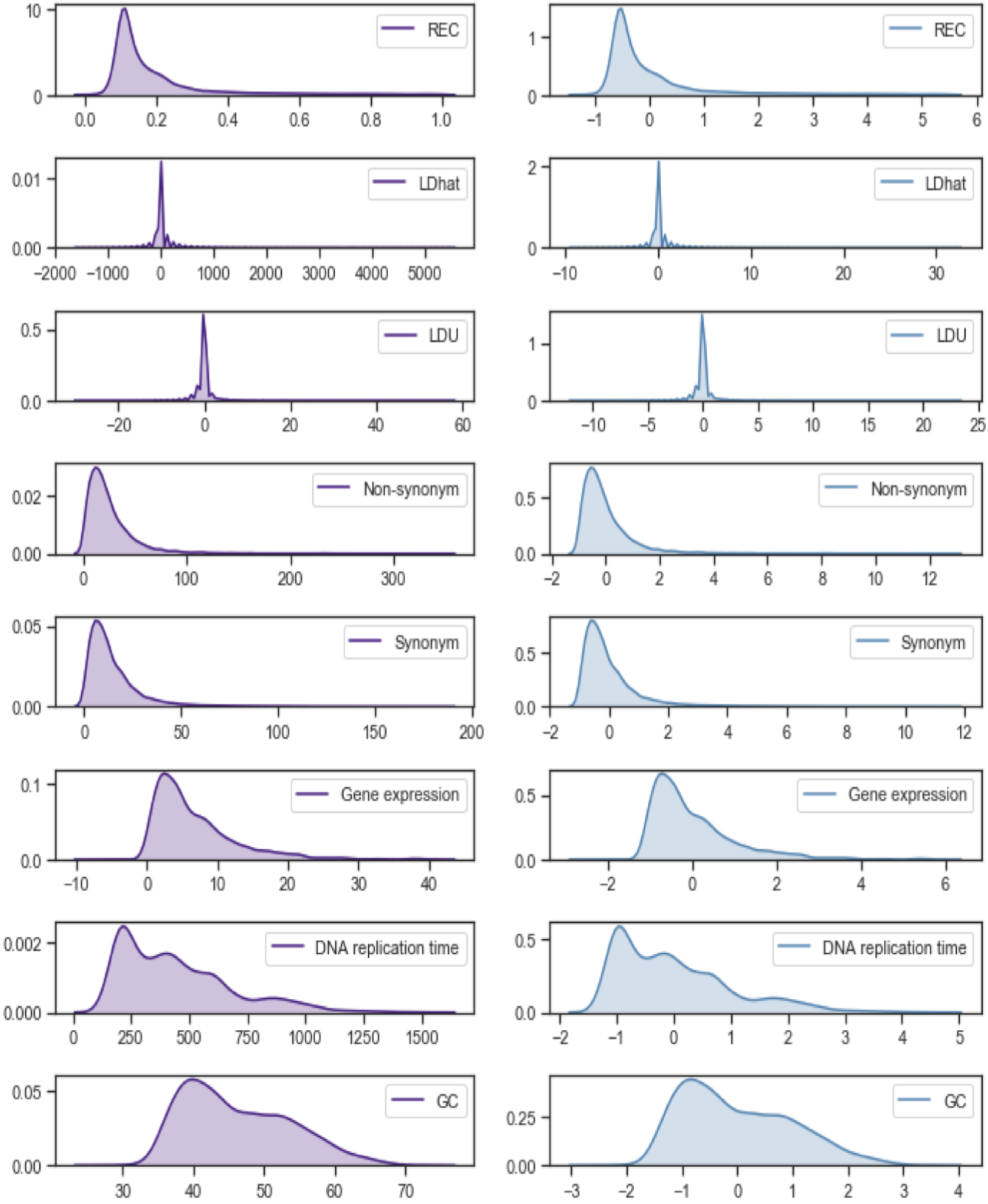


Figure 4.6: **Feature scaling performed per feature**. On the left side shows the original scale of the distribution, while the right side displays the new scale distribution.

4.3.4.6 Resampling for imbalanced dataset

The data imbalance problem is shown clearly in Table 4.5 in the training data. For example, the NDNE gene group has nearly eight times as many genes as the END gene group. The training and testing data are derived from the given gene data using data points from the majority and minority classes. The random undersampling technique was applied to the training data to generate a resampled training set to which the different machine learning algorithms were applied, whereas for a given prediction task, the testing data was kept constant between the different ML techniques for a fair comparison. All of the training data points from the MNC gene group (minority class) were used. Genes at this point were randomly removed from the NDNE, CNM and END gene groups (majority training data) until the desired balance was achieved. The number of genes selected for estimating the different machine learning algorithms was 608 for each gene group (see Table 4.9). Subsequently, a classification algorithm was applied to generate a prediction model examined on the test data to evaluate its efficacy. The steps were repeated for each ML algorithm. Intuitively, one of the advantages of undersampling is that it reduces the overall training data set size, thereby saving memory and speeding up the classification process.

Table 4.9: Number of genes in the re-balanced set by group

Gene group	Undersampling*
NDNE	608
MNC	608
CNM	608
END	608
Total	2,432

Note. *Data artificially re-balanced via random undersampling technique.

4.3.4.7 Supervised machine learning algorithms

All classifiers were trained for 14 gene-specific metrics. During the five-fold cross-validation process for each fold, the resampling technique was applied. Custom scripts were used for pre-processing the data, and `scikit-learn` from the Python libraries was implemented to perform all classification tasks. The average accuracy of each of the performance measures across the five iterations is shown in Table 4.10. The accuracy information is also presented for unbalanced data and the resampling data.

The accuracy averages for all the supervised ML algorithms are misleading in the imbalanced data. For example, ∼70% of the genes are related to the NDNE group; thus, most of the results will overfit this group as the majority class can still achieve very high accuracy. In this situation, the classifier is more sensitive to detecting the majority class patterns (NDNE gene group) but less sensitive to detecting the minority class patterns (MNC gene group). Moreover, ML algorithms are designed to maximise overall accuracy.

Of the results shown below (see Table 4.10), the Random forest (RF) and Gradient tree boosting (GTB) algorithms achieved the highest average accuracy for the balanced data (48.9% and 47.8% respectively). In addition, three of the algorithms have the highest test accuracy; however, this implies that these models are facing an overfitting problem since their training accuracy is around 100%, while their test accuracy is only around ∼40% to ∼48%. Therefore, these algorithms cannot be trusted in terms of their generalisation ability. For instance, the RF algorithm tends to fit all samples perfectly in the training data set when the model does not limit the maximum depth. Therefore the model can keep growing until it has exactly one leaf node for every single observation, perfectly classifying all of the genes. In the case of the K–nearest neighbour (KNN) trained model, there is no parameter to estimate. A bagging classifier overfits the trained model in the same way as the RF algorithm. All the supervised ML algorithms calculated on average around 47% of the genes into the four gene groups on the test data using the balanced groups (Table 4.10). This demonstrates that the features used do not describe all the gene group properties. However, these scores still provide important information to highlight potential candidates for genome filtering. The better performing algorithms were RF and GTB; therefore, those models were selected for parameter tuning.

Table 4.10: 5-fold cross-validation average for balance and unbalanced data by different machine learning algorithms

Machine learning algorithm	Unbalanced data		Sampling technique	
	MLA train accuracy mean	MLA test accuracy mean	MLA train accuracy mean	MLA test accuracy mean
***Gradient boosting**	0.766	0.715	0.735	0.478
Logistic regression cross-validation	0.711	0.710	0.487	0.479
Multi-layer perceptron	0.739	0.707	0.594	0.472
Support vector clustering	0.723	0.707	0.555	0.487
Linear discriminant analysis	0.707	0.706	0.479	0.467
Bagging classifier	0.979	0.699	0.983	0.455
***Random forest**	0.968	0.695	0.997	0.489
K-nearest neighbour	-	0.682	1.000	0.405
Quadratic discriminant analysis	0.680	0.675	0.462	0.441
Gaussian naive Bayes	0.646	0.644	0.440	0.436
Multinomial	0.564	0.562	0.488	0.479

Note. *The best-performing learners.

4.3.4.8 Bayesian optimization for parameter tuning

Hyperparameter tuning was carried out for the RF and GTB models using Bayesian optimisation. This technique evaluated the objective function by selecting the next input values based on those that have done well in the past. The objective function was validated during the training process using a specific set of hyperparameters as seen in Figure 4.7. For this process, the library `Hyperopt` in Python was implemented. First, the objective function was set by the domain space to search. Second, the surrogate model was constructed and the next hyperparameter values to evaluate were chosen. Third, the outcomes were stored from evaluations of the objective function consisting of the hyperparameters. This process was repeated until max iteration (100 times), and in each iteration, the resampling technique was simultaneously implemented.

Nine parameters were selected to be optimised for the GTB model, and five parameters were optimised for the RF model (see Figure 4.7). The mean validation was reported over five randomly initialised runs for each strategy on a withheld validation set. The results are presented in Figure 4.7 and contrasted with the average results achieved using the best parameters found by the optimisation method. Each algorithm was repeated 100 times. For the GTB classifier, to avoid overfitting on the training data, the optimal tree depth was set to five, and the weight of the regularisation term was set at 0.893. It also was found that increasing the learning rate degrades model accuracy. The optimal value of the learning rate was found to be 0.094 for all iterations. The maximum number of tree leaves for base learners was set to 20, and the best result was achieved with the aggressive subsampling (.782) of the training data; ∼44% to ∼55%. The best result achieved after tuning the model was 53.29% on the test set. As expected, RF does not overfit the training data after finding the optimal tree depth to be 5. Moreover, to reduce uncertainty in the model, a suitable function to measure the quality of the split was found to be Gini. The results also indicate that the number of features for the best separation was 13 and the number of trees in the forest was 226. This approach achieved an accuracy of ∼53% on the test data.

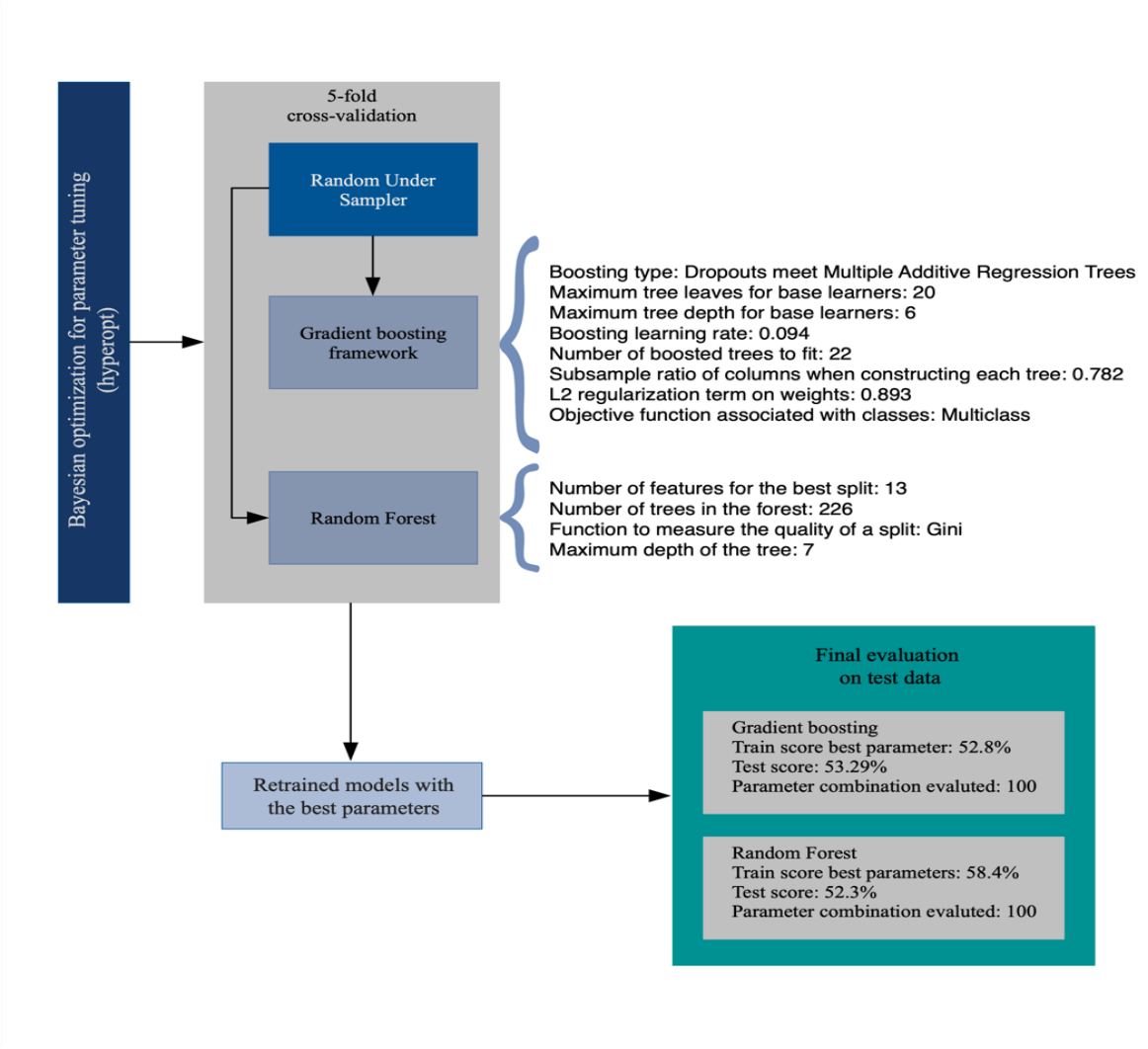


Figure 4.7: **Tuning parameters process**. First, the best models performances are selected, then an undersampling technique and the estimation of the model based on the hyperparameters is performed during the five-fold cross-validation process. Finally, the set of hyperparameters are optimise for the model to be assed in the test data.

To evaluate robustness and overfitting of the two classifiers, Table 4.11 and Table 4.12 summarise the predicted performance on the training, validation and test sets for both GTB and RF. Three metrics were used: precision, recall and F_1-score. The results show that both the GTB and RF methods do not overfit the training data and perform favourably in both validation and test sets.

The metrics evaluation for the GTB classifier on the test data were different among the groups, as seen in Table 4.11. For example, the NDNE gene group has a higher frequency than the other gene groups. 86% of the genes were correctly identified by the ML classifier (∼2,700 of the 3,169 genes). With the recall measure, 59% of the genes were correctly identified out of all the genes that belong to this class (1,870 of the 3,169 genes). The F_1-score reflected that 70% of the genes were correctly labelled and shows how robust the model was to identify genes that correspond to that class (2,218 of the 3,169 genes). The gradient boosting classifier tended to be weak for the MNC, CNM and END gene groups, with a precision of 18% (39 of the 214 genes), 25% (160 of the

638 genes) and 19% (75 of the 341 genes), respectively. Similarly, recall measures are low for these groups in classifying genes correctly. These results show that the GB classifier has poor sensitivity and performance for correctly classifying the genes.

Table 4.11: Comparison of metrics measures for gradient boosting classifier

Gene group	Training				Validation				Test			
	Precision	Recall	F_1-score	Genes	Precision	Recall	F_1-score	Genes	Precision	Recall	F_1-score	Genes
NDNE	60%	75%	67%	608	50%	62%	56%	608	86%	59%	70%	3,169
MNC	63%	72%	67%	608	52%	60%	56%	608	18%	65%	28%	214
CNM	67%	38%	48%	608	40%	22%	29%	608	25%	22%	24%	638
END	61%	62%	61%	608	48%	51%	49%	608	19%	43%	27%	392

The RF classifier was evaluated on the test data in the same way as the GTB classifier (see Table 4.12). The metrics for the NDNE gene group achieved the highest precision (86%, ∼2,700 of the 3,169 genes), recall (59%, ∼1,870 of the 3,169 genes) and F_1-score (70%, ∼2,218 of the 3,169 genes) values. The CNM group was one of the most difficult for classifying genes, with a precision of 23% (∼147 of the 638 genes). The recall measure was 20% (128 of the 638 genes) and the F_1-score was 21% (∼134 of the 638 genes). The END group was also difficult for predicting genes, with a precision of 21% (∼83 of the 392 genes), and the F_1-score was 30% (∼118 of the 392 genes). However, the recall measure accurately recognised 48% of all the genes in this class. For the MNC gene group, the RF classifier approximately identified just one of four genes that belong to this class, similar to the END group. The recall measure yielded the highest value for MNC; 64% of the genes in this group were accurately recognised. Therefore, as evaluated by precision, recall and the F_1-score, the random forest algorithm exhibited poor performance in gene classification.

Table 4.12: Comparison of metrics measures for random forest classifier

Gene group	Training				Validation				Test			
	Precision	Recall	F_1-score	Genes	Precision	Recall	F_1-score	Genes	Precision	Recall	F_1-score	Genes
NDNE	65%	76%	70%	608	52%	61%	56%	608	86%	59%	70%	3,169
MNC	68%	78%	72%	608	52%	62%	57%	608	17%	64%	26%	214
CNM	90%	57%	70%	608	44%	24%	31%	608	23%	20%	21%	638
END	70%	74%	72%	608	47%	52%	50%	608	21%	48%	30%	392

Figure 4.8 shows feature impact within the trained model for the gene groups. This feature importance was ranked in descending order, indicating the value's effect associated with the higher (in red) or lower prediction (in blue). The correlation is shown on the X-axis where REC residuals had higher importance and were negatively correlated with the gene groups, most likely because the distribution of this feature is highly skewed. pLI was positively correlated with the gene groups, which shows that those genes are likely to be essential genes; However, it may also suggest that a degree of separation between LoF and recessive genes, will be a substantial overlap [184].

In summary, the results of the ML models showed that REC, pLI, Nonsynonymous, GIMS, RVIS have the most significant impact on prediction. This suggests that genes in this state have a high degree of conservation and are related to the gene's essentiality, which implies being highly intolerant of variation due to loss of function. Furthermore, these genes could explain that the

strength and consistency of purifying selection act against functional variation. These genes may encode certain regulators of core cellular functions [196].

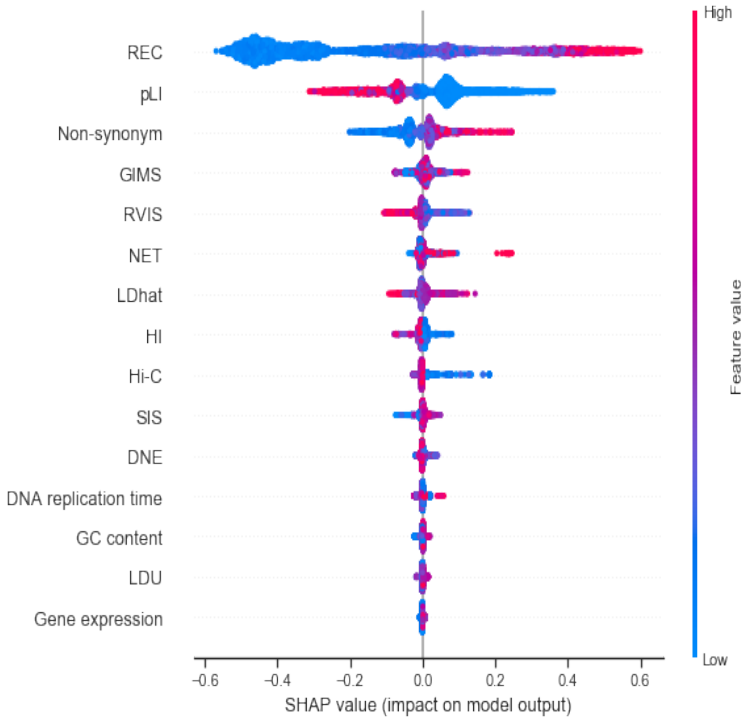


Figure 4.8: **Feature importance** describes the prediction of an gene by computing the contribution of each feature to the prediction. The position on the y-axis is determined by the feature and on the x-axis by the prediction value. The colour represents the value of the feature from low to high. Overlapping points are jittered in y-axis direction. The features are ordered according to their importance.

4.3.4.9 Structure learning

The BGGM was used for determining conditional relationships between features. This model was developed by identifying non-zero off-diagonal elements in the inverse-covariance matrix. The learning structure determined which partial correlations were non-zero. The node names are shown in Table 4.13. For prior distributions of the precision matrix, the G-Wishart distribution $W_G(3, I_{14})$ was assigned. The function was run for 10,000 iterations with 7,000 as burn-in.

The model estimated that almost half of the nodes shared connectivity of ≈0.60%, which is considered a dense network. Figure 4.9 shows the graphical structures of BGGM with 0.95 thresholds for the analytical solution. The estimated structure had strongest connections in a positive direction between the LDU and LDhat nodes (3-4), SIS and DNE scores (7-14), pLI and HI features (1-5), HI and REC scores (5-13) and RVIS and REC (2-14). The nodes that had a conditional dependence structure in the negative direction are the RVIS and SIS nodes (2-7), DNA replication time and Hi-C status features (8-11), Gene expression and DNA replication time nodes (10-11) and GIMS and SIS scores (6-7). These results appear to be strongly consistent with the application of the supervised ML techniques.

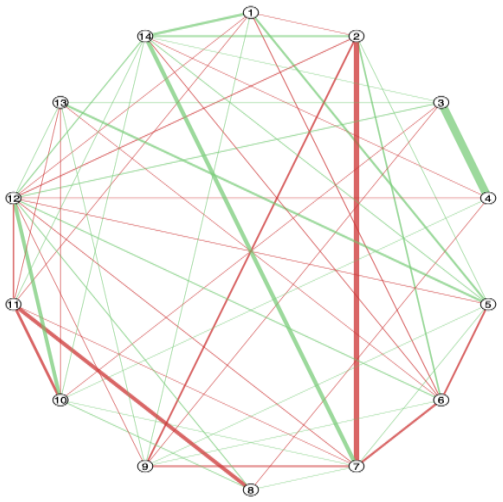


Figure 4.9: **Graphical structure of gene property features**. This figure shows partial correlations (edges) between connected nodes with posterior means and 95% credible intervals. The partial correlations are pairwise relationships in which all other variables have been controlled for. When there is evidence for a non-zero, it indicates a direct association between two variables. Green lines indicate positive correlations and red lines indicate negative correlations. Line width reflects edge strength (effect size).

Table 4.13: Node Names

Feature	Node
pLI	1
RVIS	2
LDU (residuals)	3
LDhat (residuals)	4
HI	5
GIMS	6
SIS	7
Hi-C	8
Non-synonym	9
Gene expression	10
DNA replication time	11
GC content	12
REC	13
DNE	14

In Figure 4.10, the graph at the top left has the highest posterior probability with links for which the estimated posterior probabilities are greater than 0.5. This graph follows the same Pearson correlation pattern showed earlier. The graph at the top right gives the estimated posterior probabilities of all the graphs; it indicates that the algorithm found more than 600 different graphs. The graph at the bottom left shows the estimated posterior probabilities of the size of the graphs, meaning that there is 64% of connectivity between the features. This graph also suggests that the algorithm estimated mainly graphs with measures between 60 and 68 links. At the bottom right is the trace of the algorithm based on the size of the graphs. These results suggest the adjacency matrix of the selected graph estimated by posterior probabilities of all possible links found 54 significant interactions between features.

The BGGM structure was estimated using posterior sampling and a region of practical equivalence

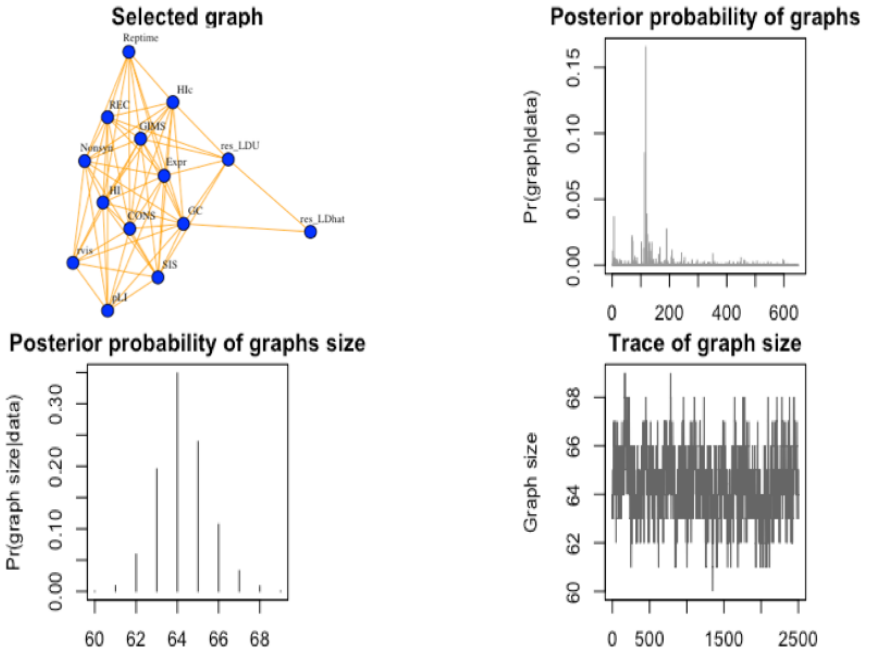


Figure 4.10: **Graphical summary of gene property features using BGGM**. At the top left is the inferred graph with the highest posterior probability. The figure at the top right gives the estimated posterior probabilities of all visited graphs. The figure at the bottom left gives the estimated posterior probabilities of all visited graphs based on the sizes of the graphs. The figure at the bottom right gives the trace of our algorithm based on the size of the graphs.

with the most conservative threshold. The model was computed by first calculating the edges, and then the posterior distributions were subtracted using the region of practical equivalence (± 0.1). The error bars correspond to 95% credible intervals. This estimated structure had fewer connections, only 22%. For example, Figure 4.11 shows that in the majority of the edges from REC, Hi-C, Gene expression and DNA replication the conditional dependence was practically equivalent to zero (Supplementary Table 2). These results suggest that those features have negligible relevance for determining conditional correlations between variables. This discrepancy could be due to the inflated false positive rate of $_1$-regularised estimation [134].

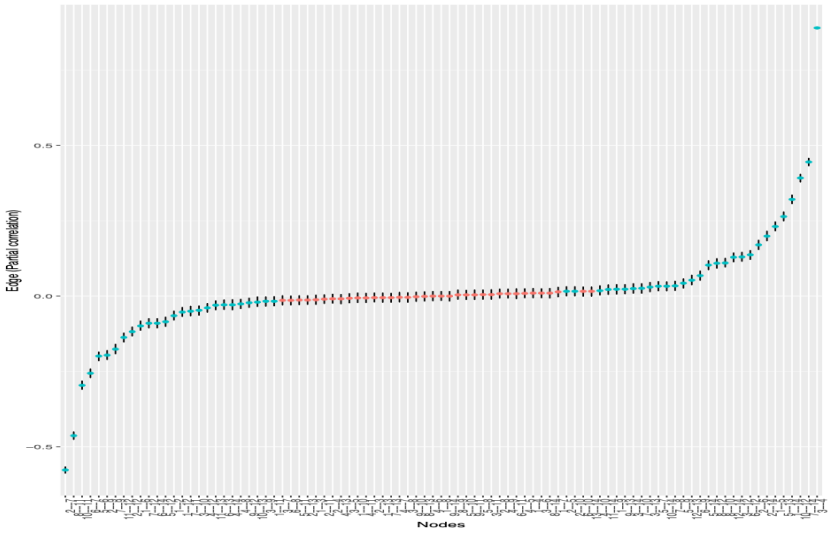


Figure 4.11: **Graphical summary of gene property features using BGGM**. The figure displays the conditional dependencies for the posterior means and 95% credible intervals. The red points denote intervals that excluded zero.

4.4 Discussion

For many years, the advent of next-generation sequencing (NGS) has driven progress in identifying disease causal genes and variants in the human genome for rare Mendelian disorders. Despite the apparent utility of high throughput technologies, hundreds if not thousands of causal genes are still unrecognised [52, 164, 166]. The integration of gene discovery methods is required for an understanding of the mechanisms of human phenotypes. Here, quantitative scores related to human disease genes, evolutionary conservation, variant intolerance genes, gene interaction protein network, and DNA sequence context have been evaluated to improve recognition of genes that are likely contributing to monogenic phenotypes.

To this end, the present study has introduced some of the most useful supervised learning techniques for modelling Bayesian classifiers, logistic regression, discriminant analysis, classification trees, nearest neighbour, neural networks, support vector machines and ensembles of classifiers. All of these supervised methods used a training data set where the classification was known a priori using the Spataro *et al.* [52] groups to develop a classifier, and this trained model was applied to classify the test data.

The combined analysis of genomic data for the gradient boosting and random forest models showed that more than ∼50% more of the variance is explained than taking the scores individually. In addition, the results from BGMM are consistent with the results of the supervised ML model. That is, the connectivity between the gene metrics used was 40%. Therefore, the features used are not sufficiently reliable for predicting causal genes for most monogenic disorders.

Although ML classifiers have been used to recognise potential Mendelian disease genes in the NDNE, CNM and END groups, the approaches do not identify a clear relationship between genes. This poor classification can be explained by the fact that essential genes are expected to overlap in monogenic

disorders [169]. Both classifiers have shown that more than 70% of the genes were found in the NDNE groups. In particular, this group has many genes, and a considerable number of them are likely to be undetected Mendelian genes which lead to inconsistent results and misinterpretations of underlying disease causal mutations [53].

The main result suggests that all subsets of MNC and CNM genes are under similar evolutionary pressures, which may be a substantial overlap between those groups. Consistently with previous studies [52, 216], genes that are classed in the human diseases group according to hOMIM follow a dominant transmission pattern were more conserved and enriched of rare frequency variants than recessive genes. These results are consistent with the hypothesis that dominant genes will be under stronger purifying selection. Moreover, Spataro *et al.* [52] recognised that more than 23% of genes linked to a Mendelian disorder are also associated with at least one complex after compiling 887 Mendelian genes. Furthermore, the intersection between disease genes explains specific biological features, indicating a prominent role of MNC genes in the aetiology of complex disease. Indeed, CNM genes present higher functional importance in the protein network, a tendency towards higher expression levels and are enriched in relevant biological categories.

Identifying genes implicated in monogenic disease from the analysis of the supervised ML models shows that the major features appear to be consistent with previous studies; that is, pLI and RVIS provide the most information for the classification shown in Table 4.1. These results suggest that the tolerance of LoF variation quantifies the degree of essentiality that aids the gene prediction of monogenic disorders [167, 52, 53]. By contrast, LDU residuals and features related to the local sequence context of the gene, such as GC content and gene expression, are the least informative features in the supervised ML model. These results reflect that detecting selection is difficult, as signatures are confounded by processes such as recombination and drift and the effects of changing demography over time [42]. In the Alyousfi *et al.* [196] study, the REC variable showed a small contribution to the combined essentiality-specific pathogenicity prioritisation (ESPP) gene score used to predict disease genes. The authors also found that SIS and NET explained a high proportion of the variance from the ESPP score. Here, in this analysis, the variable REC was very important for this model; however, this result may not be robust because of the lack of information from the original source [184]. Despite the use of imputation technique used for the REC feature, the estimation is still biased. Finally, NET and SIS had a much smaller effect on the model. The NET cannot differentiate between known disease-associated genes [167] while SIS had several missing values from the original source [71]. Interestingly, in the supervised ML analysis, the gene expression feature has the most negligible impact in the model; this result might reflect the fact that those genes could be particularly prone to splicing mutations, which are more challenging to recognise and assess [217].

Overall, pre-processing and balancing the genomic data was not sufficient to yield a robust estimate from the performance of the tuned machine learning model. Moving forward, the predictive value of ML-based classifier ensembles and the BGMM approach provides evidence of the explanatory power of using gene-level metrics. In particular, genes found to be contributing to Mendelian disorders tend to be complex in structure. For monogenic diseases, one or several validated causal variants

are associated with some variants from complex disorders [169]. Furthermore, in complex disorders, many genes may act together and modify the effect of each other, jointly contributing to disease development [52]. Despite this, it might be reasonable to separate genes for complex disorders from gene prioritisation for monogenic disorders and reformulate. For example, the Abramovs *et al.* [218] study developed a gene variation intolerance rank for prioritising disease genes. They divided Mendelian disorders into dominant and recessive Mendelian genes. Their results revealed that the most intolerant genes are potential genes related to severe dominant disorders.

Although by no means comprehensive, the preceding examples represent both the basic principles and common challenges of using combined omics data to recognise human disease genes. The difficulties of combining diverse genic properties in this way are evident, as are the problems of modelling highly unbalanced data where the majority class (NDNE) is likely to contain numerous misclassified genes that are the monogenic genes research set out to discover. However, genes positioned towards the end of the essentiality spectrum likely represent the best candidates for unrecognised monogenic disease genes. Along with the impact of variability in quality and completeness of individual gene-specific scores, an incomplete understanding of the gene groups make it challenging to classify them. This study is limited by the accuracy of the genetic information currently available for human diseases and the incomplete knowledge regarding the true susceptibility/causal variants and their corresponding genes. Identifying new genes, which have not yet been classified to the group of genes already known to be related in MNC, is the rationale behind this study. Therefore, it is expected to include misassignments. Another challenge is the difficulty in recognising essential genes, given that the inactivation of an essential gene is fatal. For this reason, integration of new functional genomic data across multiple levels of gene properties may reveal both candidate causal variants and one or more target genes for downstream study.

Chapter 5

Robust predictions to identify disease genes using unsupervised machine learning

5.1 Introduction

Causal variant identification often begins by understanding their function at a molecular level. Clustering genes can recognise these causal variants according to the similarity of their genomic and functional properties [177]. Two factors often link genes that underlie human disease; a higher probability of physical interaction between their products and higher expression profile similarities for their transcripts [219]. As a result of the interplay between gene properties, Spataro *et al.* [52] demonstrated that the biological properties and the evolutionary history of human disease genes are different compared to non-disease genes and to putatively essential human genes.

Gene property data can be represented as a matrix, with each row corresponding to a gene and each column denoting a particular property. In this case, the properties are related to genomics, functional and biological information describing the gene. Each entry of the matrix is a numeric representation of the gene describing a given property for that particular gene [220]. This structure can be used to compare a gene's profile to produce a cluster of known/candidates disease genes distinct from non-disease genes. Thus, this structure falls into the category of unsupervised machine learning (ML) methods. Unsupervised learning algorithms are used when observation labelling is not available and the analysis aims to discover hidden patterns between genes properties and genes. Moreover, these methods may help to identify potentially novel patterns of genomic elements [84]. In addition, the unsupervised ML method can partition genes into groups and assign a label to each partition based on the functional and genomic information. The goal is then to separate the genes into groups or clusters that are more similar to each other and dissimilar between gene groups [211].

An unsupervised ML method is proposed for grouping genes based on their genomic and functional similarity properties. The proposed method integrates statistical analysis to achieve simultaneous consensus clustering and outlier detection. The method enables gene clustering by applying a Gaussian mixture model (GMM) resulting from the underlying structure of the properties of the genes without a priori information and simultaneously eliminating the negative effect of outlying data.

The simultaneous Gaussian mixture clustering with outlier removal (GMM-non) uses the selected features discussed in Chapter 4 (see Table 4.3 [167]). These gene metrics measure the degree of gene essentiality by quantifying the tolerance of loss of function (LoF) variation and the degree of conservation of genes under selection. By integrating these features, distinct gene subtype classifications might be made according to the degree of gene essentiality (see Table 5.1). In this study, the proposed multi-objective clustering technique is applied over the reduced dimensionality gene set to categorise known and novel Mendelian disease genes and separate them from complex disease and non-disease genes.

Table 5.1: Features selected

Feature	Name	Properties	Literature
pLI	pLI is the probability of loss-of-function variation	Essential and conserved genes	Samocha *et al.* (2014) [178], Lek *et al.* (2016) [177]
RVIS	Residual variation intolerance score	Essential and conserved genes	Petrovski *et al.* (2013) [173]
DNE	Gene constraint *de novo* excess score	Essential and conserved genes	Samocha *et al.* (2014) [178], Hsu *et al.* (2016) [180]
SIS	Substitution intolerance score	Essential and conserved genes	Aggarwala *et al.* (2016) [71]
HI	Deletion-based haploinsufficiency score	Haploinsufficient genes	Haung *et al.* (2010) [183]
NET	Gene position in network score	Haploinsufficient genes	Khurana *et al.* (2013) [181], Hsu *et al.* (2016) [180]
GHIS	Genome-wide haploinsufficiency score	Haploinsufficient genes	Steinberg *et al.* (2015) [182]
REC	Recessive score	Haploinsufficient genes	MacArthur *et al.* (2012) [184], Hsu *et al.* (2016) [180]
GIMS	Gene-level integrated metric of negative selection score	Genes under selection	Sampson *et al.* (2013) [186]
GDI	Gene damage index score	Genes under selection	Itan *et al.* (2015) [186], Hsu *et al.* (2016) [180]
LDU (residuals)	Linkage disequilibrium residuals from LD maps in LDU	Linkage disequilibrium	Vergara-Lope *et al.* (2019) [192], Lonjou *et al.* (2003) [197]

Note. Features were selected based on the systematic review in 'Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data' by Alyousfi D, Baralle D, and Collins A. (2020) *Briefings in Bioinformatics* [196].

5.1.1 Unsupervised machine learning algorithm

Unsupervised ML algorithms train a machine to discover hidden structures and patterns in unlabelled data [221]. The unsupervised methodology 'learns' about a set of examples by considering the specific characteristics of features. Because the method is unsupervised, the output classes must then be semantically interpreted [211]. In this context, the learning model aims to resolve genes with a high potential for containing pathogenic disease-related variation and discern relationships among genes based entirely on their genomic and functional properties.

5.1.2 Clustering approaches

Clustering techniques are known as unsupervised ML algorithms since they do not require prior knowledge about the clusters. This analysis is used to uncover latent groups of homogenous observations. The performance of clustering techniques is highly dependent on the effectiveness of the clustering algorithm [222].

The standard clustering-based algorithms are hierarchical methods and k-means clustering. These algorithms divide the entire unlabelled data into relatively homogeneous clusters in such a way as to maximise data similarity within the cluster and data dissimilarity outside the cluster [223]. Hierarchical methods build a tree-like structure wherein instances or subgroups are merged until all the genes are in a single cluster based on their genomic and functional profiles (similarity matrix). This agglomerative method merges the closest profiles, building an average profile of the joined profiles, and continues iteratively until the entire tree is built. However, it is well known that hierarchical clustering algorithms are susceptible to outliers and display highly skewed dendrograms, which may not reflect the real underlying data structure [83, 224, 223].

The k-means algorithm separates a dataset into k clusters by selecting representative gene profiles. The method defines an initial centroid randomly for each k cluster using the totality of gene profiles. The k centroids are then repeatedly updated by assigning genes to the closest centroid; the algorithm then iterates until the new centroid and the previous centroid are within some defined threshold. The k-means algorithm is more sensitive to outliers since it uses the mean of cluster data points to find the cluster centre [225].

Model-based clustering, on the other hand, predicts probabilistic distributions in the data. The model parameters can be estimated using maximum likelihood algorithms such as the expectation-maximization (EM) algorithm. A GMM implements the EM algorithm to fit a mixture of Gaussian models assuming that the probability distributions are multivariate Gaussian distributions with unknown parameters. The model parameters for Gaussian distributions (mean and variance) are considered latent variables, and the goal of EM is to compute maximum likelihood estimates of the parameters based on the data. EM clustering alternates between two steps: (E) expectation and (M) maximisation. In the E step, the current estimates of the model parameters are used to compute the posterior probability – often referred to as the responsibility – of the model parameters for every instance. In the M step, the responsibility is used to re-estimate the model parameters. These two steps are repeated until the model parameters converge. The inferential goal is then to estimate the weights and the location of the components' densities for the k groups [222, 226, 120]. Similar to clustering-based techniques, GMM is not able to deal with outliers properly. The outliers are represented with low probability, which may negatively impact the performance of Gaussian-based clustering algorithms.

5.1.3 Outlier detection

Outliers are observations that deviate significantly from the majority of the patterns in the data, which suggests that they may be generated by a different process [227]. Observations having integrated squared error greater than a threshold are also termed outliers. Cluster analysis is analogous to outlier detection as the two techniques address highly related tasks. Clustering finds the majority of patterns in a data set and organises the data accordingly, whereas outlier detection deals with recognised points belonging to none of the clusters [228, 86]. Therefore, outlier detection can be used to detect and remove those extreme observations from the dataset.

5.2 Methods

5.2.1 Simultaneous Gaussian Mixture clustering-based outlier detection algorithms

5.2.1.1 Definition of gene-classes

The GMM-non method was applied to classify disease genes as distinct from non-disease genes. This proposed method uses knowledge from the hypothetical model of Pengelly *et al.* [53] (as described in Chapter 1) to describe the degree of essentiality. The four gene groups were defined as Spataro *et al.* [52] gene groups; Mendelian non-complex (MNC), Complex non-Mendelian (CNM), Essential non-disease (END), and Non-disease non-essential (NDNE) (as discussed in Chapter 4). Briefly, the genes groups constructed by Spataro *et al.* [52] showed that, relative to non-disease genes, human disease genes have specific evolutionary profiles and protein network properties. Their results also suggest that genes linked to Mendelian disorders play an important role in driving susceptibility to complex disease.

5.2.1.2 Gene data and dimensionality reduction

The simultaneous GMM-non method consists of three consecutive stages. It begins with a data preparation step, followed by dimensionality reduction and finally cluster assignment. The analysis is shown schematically in Figure 5.2. In the first stage, 11 gene-specific metrics were used as a baseline data set for constructing the gene classification (see Table 5.1). These features have already been pre-processed by data normalisation, imputing missing values, and undertaken data transformation (as described in Chapter 4). Once the appropriate gene-specific metrics for gene clustering had been selected, in the second stage, two methods for dimensional reduction were examined prior to clustering the gene data; principal components analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE). These algorithms represent opposite ends of the spectrum ranging from projection/embedding algorithms that prioritise preservation of the global structure (PCA) to those that prioritise preservation of local structure (t-SNE) [229, 125].

In the third stage, the GMM-non method was first applied to draw the probability distribution from the gene data by summing mixture components until convergence was reached. Each mixture component represented a multivariate Gaussian distribution with four dimensions (gene groups), a multivariate mean (a 4-dimensional vector) and a covariance matrix (a 4×4 matrix). In order to map comparison of the methods, the number of mixture components was chosen (the key fitting parameter of a GMM, usually denoted k) by checking the variance structure of each cluster. This k parameter was found by comparing the results of Bayesian information criterion (BIC) and Akaike information criterion (AIC). The fitting procedure attempts to find the model that maximises the gene data's likelihood given the GMM. Secondly, the outlyingness factor is computed for each vector during the iteration process, and those data points (potential to be outliers) are removed. Thus, the GMM clusters are modified in each iteration until the maximum distance to the partition centroid is found.

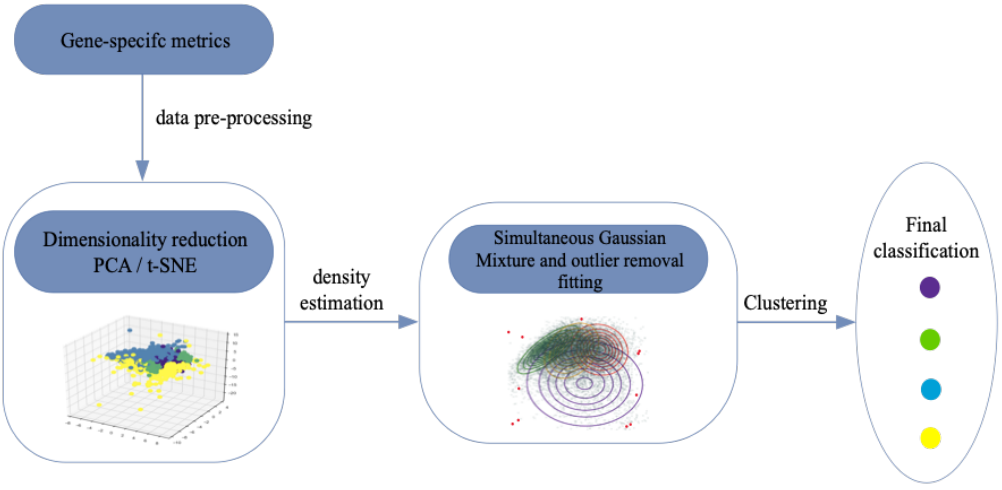


Figure 5.1: **Suggested unsupervised behavioural mapping method flow chart**. Analysis for gene clustering. The analysis is aggregated into two major segments; dimension reduction and unsupervised ML model. During the first stage, the gene data were pre-processed and then PCA was undertaken to reduce the dimensionality of the data. The simultaneous Gaussian Mixture clustering with outlier removal method was applied during the second stage to generate clusters.

5.2.1.3 Gaussian mixture model

Model-based clustering was carried out as the second stage of model construction. A GMM for the gene data X was assumed, with n observations and D variables. In particular, it was assumed that the data in each gene group was derived from a multivariate Gaussian distribution and the combined gene data was from a convex combination of multivariate Gaussian distributions. This distribution is used in Gaussian mixture modelling for G clusters, and the likelihood is given by:

$$P(X, Z|\mu_k, \Sigma_k) = \prod_{i=1}^{n} \prod_{k=1}^{G} \pi^{I(z_i=k)} \mathcal{N}(\mu, \Sigma)^{I(z_i=k)}, \tag{5.1}$$

where π_k is the probability that an observation belongs to cluster k, $\phi_k = \mathcal{N}(\mu, \Sigma)^{I(z_i=k)}$ is the normal probability density function centered at μ_k with variance–covariance matrix Σ_k, and z_{ik} defines the probability that the n^{th} observation belongs to the k^{th} cluster [230]. Thus, from the initial clusters, maximum likelihood estimation (MLE) of ϕ_k was carried out via the EM algorithm. Until convergence occurs, the EM algorithm iterates between the E-step, which computes z_{ik}, the conditional probability that observation i belongs to group k given the current parameter estimates by:

$$\log(P(X, Z|\mu_k, \Sigma_k)) = \sum_{i=1}^{n}\sum_{k=1}^{G} z_{ik}\left[\log(\pi_k) + \log(\mathcal{N}(\mu_k, \Sigma_k))\right], \tag{5.2}$$

with each z_{ik} replaced by its current conditional expectation $\hat{z_{ik}} = E(z_{ik}|\mathbf{x}_i, \phi_1, \ldots, \phi_G)$, and the M-step, where the new parameters z^*_{ik} are maximized with respect to the parameters (see Figure 5.2)

5.2.1.4 Distance-based outlier removal

For the third stage, the outlyingness factor is computed from gene data to determine which genes may be classified as outlying. Let us assume that the factor depends on the distance from the cluster centroid of k groups from the gene data X. Then the algorithm's iterations first start by finding the vector with maximum distance to the partition centroid:

$$d_{max} = \max_i\{\|\mathbf{x}_i - \mathbf{c}_{pi}\|\}, \quad i = 1, ..., I. \tag{5.3}$$

For GMM clustering, the Euclidean distance was used as a multivariate distance metric to measure the distance between each centroid and a point (vector). The centroid point corresponds to a coordinate X in the two-dimensional space. Thus, the centroid of the GMM was defined as the point with the maximal density to represent its cluster by:

$$c_{pi} = argmax \sum_{k=1}^{G} \pi_i \mathcal{N}(X|\mu_k, \Sigma_k). \tag{5.4}$$

The outlyingness factor was then calculated for each vector as:

$$o_i = \frac{\|\mathbf{x}_i - \mathbf{c}_{pi}\|}{d_{max}}, \tag{5.5}$$

where its value lies between 0 and 1 after being normalised. Finally, a threshold was set to determine all points regarded as outliers and removed from the gene data. Points with an outlier factor higher than the threshold are treated as outliers and thus removed; that is, the vectors $o_i > T$. At the end of each iteration, the GMM algorithm was run using the previous $\mathbf{c}_{pi}$ as the initial point, and a new solution was then fine-tuned for the reduced dataset. Clusters are updated at each iteration

using the new $\mathbf{c}_{pi}$ and the remaining dataset without the removed outliers. This iterative process continues until the change in log-likelihood from GMM is less than some threshold and convergence is declared (see Figure 5.2).

Proposed algorithm GMM-non

$C \leftarrow$ Run GMM until convergence with multiple initial solutions, collect the best C

$$c_{pi} = arg\ max \sum_{k=1}^{G} \pi_k N(X|\mu_k, \Sigma_k)$$

for j ← 1, … , I do (number of iteration)

$$d_{\max} = \max_i \{\|x_i - c_{pi}\|\}, \quad i = 1, \dots, I$$

for $i \leftarrow 1, \dots, N$ do (number of observations)

$$o_i = \frac{\|x_i - c_{pi}\|}{d_{\max}}$$ (compute outlier factor for each point)

if $o_i > T$ then

$X \leftarrow X/\{x_i\}$ (remove all data points respect to the threshold, and update the data)

end if

end for

$(C, P) \leftarrow GMM(X, C)$ (run GMM over the updated dataset and update each cluster)

end for

Figure 5.2: **Pseudo-code of the proposed method for simultaneous Gaussian mixture clustering outlier removal**. The pseudo-code explains the proposed method, which employs both clustering and outlier discovery to improve the estimation of the centroids of the generative distribution. First, the GMM is run until convergence and the best centroid is chosen. The algorithm continues for I iterations in each cluster to compute the outlyingness factor, then all data points exceeding the threshold are removed, and the data updated. Lastly, the GMM is run over the updated data and the cluster.

5.2.2 Data and code

The GMM-non construction and analysis were conducted in Python version 3.7.3 (https://www.python.org/), `awk` and R version 3.2.2 (https://www.r-project.org/) using custom-written scripts.

5.3 Results

5.3.1 Data

The functional and genomic metrics were described in Chapter 4, using 11 gene-level metrics (Table 4.1) to distinguish disease genes from non-disease genes. A total of 14,708 genes were allocated into the categories NDNE, MNC, CNM and END using the proposed groups of Spataro *et al.* [52] and the hypothetical model of Pengelly *et al.* [53] (as discussed in Chapter 1). Spataro gene groups were taken for comparison against the new classification provided for the proposed methodology. Here, Spataro's gene groups were further revised to include additional genes implicated in Mendelian disorders using an updated list from the genemap2.txt OMIM database https://omim.org/downloads/

(OMIM, 2020). The combined set of genes known to be involved in monogenic disorders comprised 3,350 genes for this category. In addition, to update the MNC gene group, an updated set of Mendelian gene names of disease genes was included from the HUGO Gene Nomenclature Committee (HGNC) database (https://www.genenames.org/download/statistics-and-files/) and the GeneCards database of human genes (https://www.genecards.org/cgi-bin/listdiseasecards.pl?type=full). Spataro's gene groups were used as a benchmark for showing the distribution of the genes within the gene groups. The genes in each group are unbalanced and are distributed as follows: 8,986 NDNE genes, 3,362 MNC, 1,588 CNM and 891 END.

5.3.2 Dimensionality reduction

Dimensionality reduction was used for data visualisation, reducing the number of data inputs to a manageable size, preserving the dataset's integrity as much as possible. It was implemented in the pre-processing data stage, and there were applied two different dimensionality reduction methods: PCA and t-SNE.

The results of applying PCA and t-SNE dimensionality reduction to 11 features extracted from the functional and genomic data are given in Figure 5.3. Although these methods can separate gene groups, PCA and the embedded feature space (t-SNE) overlap among the four gene groups. The results suggest that the dimensionality reduction methodologies applied could not identify an internal structure or a noticeable pattern from the set of p=11 features. For the MNC gene group, however, it can be observed that some genes have the highest PCA units along with the CNM group. Although t-SNE may improve the translation of multidimensional data into low dimensions upon nonlinear dimensionality-reduction, in this study, PCA was used for subsequent analysis because the pairwise distance between two genes is low, suggesting that genes are similar and that linear mapping is preserved in the data over short distances [231, 232].

PCA components determined to be orthogonal transform the 11 correlated features into a set of linearly uncorrelated features. As there is no significant test to elect principal components, researchers have been forced to devise some rules throughout time. Therefore, the optimal number of dimensions was chosen based on the shuffling procedure described by Berman *et al.* [233]. This procedure quantifies the cumulative explained variance ratio as a function of the number of components. The procedure indicated that between 1 and 20 principal components contained variance above sampling error and should be retained, explaining 50%–70% of the data variance, which implies that the eigenvalues greater than one can be considered as having a significant impact on the target variable. Figure 5.4 shows the cumulative proportion of explained variance. Therefore, two components were retained for the subsequent analysis as the first component explained ∼42% of the total variance, and the second component explained ∼11% of the total variance. The cumulative percentage variance described for the first two components was ∼53%.

The 2-dimensional PCA results of the 11 features are given in Table 5.2. These results show the importance of each feature reflected by the magnitude of the corresponding characteristic vectors.

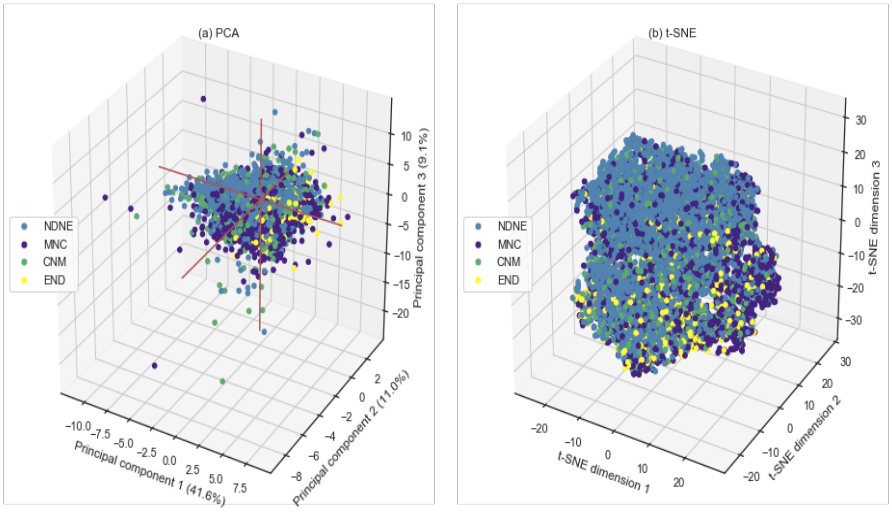
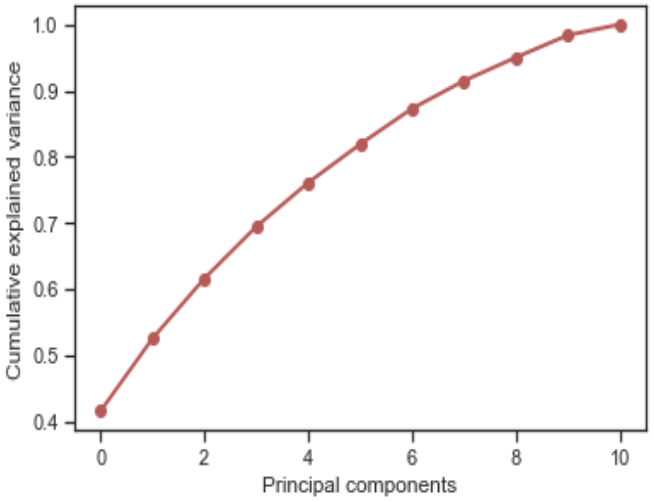


Figure 5.3: **PCA (a) and t-SNE (b) plot of a three-dimensional gene cloud**. Comparison between 3-dimensional PCA (a) and t-SNE (b) spatial display of 11 features from genomic and functional data using 14,708 genes. For PCA, the percentage variation explained by each component is indicated in parentheses and the PC are ordered by their ability to explain the variance. Gene groups proposed by Spataro *et al.* (2017) are color-coded as follows: NDNE in blue, MNC in purple, CNM in green and END in yellow. The gene distribution is 8,986 NDNE, 3,362 MNC, 1,588 CNM and 891 END.



Principal components	Variance Explained
1	0.416
2	0.110
3	0.091
4	0.078
5	0.067
6	0.058
7	0.054
8	0.042
9	0.035
10	0.034
11	0.016

Figure 5.4: **Explained variance ratio**. The curve that quantifies the 11-dimensional variance. The first two-dimensional projection describes more than ∼53% of the total variance of the data, and three components describe close to 63% of the variance.

The important features are REC and SIS, explaining 13.41% and 10.52% of the total variance, respectively. linkage disequilibrium unit (LDU) (∼4%) and pLI (∼8%) showed the smallest proportion of variance explained in the 2-dimensional space, but these features have a moderate impact on the second principal component weight. Interestingly, residual variation intolerance score (RVIS) had the most weight in the second principal component. These results suggest that the first component could be related to explaining intolerance with LoF, while the second component could be associated with the selection process.

Table 5.2: Distribution of the gene-specific mean among gene groups

Feature	Principal component loadings		Variance explained	Proportion of variance explained
	PCA 1	PCA 2		
REC	−0.32	0.10	6.20	13.41
SIS	0.33	0.23	4.86	10.52
HI	−0.34	0.03	4.62	9.99
NET	−0.40	−0.21	4.44	9.60
GIMS	−0.30	0.35	4.34	9.39
GHIS	−0.28	0.36	4.24	9.17
RVIS	−0.15	0.67	4.20	9.09
GDI	−0.35	−0.18	3.84	8.32
DNE	0.38	0.12	3.81	8.25
pLI	0.24	0.31	3.66	7.93
LDU (residuals)	0.05	0.22	2.00	4.33

Figure 5.5 displays all the component lengths, showing their magnitudes. The first principal component can be attributed to a measure of haploinsufficiency by recessive (REC), Haploinsufficiency (HI) and gene position in networks (NET), while the second component corresponds to a measure of essential genes and intolerance to LoF by substitution intolerance score (SIS), genome-wide haploinsufficiency (GHIS), gene constraint *de novo* excess (DNE), loss intolerance probability (pLI), but which at the same time is inversely related to the measure of genes under selection by gene-level integrated metric of negative selection (GIMS), RVIS, gene damage index (GDI) and LDU. For REC, SIS, HI, NET, pLI, DNE, SIS and GHIS are oriented in the same direction, indicating a correlation. The directions for GIMS, RVIS, GDI and LDU are opposite, indicating that the second principal component might be related with genes that tend to be intolerant to LoF, reflecting an intense impact of selection and *vice versa*. Following the gene-groups, CNM probability seems to be influenced by PCA2, which goes beyond the PCA1 trend reflecting the scheme shown for the END groups. This result suggests that it could be applied to GWASs, where both axes define CNM probability and might define a classifier based on these components.

5.3.3 Clustering approaches

Once the reduced gene space was formed, hierarchical and k-means clustering methods were carried out over this reduced gene space. The hierarchical clustering model was run first. This unsupervised approach identifies 2 gene groups (Figure 5.6 (a)). The hierarchical clustering performance is displayed in Figure 5.6 (b). The Figure shows hierarchical clusters under the two-dimensional projection space for four groups. This approach allocated 4,535 genes to NDNE, 4,630 genes to MNC, 3,859 genes to CNM and 1,803 genes to END. Comparing the hierarchical clusters and the Spataro gene groups, it was observed that the cluster assignment was not robust for END and CNM. These results suggest that the algorithm tends to concentrate on local clusters instead of the global expression pattern. Due to this local effect, errors in the early stage of cluster assignment increased the final result.

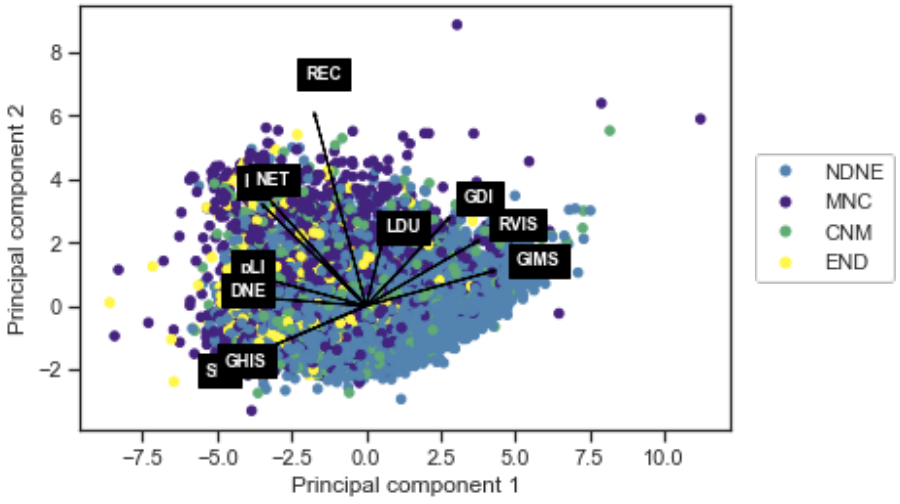


Figure 5.5: **Principal component analysis of functional and genomic data for feature importance**. Two-dimensional projection of impact and importance of 11 features. The vector loadings for each of 11 features under the two-dimension projection represent how important they are in describing the distribution of the functional and genomic data; specifically, they represent the measure of the variance of the gene data. Colours indicate Spataro's gene groups: NDNE in blue, MNC in purple, CNM in green and END in yellow.

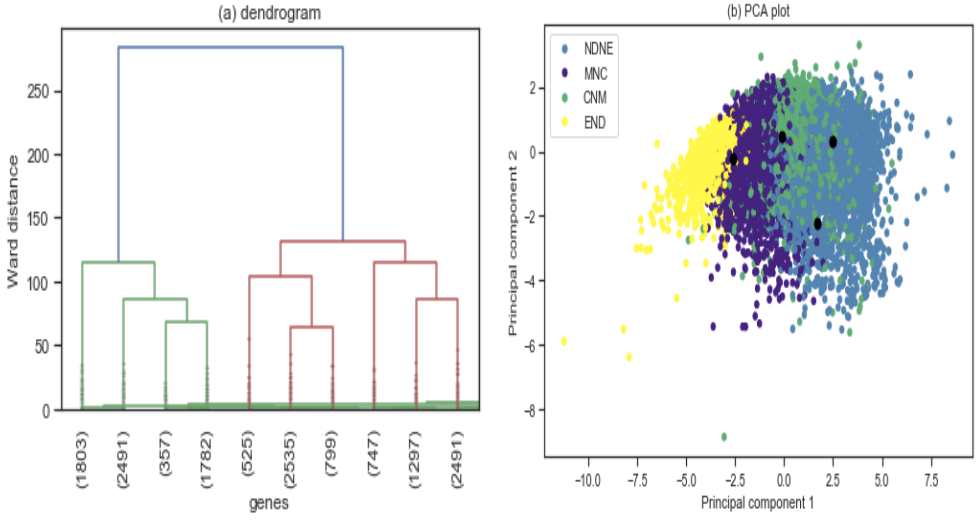


Figure 5.6: **Hierarchical clustering of functional (a) and genomic data and principal components analysis plot for hierarchical clustering (b)**. The dendrogram (a) was obtained by carrying out hierarchical clustering using ward distances and average linkage strategy. Each leaf represents one of the 14,708 genes and colours indicate the selected clusters. The vertical axis represents the distance between nodes of the tree. In the PCA plot (b) for functional and genomic data, colours indicate the four classes from the hierarchical approach in the gene data: 4,535 NDNE, 4,630 MNC, 3,859 CNM and 1,803 END.

Secondly, the k-means clustering was applied to assigned genes into four classes (Figure 5.7). The k-means procedure yielded a partition very close to the underlying structure in the functional and genomic data. Following Spataro's gene-groups, k-means did not stratify genes according to END, MNC and CNM. The approach assigned 4,143 genes to NDNE, 5,662 genes to MNC, 3,716 genes to CNM and 1,306 genes to END. These results may imply that the similarity between the END, MNC and CNM gene groups become isometric when more genes are involved and noisy genes can

further distort the relationship.

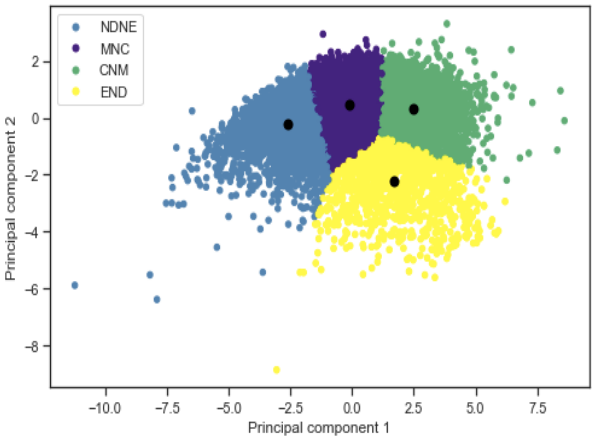


Figure 5.7: **Principal components analysis plot for k-means clustering**. The scatter plot with the two first largest components from the PCA for the functional and genomic data. Colours indicate the four classes in the gene data: 4,143 NDNE genes, 5,662 MNC genes, 3,716 CNM genes, and 1,306 END genes.

5.3.4 Application of the proposed approach

The determination of the number of GMM-non components was based on the likelihood function using BIC and AIC. The optimal number of clusters is the value that minimises the BIC and AIC. Following these techniques, Figure 5.8 suggests that a four-component Gaussian mixture was more appropriate to represent the gene data. This result is consistent with the gene groups proposed by Spataro *et al.* [52].

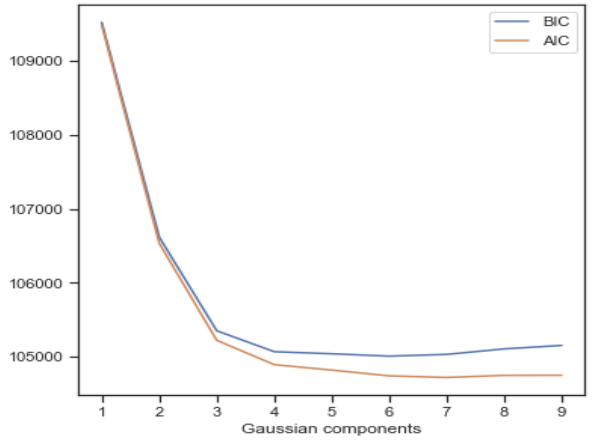


Figure 5.8: **Principal components analysis plot for k-means clustering**. This graph displays the optimal Gaussian components.

Once the appropriate number of GMM-non components was set, the proposed algorithm was applied to functional and genomic data, and distinct gene classes can be identified. Briefly, the GMM-non method seeks to model the relationship of the genes via a mixture of Gaussian density distributions that can conceptually be visualised as a collection of 4-dimensional Gaussian distributions overlaying

a two-dimensional projection. The peaks of each distribution form where points are most dense and the ellipses of the distribution conform to nearby points (see Figure 5.9 (a)). The maximum likelihoods for distribution parameters of GMM are updated using a series of successive steps in the expectation-maximisation algorithm. The genes in a GMM cluster will always meet the outlyingness factor assumption on the outlier removal step. Once this is completed, every gene receives a label with membership probabilities, indicating the distribution to which it most likely belongs. The results of the proposed algorithm can be seen in Figure 5.9 (b). Genes are coloured according to the different gene classes: NDNE in blue, MNC in purple, CNM in green and END in yellow. Genes coloured maroon were determined to be outliers and were removed from the data. The distribution of the genes within the gene classes found by the GMM-non model with outlier removal was 3,630 in NDNE, 5,865 in MNC, 3,808 in CNM and 1,510 in END, excluding 14 potential outlier genes. The proportion of outlier genes was relatively small within the gene classes: 6 in NDNE, 6 in MNC and 2 in CNM, indicating that the proposed methods can detect genes related to various outlier genes.

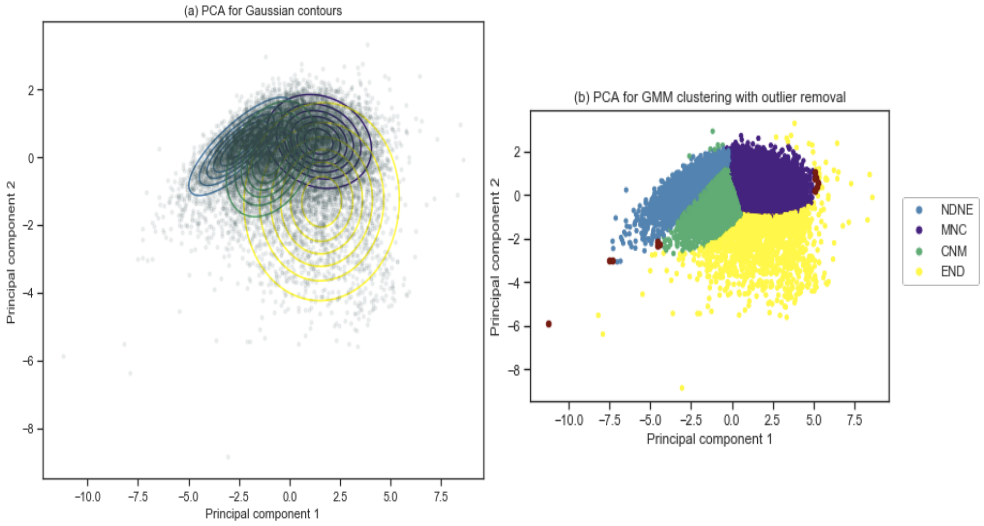


Figure 5.9: **PCA for Gaussian contours (a) and GMM clustering with outlier removal (b)**. In (a), each GMM ellipse is identified with a different colour: NDNE in blue, MNC in purple, CNM in green and END in yellow. It can be seen that there is an overlapping area between the GMM ellipses. In (b), the GMM clustering model with outlier removal was applied in 11 features under the two-dimensional projection using 14,813 genes (after the outlier removal step). The distribution into gene classes is 3,630 in NDNE, 5,865 in MNC, 3,808 in CNM and 1,510 in END. Colour coding in (b) is consistent with sequence colouring in (a), adding maroon, which identifies the outlier genes to be removed.

The results given by GGM clustering with outlier removal (Figure 5.9 (b)) might suggest that the END gene-class has positive loadings for the first component and negative loadings for the second component, indicating that these genes tend to be intolerant of LoF and might be subject to a strong selection effect. By contrast, the NDNE group has negative loadings in the first component and positive loadings in the second component, implying that these genes can be tolerant of mutation since they are exposed to high recombination events and are weakly impacted by selection. Genes enriched for both MNC and CNM have an intermediate loading for both the first and second components involving recombination and selection processes, which might preserve damaging but non-lethal variants [53].

5.3.5 Clustering performance

According to current understanding, the proposed method can capture the heterogeneity of gene relationships through their functional and genomic properties under a two-dimensional projection. However, functional and genomic scores cannot allocate all genes into their known classes due to extensive overlap and similarities between gene groups. However, the analysis aims to aid recognition of potential but currently unrecognised Mendelian disease genes. Genes currently assigned to classes other than the Mendelian gene group but classified by these methods towards the essential end of the gene spectrum are potential Mendelian disease candidates and recognition of even a small number of potential novel Mendelian genes is valuable. Figure 5.10 shows a comparison of the distribution of genes in two-dimensional space between gene groups given by Spataro and gene classes estimated by GMM-non. According to gene groups (Figure 5.10 (a)), there was no clear separation between those groups, although the END gene group seems to be separated from the NDNE group. In contrast, for genes distributed within gene classes predicted by the proposed methodology (Figure 5.10 (b)), there is segregation between these classes. Both NDNE and MNC have a visible separation showing different gene characteristics, while the CNM class overlaps with the other gene classes. Genes in both MNC and END show a strong relationship indicating that genes at this end share similar properties explained by the scores.

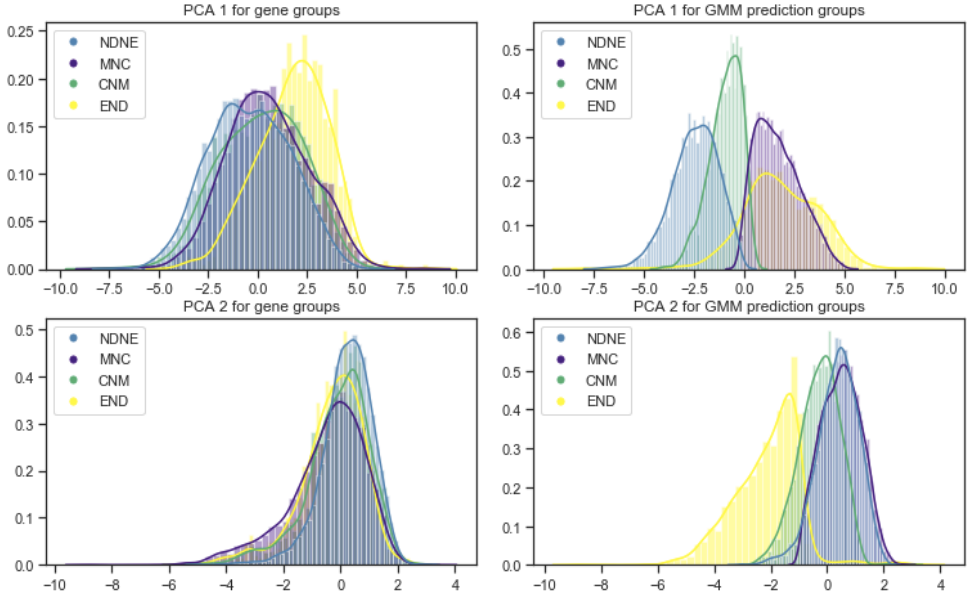


Figure 5.10: **Comparison of gene distribution between gene groups**. In (a) the distributions of the genes are coloured according to Spataro's gene groups; in (b) they are coloured by GMM-non prediction groups. Both gene groups are graphed under a two-dimensional projection. Colours indicate the groups: NDNE in blue, MNC in purple, CNM in green and END in yellow.

The precision, recall and F_1-score metrics (see Table 5.3) were computed using a combination of the OMIM-Mendelian disease list [19] and Spataro's groups as a benchmark to evaluate the robustness of the method of GGM clustering with outlier removal. Approximately 30% (precision) of the genes in NDNE were identified correctly as belonging to this class; however, the proposed GMM-non model can recognise 80% (recall) of the genes as belonging to NDNE, while just one of four genes

were allocated to the MNC class. The CNM and END have the lowest precision, recall and F_1 score. This indicates that the model is prone to producing more false negatives in these classes than in the NDNE and MNC groups.

Table 5.3: Performance of GMM clustering with outlier removal model

Gene group	Precision	Recall	F_1-score	Genes
NDNE	32%	80%	46%	3,630
MNC	36%	21%	26%	5,865
CNM	26%	11%	15%	3,808
END	20%	12%	15%	1,510

A total of 5,865 of the 14,813 genes were identified belong to MNC by the proposed model. 3,808 genes were placed in CNM, 3630 in NDNE and 1,510 in END. Due to the similarities and shared properties between MNC, CNM and END genes, the proposed model cannot classify all genes precisely into their proper categories. However, it must be noted that the Spataro classification is based on current knowledge and is likely to include genes that are unrecognised Mendelian candidates (presumably mostly classified as NDNE). Comparing the gene classes given by the proposed model and the benchmark gene group (see Table 5.4), 1,750 genes that are well known to be related to Mendelian disorders were found in CNM and END. Moreover, 388 and 3017 were wrongly classified as NDNE when they are MND or CNM. Many genes that are classified in the NDNE group by Spataro *et al.* [52] study were placed into the MND (3,440) and CNM (2,253) groups.

Table 5.4: Gene distribution by GMM-non model prediction and OMIM/Spataro groups

OMIM/Spataro gene group	GMM model prediction			
	NDNE	**MNC**	**CNM**	**END**
NDNE	2,906	3,440	2,253	381
MNC	388	1,218	1,000	750
CNM	307	666	415	198
END	29	541	140	181

The mean of the scores for each score by Spataro and GMM-non gene group is given in Table 5.5. It reports that the directions of the scores were preserved between all groups, which is consistent with the model proposed by Pengelly *et al.* [53]. For example, the END group presents the highest degree of essentiality, while the NDNE group appears to be the least essential. The MNC and the CNM groups are found to be at intermediate levels of degree of essentiality. The means of the metrics were found to be statistically different between the classification by the GMM-non model and the benchmark gene groups (Supplementary Table S4.1). Interestingly, the pLI score from both gene groups in END was not statistically different, with similar results for LDU residuals from both NDNE and CNM.

Table 5.5: Means of the metrics by GMM-non model prediction and OMIM/Spataro groups

Feature	NDNE		MNC		CNM		END	
	OMIM Spataro	GMM prediction	OMIM Spataro	GMM prediction	OMIM Spataro	GMM prediction	OMIM Spataro	GMM prediction
Number of genes	8,986	3,630	3,362	5,865	1,588	3,808	891	1,510
pLI	0.252	0.031	0.317	0.513	0.364	0.119	0.589	0.566
RVIS	0.082	0.618	−0.219	−0.525	−0.079	0.204	−0.425	−0.301
DNE	0.696	−0.183	1.074	1.587	0.942	0.472	1.753	1.619
SIS	−0.072	−0.853	0.194	0.715	0.096	−0.275	0.601	0.432
HI	0.258	0.118	0.358	0.381	0.330	0.237	0.478	0.598
NET	0.498	0.186	0.657	0.679	0.574	0.603	0.747	0.856
GHIS	0.515	0.456	0.530	0.581	0.522	0.488	0.567	0.542
REC	0.149	0.109	0.269	0.161	0.194	0.175	0.239	0.502
GIMS	0.516	0.745	0.426	0.258	0.452	0.605	0.314	0.358
GDI	4.352	5.974	4.787	2.424	4.521	5.905	3.303	4.494
LDU (residuals)	−0.047	−0.075	−0.084	−0.465	0.204	0.317	−0.291	0.768
Principal component 1	−0.400	−2.466	0.510	1.677	0.196	−0.962	1.757	1.840
Principal component 2	0.232	0.370	−0.487	0.497	−0.105	−0.276	−0.302	−2.115

5.4 Discussion

This chapter examines the efficiency of recognising Mendelian disease genes from complex disease and non-disease genes through essentiality-specific information sharing across genes. Here, different clustering methods were applied to separate genes into four groups such as $k-means$, hierarchical clustering and the proposed GMM-non model. The latest showed improvements in efficiency for detecting MNC genes. This proposed GMM-non model would therefore be effective for selecting Mendelian-related genes that are positioned towards the essential-gene end of the spectrum as defined by Pengelly *et al.* [53]. However, the model for selecting MNC genes with intermediate and low essentiality could be a subject of further research. Another important subject would be adding a gene-level mixture structure; more integration of additional functional and genomic data of the association specifically with MNC and CNM would help provide a more formal basis for evaluating false positives and true positives in gene selection.

Different methods have been developed to separate genes, as an example, the study conducted by Alyousfi *et al.* [196] developed the essentiality-specific pathogenicity prioritisation (ESPP) score based on PCA. The score integrates eight individual gene-specific scores, which have different properties and assumptions. The ESPP score was calculated as the weighted sum of each of the component scores. For the gene groups, the distribution of ESPP score was considered to located the genes following Spataro *et al.* [52] gene groups. In this study, the proposed GMM-non model also integrates gene-specific metrics according to their essential properties. Although these methodologies are differently measured, the results are similar. For example, genes currently classed as NDNE with a particularly high ESPP are related to monogenic disorders scores. In this context, those genes were classified into the monogenic disease by GMM-non. Both analyses showed that candidates genes to be potentially essential classified as NDNE (for example, SUPT6H, FRY) and genes known to contain CNM variation but have properties that suggest that they are also candidate monogenic disease genes (for example, RYR3, DIP2C).

Although PCA has its own structure to elucidate population structure or separate hyperplane ranking strategy, it does not necessarily represent important discriminant directions to separate

sample groups. In this context, it was decided to develop a new method that includes a robust mathematical structure. The GMM-non is an alternative methodology for looking at the data and recognising genes with high potential for pathogenic disease-related variation. The GMM-non classifier is suitable for the classification part, but the initial dimension reduction is performed by principal component analysis. Then, the reduced data (which contained only two components) dealt with the two–variables together and fitted a two-component bivariate GMM-non to the data. Later, the classifier was obtained using an EM algorithm and simultaneously outlier removal.

The analysis of GMM-non showed that genes allocated to MND are expected to include misassignment between END and CNM due to overlapping the essentiality level properties across these groups. Moreover, ∼38% of the genes that are placed in the NDNE groups [52, 196] were assigned to the MND group by using the proposed model. For example, genes in these MND groups, AADAT (Aminoadipate Aminotransferase) was identified by the proposed model related to MND, but it was placed in NDNE. According to OMIM [19], *AADAT* is a protein-coding gene which is associated with Huntington disease (autosomal dominant) (Rappaport *et al.* [234]). Another example is *ANKRD44* (Ankyrin Repeat Domain 44), which tends to be involved in developing Glass syndrome (autosomal dominant and fatal disease). The gene CT62 (cancer/testis associated 62) related to cancer was found in the MND groups when in reality it belongs to CNM [234]. New genes were identifying into MND using the proposed model. Although these genes have not yet been labelled, they are already known to be involved in MND were

The analysis of CNM groups given by the proposed model found that ∼25% of the total genes in the NDNE groups were assigned to CNM. Comparing genes in the CNM group, the model has distinguished 15 genes well known to be related to inflammatory bowel diseases [234]. However, there are ∼30% of the genes that are well known to be Mendelian-related genes that were wrongly assigned into CNM group. This is the result of the shared properties between both MND and CNM groups. All results above indicate that CNM genes tend to present higher expression levels and are enriched in specific relevant protein function categories compared to CNM genes. According to Spataro *et al.* [52] study, genes that overlap into the MND and CNM groups might be involved in essential metabolic processes. They might also enrich protein products in essential biological functions, such as cytoskeletal proteins, extracellular matrix components, enzymes, proteins involved in cellular communication, transporters and transfers/carriers, and proteins involved in the immune and defence system. Thus MND genes seem to have more relevant functional roles among those associated with at least one complex disease.

While using the proposed model shows promise for identifying condition-specific MND genes from non-disease genes, several challenges still have to be tackled. The GMM-non assumed four components across genes. In some cases, using only one cluster for a particular direction of gene properties may be rather restrictive for identifying plausible considerable heterogeneity among disease genes. The complexity of disease-gene relationships and the diversity of gene properties limit the ability of individual and integrated scores to discriminate certain gene classes fully. For example, MacArthur et al. [12] developed their gene score based on human–macaque conservation and proximity to known recessive genes in protein interaction networks. Although their score, which describes the

probability of a gene containing recessive variation, provides a degree of separation between loss of function tolerant and recessive genes, there is a substantial overlap. These scores do, however provide useful information to rank potential candidates in a genome filtering context. Furthermore, with the continued and dramatic rise in the number of genomes sequenced, a greater understanding of gene properties and functions is likely to improve the recognition of genes likely to contain monogenic disease variation. Given a sequenced genome for which there are several potential functional candidate variants in different genes access to the available scores provides a basis for ranking candidates objectively.

To improve the model's performance, an effort to integrate additional genomic and functional gene properties alongside improving gene classification given developing knowledge would be a worthwhile basis for future studies. Therefore, this methodology can be extended to involve multiple components, possibly with a selection of the number of Mendelian-related genes based on autosomal recessive or dominant genes [218]. Another restriction in the context of GMM-non is that no interaction or correlation is assumed among genes. According to studies by Cacheiro *et al.* [169] and Spataro *et al.* [52], the integration of protein–protein interaction data at gene level can quantify the correlation between genes that tends to be associated with essential genes. Further, in the context of the proposed model, the impact of allowing correlation among the genes can control the misassigned genes in the clustering model [235]. Although the gene data underwent a pre-processing step and noise removal (outlier removal), the proportion of bias is expected to be high due to the similarities and shared properties across the genes. This is especially the case for genes with a higher degree of essentiality and are related to MND. However, the quantity and type of bias created by the model is not yet known and requires further investigation on the complexity of disease–gene relationships and the diversity of gene properties (i.e., very close and overlapping). The proposed model may not be appropriate for every relationship between pairs of genes. For example, the resulting clusters are near each other, often have no distinguishable gap between them and hence could be merged into a larger nonlinear cluster [121]. Strategies exist for merging GMM clusters (e.g., [236]), but these were not yet incorporated into this analysis.

After screening disease-related genes among different clusters given by GMM-non, the proposed model shows an improvement on classifying genes in the MND group. However, it will be necessary to identify gene patterns that belong to the same molecular pathway related to disease biology to obtain better clusters [53]. At the same time, an important topic for future studies to improve gene classification would be developing a two-way model-based clustering of genes in the context of Mendelian disorder analysis as an extension of the proposed model-based method in this chapter.

Chapter 6

Polygenic risk score to quantify the cumulative effect of low-penetrance alleles on breast cancer and breast cancer subtypes

6.1 Introduction

6.1.1 Breast cancer

Breast Cancer (BC) is one of the most common cancer diagnosed worldwide among women with an estimated ∼2 million new cases in 2018, thus making it the second leading cause of global mortality after lung cancer [237] (see Figure 6.1). BC is the primary cause of mortality among women aged 45–55. Although BC aetiology is still unknown, the predisposition to BC is driven by multifactorial and genetic factors. Approximately 10% of BC cases have a strong genetic basis [238]. Rare high-penetrance variants located in two major predisposition genes, *BRCA1* and *BRCA2*, confer a high risk of developing BC; however, these variants account only for approximately 5–10% of BC cases[239]. In contrast, both intermediate-risk variants (e.g., in *PALB2*, *ATM*, and *CHEK2*) [240] and commoner variants (mostly single nucleotide polymorphisms (SNPs)) confer lower risks of BC [241].

The advent of genome-wide association studies (GWAS) enabled the identification of more than 100 low-penetrance and moderate-penetrance variants associated with BC [100]. These studies were highly successful at discovering variants associated with both familial and sporadic[1] BC; however, the variants identified to date collectively explain around 15 to 30% of the heritability of BC

[1]Sporadic BC occurs in women who carry a high penetrance BC susceptibility gene mutation but do not have a family history of BC.

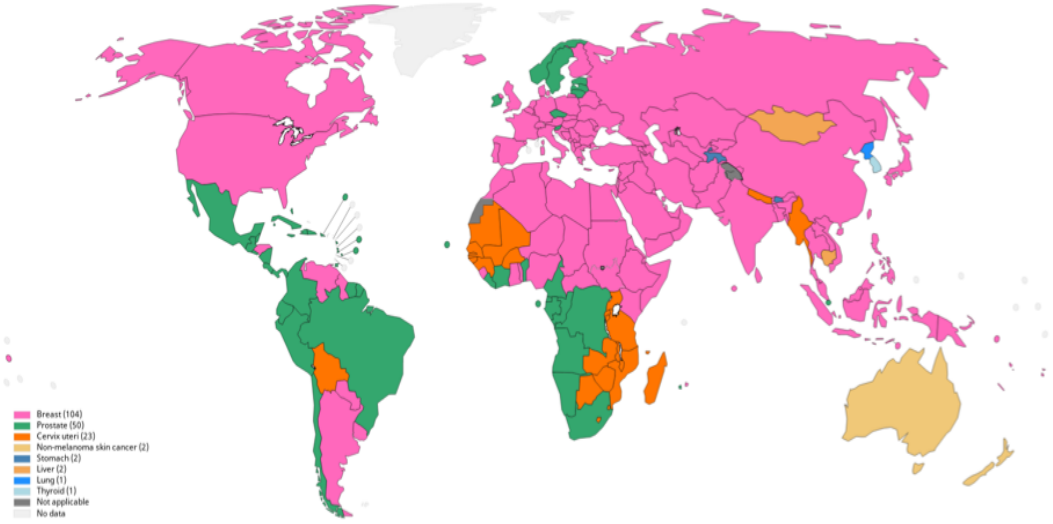


Figure 6.1: **Estimated incidence rate of top cancer per country age-standardized for both sexes in 2018**. IData source: GLOBOCAN 2018. Graph production: IARC (http://gco.iarc.fr/today) World Health Organization.

cases in the general population [242, 243]. Nevertheless, the majority of individuals afflicted with this disease do not harbour any such pathogenic variants. Instead, the inherited susceptibility of breast cancer has a considerable polygenic component, driven by the cumulative effect of numerous common variants scattered across the genome [101].

6.1.2 Polygenic risk score on BC

Multiple common variants in BC liability confer a small risk individually, but their combined effect is substantial and associated with much larger relative risk. Genome-wide heritability estimates that all common variants explain only about 45% of the familial contribution to BC on genome-wide SNP arrays, and the remaining BC cases are of unknown genetic aetiology [244, 100, 241]. Thus, the BC risks associated with SNPs combine multiplicatively and hence their joint effect can be conveniently represented as a polygenic risk score (PRS). Based on this, several studies have focused on developing efficient PRS for BC to improve the diagnosed [245]. PRS therefore might be an essential tool for interpreting human genomes in complex phenotypes for which effect sizes of genetic variants are individually small [246]. The construction of PRS focuses on estimating a small fraction of variation in the phenotype, which might elucidate the BC risk from the accumulation of many small genomic contributions [247]. For example, Mavaddat *et al.* [241] developed an optimised PRS for the prediction of subtype-specific breast cancer using *BRCA1* and *BRCA2* from the largest available GWAS dataset. The authors estimated a PRS of BC used 303 genetic variants, showing that women had an increased risk of developing estrogen receptor-positive (or ER+) breast cancer in the top 1 percentile of the PRS while women in the lowest percentile decreased the risk sixfold. Although the PRS had a receiver-operator curve (AUC) = 0.630, it demonstrated that breast cancer PRS might reach sufficient information to identify a high-risk subgroup of women who could be advised for a specific preventive intervention [241]. Another recent study conducted

by Läll *et al.* [248] provided a PRS for BC, which enables the different strata in BC incidence to be identified and its potential for personalised risk assessed based on GWAS studies. Similarly, Shi *et al.* [249] established a PRS based using 77-SNP PRS and non-genetic risk factors for young-onset BC. Despite extensive research, the implementation of polygenic scores in breast cancer is still challenging due to the limited data for quantifying the cumulative effect on this disease of variation across the whole genome.

6.1.3 An overview of polygenic risk score

In recent years, PRS has generated much excitement due to its potential applications in precision medicine. A PRS can represent individual genetic predictions of phenotypes; in other words, it can predict a person's risk of developing a particular disease. Along with the accumulation of genomic loci related to common and complex diseases, it can quantify polygenic risk using allele frequencies. As a usual practice, PRS has been calculated as a weighted sum of several risk alleles carried by an individual using GWAS studies. The risk alleles and their weights are defined by SNPs and their measured effect sizes. Effect sizes are typically estimated as standardised (regression) coefficients, also called beta coefficients for quantitative traits or as odds ratios for categorical binary traits. Frequently, PRS is calculated using a set of SNPs with different p-value thresholds for disease association, and then a PRS series is estimated for a particular disease or trait. Once the PRS has been calculated in one cohort, it is essential to evaluate its predictive performance in another external cohort, not used to construct the PRS [250].

However, there are some methodological concerns in the calculation and validation of the PRS [251, 98]. For example, although the construction of PRS by including a more significant number of SNPs may have higher predictive precision, it is argued whether the inclusion of those SNPs with close to zero effects in the PRS is valid [252]. Another example, Linkage disequilibrium (LD), the correlations between nearby SNPs, which leads to over-representation of high LD regions in calculating PRS, potentially reduces the predictive performance of PRS. To reduce the effect of LD, 'clumping/pruning and thresholding' methods have been used [253]. In contrast, other methods evaluate the best prediction across the genome by explicitly modelling the correlation structure between variants without identifying a subset of SNPs for prediction; the Bayesian approaches are the most widely used implementation [254]. Many novel risk scoring methods are being developed and may be more powerful compared to the current methods.

An alternative methodology to compute a PRS has been developed by Smyth *et al.* [96]. Their PRS is based on surprisal theory for measuring differences in the global genome structure between cases and controls. This approach measures the information provided by sub-sequences surrounding individual loci, where sequences with lower surprise represent the frequency of alleles. Those sequences appear more common and therefore tend to be encoded with common sequences. Conversely, loci with specific functions are encoded with rarer and higher surprisal sequences. In this study, a PRS was developed following this global PRS structure of Smyth *et al.* [96] using the same reference

population (Wellcome Trust Case Control Consortium (WTCCC)). This approach measures an alternative PRS that quantifies the cumulative effect of low penetrance genetic variants on BC. This approach can be applied broadly across all sequences, irrespective of overall sequence similarity, and independent of the functional relationships used to group them. The method can also analyse unconnected genomic loci such as those harbouring single nucleotide variants (SNVs) or somatic mutations or evaluate different molecular sequences.

6.1.4 Relative genome information

Maximal genotype density base on whole-genome sequencing (WGS) were used in this study for characterisation of the cumulative effect of numerous polygenic variants in BC for different types of mutations. Following the global measure of Smyth *et al.* [96] based on surprisal theory, the differences in the genome structure between cases and controls were measured. This global structure based on surprisal analysis allows to define a balanced as a reference population (control genomes) to all the types of genomes; that is, the steady-state is the standard reference to which different types of measured genomes can be compared in their global structure. The deviations of these genomes represent a perturbation in the genome [255].

The quantification of those deviations is defined as relative genome information (RGI), which estimates how unusual a genome is related to the reference genome. A DNA sequence can be represented as a string of the letters A, C, G, T corresponding to the individual nucleotides in the genome. Since all the possible alleles in each locus code naturally lead to the likelihood of the allele pair, different allele frequencies reflect different aspects of the genome. The RGI follows the natural information theory measure of the surprise of observing a specific genome given the probability of finding each allele pair in the reference population. This decomposition allows us to distinguish and thus separately quantify the fraction of the relative accuracy[2] that is attributable to differences between populations and the fraction attributable to alleles frequencies. Thus, a person with a higher RGI has a more unusual genome, having either fewer common alleles more often than expected or having some particularly rare alleles. Likewise, a lower RGI has more common alleles and, therefore a less surprising genome.

Formally, surprisal analysis can be expressed as maximal entropy discussed in the Introduction. This entropy measures the complexity of an ensemble X, of samples (nucleotides) x, where each sample occurs with a probability p(x). This self-information associated with each sample is called surprisal and is defined by $S(x) = -log_2(p(x))$ where $S(x)$ is estimated in bits[3]. For the complete set of samples is the sum of all surprisals $RGI(x) = \sum_{i=1}^{n} S(x_i)$. Entropy is therefore defined as the average information per of samples or the expectation value of all surprisals as $EIL(x) = E(S(x)) = -\sum_{i=1}^{n} p(x_i) log_2(p(x_i))$ where $\sum_{i=1}^{n} p(x_i) = 1$ and $EIL(x)$ is also measured in bits.

[2] A measure of positional consistency between a data point and other near data points. Relative accuracy compares the scaled distance of objects on a map with the same measured distance on the ground.

[3] where logarithms are taken to base 2 so that information is measured in bits

In this study, a PRS is developed based on the surprisal theory to evaluate the important impact of a polygenic component on the risk of developing BC. This PRS quantifies the cumulative effect of low-penetrance alleles on BC risk. The term 'polygenic risk scores' will cover the sum of allele likelihood to provide individual risk measures based on the RGI in different types of mutations in BC from Prospective Study of Outcomes in Sporadic versus Hereditary breast cancer (POSH) cohort and unaffected individuals from WTCCC dataset.

6.2 Methods

6.2.1 Study cohort

The dataset used for developing the polygenic score comprises 2,064 breast cancer-affected cases and 5,195 control subjects of European ancestry. SNP genotypes of early-onset breast cancer (EOBC) cases were collected from POSH. This cohort recruited 3,021 women aged 40 years or younger with breast cancer diagnosed with invasive breast cancer between 1 January 2000 and 31 December 2007. The study was conducted in 127 hospitals across England, Northern Ireland, Scotland and Wales. The exclusion criteria included prior history of invasive malignancies apart from non-melanomatous skin cancer and not being available for follow-up or refusing consent to retain data. Written informed consent was obtained at study entry [256].

Existing whole-genome data was used from healthy controls in Phase 2 from WTCCC data. This dataset comprises individuals derived from the UK 1958 British and the UK National Blood Service. The 1958 British birth cohort is a sequential sample of live births in England, Wales and Scotland during one week in 1958 who were followed up in 2002–2004 when they were 44–46 years old. The UK Blood Service Control Group cohort is collected from individuals in the age range of 18–69 who have donated blood to the UK National Blood Services (NBS) Collection. Individuals from the NBS cohort have been screened by standard processes used to exclude blood donors from dissident groups (WTCCC, `http://www.wtccc.org.uk/`) [257].

6.2.2 Genotyping

The POSH cohort consists of approximately 3000 patients, from which 2,503 SNPs genotypes were obtained from blood samples in six separate batches using different genotyping arrays (Figure 6.1). Genotyping of the first batch was conducted at the Mayo Clinic, Rochester, MN, on 274 patients with positive family history and triple-negative breast cancer. These patients had little or no tumour expression for the absence of oestrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 (*HER2*). Genotyping of the second batch was carried out at the Genome Institute of Singapore, National University of Singapore, on 300 samples with survival extremes of either early distant metastasis or death (n=200) or long-term event-free survival (n=100) [258]. The third batch consisted of 308 patients from the Cambridge dataset with a positive family history.

These three batches were genotyped on the Illumina (San Diego, CA, USA) 660-Quad SNP array [259]. The fourth batch included 377 individuals selected from the Haiman dataset, who were genotyped using the Illumina HumanOmni5Exome-4v1-1+exome SNP array. These patients were defined as having high risk by onset age, second diagnosis and family history. The 1,088 samples in the fifth batch were randomly chosen from the POSH cohort and were genotyped employing the Consortium-OncoArray [100]. The 156 patients of the sixth batch had mostly triple-negative breast cancer. These patients were genotyped using the Illumina Global Screening Array v2.0.

The WTCCC comprises two independent sets of controls (disease-free): 2,699 individuals from the 1958 British Birth Cohort and 2,501 individuals from the NBS Collection. Genotyping of both sets was conducted using the Illumina 1.2M chip Illumina [260]. A summary of all data sets is given in Table 6.1.

Table 6.1: Data characteristics and genotyping methods

Dataset	Inclusion criteria	samples	SNP chip/genotyping array
MAYO POSH cases	Cases with little or no tumour expression for ER, PR and HER2	274	Illuminia 660-Quad SNP array
GIS POSH cases	Cases with survival extremes; early relapse (n=200) or long term survival (n=100)	300	Illuminia 660-Quad SNP array
Cambridge POSH cases	Cases with positive family history	308	Illuminia 660-Quad SNP array
Haiman POSH cases	Cases with high risk cases defined by onset age, second diagnosis and family history	377	Illumina Human Omni5 Exome SNP array
Oncoarray POSH cases	Random samples from POSH	1088	Consortium-OncoArray
GSA POSH cases	Cases with mostly triple-negative breast cancer	156	Illumina Global screening array
1958 British birth cohort WTCCC control	Control, disease-free	2699	Illuminia 1.2M chip
National blood service WTCCC control	Control, disease-free	2501	Illuminia 1.2M chip

Note. Post-quality control analysis in caucasian population.

6.2.3 Quality control

Quality control (QC) filtering was undertaken before analysis of the whole-genome data using standard procedures for GWAS in order to minimise potential false findings [261]. Figure 6.2 shows the QC pipeline used in this study in detail. QC for each dataset was implemented using Plink v1.90p 64-bit [143] and custom scripts using awk, and R version 3.6.3 (http://www.r-project.org/). Autosomes were only considered, and SNPs were excluded from each data set if they failed any subsequent QC filters. SNPs with a low minor allele frequency (MAF) of less than 1% were excluded. Samples and SNPs with missing genotypes greater than 10% were removed. In addition, SNPs were identified with a significant deviation from the Hardy-Weinberg equilibrium (HWE) (cases $p < 10^{-10}$ and for control $p < 10^{-4}$) were rejected for subsequent work. The significance level in HWE for cases was higher than for the controls, as it is more likely to have poor genotyping than disease association. SNPs were excluded if they were not in the haplotype reference consortium (HRC) or had alleles that do not match HRC. SNPs with MAF difference > 0.2 were compared to the HRC. AT/GC SNPs with MAF > 0.4, duplicate SNPs and indels were removed for subsequent filter procedures.

Further rigorous QC was also carried out for each individual in the datasets. Heterozygosity in each sample was calculated, and individuals with outlying levels of heterozygosity were also excluded. Excess heterozygosity may indicate sample contamination or admixture, and reduction suggests

inbreeding or deletions. Samples with a discrepancy between reported and inferred sex based on X chromosome homozygosity were also identified and removed [261]. SNP chip annotation files (http://www.well.ox.ac.uk/~wrayner/strand/) were employed to remove custom SNPs that were not in the strand file, update SNP locations to genome reference consortium 37 (GRCh37/hg19) and flip genotypes to the positive strand [262]. Non-AT/GC custom SNPs were retained in the Cambridge and Haiman datasets, which have a large number of custom SNPs.

6.2.4 Imputation

Imputation on the genotyped samples was carried out to increase resolution. Because cases and controls were genotyped separately using different arrays, imputation was carried out in separate batches according to the genotype array. To minimise false positives due to differential measurement error whereby some SNPs are measured almost correctly (through actual genotyping) in one batch but measured imperfectly (through imputation based on nearby measured SNPs). In other batches imputed, SNPs were quality controlled as genotyping call rate $< 99\%$; MAF $< 1\%$; missing genotypes $> 10\%$; extreme deviation from HWE ($p \leq 1 \times 10^{-10}$ in cases, $\leq 1 \times 10^{-4}$ in controls); SNPs with one or more discordant genotype between duplicate samples; and GWAS significance between batches ($p < 5 \times 10^{-8}$). SNPs were imputed using the Sanger imputation server (https://imputation.sanger.ac.uk/) and EAGLE2 for pre-phasing into the Haplotype Reference Consortium (r1.1). This imputation employs the largest reference panel of human haplotypes and positional Burrows-Wheeler transform (PBWT).

6.2.5 Merging datasets

Before integrating the genotypic data from POSH and HapMap, individual datasets were merged in different pairwise combinations of cases (n=15) and further QC steps were completed. All known duplicate genotypes were established using pairwise measures of identity by state (IBS). The evaluation of genotyping concordance between duplicate samples per each SNP and sample was performed using Plink v1.90p 64-bit [143], and SnpSift [263]. SNPs with discordant genotypes between duplicate samples were removed.

6.2.6 Population stratification and relatedness

Population stratification might cause false positives in association studies, and therefore evidence of ethnic admixture was excluded by performing multi-dimensional scaling (MDS) in Plink v1.09 [143]. Ancestry was inferred using genome-wide autosomal SNPs in linkage equilibrium (independent SNPs) to minimise underlying differences between case and control groups. To avoid confounding bias of relatedness pairwise values of IBS were assessed by keeping only the LD independent SNPs. Thus, autosomal SNPs with maximum pairwise $r_2 \leq 0.5$, a window size of 50kb and a step size of 5 SNPs were selected using LD-based pruning in Plink v1.09 [143]. SNPs selected after pruning were

used to calculate the IBS distance between each pair of individuals that passed QC. For sample pairs with evidence of relatedness ($PI_{HAT} \geq 0.125$, equivalent to third-degree relatives), the sample with the lowest genotyping rate for all SNPs passing QC was excluded. These data were then merged with data from the founders or unrelated individuals in the HapMap sample [261]. The HapMap data were used to describe reference population genotypes against the genotype data of the cases and the controls for the African, East Asian and Caucasian populations [264].

6.2.7 Testing for batch effects

Systematic differences among the composition of individuals within each batch (i.e., the case to control ratio or race/ethnicity of individuals on plates) can result in batch effects. Thus, MDS analysis was also performed to recognise whether samples clustered according to the genotyping batch or if any variation was due to batch effects [261]. Dummy variables for cases and controls analysis were assigned to compare all pairs of batches using logistic regression with correction for population stratification, including the first 5 'dimensions from the MDS analysis. The genomic inflation and significant SNPs ($p < 10^{-8}$) were identified. SNPs with MAF difference > 0.15 between batches and > 0.1 were recorded and then compared to the HRC reference to be removed. SNPs were only retained if they passed the QC steps in every pairwise combination.

6.2.8 LD-pruned SNPs

LD describes a non-random association between alleles at different loci on the same chromosome in a given population. SNPs are in LD when the frequency of association of their alleles is higher than expected under random assortment. LD concerns patterns of correlations between SNPs, which makes identifying the contribution from causal independent genetic variants extremely challenging. Due to the presence of LD, SNPs were LD-pruned before the construction of PRS using Plink v1.90p 64-bit [143] to select a set of approximately independent SNPs [265]. The LD pruning was used as a first step to select a subset of independent SNPs. The algorithm uses the first SNP following the genome order and calculates the correlation. The SNPs were removed when a large pair correlation was found, and the SNP with largest MAF was retained. The strength of LD between SNPs was set using a window size of 50kb of the chromosome with a step size of 5 SNPs. SNPs that are approximately uncorrelated were selected based on an r^2 threshold < 0.5. Hence, the LD-clumping method was used as a second step to increase the resolution and retain the most important SNPs in each LD block. This algorithm sorts SNPs based on their marginal p-values and iteratively removes SNPs in LD with a higher-ranked SNP based on this ranked list. This reduces the correlation between the remaining SNPs while retaining SNPs with the most robust statistical evidence associated with the phenotype. In this study, the threshold was set to $p < 10^{-5}$ with a 500 kb window [143]. Lastly, the independent SNPs found were merged and used for subsequent analyses.

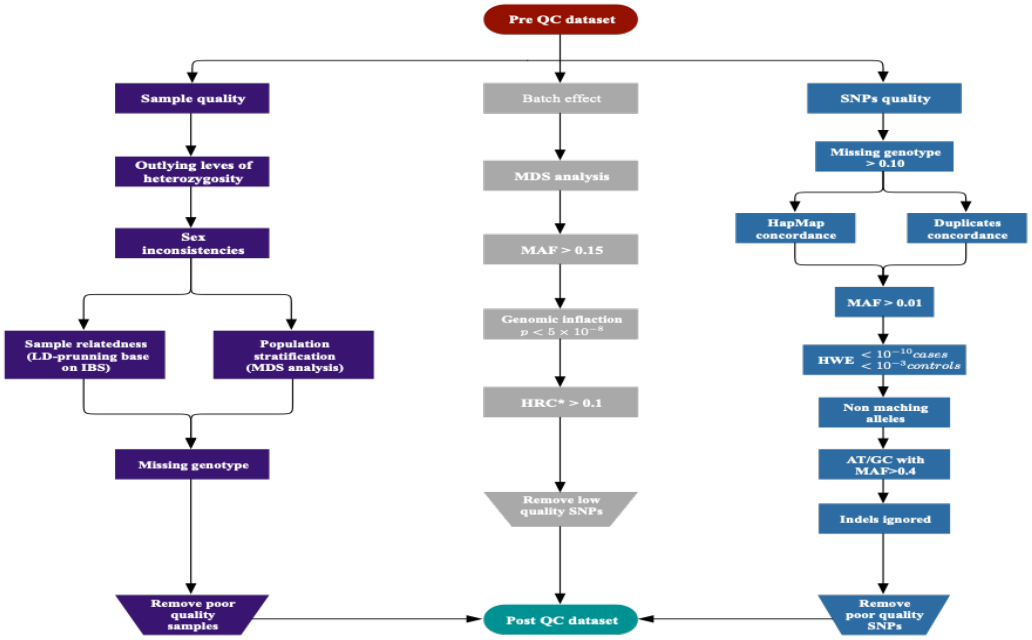


Figure 6.2: **Flowchart overview of the complete quality control process**. This figure shows each step of the QC procedures that should be performed prior to GWAS data analysis. Each topic is discussed in detail in the corresponding section of the text. Squares represent steps, ovals represent input or output data, and trapezoids represent filtering of data. Figure modified from "Quality Control Procedures for Genome-Wide Association Studies" by Turner S. *et al.* (2011) [261] Current Protocols in Human Genetics.

6.2.9 Relative genome information

PRSs are usually constructed from a weighted sum of allGenome-Wide Without loss of generality, in this study, the relative genome information (RGI) was defined using the WTCCC dataset as a reference population for PRS construction. Thus, the procedure first uses the reference population to estimate the RGI as the probability measure on the space of all genomes, then uses the estimated probability measure to assess how unusual this RGI of an individual's genome is compared to the reference population.

Formally, the RGI estimate is given according to surprisal theory [96]: the uncertainty in a random genome X is quantified taking a certain value x based on its probability of occurrence $\prod(x)$. Let L denote the set of locations in the genome (loci), and let $\Omega = A, C, G, T$ be all the possible alleles at each locus $I \in L$. Let $\prod_l(\lambda, \mu)$ be the likelihood of the unordered allele pair $(\lambda, \mu) \in \Omega \times \Omega$ at locus $l \in L$ in the reference population. Then $X_l \in \Omega^{2L}$, where the space of all possible genomes is Ω^{2L} with likelihood π_l over all $I \in L$. The relative local information (RLI) is derived as the surprisal measure in bits when the base of the logarithm is 2 [266]:

$$I_l(X_l) = -\log_2 \prod_l (X_l) \tag{6.1}$$

at each locus $I \in L$ in the genome X. Hence the RGI for each genome X of interest is defined as

$$\mathbf{I}(X) = \sum_{l \in L} I_l(X_l). \tag{6.2}$$

Since chromosome size is highly variable, it is necessary to normalize the RGI by the number n of loci genotyped. Thus, the comparison between sequences of different lengths or different chromosomes tends to be equivalent up to normalising. Hence, the expected information per locus (EIL) is defined as follows:

$$\mathbb{E}_n(\mathbf{I}_l) = \frac{1}{n} \sum_{l \in L} I_l(X_l) \tag{6.3}$$

The probability π is unknown and estimated from a reference sample of similar ethnic backgrounds to the cases. Here, the WTCCC cohort was used to determine the likelihood of the $\prod$ for each locus over all available genotypes in the reference at that locus [96]. RLI, RGI and EIL were constructed using custom-written scripts in Python.

Once π is estimated, the RGI was calculated for each genome in each of the remaining six (test) samples (POSH cases). Estimating RGI takes $O(n(m+N))$ computational time for N case genomes, where n is the number of loci and m is the number of genomes in the control population. This was carried out on IRIDIS 4 at the University of Southampton computing cluster and required on average ∼23 hours of run time on a 16-processor node per sample.

6.2.10 Polygenic component

Once the probability of the RGI and therefore EIL were obtained, the POSH study cases were divided into three significant subtypes of cancer; *BRCA1* mutation carriers, *BRCA2* rare pathogenic variant carriers and polygenic components. In this study, the polygenic component was defined as the combination of intermediate-risk and commoner variants (SNPs joint effect). Therefore, the polygenic component stringently excluded people who carried either the *BRCA1* or the *BRCA2* mutation.

6.2.11 Statistical Analysis

Analysis of significant differences between groups was tested using Wilcoxon rank-sum tests. Two-sided tests were used when testing the null hypothesis of no difference in EIL between cases and controls against the alternative hypothesis that EIL differs in cases and controls; one-sided tests were used when testing the null hypothesis of no difference in EIL between cases and controls against the alternative hypothesis that EIL is higher in cases. Logistic regression was performed to determine the BC mutations' contribution; this accounts for the variation in different types of mutations and controls as dependent variable and EIL as an independent variable.

6.2.12 Data and code

Quality control analyses were performed using Plink v1.90p 64-bit [143]. The PRS construction and analysis were conducted in Python version 3.7.3 (https://www.python.org/), `awk` and R version 3.2.2 (https://www.r-project.org/) using custom-written scripts.

6.3 Results

6.3.1 Characteristics of patient cohort

Table 6.2 shows the demographics of the POSH participants recruited in each batch. The proportion of women aged 31–40 was 86–91%; fewer women were in the 21–31 age group (9–14%) and fewer than 1% were under-20. The majority of patients (97%) were ethnically Caucasian (self-reported) and the proportion of women that reported Asian, Black or other ethnicity was around 3%. Approximately half of the POSH population had family history information, with a higher proportion of cases having a family history in the Cambridge and Haiman batches. 337 (14%) of 2,503 patients included in the POSH study had either a *BRCA1* or *BRCA2* mutation, and 2,166 (87%) carried a polygenic component.

Table 6.2: Demographic characteristics of POSH cohort

Dataset	**Mayo**	**GIS**	**Cambridge**	**Haiman**	**Oncoarray**	**GSA**	**Total**
Cases	**274**	**300**	**308**	**377**	**1088**	**156**	**2503**
Age at diagnosis							
≤20 years		1 (<1%)			1 (<1%)		2 (<1%)
21 - 30 years	30 (11%)	37 (12%)	29 (9%)	47 (12%)	123 (11%)	20 (13%)	286 (11%)
31 - 40 years	244 (89%)	262 (88%)	279 (91%)	330 (88%)	964 (89%)	136 (87%)	2215 (88%)
Mean age	35.1	35.2	35.6	35.3	35.3	34.8	35.5
Stated ethnicity							
Asian		1 (<1%)	7 (2%)	8 (2%)	1 (<1%)	12 (8%)	29 (1%)
Black			7 (2%)	14 (4%)		11 (7%)	32 (1%)
Caucasian	274 (100%)	296 (99%)	288 (95%)	348 (93%)	1086 (99%)	128 (82%)	2420 (97%)
Other			2 (1%)	3 (1%)	1 (<1%)	1 (<1%)	7 (<1%)
Missing		3	4	4		4	15
Family history							
Yes	57 (21%)	87 (30%)	285 (93%)	341 (91%)	300 (28%)	60 (40%)	1130 (45%)
No	213 (79%)	205 (70%)	20 (7%)	33 (9%)	755 (72%)	88 (60%)	1314 (53%)
Missing	4	8	3	3	33	8	59
Cancer status							
Polygenic component	232 (85%)	259 (86%)	235 (76%)	343 (91%)	992 (91%)	105 (67%)	2166 (87%)
BRCA1 or 2	42 (15%)	41 (14%)	73 (24%)	34 (9%)	96 (9%)	51 (33%)	337 (13%)

6.3.2 Quality control

After QC, this study included 2,064 unique cases from POSH cohort and 5,195 individuals from WTCCC data that passed quality control filters, and 40,545 SNP markers overlapped between

these platforms. Downstream analysis was performed, which prepared raw genetic data on both the POSH and the WTCCC cohort, completed pre-imputation quality control, phasing, imputation, post-imputation quality control, population stratification and batch effect analysis. The steps of the analysis are detailed herein.

Table 6.3 summarises the SNPs and samples removed from the POSH cohort at each stage of the QC procedure on the raw data across all six batches. It shows that a little more than one-third of SNP cases (32%, 7,731,260 SNPs) were removed, rejecting SNPs with ≥10% missing genotypes (5%, 362,898 SNPs), significant deviations from Hardy-Weinberg equilibrium $p < 10^{-10}$ (< 1%, 4,357 SNPs) and allele frequencies less than 10% (27%, 2,119,800 SNPs).

For the observed data on the WTCCC, approximately 97% of the SNPs passed quality control filtering. The fraction for SNPs was about 2% (29,295 SNPs) without the MAF > 0.01. Approximately 1% of these SNPs (24,640) were removed for genotype frequencies observed to have a significant deviation from HWE ($p \leq 0.001$). Most samples had very low rates of missing data, and 9,392 SNPs with ≥ 10% missing genotypes were excluded.

The overall genotyping rate was above 0.997 across all batches on the raw data, showing that the samples had complete genotypic data with few missing genotypes. Table 6.3 also shows that further sample QC was implemented to remove SNPs with discordance, sex differences in allelic frequency, outlying rate of heterozygosity, duplication, related samples, and divergent ancestry for both the POSH and WTCCC datasets. Once the data was cleaned and merged from all POSH batches and WTCCC cohorts, a total of 2,772,602 SNPs remained, of which 892,217 were genotyped in both cases and controls for subsequent analysis. Following the merge of all POSH case and control data, 2,064 of 2,503 remained as unique cases and for unaffected individuals, 5,195 of 5,200 were retained for PRS construction (Table 6.4).

6.3.3 Population stratification

Population diversity from the POSH study was examined using MDS. Patients reported their ancestry as Caucasian, Black or Asian. The inferred ancestry was based on pairwise IBS across 57,258 autosomal SNPs in LD and genotyped across all cohorts (6 case batches, controls and reference populations from HapMap). The differences between self-reported ancestry tended to be close to inferred ancestry (Figure 6.3). There is a clear overlap between most POSH cases and WTCCC, indicating appropriate controls for the subsequent analysis.

6.3.4 Testing for batch effects

To further test whether the results are robust, the dummy case versus control analyses between all batches was performed (Table 6.5). The results suggest that the close agreement between observed and expected $p < 10^{-8}$ over the entire distribution, from least to most significant, and genomic inflation factors below 1.05 show that there is no sign of systematic bias between batches or residual

Table 6.3: SNPs quality control summary

QC criteria	POSH cases						Controls	
	Mayo	GIS	Cambridge	Haiman	Oncoarray	GSA	WTCCC 1958 *BBC*[1]	WTCCC *NBS*[2]
Standard filters								
Raw	559,348	559,348	594,360	4,763,767	494,444	759,993	954,144	954,144
Missing genotypes	363	629	2,489	341,607	10	17,800	12,697	11,943
Hardy-Weinberg	69	71	1,541	2,559	0	117	5,320	4,072
Minor allele frequency (MAF)	25,486	24,644	35,208	1,712,899	91,550	230,013	14,600	14,695
Remaining	533,430	534,004	555,122	2,706,702	402,884	512,063	921,527	923,434
SNPs passing standard filters								
In strand file	533,313	534,004	554,437	2,615,815	402,824	512,063		541,469
MAF difference >0.2	172	183	656	6,840	210	1,026		274
AT/GC MAF >0.4	294	293	721	10,772	2,036	551		591
Duplicates	2	2	4	57,437	12	1,180		1
Non-matching alleles	1,228	1,235	1,334	6,739	909	595		1,267
Not in haplotype reference consortium (HRC)	107	154	798	28,163	2,752	12,633		351
Indels ignored	0	0	0	2071	5,602	207		0
Skipped	64	65	13	227	16	280		5
Remaining	531,446	532,072	550,911	2,503,566	391,287	495,591		538,980
Custom SNPs passing standard filters								
All	117	0	685	90,887	60	0		394,076
Remove AT/GC			36	29,690				14,045
MAF difference >0.2			18	155				347
AT/GC MAF >0.4			0	0				7
Duplicates			0	683				1
Non-matching alleles			2	112				874
Not in HRC			185	7,942				832
Skipped			0	241				590
Remaining			444	52,064				377,380
Merge strand and *custom*[3]	531,446	532,072	551,355	2,555,630	391,287	495,591		916,353
Merge pairs								
Discordant	776	1,610	2,375	7,201	2,381	1,771		
MAF difference, batches >0.15 and HRC >0.1	1	1	30	2	0	50		
	|		|		|			916,353 (Both)
Remaining	530,669	530,461	548,950	2,548,427	388,906	493,770		902,868 (BBC)
								904,723 (NBS)
Merge pairs for Mayo, GIS, Cambridge, Haiman, Oncoarray, GSA and WTCCC								
Remaining					2,748,467			**916,353**
SNPs in both					**89,217**			
SNPs remaining					**2,772,602**			

1 1958 BBC – 1958 British Birth Cohort.
2 NBS – National Blood Service.
3 Merge strand and custom: SNPs may overlap between strand and custom after updating identifiers, therefore total SNPs can be less than sum of strand plus custom.

Table 6.4: Sample quality control summary

QC criteria	POSH cases						Controls
	Mayo	GIS	Cambridge	Haiman	Oncoarray	GSA	WTCCC
Raw	274	300	308	377	1,088	156	5,200
Missing genotypes >10%	0	2	0	0	0	0	0
Sex check	0	1	1	2	0	0	0
Heterozygosity	4	4	6	3	6	5	0
Remaining	**270**	**293**	**301**	**372**	**1,082**	**151**	**5,200**
Merge pairs for Mayo, GIS, Cambridge, Haiman, Oncoarray and GSA =						**2,068**	
Relatedness						4	5
Remaining						**2,064**	**5,195**

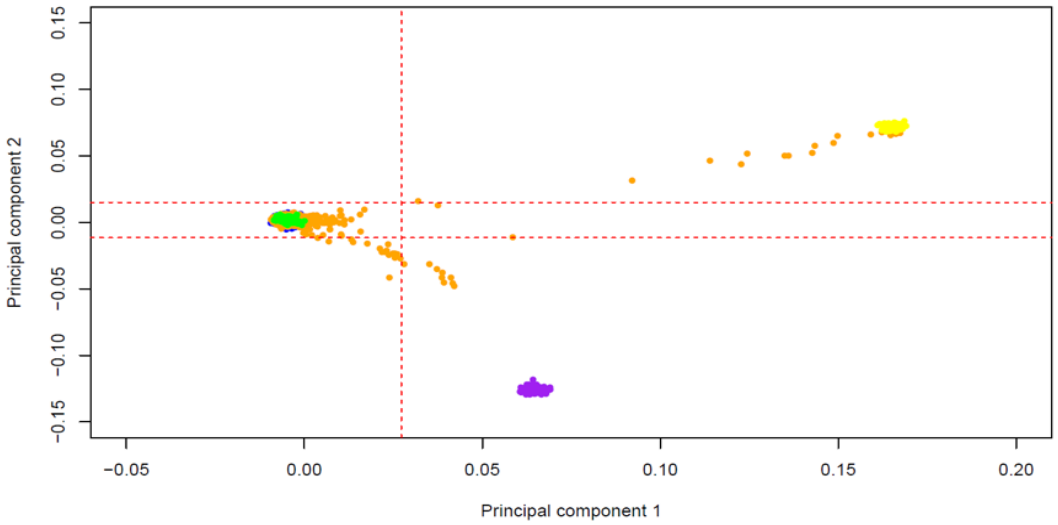


Figure 6.3: **Multidimensional scaling plot of Caucasian populations from POSH study and WTCCC individuals**. MDS for whole-genome genotype data for POSH study and WTCCC individuals. Colours indicate cohort (POSH cases=orange, WTCCC controls=blue, HapMap CEU=green, HapMap YRI=yellow, HapMap ASI = purple).

evidence of population stratification. One genome-wide significant SNP was removed to mitigate incorrect associations with the phenotype due to a batch effect. Therefore, it can be concluded that systematic bias is unlikely to impact the results.

Table 6.5: Dummy case control analysis between all pairwise batches

Cases	Controls	Duplicates	GIF*	SNPs tested
Haiman (237)	Cambridge (231)	205	1.0188	507,542
GIS (293)	Cambridge (301)	0	1.0089	526,754
Mayo (270)	Cambridge (301)	0	1.0174	526,995
Oncoarray (1073)	Cambridge (301)	7	1.0127	147,483
GSA (149)	Cambridge (283)	18	1.0122	113,197
Haiman (345)	GIS (293)	27	1.0094	502,790
GSA (149)	GIS (293)	0	1.0160	112,136
Mayo (270)	GIS (293)	0	1.0117	527,620
Oncoarray (1071)	GIS (293)	9	1.0122	146,943
GSA (151)	Haiman (372)	0	0.9700	387,642
GSA (149)	Mayo (267)	3	1.0009	112,075
Oncoarray (1039)	GSA (149)	41	0.9723	379,989
Haiman (357)	Mayo (270)	15	1.0098	502,291
Haiman (372)	Oncoarray (1005)	75	0.9995	332,384
Oncoarray (1076)	Mayo (270)	4	0.9879	146,896

Note. *P$<5\times\mathbf{10^{-8}}$

Because the POSH cohort data were genotyped in six separate batches, an MDS analysis was performed to detect potential batch effects (see Figure 6.4). The scatterplot of the first two dimensions from MDS of the IBS metric indicated a considerable batch effect in the SNP data collected from the POSH study, as most samples were in six small clusters. Due to this limited overlap in SNPs genotyped across all batches, the sample size and therefore power was reduced and the batch effect was increased. It was necessary to remove technical variations not to skew the accuracy analysis to capture the 'true' biological variation of their disease phenotype.

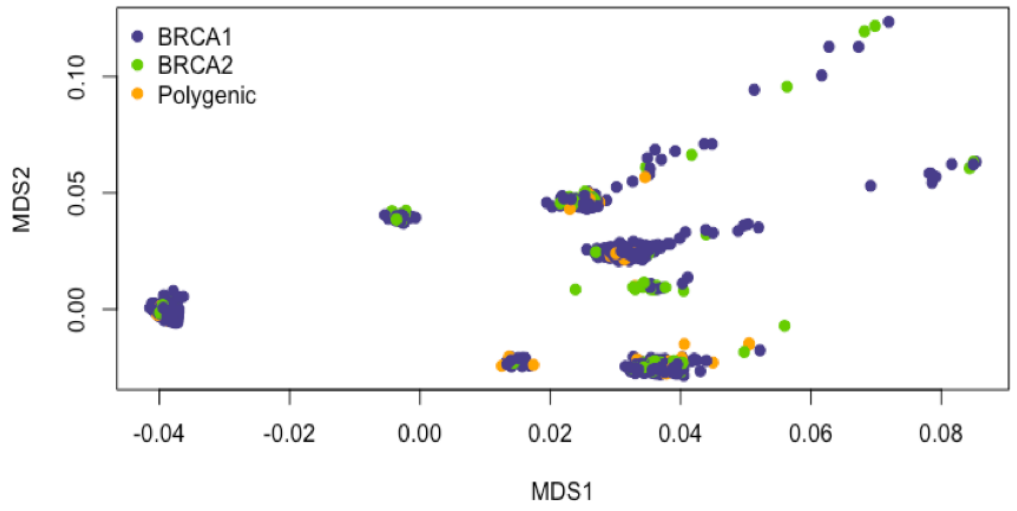


Figure 6.4: **Multidimensional scaling plot of POSH study by BC risk in mutation**. Genotype data of samples from POSH study. Plotting under two dimensions.

Therefore, imputation was performed to fill in the genomic gaps for both POSH and WTCCC datasets, increase statistical power, and standardise the datasets. As a result, batches genotyped with different arrays can be combined, increasing the resolution of overlap in genotypic data and reducing batch effects. Here, SNPs were imputed in separate batches according to genotype array. In some cases, the power to replicate SNPs was limited due to small effect sizes, rare risk alleles or where SNPs were only genotyped in a subset of patients.

Following imputation, the 7,259 individuals with genotype data from both the POSH cohort and WTCCC were imputed from EAGLE2 for pre-phasing into the HRC and positional PBWT for imputation (see Table 6.6). Specifically, the Haiman batch resulted in the most imputed variants across the autosomes with a total of **6,461,507** variants, followed by the GSA cohort with approximately four million autosomal variants. In contrast, the MAYO and Cambridge batches had similar imputed variants of around three million. The Oncoarray and GSA batches had about two million imputed variants. The WTCCC reference population had slightly fewer imputed variants than the POSH batches cohort. This reference had around four million imputed variants.

Once all the chromosomal segments were imputed, a post-imputation quality control check was run, where poorly imputed variants were filtered out (see Table 6.7). These were: first, rare SNPs with minor allele frequency less than 2%; and second, SNPs with MAF $< 1\%$, SNPs with $> 10\%$ missing genotypes, SNPs with significant deviations from Hardy-Weinberg equilibrium ($p <= 1 \times 10^{-10}$ in

Table 6.6: Sample imputation quality control summary

Criterion	Mayo	Gis	Cambridge	Haiman	Oncoarray	GSA
Input	530,669	530,461	548,950	2,548,427	388,906	493,672
MAF difference >0.2	1	0	4	19	0	15
Imputation input	530,668	530,461	548,946	2,548,408	388,906	493,657
Genotyping rate	0.999134	0.999619	0.998985	0.998177	0.999446	0.997228
Imputation output and filtering						
Filtered (MAF)>1%, info>0.99)	3,951,850	4,029,602	3,763,866	6,461,507	2,928,788	2,054,255
Missing genotypes	0	22	0	0	0	0
Hardy-Weinberg	45	56	208	1,384	25	10,510
Duplicates	618	696	538	1,198	346	244
Remaining	3,951,187	4,028,828	3,763,120	6,458,925	2,928,417	2,043,501
Merge pairs						
Triallelic	148	151	2+141	3+170+43	105	137
Non matching alleles	9	2	6	24	2	2
Discordant	323,541	673,001	535,330	2,201,578	809,870	371,575
P<5×10^{-8} between batches	0	0	2	0	0	0
Remaining	**3,627,489**	**3,355,674**	**3,227,639**	**4,257,107**	**2,118,440**	**1,671,787**

cases ≤ 0.001 in controls), SNPs with one or more discordant genotypes between duplicate samples, SNPs with GWA significance between batches ($p <= 5 \times 10^{-8}$). Lastly, the resulting SNPs were merged for the POSH and WTCCC datasets, accounting for 5,830,116 variants. The genotyping rate for the imputed data was above 0.86.

Table 6.7: Sample merge summary

Merge case batches	5,830,116
Hardy-Weinberg	84
MAF	574
Remaining	5,829,458
Merge with controls, no SNPs in common	4,132,794

For the merged imputed data, all related individuals were removed based on IBD estimation, QC, LD pruned with $r_2 < 0.5$, a window size of 50kb, step size of 5 SNPs and MAF > 10% to include only common variants. SNPs selected after pruning were used to calculate IBS distance between each pair of individuals that passed QC. A total of **40,545** SNPs were retained from both datasets after imputation not to skew the accuracy analysis. MDS was performed again to test the batch effect on the POSH imputed dataset. Figure 6.5 presents plots for the first two components, 1 and 2, coloured by genotyped array. The distribution structure is well evident from these MDS analyses, and it can be inferred that there is no batch effect.

The basic descriptive statistics of the BC patients according to pathogenic mutations and controls are presented in Table 6.8. Of the 7,259 individuals used for the analysis, 5,195 (∼72%) were unaffected individuals, and 1,790 (∼25%) had a polygenic component moderate risk or commoner variants. This outcome was more prevalent in carriers of *BRCA1* with 169 (∼2%) and *BRCA2* with 105 (∼1%).

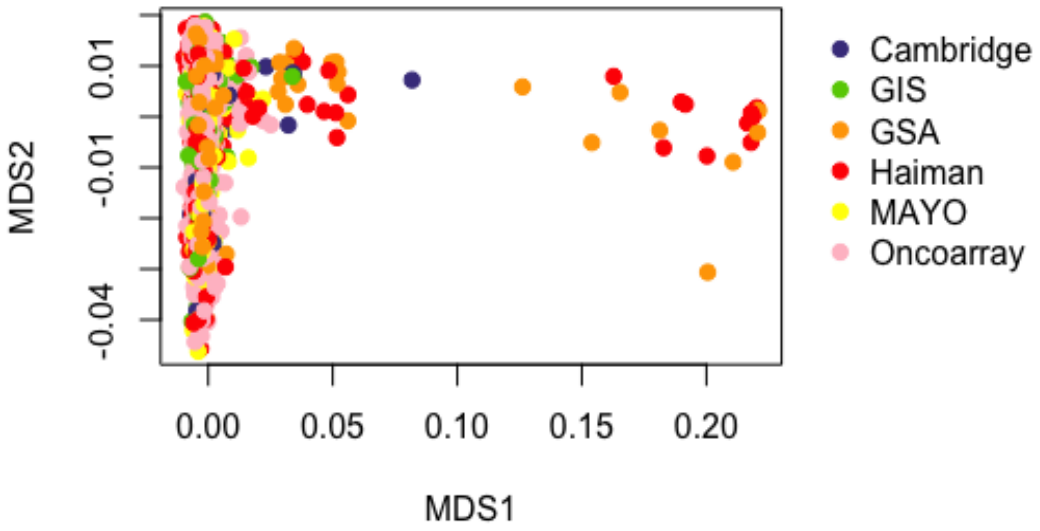


Figure 6.5: **Multidimensional scaling plot of POSH batches by BC risk in mutation**. Genotype data of samples from POSH study. Plotting under two dimensions. Population cluster designations are labelled on the plot.

Table 6.8: Descriptive statistics of expected information per locus

Type	Samples	Mean/SD
Control	5195 (71.6%)	1.0098 (± 0.003)
BCRA1	169 (2.3%)	1.1550(± 0.087)
BCRA2	105 (1.4%)	1.1390 (± 0.019)
Polygenic component	1790 (24.7%)	1.1400 (± 0.042)
Total	7259 (100%)	-

6.3.5 Polygenic risk score association with breast cancer prevalence

Significant differences of EIL were only observed in *BRCA1* mutation and polygenic component carries. (p-value = 0.009, two-sided Wilcoxon rank-sum test) (Table 6.8). In Figure 6.6 (A) shows that the *BRCA1* mutation carries elevated EIL in comparison to *BRCA2* mutation and polygenic component patients. In contrast, the polygenic component carriers had a slightly heavier tail than the distribution of *BRCA1* and *BRCA2* mutation carriers indicating a greater proportion of samples with higher EIL (see Figure 6.6 (B)). The greatest EIL was observed in carriers with polygenic component mutations (orange line Figure 6.6 (B)), while the risk levels associated with *BRCA1* and *BRCA2* mutations (purple and green line, respectively Figure 6.6 (B)) are also showing elevated EIL.

To investigate further the relationship between BC risk in mutations odds and EIL, logistic regression was conducted. However, the cases and control samples are not directly comparable due to

Table 6.9: EIL two-sided Wilcoxon rank sum test within BC risk in mutation carriers and controls

Type		Wilcoxon rank-sum test	p-value
Polygenic component	*BRCA1*	-2.61	0.009
	BRCA2	-0.58	0.560
BRCA1	*BRCA2*	-1.21	0.226

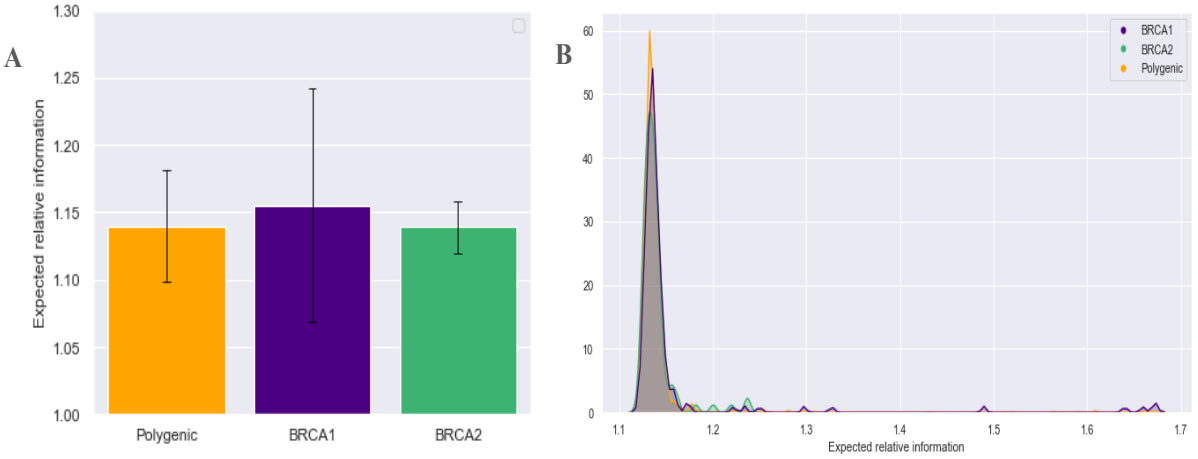


Figure 6.6: **Breast cancer risk associated with the increased genome-wide disorder**. A) Expected information per locus (EIL) within BC in mutation carriers. Median ±95% confidence intervals are shown. (B) EIL for BC risk in mutation carriers density distributions. The plot shows the area under the curve of a density function representing the probability of carrier and noncarrier mutations between the ranges of EIL values.

the batch effect for using different genotyping arrays. The sub-structure in control data will lead to a systematic bias that influences the results of association testing. Therefore, it was no possible to make comparisons between these populations.

In addition, a logistic regression model was performed in order to estimate the relationship between EIL for BC cases (see Table 6.10). According to the logistic regression results, there is a significant association in EIL among the BC cases. Consistent with the heavy–tailed nature of the polygenic component and *BRCAs* distribution (Figure 6.6 (B)). In particular, the highest odds ratio for EIL was between *BRCA2* and the polygenic component having an odds ratio greater than 12. In comparison, the odds ratio for EIL of the polygenic component was 7.82 times higher than *BRCA1*. The results also show that EIL is slightly elevated in patients for *BRCA1* compared to *BRCA2* with an odds ratio of 1.53. Overall, these results indicate that EIL is significantly elevated in BC cases, with the highest EIL for the polygenic component conferring a substantially increased risk.

Table 6.10: Logistic regression model for cases

Mutation type	Odds ratio	coefficient	std. err	z	p>z	[0.025	0.975]
Polygenic vs *BRCA1*	7.82	2.06	0.07	29.30	0.00	1.92	2.19
Polygenic vs *BRCA2*	12.06	2.49	0.09	28.23	0.00	2.32	2.66
BRCA1 vs *BRCA2*	1.53	0.43	0.11	3.93	0.00	0.21	0.64

In order to investigate the genetic basis with particular genomic loci, the EIL was assessed on individual chromosomes. The EIL was consistently elevated in polygenic component cases in comparison with the carriers of *BRCA1* and *BRCA1* along the whole chromosomes (Figure 6.7) ($p < 0.05$, one-sided Wilcoxon rank-sum test), indicating that differences in EIL are distributed throughout the genome. It was observed that *BRCA1* and polygenic component carriers are significantly different for autosome chromosomes 1–22. *BRCA1* and *BRCA2* were different on chromosomes 3, 5 and 13. However, there was no evidence of heterogeneity in the effect of EIL between polygenic cases and *BRCA2* carriers across chromosomes.

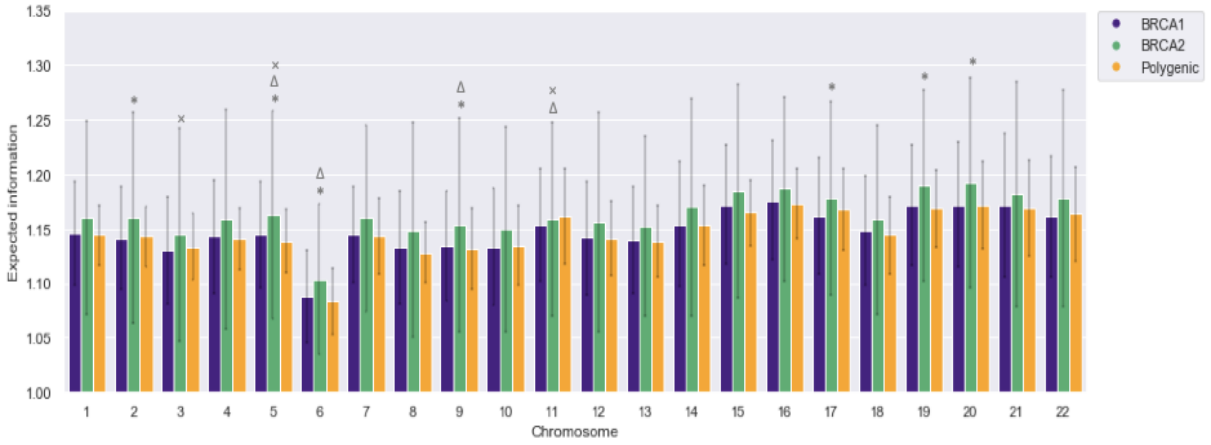


Figure 6.7: **BC risk in mutations in specific regions of the genome**. Expected information per locus (EIL) by chromosome. In all panels, median ±95% confidence intervals are shown. Two-sided Wilcoxon rank-sum test is represented with stars which indicate significant changes between *BRCA1* and polygenic component; triangles denote *BRCA1* and *BRCA2* statistical differences; x's represents *BRCA2* and polygenic component dissimilarities. There were no significant differences observed between *BRCA2* and polygenic component.

The results among BC cases may be biased due to the structured population in the data set used and thus risk of false-positive findings. Although the data was treated to mitigate this batch effect, the sample only considered approximately ∼40,000 SNPs which may not be representative, or it is not enough power to find a statistically significant difference between BC cases. More insight will be detailed in the conclusion section below.

6.4 Discussion

The main conclusion of this work reflects a sub-structure in both data set used, POSH and WTCCC, that were not expected. Briefly, batch effect in a population generates structure in genetic variation, correlated most strongly with false-positive findings. Despite clumping/pruning methods have been used to mitigate the batch effect, the relevant evidence is insufficient to confirm the results found in this study. Because the allele frequencies might differ systematically between intra-batch and intra-batch variations, therefore cannot be possible to identify whether the differences come from genetic drift or the ascertainment of genotype variants. Hence, there could be a risk that differences at null SNPs may not generate an association between the PRS and BC subtypes.

The population from the POSH study was well characterised, although a fundamental limitation of the present study is derived from the data genotyping. The SNPs used to construct the PRS were in both data, WTCCC and POSH study, which may not be directly comparable among batches due to the difference in the genotyping arrays. The differences in genotyping SNPs for BC cases in six separate batches are limited overlap of SNPs genotyped across all those batches in the data, which reduced the sample size to $\sim$ 40,000 SNPs less statistical power. However, the findings found in this study were not very satisfying, noteworthy the problems observed in the datasets had demonstrated that in the Smyth *et al.* [96] study had not recognised a substructure in the datasets that might conduct false-positive results. It is of note that should be considered for further future analysis. Moreover, the construction of the PRS was based on WTCCC; it was not possible to show the efficiency of the PRS due to its data substructure. Therefore, applying this PRS to a new reference dataset will help quantify the risk of developing a particular disease. Finally, another possibility for finding spurious results might be that many non-*BRCA1/2* cases are not certainly polygenic. They could be more related to environmental factors or sporadic BC; these results may indicate that low and moderate penetrance genes/loci can only explain a minor fraction of non-*BRCA1/2* BC.

Currently available methods to compute PRS have used either parametric or Bayesian approaches to determine the BC risk. Despite this, the methodology proposed here is considered to be a robust alternative to measuring BC risk. This approach introduced the estimation of probabilistic expectations given a rigorous mathematical formulation in the form of surprisal theory [267, 96]. Thus, this theory focused on the intuitive notion that allele frequencies are more difficult to process when they are less likely to occur in their genome. More formally, the surprisal theory proposed that allele frequencies are proportional to its negative log probability on all previous frequencies.

Nevertheless, this research shows that the PRS developed for women who carry *BRCA1/2* mutations or intermediate-risk/common variants supported the hypothesis that the BC cases include a strong inherited polygenic component. The PRS also showed that the non-*BRCA1/2* breast cancer patients have particularly high EIL compared to *BRCAs* cases. The EIL measured the cumulative effect of low-penetrance allele frequency on disease risk in BC cases compared to controls. The results obtained in this study for BC carriers with different mutations showed higher EIL values, implying a surprising genome. The results also show that the polygenic component carriers tend to have similar allele frequencies for the risk of BC to highly penetrant genes. By contrast, the reference population showed lower EIL, implying more common alleles and a less unusual genome.

As genetic susceptibility in *BRCA1* and *BRCA2* mutation is already known to confer a high risk of developing BC, the combined effects of risk-modifying variants could lead to much more significant differences in the absolute risk of developing BC as compared with the general population [268, 269, 270]. Consistent with this, the RGI cumulative effects demonstrated that women with increased BC risk have a significant polygenic component involving variation at thousands of markers distributed throughout the genome [96]. This study also has shown that the SNPs by which modify BC risk in carriers of the polygenic component mutation have somewhat higher EIL than *BRCA1* and *BRCA2*

carriers (see Table 6.10). Conversely, *BRCA2* conferred lower risk to BC compared to *BRCA1* and polygenic component.

The present study's results align with recent BC cases findings compared with noncarriers with BC risk-increasing alleles. For example, Mavaddat *et al.* [99] developed a PRS derived from 77 SNPs showing a strong effect of the score in predicting future BC cases. Sieh *et al.* [271] also developed a genomic risk score using allele frequencies and effect sizes of 86 SNPs, inferring a distribution of breast cancer risks for BC cases. L "all *et al.* [248] found the strongest association with prevalent BC status in patients compared to controls based on PRS. More recently, Jia *et al.* [272] constructed a polygenic risk score using GWAS to identify risk variants for eight common cancers. The authors estimated that female breast cancer patients had a higher mean value of PRS than non-BC patients (0.628, 95% CI = 0.620 to 0.637). Mavaddat *et al.* [99] did not observe a strong effect between the SNPs that modify breast cancer risks in *BRCA1* and *BRCA2* mutation carriers compared to the general population, consistent with the results from EIL presented here in different types of mutations. Despite the differences in methodology, Mavaddat *et al.* [99] demonstrated that there are inherited components associated with BC risk, which compares closely with the findings of this study.

SNPs associated with BC risk in different types of mutations were observed in *BRCA1* and polygenic component carriers. Their distributions tend to overlap, which indicate that these genes may share similar behaviour (see Figure 6.6 (B)). This could explain why mutations in these genes lead to functional similarities and specific hereditary predisposition to BC. Moreover, BC risk for *BRCA1*, *BRCA2* and polygenic mutation may be influenced by different loci (see Figure 6.7). These findings are consistent with the fact that both *BRCA1* and *BRCA2* are associated with maintenance of chromosome stability and recombination-mediated double-strand break repair of DNA [273]. Additionally, the polygenic component was defined as small multiplicative effects on BC risk of non-*BRCA1/2* cases, which may indicate that SNPs in these genes are acting in a common pathway [274, 93]. However, the links between these genes are still not well understood.

In summary, the findings demonstrate that RGI global measures of genome variation enable the polygenic basis of BC to be quantified more efficiently than GWAS. In addition, the result showed that an efficient PRS estimate might identify higher-risk strata in BC risk levels (see Table **??**). Thus, the individual's genome information can be used as a predictor of breast cancer susceptibility. However, the PRS is still a proxy of real genetic risk and is not uniquely defined. More efficient polygenic predictors could therefore be developed integrating molecular subtype distribution between the polygenic component and sporadic tumours to make more accessible the identification of BC [275].

Furthermore, stratifying BC by molecular subtypes based on factors including estrogen, progesterone or human epidermal receptor status [275, 273] may improve the estimation of RGI for BC risk. In addition, incorporated signatures based on transcriptome as well as genome profiling have proven suitable for predictions of *BRCA1/2* and polygenetic component status [275]. Furthermore, the PRS developed in this study can also be extended straightforwardly to include other specific

genetic variants such as Mendelian, or if deemed helpful in the future, a combination of Mendelian variants and PRS levels would require further study.

Chapter 7

Conclusions and future work

7.1 Thesis summary

Recent rapid technological advances to reveal genomic insights have increased the ability to carry out high-throughput studies characterised by large datasets. Big data management has become a significant aspect of genomic research, including the study of human diseases. Now, the challenge is to identify, within the massive amount of data obtained with next-generation sequencing (NGS), what is of genomic relevance. In this context, computational methods are being developed to incorporate genotypic and phenotypic knowledge to better understand genome function. Bringing together computational genomics expertise and research knowledge will provide significant and novel insights into biological mechanisms and aetiology of disease.

The central goal of this thesis was to develop and apply statistical methods and mathematical tools that can boost the analysis of genome function, Linkage disequilibrium (LD) structure and disease gene prediction. To that end, this thesis has presented different approaches to using different types of data to transform the information gained from newly generated data and produce knowledge.

At a first level, high-resolution LD patterns from the whole-genome sequencing (WGS) Wellderly data were constructed to reveal the LD structure of functional elements within genic and subgenic sequences. Next, application of the mathematical Malécot–Morton model was introduced to construct LD map distances in linkage disequilibrium units (LDUs) from 454 individuals from the Wellderly cohort [149]. The ratio of corresponding map lengths, kilobases/LDU, was defined to describe the extent of LD within genome components. The findings demonstrated that significant differences between exonic, intronic and intergenic components at fine-scale LD structure provide important insights into genome function. In addition, patterns of LD were observed that vary across the gene profile, although the functional implications are not fully understood.

On another level, this thesis introduces the application of supervised machine learning (ML) algorithms to identify which determinants are acting on monogenic disease-causing genes based on evolutionary and functional properties at the gene level. These metrics were integrated using LD

maps constructed in Chapter 3 and publicly available biological data. The Gradient tree boosting (GTB) and Random forest (RF) models were run in a multi-stage procedure to reveal which functional properties are closely associated with the degree of essentiality of the genes that are interesting and relevant to be linked to diseases. Bayesian inference in Gaussian graphical models (BGGM) was implemented in such a way that putative features were selected for classifying genes. These findings enable analyses to be optimised by genotype to improve recognising genes that are likely contributing to Mendelian phenotypes.

Following on from this, in chapter 5, a hybrid approach was proposed called simultaneous Gaussian mixture clustering with outlier removal (GMM-non) to classify and stratify genes according to their degree of essentiality into four groups; Non-disease non-essential (NDNE), Mendelian non-complex (MNC), Complex non-Mendelian (CNM) and Essential non-disease (END). This approach recognises aberrant genes in order to remove them during the clustering process. The iterative process is improved simultaneously until algorithm convergence is reached. The success of this framework has revealed subtle differences in gene patterns that characterise them into different gene groups. It is clear from these analyses that the features used provided significant information gains, particularly in genes associated with Mendelian disorders. However, further analysis and discovery of patterns and associations of these gene groups must be made based on integrating molecular pathways such as protein–protein interactions or gene expression patterns to obtain non-overlapping clusters.

Due to an explosion in the search for polygenic risk scores, surprisal theory was applied in this work to construct a polygenic risk score (PRS) for Breast Cancer (BC). This PRS enabled the cumulative effect of low-penetrance genetic variants in known BC susceptibility genes to be quantified. Surprisal theory was used to measure differences in global genome structure between cases and appropriate controls. The quantification of those deviations is defined as relative genome information (RGI), which estimates how unusual a genome is compared to the reference genome. In order to test its ability to predict disease risk, RGI was used to compare approximately 40,000 SNPs of 7,259 genomes from the Prospective Study of Outcomes in Sporadic versus Hereditary breast cancer (POSH) and Wellcome Trust Case Control Consortium (WTCCC) datasets. The results suggest that RGI is significantly higher in BC cases, with a polygenic component conferring a substantially increased risk compared to controls. These results also indicate that BC cases comprise a strong inherited polygenic component.

Overall, throughout this thesis, it was sought to use mathematical and statistical models to generate accurate predictions and facilitate scientific exploration. A new statistical method was not purely developed, but rather the capability of these methods to address insight into genomic signatures, linkage disequilibrium patterns and disease gene prediction was demonstrated.

7.2 Study limitations

Although the mathematical tools and statistical strategies were successfully applied, this study was limited by various general factors, as described in the following.

Machine learning limitations

A reliable set of ML algorithms was not found for this type of data. Although the dataset was preprocessed to be balanced and imputed, the features' combined effect only explains around 50% of the variance in gene properties that would yield gene predictions within a reasonable error margin. The classifiers considered and tested, including gradient boosting and random forest, were unable to efficiently and adequately distinguish between gene groups. One possible explanation for this inability to separate genes could stem from more variance amongst samples within gene groups than between genes belonging to different groups. This would make it difficult to separate each group and would drive poor performance of the ML algorithms. Another possibility is that the features selected in Chapter 4 were not accurately informative because the variance explained by them was not suitable to describe the gene groups used. The scope of the data is limited to measure gene properties and may not include important insights for the disorders studied.

Furthermore, this study is limited by the currently available genetic information for human diseases and the incomplete knowledge of the true susceptibility/causal variants and their corresponding genes. Moreover, the similar properties of disease-gene relationships restrict the ability of individual and integrated measurements to segregate certain gene classes. Therefore, a portion of the HD genes may be misassigned to the corresponding disease group.

Polygenic risk score limitations

The limitations of the polygenic risk score detected in this study were that the PRS was highly sensitive to the batch effect, meaning that variability in a PRS could be heavily influenced by allele frequency differences, differences in estimated effect sizes, and differences in population structure across different batches. The British dataset's population structure and POSH cohort were genotyped differently, making the SNPs not directly comparable between cases and controls. This lack of the polygenic risk score depends on the control dataset's population structure requires very complex attention and further analysis. Although in this study, the batch effect was controlled, the sample size was reduced. Therefore, the statistical power of the PRS was impacted due to the use of a modest sample size. Here, the PRS was calculated with ∼40,000 SNPs, making genetic association signals less clear and increasing heterogeneity within genetic levels. Although no power test was conducted, the power to detect an association was not enough to satisfy the whole genome-wide significant threshold (1×10^6). The reduction of these SNPs may only sample about 10

7.3 Future work

There are many encouraging areas of future work that can be explored using the ideas presented in this thesis as a starting point. The methods presented here could be improved in different ways. Some of the items listed below may help improve the efficacy of machine learning methods or PRS construction or simply may suggest interesting topics to explore further.

Using multi-omics data to predict disease genes

This work observed that ML-based approaches achieved lower prediction performance for clustering genes into the four gene groups NDNE, MNC, CNM, and END, due to the lower interconnectedness among these genes. This showed the weakness of the classifiers selected both in terms of distributions of complex and Mendelian disease genes and complex disease genes. However, this problem can be addressed by efficient integration of genomic data and omics data.

Multi-omics data characterise different stages of cellular activities, and analysing omic data may improve the accuracy of computational prediction. However, for disease gene prediction, most algorithms still focus on genomic data, and only a few algorithms have used multi-omics data in their studies [276]. Therefore, a critical area of future work is to extend the ML algorithms and the proposed GMM-non used in Chapters 4 and 5, integrating genomic data alongside other omics such as transcriptomic or proteomic data for disease gene prediction.

Expanding learning techniques

One potential future direction would be to extend the ML methods developed in this thesis to directly estimate latent structure while simultaneously using the structure of positive-unlabelled (PU) learning. This is because, for a specific disease, there are only a few known disease-associated genes from the human genome, and the rest of the genes are waiting for further analysis and remain unlabelled. Therefore, unlike traditional machine learning methods, treating the rest of the genes as an unlabelled set rather than a negative training set is a suitable strategy for this problem. Thus, unbiased PU learning algorithms [277] can be adopted for disease gene prediction algorithms. Hou *et al.* [278] developed a generative adversarial network (GAN)-based PU learning algorithm, which can open our minds to alternative ideas. GAN is a generative algorithm that is designed as a two-player game. One of the players in the generator, which synthesises fake data from random noise. The other player is the discriminator, which examines both fake and accurate data to determine whether they are real or not. Their design provides a more robust approach to adapt PU learning to deep neural networks without overfitting.

Incorporating information to PRS

One exciting direction for future work in the development of PRS would be to incorporate information about underlying linkage disequilibrium structure. This improvement may include information about the relationship of genotypes to gene expression profiling of breast tumours, a rapidly developing knowledge base, which would have beneficial effects in optimally weighting and calibrating the calculation of the PRS [279]. In fact, it has been demonstrated that the integration of gene expression database to generate PRS in psychiatric disorders may improve the predictive power of genotypic data by over threefold [280]. Furthermore, classifying subtypes of BC by estrogen, progesterone or human epidermal receptor status may improve the prognosis of BC [281].

Particular contributions

This thesis significantly contributed to constructing high-resolution linkage disequilibrium (LD) maps using whole-genome sequence data. Major differences between exonic, intronic and intergenic components confirm that fine-scale LD structure provides significant insights into genome function, which traditional linkage maps cannot explain. Moreover, this thesis proposed a robust prediction to identify disease genes using a simultaneous Gaussian mixture clustering outlier removal model. This approach improved the efficiency of recognising Mendelian disease genes from complex disease and non-disease genes through essentiality-specific information. Finally, this thesis demonstrated that the Smyth *et al.* [96] study had not recognised a substructure in the datasets for control that might conduct false-positive breast cancer diagnoses. Despite this, this thesis also has shown an alternative PRS based on the surprisal theory to evaluate the cumulative effect of low-penetrance alleles on BC risk.

Appendix A

Machine learning algorithms

This appendix details the supervised machine learning (ML) approaches used in this study to select the most significant features to distinguish between genes that have an association with Mendelian/-Complex disorders and those that do not. These machine learning methods included Gaussian naïve Bayes (GNB), k nearest-neighbour (k-NN), support vector clustering (SVC), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression cross-validation (LRCV), multi-layer perceptron (MLP), decision tree (DT), random forest (RF), and AdaBoost (AB).

Gaussian naïve Bayes classification

Gaussian naïve Bayes (GNB) classification is a supervised learning algorithm that uses Bayes' theorem as a framework for classifying observations into one of a pre-defined set of classes. GNB classifier takes a probabilistic approach for calculating an observation that belongs to a particular class based on the conditional independence assumption, thereby (naively) the covariance is not considered among the features. The posteriori Y is given by the values of features $X = (X_1, ..., X_p)$ into k classes. Y is modelled according to Bayes theorem as:

$$\hat{P}(Y = k|X_1, ..., X_p) = \frac{\pi(Y = k) \prod_{j=1}^{p} P(X_j|Y = k)}{\sum_{k=1}^{k} \pi(Y = k) \prod_{j=1}^{p} P(X_j|Y = k)}, \tag{A.1}$$

where $\pi(Y = k)$ is the prior probability that the class index is k. For each feature, the algorithm estimates a separate Gaussian distribution for each class, and observations are categorised to the class with the maximum posterior probability given the features values [282].

K-nearest neighbour classifier

The k-NN algorithm is a nonparametric supervised classifier in which the distance as a basis to weight the contribution of each k neighbour in the class assignment process. The k-NN algorithm clusters a set of data points into groups and classifies new data based on a measure of similarity using the Euclidean distance. The k-NN is entirely based on data-driven learning. Technically stated, given a value k and a feature vector to classify Y, locates the k nearest neighbours of Y in the sample set and uses the categories of neighbours to determine the class of I. k-NN computes the distances between a new observation and all the observations in the set use for learning, and thus choose the k observations from the learning set are the closest to the new observation. Finally, k-NN classify the new observation to the group associated with the most significant number the k observations. Once the k nearest neighbours are located, the class of the new observation is identified by using a voting algorithm.

The formal k-NN classifier algorithm is as follows:

$$argmin(d_e(t,n,k)) \implies identifyP \tag{A.2}$$

where t is the training data, n is the object to be classified, P is the assigned class of the new observation, k is the number of closest neighbours to be considered, and de is the Euclidean distance given by:

$$d_e(t,n,k) = \sqrt{\sum_{i=1}^{L}(t_{i,k} - n_{i,k})^2} \tag{A.3}$$

where L is the length of each data vector [116].

Support vector clustering

Support vector clustering (SVC) consists of mapping data points into a sphere with minimal radius by means of a Gaussian kernel function from high dimensional feature space. This sphere represents a set of contours which enclose the data points in input space. These contours are defined as cluster boundaries and points enclosed by each separate contour are associated with the same cluster. The SVC training phase contains to fix the width parameter of the Gaussian kernel, calculate kernel matrix, calculate Lagrange multipliers, select support vectors and calculate the radius of the sphere in the high dimensional feature space [283].

Following Ben-Hur *et al.* [283], denote a data set as $\{X_i\} \subseteq \chi$ of N points, with $\chi \subseteq R^d$, the data space. Using a nonlinear transformation Φ from χ to some high dimensional feature space to find

a minimal radius sphere R that comprises most of the data points in the feature space, described by the constraints:

$$||\Phi(X_j) - \mathbf{a}||^2 \leq R^2 + \xi_j \tag{A.4}$$

where . is the Euclidean norm and $\mathbf{a}$ is the center of the sphere and $\xi_j > 0$ is a soft constraints. Therefore, Lagrange multiplier is estimated as:

$$max \sum_{j=1}^{N} K(x_j, x_j)\beta_j - \sum_{i=1}^{N}\sum_{j=1}^{N} \beta_i \beta_j K(x_j, x_j) \tag{A.5}$$

such that $\sum_{j=1}^{N} \beta_j = 1$ and $0 \subseteq \beta_j \subseteq C, \forall j = 1, 2, \ldots, N$. Where β_j are Lagrange multipliers and $K(x_i x_j) = \Phi(x_i \bullet \Phi(x_j))$ is kernel function. It is demonstrable that only those points with $0 \subseteq \beta_j \subseteq C$ lie on the boundary of the sphere and are called support vectors (SVs).

Gaussian kernel is the next form:

$$K(x_i, x_j) = e^{-q||x_i - x_j||^2} \tag{A.6}$$

with width parameter q.

The distance of its projection in feature space from the center of the sphere at each x point:

$$R^2(x) = ||\Phi(X_j) - \mathbf{a}||^2 = K(x, x) - 2\sum_{j=1}^{N} K(x_j, x_j)\beta_j - \sum_{i=1}^{N}\sum_{j=1}^{N} \beta_i \beta_j K(x_j, x_j) \tag{A.7}$$

Let Ω is the set of all of SVs, then the radius of the sphere is:

$$R = \{R(x_i) | x_i \in \Omega\} \tag{A.8}$$

The contours that enclose the points in data space are defined by the set:

$$\{x | R(x) = R\} \tag{A.9}$$

Finally, the cluster assignment is leading by the definition of the adjacency matrix element A_{ij} between pairs of points x_i and x_j whose projections lie in or on the sphere in feature space:

$$A_{ij} = \begin{cases} 1, & \text{if R}(\lambda x_i + (1-\lambda)x_j) \leq \text{R} \ \ \forall\, \lambda \in [0,1] \\ 0, & \text{otherwise} \end{cases}, \tag{A.10}$$

where data points pair are given by (x_i, x_j), which belong to different clusters, any path that connects them must exit from the sphere in feature space, i.e., $\lambda \in [0, 1]$, such that $R(y) > R$, where $y = \lambda x_i + (1 - \lambda)x_j$.

Linear discriminant analysis

Linear discriminant analysis (LDA) is a dimensionality reduction technique based on Fisher's linear discriminant, which generates a linear projection matrix used to improve classification accuracy. In particular, the redundant and dependent features are removed by transforming the features from higher dimensional space to a space with lower dimensions. Besides, LDA determines linear decision boundaries by maximising the proportion of intra-class and inter-class variability. Thus, eigenvalue decomposition (EVD) estimates these linear boundaries. EVD assumes that the scatter matrix is non-singular, and all the variables are normally distributed and the covariance matrices are identical. Linear discriminant analysis is modelled as a multivariate Gaussian distribution with density:

$$P(X|y = k) = \frac{1}{(2\pi)^{\frac{d}{2}}|\sum_k|^{\frac{1}{2}}} e^{(-\frac{1}{2})(X-\mu_k)^t \sum_k^{-1}(X-\mu_k)} \tag{A.11}$$

where d is the number of features.

Formally stated, the intra groups (S_i) and inter groups variances are expressed by:

$$S_i = \sum_{i=1}^{C} \sum_{x \in C_i} (x_i - \mu_i)(x_i - \mu_i)^T \tag{A.12}$$

$$S_b = \sum_{i=1}^{C} \sum_{x \in C_i} L_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{A.13}$$

where x_i, L_i and μ_i are represented by the biological and functional properties of the genes at each ith class; μ is the mean of all classes and T is the transpose operator. Therefore, the W^* eigenvector is the maximization between intra and inter groups:

$$W^* = argmax_w \frac{|W^T S_b W|}{W^T S_i W} \tag{A.14}$$

i.e. W of $S_i^{-1} S_b$ [116].

Quadratic discriminant analysis

Quadratic discriminant analysis (QDA) is a variant of LDA, in which the assumption is that the observations are drawn from a Gaussian distribution which are not follow the same covariance matrix. That is, QDA has different feature covariance matrices for different classes, leading to a quadratic decision boundary. QDA is especially beneficial due to the prior knowledge that individual classes exhibit distinct covariances [284].

$$P(X|y=k) = -\frac{1}{2}(log|\sum_{k}| - (X-\mu_k)^t \sum_{k}(X-\mu_k) + log(\pi_k) \tag{A.15}$$

Logistic regression cross-validation

Logistic regression takes a cost function defined as a logistic function or a Sigmoid function. The hypothesis of logistic regression to limit the cost function between 0 and 1. It models the conditional probability as:

$$P_w(y=\pm 1|\mathbf{x}) = -\frac{1}{1+e^{yw^t x}} \tag{A.16}$$

where x is the data, y is the class label, and $w \in R^n$ is the weight vector. Given two-classes training data $\{x_1, y_i\}_{i=1}^{l}, x_i \in R^n, y_i \in (1, -1)$, logistic regression minimizes the following regularized negative likelihood.

$$P_w = C\sum_{i=1}^{l} log(1+e^{yw^t x)} + \frac{1}{2}w^t w \tag{A.17}$$

where $C > 0$ is a penalty parameter. This probability is referred to as the primal form of logistic regression, as one may instead solve dual problem. Thus, it is necessary to find the parameter vector weight w that minimize a cost function in order to get the best predicted output.

$$l(\mathbf{w}) = P(\mathbf{w}) \tag{A.18}$$

Dual formulation can be solved by implementing different solvers such as Newton methods, stands for Limited-memory Broyden-Fletcher–Goldfarb, library for large linear classification or stochastic average gradient descent [285].

Multi-layer perceptron

Multi-layer Perceptron (MLP) is a supervised learning algorithm with at least three layers of nodes, an input layer, certain number of intermediate layers, and an output layer. Each node in a given layer is connected to every node in the adjacent layers.

The input layer consists of a set of neurons $\{x_i|x_1, ..., x_m\}$ representing the input features, where m is the number of dimensions for input. Each neuron in the hidden layer transform the value from the previous layer with a weighted linear summation, followed by a non-linear function $g(.) : R \rightarrow R$. While, the output layer receives the values from the last hidden layer and transforms them into output values

MPL learns from a non-linear function approximator on the training dataset $(x_1, y_1), ..., (x_n, y_n)$ where $x_i \in R^n$ and $y_i \in (0, 1)$. The function is defined by:

$$f(x) = W_2 g(W_1^t x + b_1) + b_2 \tag{A.19}$$

where W_1, W_2 represent the weights of the input and hidden layer; and b_1, b_2 correspond the bias added to the hidden layer and the output layer. The activation function is set by the hyperbolic tan. It is given as:

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{A.20}$$

For more than two classes, $f(x)$ is a vector of size n which passes through the softmax function, which is written as:

$$softmax(z)_i = \frac{e(z_i)}{\sum_{l=1}^{k} e(z_i)} \tag{A.21}$$

where z_i represents the ith element of the input to softmax, which corresponds to class i, and K is the number of classes. The result is a vector containing the probabilities that sample belong to each class. The output is the class with the highest probability.

Finally, MLP uses different loss functions depending on the problem type. The loss function for classification is Cross-Entropy [286].

Bagging classifier

A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individuals predictions (either by voting or averaging) to reach a final prediction.

A learning set of L consists of data $\{(y_n, x_n), n = 1, ..., N\}$ where $y's$ are either class labels or a numerical response. Assuming that the learning set to form a predictor is $\varphi(x, L)$. The sequence of learning sets are given by L_k for N independent observations which follow the same L underlying distribution. Therefore, the average is taken as $\varphi_a(x) = E_{L\varphi(x,L_k)}$ where A in φ_A denotes aggregation. Thus, $\varphi(x, L)$ predicts a class $j \in (1, ..., J)$, then one method of aggregating the $\varphi(x, L_k)$ is by voting. Let $N_j = nr\{k : \varphi(x, L_k) = j\}$ and take $\varphi_A(x) = argmax_j N_j$, that is, the j for which N_j is maximum. Finally, taking repeated bootstrap samples $L^{(B)}$ from L, and form $\varphi(x, L^{(B)})$, the classifier is written by:

$$\varphi_B(\mathbf{x}) = av_b\varphi(x, L^{(B)}) \tag{A.22}$$

where av_B is the average of $\varphi(x, L^{(B)})$ [287].

Appendix B

Supplementary data

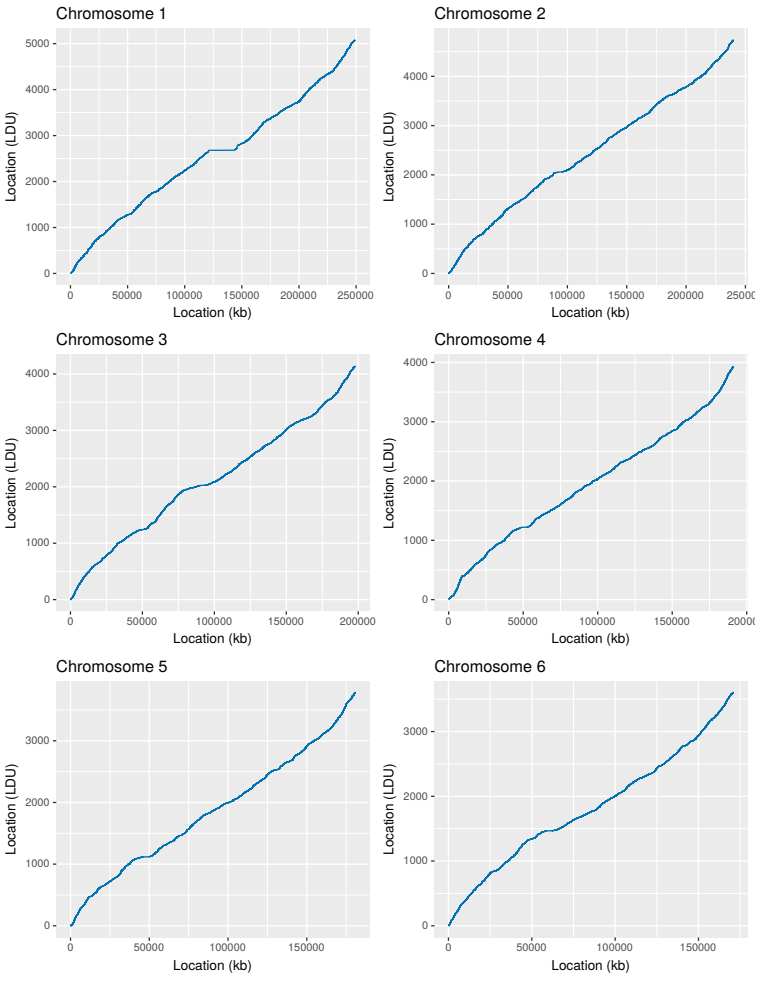


Figure B.1: LDU maps of chromosomes 1 to 6

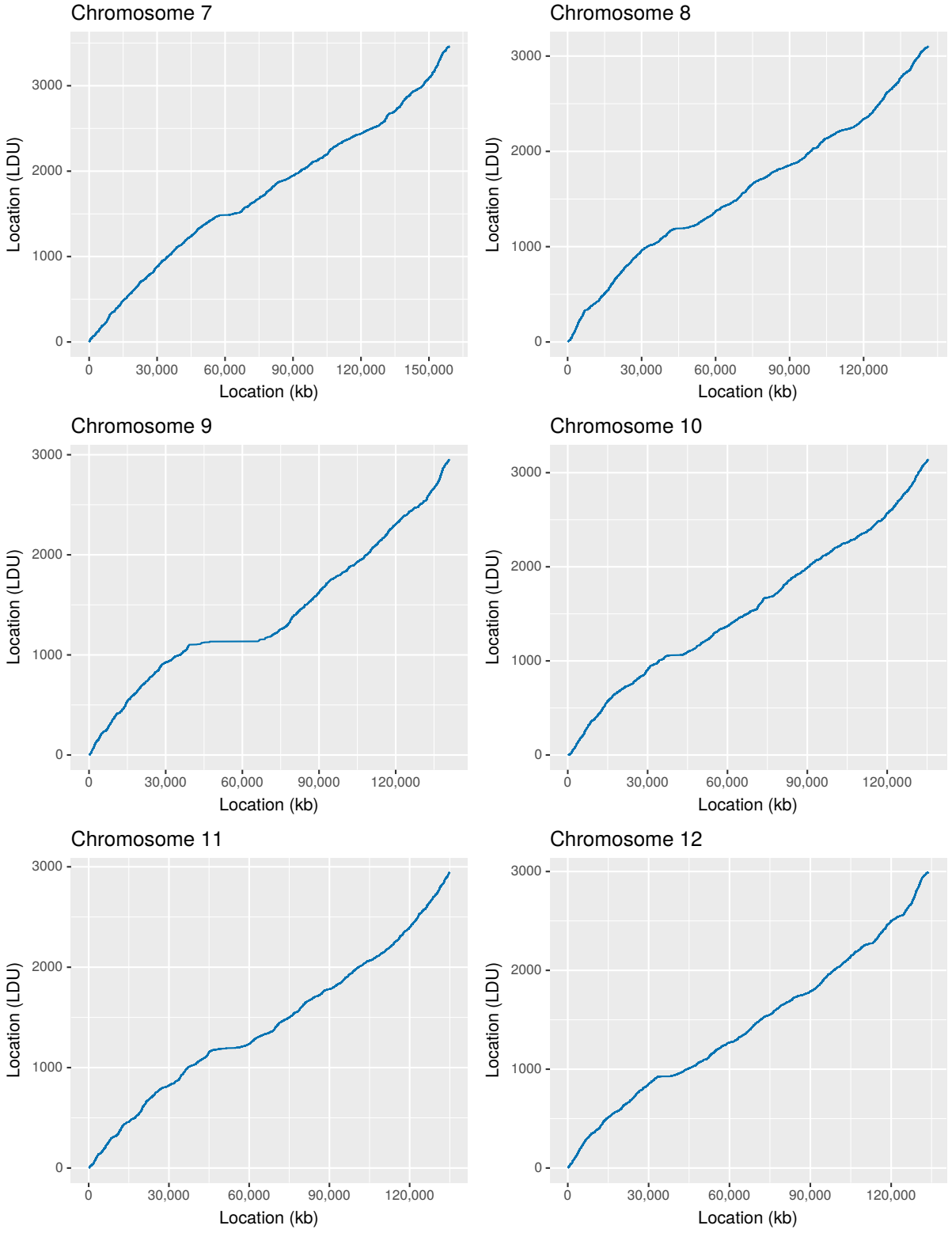


Figure B.2: LDU maps of chromosomes 7 to 12

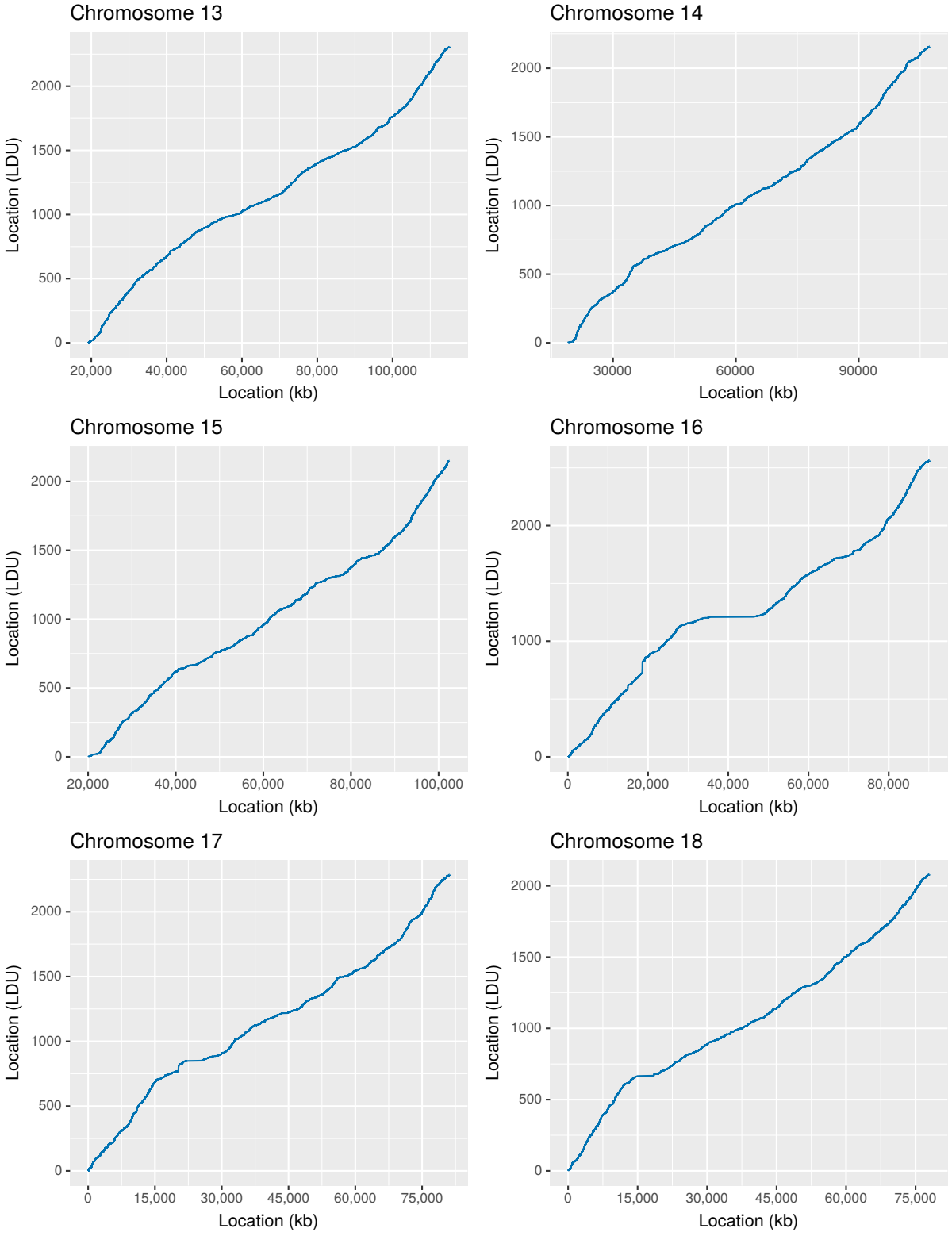


Figure B.3: LDU maps of chromosomes 13 to 18

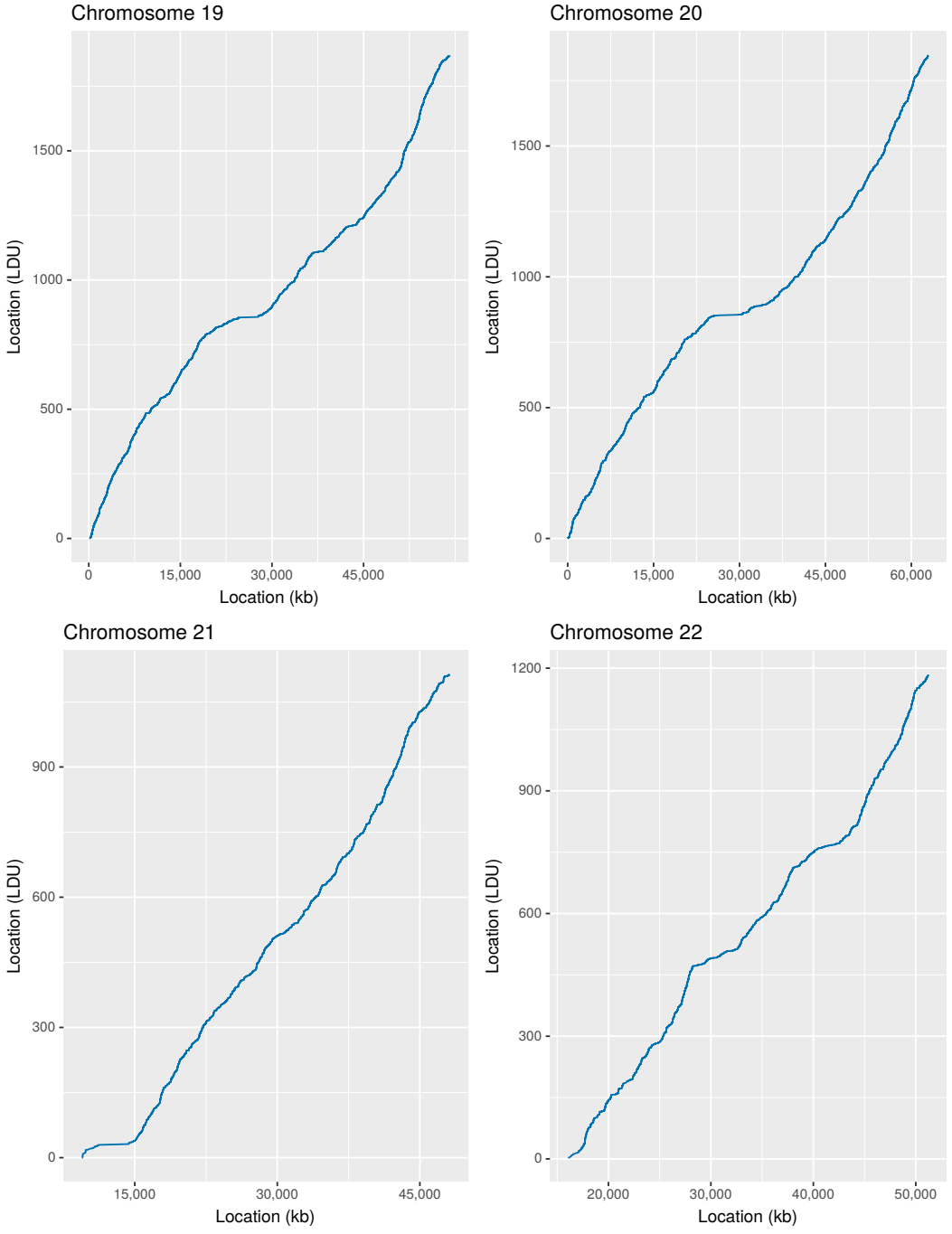


Figure B.4: LDU maps of chromosomes 19 to 22

Table B.1: Number of genome regions in each category by chromosome

Chromosome	**Genic regions**	**Gene Exons**	**Gene Introns**	**Non-coding RNAs**	**Intergenic regions**
1	1,804	21,642	19,493	859	1,803
2	1,084	15,645	14,338	637	1,083
3	944	12,793	11,600	580	943
4	681	8,298	7,494	367	680
5	760	9,308	8,409	481	759
6	916	10,499	9,389	488	915
7	811	9,897	8,932	456	810
8	621	7,277	6,558	385	620
9	691	8,478	7,650	396	690
10	661	8,829	8,037	416	660
11	1,160	11,991	10,591	446	11,59
12	922	12,107	11,026	448	921
13	287	3,768	3,422	306	286
14	539	6,470	5,798	318	538
15	521	7,483	6,844	358	520
16	723	8,926	8,008	370	722
17	1,013	12,386	11,094	467	1,012
18	247	3,263	2,967	167	246
19	1,295	12,439	10,930	457	1,294
20	480	5,285	4,690	259	479
21	193	2,168	1,939	152	192
22	389	4,469	3,980	227	388
Total	16,742	203,421	183,189	9,040	16,720

Table B.2: Physical size of genome regions **Kb!** (**Kb!**) across the autosomal chromosomes

Chromosome	**Whole chromosomes**	**Genic regions**	**Gene exons**	**Gene introns**	**Non-coding RNAs**	**Intergenic**	**Centromeric heterochromatin**
1	249,153.02	106,391.14	6,966.09	95,072.08	31,857.81	119,817.23	22,637.32
2	239,845.03	93,819.86	4,842.43	87,935.59	32,003.58	139,486.84	6,486.72
3	197,820.58	85,036.44	4,092.21	80,299.69	27,636.41	108,435.28	4,060.60
4	191,019.76	64,694.37	2,887.17	61,022.58	17,318.02	122,555.79	3,645.07
5	180,702.67	63,226.23	3,309.01	59,097.75	23,953.80	113,373.81	3,995.81
6	170,771.73	65,599.37	3,553.01	61,101.57	19,085.66	100,231.21	4,770.63
7	159,105.07	70,410.39	3,228.88	65,213.91	19,080.67	82,362.62	5,972.54
8	146,135.37	54,619.05	2,468.18	52,250.19	20,968.91	86,526.18	4,955.14
9	141,040.77	47,100.87	2,740.57	43,702.30	12,844.84	72,221.93	21,580.89
10	135,414.19	61,181.39	2,833.16	57,160.70	15,859.62	69,521.25	4,645.33
11	134,756.10	57,196.76	3,795.30	52,874.20	14,080.35	73,377.12	3,513.45
12	133,755.84	58,557.90	3,692.46	54,191.12	13,161.68	70,549.56	4,529.32
13	95,940.79	31,424.37	1,316.66	29,831.53	11,994.84	63,920.46	-
14	88,238.10	35,089.01	2,161.38	32,560.97	10,400.25	51,529.84	-
15	82,476.10	39,306.01	2,333.19	37,124.44	13,575.37	42,420.39	-
16	90,096.83	34,243.66	2,657.50	31,142.67	8,516.88	42,453.93	13,347.78
17	81,152.95	41,358.90	3,716.92	36,888.72	9,603.98	36,091.02	3,597.06
18	78,004.28	27,429.14	1,185.86	26,293.07	7,693.70	46,740.84	3,676.96
19	59,003.31	28,400.76	3,932.01	24,335.12	6,697.98	25,188.24	5,385.40
20	62,903.17	26,602.88	1,714.88	24,277.99	6,146.94	32,068.45	4,167.93
21	38,604.75	11,626.38	681.67	10,764.19	6,035.17	25,552.46	-
22	35,169.19	17,952.38	1,416.09	16,096.51	5,615.93	17,013.45	-
Total & %coverage	2,791,109.60	1,121,267.26 (40.17)	65,524.61 (2.35)	1,039,236.88 (37.23)	334,132.38 (11.97)	1,541,437.90 (55.23)	120,967.95 (4.33)

Table B.3: LDU size of genome regions across the autosomal chromosomes

Chromosome	Whole chromosomes	Genic regions	Gene exons	Gene introns	Non-coding RNAs	Intergenic	Centromeric heterochromatin
1	5,078.92	2,209.43	137.00	1,991.46	610.26	2,858.10	6.00
2	4,736.82	1,564.72	76.14	1,465.07	561.48	3,145.75	24.72
3	4,138.09	1,721.74	80.80	1,633.80	523.42	2,399.74	9.82
4	3,936.59	1,262.07	53.50	1,186.58	355.96	2,658.18	2.16
5	3,785.07	1,241.02	56.01	1,162.17	460.56	2,537.92	3.00
6	3,604.75	1,310.61	74.46	1,222.90	363.34	2,277.70	10.71
7	3,460.39	1,455.49	71.52	1,346.96	350.78	1,978.58	15.58
8	3,101.18	1,127.98	45.54	1,077.69	395.37	1,957.49	13.11
9	2,953.02	1,073.42	64.78	998.06	266.52	1,857.18	18.00
10	3,140.77	1,310.58	55.02	1,234.98	325.28	1,822.07	6.27
11	2,943.56	1,249.33	78.86	1,158.70	303.48	1,655.63	5.61
12	2,990.16	1,216.67	83.90	1,120.99	253.33	1,761.32	7.16
13	2,309.50	708.45	27.19	674.65	254.68	1,583.06	-
14	2,158.42	713.46	57.02	646.05	205.78	1,413.62	-
15	2,151.48	845.87	44.77	791.93	294.94	1,297.06	-
16	2,562.56	1,054.57	56.80	990.89	250.18	1,487.36	20.10
17	2,287.03	1,041.98	94.77	935.15	268.40	1,235.20	5.50
18	2,079.02	749.03	36.98	710.61	159.85	1,311.23	13.18
19	1,869.27	903.86	116.72	782.73	198.36	924.66	38.60
20	1,846.33	724.34	52.78	665.78	158.68	1,113.93	4.25
21	1,110.60	355.63	15.25	334.36	204.02	728.87	-
22	1,184.17	457.55	33.76	406.28	134.50	721.05	-
Totals /% coverage	63427.68	24297.76 (38.31%)	1413.55 (2.23%)	22537.79 (35.53%)	6899.19 (10.88%)	38725.67 (61.05%)	203.77 (0.32%)

Table B.4: Extent of LD in Kb (Kb/LDU) for genome regions across the autosomal chromosomes

Chromosome	*Whole chromosomes	Genic	Gene exons	Gene introns	Non-coding RNAs	**Intergenic	Centromeric	Gene exons + Non-coding RNAs
1	49.06	48.15	50.85	47.74	52.20	41.92	3772.89	51.96
2	50.63	59.96	63.60	60.02	57.00	44.34	262.41	57.79
3	47.80	49.39	50.65	49.15	52.80	45.19	413.5	52.51
4	48.52	51.26	53.97	51.43	48.65	46.11	1687.53	49.35
5	47.74	50.95	59.08	50.85	52.01	44.67	1331.94	52.78
6	47.37	50.05	47.72	49.96	52.53	44.01	445.44	52.83
7	45.98	48.38	45.15	48.42	54.39	41.63	383.35	52.83
8	47.12	48.42	54.20	48.48	53.04	44.20	377.97	53.16
9	47.76	43.88	42.31	43.79	48.19	38.89	1198.94	47.04
10	43.11	46.68	51.49	46.28	48.76	38.16	740.88	49.15
11	45.78	45.78	48.13	45.63	46.40	44.32	626.28	46.75
12	44.73	48.13	44.01	48.34	51.95	40.05	632.59	49.98
13	41.54	44.36	48.42	44.22	47.10	40.38	-	47.23
14	40.88	49.18	37.91	50.40	50.54	36.45	-	47.80
15	38.33	46.47	52.12	46.88	46.03	32.71	-	46.83
16	35.16	32.47	46.79	31.43	34.04	28.54	664.07	36.4
17	35.48	39.69	39.22	39.45	35.78	29.22	654.01	36.68
18	37.52	36.62	32.07	37	48.13	35.65	278.98	45.11
19	31.56	31.42	33.69	31.09	33.77	27.24	139.52	33.74
20	34.07	36.73	32.49	36.47	38.74	28.79	980.69	37.18
21	34.76	32.69	44.7	32.19	29.58	35.060	-	30.63
22	29.70	39.24	41.95	39.62	41.75	23.6	-	41.79
Chromosome means / SD	42.03 / 6.40	44.54 / 7.23	46.39 / 8.20	44.49 / 7.42	46.52 / 7.61	37.78 / 6.81	858.29	46.34 / 7.27

*Includes centromeric regions
** Excludes centromeric regions

References

[1] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. Available from: https://doi.org/10.1038/nature11247.

[2] Strachan T, Read AP. Human molecular genetics 4th Edition. Garland Science; 2018.

[3] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2018 10;47(D1):D766–D773. Available from: https://doi.org/10.1093/nar/gky955.

[4] Gilbert W. Why genes in pieces? Nature. 1978;271(5645):501. Available from: https://doi.org/10.1038/271501a0.

[5] Saberkari H, Shamsi M, Heravi H, Sedaaghi MH. A fast algorithm for exonic regions prediction in DNA sequences. Journal of medical signals and sensors. 2013 jul;3(3):139–149.

[6] Liu Y, Wei X, Kong X, Guo X, Sun Y, Man J, et al. Targeted next-generation sequencing for clinical diagnosis of 561 Mendelian diseases. PLoS ONE. 2015.

[7] Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al.. Exome sequencing as a tool for Mendelian disease gene discovery; 2011.

[8] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011 may;43(5):491–498.

[9] Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2018 10;47(D1):D766–D773. Available from: https://doi.org/10.1093/nar/gky955.

[10] Haworth A, Savage H, Lench N. Chapter 4 - Diagnostic Genomics and Clinical Bioinformatics. In: Kumar D, Antonarakis S, editors. Medical and Health Genomics. Oxford: Academic Press; 2016. p. 37–50. Available from: https://www.sciencedirect.com/science/article/pii/B9780124201965000046.

[11] Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genetics in Medicine. 2018;20(10):1122–1130. Available from: https://doi.org/10.1038/gim.2017.247.

[12] Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. New England Journal of Medicine. 2013;369(16):1502–1511. PMID: 24088041. Available from: https://doi.org/10.1056/NEJMoa1306555.

[13] Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nature genetics. 2015 jul;47(7):717–726.

[14] Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences. 2015;112(17):5473–5478. Available from: https://www.pnas.org/content/112/17/5473.

[15] Zhao S, Jing W, Samuels DC, Sheng Q, Shyr Y, Guo Y. Strategies for processing and quality control of Illumina genotyping arrays. Briefings in Bioinformatics. 2017 02;19(5):765–775. Available from: https://doi.org/10.1093/bib/bbx012.

[16] Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLOS Computational Biology. 2012 12;8(12):1–11. Available from: https://doi.org/10.1371/journal.pcbi.1002822.

[17] Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, et al. Genome-Wide Association Study in East Asians Identifies Novel Susceptibility Loci for Breast Cancer. PLOS Genetics. 2012 02;8(2):1–10. Available from: https://doi.org/10.1371/journal.pgen.1002532.

[18] Mullighan CG, Goorha S, Radtke I, Miller CB, Coustan-Smith E, Dalton JD, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature. 2007 April;446(7137):758—764. Available from: https://doi.org/10.1038/nature05690.

[19] OMIM. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine;. Available from: https://www.omim.org/statistics/entry.

[20] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009.

[21] Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. Nature Reviews Genetics. 2006;7(4):277–282. Available from: https://doi.org/10.1038/nrg1826.

[22] Johanna Craig. Complex Diseases: Research and Applications. Nature Education. 2008;1(1):184. Available from: https://www.nature.com/scitable/topicpage/complex-diseases-research-and-applications-748/.

[23] Heidi Chial. Rare Genetic Disorders: Learning About Genetic Disease Through Gene Mapping, SNPs, and Microarray Data. Nature Education. 2008;1(1):192. Available from: https://www.nature.com/scitable/topicpage/rare-genetic-disorders-learning-about-genetic-disease-979/.

[24] Angural A, Spolia A, Mahajan A, Verma V, Sharma A, Kumar P, et al. Review: Understanding Rare Genetic Diseases in Low Resource Regions Like Jammu and Kashmir – India. Frontiers in Genetics. 2020;11:415. Available from: https://www.frontiersin.org/article/10.3389/fgene.2020.00415.

[25] Jamuar SS, Tan EC. Clinical application of next-generation sequencing for Mendelian diseases. Human genomics. 2015 jun;9(1):10.

[26] Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. Briefings in Functional Genomics. 2015 dec;15(5):374–384. Available from: https://doi.org/10.1093/bfgp/elv054.

[27] Collins A. The genomic and functional characteristics of disease genes. Briefings in Bioinformatics. 2015 jan;16(1):16–23. Available from: https://doi.org/10.1093/bib/bbt091.

[28] McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010 apr;141(2):210–217.

[29] Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. Trends in genetics : TIG. 2003 mar;19(3):135–140.

[30] Shawky RM. Reduced penetrance in human inherited disease. Egyptian Journal of Medical Human Genetics. 2014;15(2):103–111. Available from: http://www.sciencedirect.com/science/article/pii/S1110863014000184.

[31] Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. Cell. 2011 sep;147(1):57–69. Available from: https://pubmed.ncbi.nlm.nih.gov/21962507https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198013/.

[32] Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009 jun;106(23):9362–9367.

[33] Casillas S, Barbadilla A. Molecular Population Genetics. Genetics. 2017 mar;205(3):1003–1035. Available from: https://pubmed.ncbi.nlm.nih.gov/28270526https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5340319/.

[34] Tsumura Y, Matsumoto A, Tani N, Ujino-Ihara T, Kado T, Iwata H, et al. Genetic diversity and the genetic structure of natural populations of Chamaecyparis obtusa: implications for management and conservation. Heredity. 2007;99(2):161–172. Available from: https://doi.org/10.1038/sj.hdy.6800978.

[35] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68–74.

[36] Cox ME, Campbell JK, Langefeld CD. An exploration of sex-specific linkage disequilibrium on chromosome X in Caucasians from the COGA study. BMC Genetics. 2005;6(1):S81. Available from: https://doi.org/10.1186/1471-2156-6-S1-S81.

[37] Collins AR. Linkage Disequilibrium and Association Mapping. Humana Press; 2007.

[38] Laan M, Wiebe V, Khusnutdinova E, Remm M, Pääbo S. X-chromosome as a marker for population history: linkage disequilibrium and haplotype study in Eurasian populations. European journal of human genetics : EJHG. 2005 apr;13(4):452–462. Available from: https://pubmed.ncbi.nlm.nih.gov/15657606https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1450114/.

[39] Slatkin M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nature reviews Genetics. 2008 jun;9(6):477–485.

[40] Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genetical research. 2007 dec;89(5-6):391–403.

[41] Tapper WJ, Maniatis N, Morton NE, Collins A. A metric linkage disequilibrium map of a human chromosome. Annals of human genetics. 2003 nov;67(Pt 6):487–494.

[42] Horscroft C, Ennis S, Pengelly RJ, Sluckin TJ, Collins A. Sequencing era methods for identifying signatures of selection in the genome. Briefings in Bioinformatics. 2018 jul;20(6):1997–2008. Available from: https://doi.org/10.1093/bib/bby064.

[43] Myers S, Freeman C, Auton A, Donnelly P, McVean G. A common sequence motif associated with recombination hot spots and genome instability in humans. Nature genetics. 2008 sep;40(9):1124–1129.

[44] Jeffreys AJ, Neumann R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nature genetics. 2002 jul;31(3):267–271.

[45] Hayashi K, Yoshida K, Matsui Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature. 2005 nov;438(7066):374–378.

[46] Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012;485(7400):642–645. Available from: https://doi.org/10.1038/nature11089.

[47] Wall JD, Stevison LS. Detecting Recombination Hotspots from Patterns of Linkage Disequilibrium. G3 (Bethesda, Md). 2016 aug;6(8):2265–2271.

[48] Jacobs GS, Sluckin TJ, Kivisild T. Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps. Genetics. 2016 aug;203(4):1807–1825.

[49] Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE. A map of the human genome in linkage disequilibrium units. Proceedings of the National Academy of Sciences of the United States of America. 2005 aug;102(33):11835–11839. Available from: https://pubmed.ncbi.nlm.nih.gov/16091463https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1188000/.

[50] Pengelly RJ, Tapper W, Gibson J, Knut M, Tearle R, Collins A, et al. Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. BMC genomics. 2015 sep;16(1):666.

[51] Gibson J, Tapper W, Ennis S, Collins A. Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. Human genetics. 2013 feb;132(2):233–243.

[52] Spataro N, Rodríguez JA, Navarro A, Bosch E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Human molecular genetics. 2017 feb;26(3):489–500. Available from: https://pubmed.ncbi.nlm.nih.gov/28053046https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409085/.

[53] Pengelly RJ, Vergara-Lope A, Alyousfi D, Jabalameli MR, Collins A. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. Briefings in Bioinformatics. 2017 aug;20(1):267–273. Available from: https://doi.org/10.1093/bib/bbx110.

[54] McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science (New York, NY). 2004 apr;304(5670):581–584.

[55] Eberle MA, Rieder MJ, Kruglyak L, Nickerson DA. Allele Frequency Matching Between SNPs Reveals an Excess of Linkage Disequilibrium in Genic Regions of the Human Genome. PLOS Genetics. 2006 sep;2(9):e142. Available from: https://doi.org/10.1371/journal.pgen.0020142.

[56] Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature. 2010 oct;467(7319):1099–1103.

[57] Berger S, Schlather M, de los Campos G, Weigend S, Preisinger R, Erbe M, et al. A Scale-Corrected Comparison of Linkage Disequilibrium Levels between Genic and Non-Genic Regions. PLOS ONE. 2015 oct;10(10):e0141216. Available from: https://doi.org/10.1371/journal.pone.0141216.

[58] Bruce A. Molecular biology of the cell. 4th edition. In: General Recombination. New York: Garland Science; 2002. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26898/.

[59] Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974 oct;78(2):737–756.

[60] Webster MT, Hurst LD. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends in genetics : TIG. 2012 mar;28(3):101–109.

[61] Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked Sites. Genetics. 2012 jan;190(1):5 LP – 22. Available from: http://www.genetics.org/content/190/1/5.abstract.

[62] Hussin JG, Hodgkinson A, Idaghdour Y, Grenier JC, Goulet JP, Gbeha E, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nature genetics. 2015 apr;47(4):400–404.

[63] Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St Jean PL, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome research. 2013 dec;23(12):1974–1984. Available from: https://pubmed.ncbi.nlm.nih.gov/23990608https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3847768/.

[64] Jeff Hardin, Wayne M Becker, Lewis J Kleinsmith JH. Becker's World of the Cell 9th. Pearson Education; 2016.

[65] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetical research. 2007 dec;89(5-6):311–336.

[66] Eyre-Walker YC, Eyre-Walker A. The Role of Mutation Rate Variation and Genetic Diversity in the Architecture of Human Disease. PLOS ONE. 2014 feb;9(2):e90166. Available from: https://doi.org/10.1371/journal.pone.0090166.

[67] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013 jul;499(7457):214–218.

[68] Lodish H. Molecular Cell Biology 8th ed.; 2016.

[69] Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nature reviews Genetics. 2011 oct;12(11):756–766.

[70] Lipson M, Loh PR, Sankararaman S, Patterson N, Berger B, Reich D. Calibrating the Human Mutation Rate via Ancestral Recombination Density in Diploid Genomes. PLOS Genetics. 2015 nov;11(11):e1005550. Available from: https://doi.org/10.1371/journal.pgen.1005550.

[71] Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nature Genetics. 2016;48(4):349–355. Available from: https://doi.org/10.1038/ng.3511.

[72] Jiang P, Wang H, Li W, Zang C, Li B, Wong YJ, et al. Network analysis of gene essentiality in functional genomics experiments. Genome Biology. 2015;16(1):239. Available from: https://doi.org/10.1186/s13059-015-0808-9.

[73] Kondrashov FA, Ogurtsov AY, Kondrashov AS. Bioinformatical assay of human gene morbidity. Nucleic acids research. 2004;32(5):1731–1737.

[74] Abu-Mostafa YS, Magdon-Ismail M, Lin HT. Learning From Data. AMLBook; 2012.

[75] Zakeri P, Simm J, Arany A, ElShal S, Moreau Y. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics (Oxford, England). 2018 jul;34(13):i447–i456. Available from: https://pubmed.ncbi.nlm.nih.gov/29949967https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6022676/.

[76] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nature biotechnology. 2006 may;24(5):537–544.

[77] Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. BMC bioinformatics. 2005 mar;6:55.

[78] Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics (Oxford, England). 2006 nov;22(22):2800–2805.

[79] Radivojac P, Peng K, Clark WT, Peters BJ, Mohan A, Boyle SM, et al. An integrated approach to inferring gene-disease associations in humans. Proteins. 2008 aug;72(3):1030–1037. Available from: https://pubmed.ncbi.nlm.nih.gov/18300252https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824611/.

[80] Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics. 2012;44(4):369–375. Available from: https://doi.org/10.1038/ng.2213.

[81] Everitt B, Landau S, Leese M, Stahl D. Cluster Analysis; 2011.

[82] Lu H, Plataniotis K, Venetsanopoulos AN. Multilinear Subspace Learning: Dimensionality Reduction of Multidimensional Data, Chapman, Hall, Press Machine, C.R.C. Learning and Pattern Recognition Series. 2013 jan.

[83] Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, et al. Clustering Algorithms: Their Application to Gene Expression Data. Bioinformatics and biology insights. 2016 nov;10:237–253. Available from: https://pubmed.ncbi.nlm.nih.gov/27932867https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5135122/.

[84] Lopez C, Tucker S, Salameh T, Tucker C. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. Journal of Biomedical Informatics. 2018;85:30–39. Available from: http://www.sciencedirect.com/science/article/pii/S1532046418301308.

[85] Karmakar B, Das S, Bhattacharya S, Sarkar R, Mukhopadhyay I. Tight clustering for large datasets with an application to gene expression data. Scientific Reports. 2019;9(1):3053. Available from: https://doi.org/10.1038/s41598-019-39459-w.

[86] Christy A, Gandhi GM, Vaithyasubramanian S. Cluster Based Outlier Detection Algorithm for Healthcare Data. Procedia Computer Science. 2015;50:209–215. Available from: http://www.sciencedirect.com/science/article/pii/S1877050915005591.

[87] de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics. 2008;9(1):497. Available from: https://doi.org/10.1186/1471-2105-9-497.

[88] Peterson CB, Stingo FC, Vannucci M. Bayesian Inference of Multiple Gaussian Graphical Models. Journal of the American Statistical Association. 2015 mar;110(509):159–174. Available from: https://pubmed.ncbi.nlm.nih.gov/26078481https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465207/.

[89] Verzilli C, Stallard N, Whittaker J. Bayesian Graphical Models for Genomewide Association Studies. American journal of human genetics. 2006 jul;79:100–112.

[90] Williams D, Piironen J, Vehtari A, Rast P. Bayesian Estimation of Gaussian Graphical Models with Projection Predictive Selection. 2018 jan.

[91] Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, et al. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes & Diseases. 2018;5(2):77–106. Available from: http://www.sciencedirect.com/science/article/pii/S2352304218300680.

[92] Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Human Genetics. 2013;132(10):1077–1130. Available from: https://doi.org/10.1007/s00439-013-1331-2.

[93] Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nature reviews Cancer. 2011 dec;12(1):68–78. Available from: https://pubmed.ncbi.nlm.nih.gov/22193408https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4972490/.

[94] Skol AD, Sasaki MM, Onel K. The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. Breast Cancer Research. 2016;18(1):99. Available from: https://doi.org/10.1186/s13058-016-0759-4.

[95] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 oct;461(7265):747–753. Available from: https://pubmed.ncbi.nlm.nih.gov/19812666https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831613/.

[96] Smyth C, Špakulová I, Cotton-Barratt O, Rafiq S, Tapper W, Upstill-Goddard R, et al. Quantifying the cumulative effect of low-penetrance genetic variants on

breast cancer risk. Molecular genetics & genomic medicine. 2015 may;3(3):182–188. Available from: https://pubmed.ncbi.nlm.nih.gov/26029704https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4444159/.

[97] Ghoussaini M, Pharoah PDP. Polygenic susceptibility to breast cancer: current state-of-the-art. Future oncology (London, England). 2009 jun;5(5):689–701. Available from: https://pubmed.ncbi.nlm.nih.gov/19519208https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4931895/.

[98] Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nature reviews Genetics. 2016 jul;17(7):392–406. Available from: https://pubmed.ncbi.nlm.nih.gov/27140283https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6021129/.

[99] Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. Journal of the National Cancer Institute. 2015 may;107(5).

[100] Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017 nov;551(7678):92–94. Available from: https://pubmed.ncbi.nlm.nih.gov/29059683https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5798588/.

[101] Sawyer S, Mitchell G, McKinley J, Chenevix-Trench G, Beesley J, Chen XQ, et al. A role for common genomic variants in the assessment of familial breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2012 dec;30(35):4330–4336.

[102] Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic acids research. 2016 jan;44(D1):D20–6.

[103] Maniatis N, Collins A, Xu CF, McCarthy L, Hewett DR, Tapper W, et al. The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. Proceedings of the National Academy of Sciences of the United States of America. 2002 mar;99:2228–2233.

[104] Zhang W, Collins A, Maniatis N, Tapper W, Morton NE. Properties of linkage disequilibrium (LD) maps. Proceedings of the National Academy of Sciences of the United States of America. 2002 dec;99(26):17004–17007. Available from: https://pubmed.ncbi.nlm.nih.gov/12486239https://www.ncbi.nlm.nih.gov/pmc/articles/PMC139259/.

[105] Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nature Genetics. 2006;38(5):556–560. Available from: https://doi.org/10.1038/ng1770.

[106] Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, et al. Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proceedings of the National Academy of Sciences of the United States of America. 2004 dec;101(52):18075–18080.

[107] Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. Psychological medicine. 2016 sep;46(12):2455–2465.

[108] Asif M, Martiniano HFMCM, Vicente AM, Couto FM. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PLOS ONE. 2018 dec;13(12):e0208626. Available from: https://doi.org/10.1371/journal.pone.0208626.

[109] Breiman L. Random Forests. Machine Learning. 2001;45(1):5–32. Available from: https://doi.org/10.1023/A:1010933404324.

[110] Wolpert DH. Stacked generalization. Neural Networks. 1992;5(2):241–259. Available from: http://www.sciencedirect.com/science/article/pii/S0893608005800231.

[111] Berkhin P. In: Kogan J, Nicholas C, Teboulle M, editors. A Survey of Clustering Data Mining Techniques. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 25–71. Available from: https://doi.org/10.1007/3-540-28349-8_2.

[112] Goldberg DE, Holland JH. Genetic Algorithms and Machine Learning. Machine Learning. 1988;3(2):95–99. Available from: https://doi.org/10.1023/A:1022602019183.

[113] Hawkins DM. The Problem of Overfitting. Journal of Chemical Information and Computer Sciences. 2004 01;44(1):1–12. Available from: https://doi.org/10.1021/ci0342472.

[114] Jain AK, Murty MN, Flynn PJ. Data Clustering: A Review. ACM Comput Surv. 1999 Sep;31(3):264–323. Available from: https://doi.org/10.1145/331499.331504.

[115] Refaeilzadeh P, Tang L, Liu H. Cross-Validation BT - Encyclopedia of Database Systems. Boston, MA: Springer US; 2009. p. 532–538. Available from: https://doi.org/10.1007/978-0-387-39940-9{_}565.

[116] Lahmiri S, Dawson DA, Shmuel A. Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures. Biomedical engineering letters. 2018 feb;8(1):29–39.

[117] Pedrycz W, Chen SM. Data Science and Big Data: An Environment of Computational Intelligence. Springer, Cham; 2017.

[118] Jaynes ET. Probability Theory: The Logic of Science. Cambridge: Cambridge University Press; 2003. Available from: https://www.cambridge.org/core/books/probability-theory/9CA08E224FF30123304E6D8935CF1A99.

[119] Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. Clustering algorithms: A comparative approach. PLOS ONE. 2019 01;14(1):1–34. Available from: https://doi.org/10.1371/journal.pone.0210236.

[120] Sarkar S, Melnykov V, Zheng R. Gaussian mixture modeling and model-based clustering under measurement inconsistency. Advances in Data Analysis and Classification. 2020. Available from: https://doi.org/10.1007/s11634-020-00393-9.

[121] Ficklin SP, Dunwoodie LJ, Poehlman WL, Watson C, Roche KE, Feltus FA. Discovering Condition-Specific Gene Co-Expression Patterns Using Gaussian Mixture Models: A Cancer Case Study. Scientific Reports. 2017;7(1):8617. Available from: https://doi.org/10.1038/s41598-017-09094-4.

[122] Patel E, Kushwaha DS. Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model. Procedia Computer Science. 2020;171:158–167. Third International Conference on Computing and Network Communications (CoCoNet'19). Available from: https://www.sciencedirect.com/science/article/pii/S1877050920309820.

[123] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical transactions Series A, Mathematical, physical, and engineering sciences. 2016 apr;374(2065):20150202. Available from: https://pubmed.ncbi.nlm.nih.gov/26953178https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792409/.

[124] Vanhatalo E, Kulahci M, Bergquist B. On the structure of dynamic principal component analysis used in statistical process monitoring. Chemometrics and Intelligent Laboratory Systems. 2017;167:1–11. Available from: http://www.sciencedirect.com/science/article/pii/S0169743917300734.

[125] Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nature Communications. 2019;10(1):5416. Available from: https://doi.org/10.1038/s41467-019-13056-x.

[126] Zhou H, Wang F, Tao P. t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. Journal of chemical theory and computation. 2018 nov;14(11):5499–5510. Available from: https://pubmed.ncbi.nlm.nih.gov/30252473https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6679899/.

[127] Gaspar HA, Breen G. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. BMC Bioinformatics. 2019;20(1):116. Available from: https://doi.org/10.1186/s12859-019-2680-1.

[128] Dobra A, Lenkoski A, Rodriguez A. Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data. Journal of the American Statistical Association. 2011;106(496):1418–1433. Available from: https://pubmed.ncbi.nlm.nih.gov/26924867https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4767185/.

[129] Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004;90(1):196–212. Available from: http://www.sciencedirect.com/science/article/pii/S0047259X04000259.

[130] Dempster AP. Covariance Selection. Biometrics. 1972 jun;28(1):157–175. Available from: http://www.jstor.org/stable/2528966.

[131] Atay-Kayis A, Massam H. A Monte Carlo Method for Computing the Marginal Likelihood in Nondecomposable Gaussian Graphical Models. Biometrika. 2005 jun;92(2):317–335. Available from: http://www.jstor.org/stable/20441191.

[132] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis; 2013. Available from: https://books.google.co.uk/books?id=ZXL6AQAAQBAJ.

[133] Mitsakakis N, Massam H, D Escobar M. A Metropolis-Hastings based method for sampling from the G -Wishart distribution in Gaussian graphical models. Electron J Statist. 2011;5:18–30. Available from: https://projecteuclid.org:443/euclid.ejs/1295457468.

[134] Williams D, Rast P, Pericchi L, Mulder J. Comparing Gaussian graphical models with the posterior predictive distribution and Bayesian model selection. Psychological Methods. 2020 feb.

[135] Kravchenko-Balasha N, Simon S, Levine RD, Remacle F, Exman I. Computational surprisal analysis speeds-up genomic characterization of cancer processes. PloS one. 2014 nov;9(11):e108549–e108549. Available from: https://pubmed.ncbi.nlm.nih.gov/25405334https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236016/.

[136] Cover TM, Thomas JA. Elements of Information Theory. Wiley; 2012. Available from: https://books.google.co.uk/books?id=VWq5GG6ycxMC.

[137] Kernighan BW, Ritchie DM. The C programming language; 2006.

[138] G van Rossum. Python Software Foundation. Python Language Reference; 1995.

[139] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2012 jan;12.

[140] Oliphant T. SciPy: Open source scientific tools for Python. 2007 jan;9:10–20.

[141] McKinney W, et al. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. vol. 445. Austin, TX; 2010. p. 51–56.

[142] R Core Team. R: A Language and Environment for Statistical Computing; 2020. Available from: http://www.R-project.org/.

[143] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of

human genetics. 2007 sep;81(3):559–575. Available from: https://pubmed.ncbi.nlm.nih.gov/17701901https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/.

[144] Wall L, Christiansen T, Orwant J. Programming perl. " O'Reilly Media, Inc."; 2000.

[145] Jakkula E, Rehnström K, Varilo T, Pietiläinen OPH, Paunio T, Pedersen NL, et al. The genome-wide patterns of variation expose significant substructure in a founder population. American journal of human genetics. 2008 dec;83(6):787–794.

[146] Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature genetics. 2001 oct;29(2):217–222.

[147] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science (New York, NY). 2002 jun;296(5576):2225–2229.

[148] Lau W, Kuo TY, Tapper W, Cox S, Collins A. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics (Oxford, England). 2007 feb;23(4):517–519.

[149] Erikson GA, Bodian DL, Rueda M, Molparia B, Scott ER, Scott-Van Zeeland AA, et al. Whole-Genome Sequencing of a Healthy Aging Cohort. Cell. 2016 may;165(4):1002–1011.

[150] Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2015 update. Nucleic acids research. 2015 jan;43(Database issue):D670–D681. Available from: https://pubmed.ncbi.nlm.nih.gov/25428374https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383971/.

[151] Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. American journal of human genetics. 2000 jan;66(1):216–234. Available from: https://pubmed.ncbi.nlm.nih.gov/10631153https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1288328/.

[152] Toosi A, Fernando RL, Dekkers JCM. Genomic selection in admixed and crossbred populations. Journal of animal science. 2010 jan;88(1):32–46.

[153] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics (Oxford, England). 2011 aug;27(15):2156–2158. Available from: https://pubmed.ncbi.nlm.nih.gov/21653522https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/.

[154] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2015 11;44(D1):D733–D745. Available from: https://doi.org/10.1093/nar/gkv1189.

[155] Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. Proceedings of the National Academy of Sciences of the United States of America. 1999 12;96(26):15173–15177. Available from: https://pubmed.ncbi.nlm.nih.gov/10611357.

[156] Payseur BA, Nachman MW. Gene Density and Human Nucleotide Polymorphism. Molecular Biology and Evolution. 2002 03;19(3):336–340. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a004086.

[157] Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nature genetics. 2002 jul;31(3):241–247.

[158] Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M. A neutral explanation for the correlation of diversity with recombination rates in humans. American journal of human genetics. 2003 jun;72(6):1527–1535. Available from: https://pubmed.ncbi.nlm.nih.gov/12740762https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1180312/.

[159] Sun P, Zhang R, Jiang Y, Wang X, Li J, Lv H, et al. Assessing the patterns of linkage disequilibrium in genic regions of the human genome. The FEBS journal. 2011 oct;278(19):3748–3755.

[160] Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nature reviews Genetics. 2018 jan;19(1):51–62.

[161] Zhu L, Zhang Y, Zhang W, Yang S, Chen JQ, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC genomics. 2009 jan;10:47.

[162] Quinn JJ, Zhang QC, Georgiev P, Ilik IA, Akhtar A, Chang HY. Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. Genes & development. 2016 jan;30(2):191–207.

[163] Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. American journal of human genetics. 2015 aug;97(2):199–215. Available from: https://pubmed.ncbi.nlm.nih.gov/26166479https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4573249/.

[164] Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science translational medicine. 2017 apr;9(386).

[165] Reich DE, Lander ES. On the allelic spectrum of human disease. Trends in genetics : TIG. 2001 sep;17(9):502–510.

[166] Rao AR, Nelson SF. Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. BMC Medical Genomics. 2018;11(1):53. Available from: https://doi.org/10.1186/s12920-018-0371-9.

[167] Alyousfi D, Baralle D, Collins A. Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review. Briefings in Functional Genomics. 2018 oct;18(1):23–29. Available from: https://doi.org/10.1093/bfgp/ely033.

[168] Xu D, Gokcumen O, Khurana E. Loss-of-function tolerance of enhancers in the human genome. PLOS Genetics. 2020 apr;16(4):e1008663. Available from: https://doi.org/10.1371/journal.pgen.1008663.

[169] Cacheiro P, Muñoz-Fuentes V, Murray SA, Dickinson ME, Bucan M, Nutter LMJ, et al. Human and mouse essentiality screens as a resource for disease gene discovery. Nature Communications. 2020;11(1):655. Available from: https://doi.org/10.1038/s41467-020-14284-2.

[170] Chen H, Zhang Z, Jiang S, Li R, Li W, Zhao C, et al. New insights on human essential genes based on integrated analysis and the construction of the HEGIAP web-based platform. Briefings in Bioinformatics. 2019 aug. Available from: https://doi.org/10.1093/bib/bbz072.

[171] Kim I, Lee H, Lee K, Han SK, Kim D, Kim S. Link clustering explains non-central and contextually essential genes in protein interaction networks. Scientific Reports. 2019;9(1):11672. Available from: https://doi.org/10.1038/s41598-019-48273-3.

[172] Dickerson JE, Zhu A, Robertson DL, Hentges KE. Defining the role of essential genes in human disease. PloS one. 2011;6(11):e27368–e27368. Available from: https://pubmed.ncbi.nlm.nih.gov/22096564https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3214036/.

[173] Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genetics. 2013.

[174] Luo H, Gao F, Lin Y. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. Scientific Reports. 2015;5(1):13210. Available from: https://doi.org/10.1038/srep13210.

[175] Li J, Wang HT, Wang WT, Zhang XR, Suo F, Ren JY, et al. Systematic analysis reveals the prevalence and principles of bypassable gene essentiality. Nature Communications. 2019;10(1):1002. Available from: https://doi.org/10.1038/s41467-019-08928-1.

[176] Derks MFL, Gjuvsland AB, Bosse M, Lopes MS, van Son M, Harlizius B, et al. Loss of function mutations in essential genes cause embryonic lethality in pigs. PLOS Genetics. 2019 mar;15(3):e1008055. Available from: https://doi.org/10.1371/journal.pgen.1008055.

[177] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 aug;536(7616):285–291.

[178] Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nature genetics. 2014 sep;46(9):944–950. Available from: https://pubmed.ncbi.nlm.nih.gov/25086666https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4222185/.

[179] Ohnuki S, Ohya Y. High-dimensional single-cell phenotyping reveals extensive haploinsufficiency. PLoS biology. 2018 may;16(5):e2005130–e2005130. Available from: https://pubmed.ncbi.nlm.nih.gov/29768403https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5955526/.

[180] Hsu JS, Kwan JSH, Pan Z, Garcia-Barcelo MM, Sham PC, Li M. Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. Bioinformatics. 2016 jun;32(20):3065–3071. Available from: https://doi.org/10.1093/bioinformatics/btw381.

[181] Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS computational biology. 2013;9(3):e1002886–e1002886. Available from: https://pubmed.ncbi.nlm.nih.gov/23505346https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3591262/.

[182] Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic acids research. 2015 sep;43(15):e101.

[183] Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS genetics. 2010 oct;6(10):e1001154–e1001154. Available from: https://pubmed.ncbi.nlm.nih.gov/20976243https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2954820/.

[184] MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science (New York, NY). 2012 feb;335(6070):823–828.

[185] Templeton AR. Detecting Selection Through Its Interactions With Other Evolutionary Forces. San Diego: Academic Press; 2019. p. 303–337. Available from: http://www.sciencedirect.com/science/article/pii/B9780123860255000105.

[186] Sampson MG, Gillies CE, Ju W, Kretzler M, Kang HM. Gene-level Integrated Metric of negative Selection (GIMS) Prioritizes Candidate Genes for Nephrotic Syndrome. PLOS ONE. 2013 nov;8(11):e81062. Available from: https://doi.org/10.1371/journal.pone.0081062.

[187] Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. Proceedings of the National Academy of Sciences. 2015 nov;112(44):13615 LP – 13620. Available from: http://www.pnas.org/content/112/44/13615.abstract.

[188] Štorchová H, Stone JD, Sloan DB, Abeyawardana OAJ, Müller K, Walterová J, et al. Homologous recombination changes the context of Cytochrome b transcription in the mitochondrial genome of Silene vulgaris KRA. BMC Genomics. 2018;19(1):874. Available from: https://doi.org/10.1186/s12864-018-5254-0.

[189] Pál C, Hurst LD. Evidence for co-evolution of gene order and recombination rate. Nature Genetics. 2003;33(3):392–395. Available from: https://doi.org/10.1038/ng1111.

[190] Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome research. 2007 aug;17(8):1219–1227. Available from: https://pubmed.ncbi.nlm.nih.gov/17623807https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933511/.

[191] Wang Y, Zhao LP, Dudoit S. A Fine-Scale Linkage-Disequilibrium Measure Based on Length of Haplotype Sharing. The American Journal of Human Genetics. 2006;78(4):615–628. Available from: http://www.sciencedirect.com/science/article/pii/S0002929707637008.

[192] Vergara-Lope A, Ennis S, Vorechovsky I, Pengelly RJ, Collins A. Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding RNA and intergenic chromosome regions. European Journal of Human Genetics. 2019;27(9):1436–1444. Available from: https://doi.org/10.1038/s41431-019-0419-0.

[193] Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Human mutation. 2013 sep;34(9):E2393–402.

[194] Kang JY, Mishanina TV, Landick R, Darst SA. Mechanisms of Transcriptional Pausing in Bacteria. Journal of Molecular Biology. 2019;431(20):4007–4029. Available from: http://www.sciencedirect.com/science/article/pii/S0022283619304462.

[195] Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome research. 2012 dec;22(12):2399–2408.

[196] Alyousfi D, Baralle D, Collins A. Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data. Briefings in Bioinformatics. 2020.

[197] Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N, et al. Linkage disequilibrium in human populations. Proceedings of the National Academy of Sciences of the United States of America. 2003 may;100(10):6069–6074. Available from: https://pubmed.ncbi.nlm.nih.gov/12721363https://www.ncbi.nlm.nih.gov/pmc/articles/PMC156327/.

[198] Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database. 2011 nov;2011. Available from: https://doi.org/10.1093/database/bar049.

[199] Liu H, Cocea M. Semi-random partitioning of data into training and test sets in granular computing context. Granular Computing. 2017;2(4):357–386. Available from: https://doi.org/10.1007/s41066-017-0049-2.

[200] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software; Vol 1, Issue 3 (2011). 2011 dec. Available from: https://www.jstatsoft.org/v045/i03http://dx.doi.org/10.18637/jss.v045.i03.

[201] Tipping M. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research. 2001 jan;1:211–244.

[202] Hair JF. Multivariate data analysis : a global perspective. Upper Saddle River, N.J.; London: Pearson Education; 2010.

[203] O'brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity. 2007;41(5):673–690. Available from: https://doi.org/10.1007/s11135-006-9018-6.

[204] Papp L, Spielvogel CP, Rausch I, Hacker M, Beyer T. Personalizing Medicine Through Hybrid Imaging and Medical Big Data Analysis. Frontiers in Physics; 2018. Available from: https://www.frontiersin.org/article/10.3389/fphy.2018.00051.

[205] Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation. 1997;1(1):67–82.

[206] Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics. 2018;34(4):301–312. Available from: http://www.sciencedirect.com/science/article/pii/S0168952517302251.

[207] Yen SJ, Lee YS. Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset BT - Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 731–740. Available from: https://doi.org/10.1007/978-3-540-37256-1{_}89.

[208] Srinivas N, Krause A, Kakade S, Seeger M. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design; 2010.

[209] Bull A. Convergence rates of efficient global optimization algorithms. Journal of Machine Learning Research. 2011 jan;12.

[210] Kuhn M, Johnson K. Feature Engineering and Selection: A Practical Approach for Predictive Models. Chapman & Hall/CRC Data Science Series. CRC Press; 2019. Available from: https://books.google.co.uk/books?id=q5alDwAAQBAJ.

[211] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321–332. Available from: https://doi.org/10.1038/nrg3920.

[212] Stoneking C. Bayesian inference of Gaussian mixture models with noninformative priors. 2014 may.

[213] Bernardi G. Structural And Evolutionary Genomics: Natural Selection In Genome Evolution. Elsevier Science; 2005.

[214] Bonifer C, Cockerill PN. Chromatin mechanisms regulating gene expression in health and disease. Advances in experimental medicine and biology. 2011;711:12–25.

[215] Wu J, Cheng J, Zhao C, Lu H. Fusing multi-modal features for gesture recognition; 2013.

[216] Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural Selection on Genes that Underlie Human Disease Susceptibility. Current Biology. 2008;18(12):883–889. Available from: https://www.sciencedirect.com/science/article/pii/S0960982208006015.

[217] Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annual review of biochemistry. 2003;72:291–336.

[218] Abramovs N, Brass A, Tassabehji M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. Nature Genetics. 2020;52(1):35–39. Available from: https://doi.org/10.1038/s41588-019-0560-2.

[219] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proceedings of the National Academy of Sciences. 2007 may;104(21):8685 LP – 8690. Available from: http://www.pnas.org/content/104/21/8685.abstract.

[220] Barido-Sottani J, Chapman SD, Kosman E, Mushegian AR. Measuring similarity between gene interaction profiles. BMC bioinformatics. 2019 aug;20(1):435. Available from: https://pubmed.ncbi.nlm.nih.gov/31438841https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6704681/.

[221] Bashir S, Qamar U, Khan FH. IntelliHealth: A medical decision support application using a novel weighted multi-layer classifier ensemble framework. Journal of Biomedical Informatics. 2016;59:185–200. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415002816.

[222] Cleophas TJ, Zwinderman AH. Machine Learning in Medicine – A Complete Overview. Springer International Publishing; 2020. Available from: https://books.google.co.uk/books?id=D1TUDwAAQBAJ.

[223] Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. Journal of Biomedical Informatics. 2017;66:82–94. Available from: http://www.sciencedirect.com/science/article/pii/S1532046416301812.

[224] Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics. 2006 jan;22(7):830–836. Available from: https://doi.org/10.1093/bioinformatics/btk048.

[225] Alashwal H, El Halaby M, Crouse JJ, Abdalla A, Moustafa AA. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Frontiers in computational neuroscience. 2019 may;13:31. Available from: https://pubmed.ncbi.nlm.nih.gov/31178711https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6543980/.

[226] Bishop CM. Pattern Recognition and Machine Learning. Information Science and Statistics. Springer; 2006. Available from: https://books.google.co.uk/books?id=kTNoQgAACAAJ.

[227] Hawkins DM. Identification of Outliers. Monographs on applied probability and statistics. Chapman and Hall; 1980. Available from: https://books.google.co.uk/books?id=fb0OAAAAQAAJ.

[228] Liu H, Li J, Wu Y, Fu Y. Clustering with Outlier Removal. IEEE Transactions on Knowledge and Data Engineering. 2019:1.

[229] Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLOS Computational Biology. 2019 jun;15(6):e1006907. Available from: https://doi.org/10.1371/journal.pcbi.1006907.

[230] Fraley C, Raftery AE. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association. 2002 jun;97(458):611–631. Available from: https://doi.org/10.1198/016214502760047131.

[231] Jolliffe I. Principal Component Analysis BT - International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 1094–1096. Available from: https://doi.org/10.1007/978-3-642-04898-2{_}455.

[232] Verleysen M, Lee JA. Nonlinear Dimensionality Reduction for Visualization BT - Neural Information Processing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 617–622.

[233] Berman GJ, Choi DM, Bialek W, Shaevitz JW. Mapping the stereotyped behaviour of freely moving fruit flies. Journal of the Royal Society, Interface. 2014 oct;11(99).

[234] Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, et al. MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. Current protocols in bioinformatics. 2014 sep;47:1.24.1–19.

[235] McLachlan GJ, Bean RW, Jones LBT. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics. 2006 apr;22(13):1608–1615. Available from: https://doi.org/10.1093/bioinformatics/btl148.

[236] Hennig C. Methods for merging Gaussian mixture components. Advances in Data Analysis and Classification. 2010;4(1):3–34. Available from: https://doi.org/10.1007/s11634-010-0058-3.

[237] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018 nov;68(6):394–424.

[238] Ataollahi MR, Sharifi J, Paknahad MR, Paknahad A. Breast cancer and associated factors: a review. Journal of medicine and life. 2015;8(Spec Iss 4):6–11. Available from: https://pubmed.ncbi.nlm.nih.gov/28316699https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5319297/.

[239] Apostolou P, Fostira F. Hereditary breast cancer: the era of new susceptibility genes. BioMed research international. 2013;2013:747318.

[240] Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. New England Journal of Medicine. 2015 may;372(23):2243–2257. Available from: https://doi.org/10.1056/NEJMsr1501341.

[241] Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. American journal of human genetics. 2019 jan;104(1):21–34.

[242] Möller S, Mucci LA, Harris JR, Scheike T, Holst K, Halekoh U, et al. The Heritability of Breast Cancer among Women in the Nordic Twin Study of Cancer. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2016 jan;25(1):145–150.

[243] Peto J, Mack TM. High constant incidence in twins and other relatives of women with breast cancer. Nature genetics. 2000 dec;26(4):411–414.

[244] Yanes T, Meiser B, Kaur R, Scheepers-Joynt M, McInerny S, Taylor S, et al. Uptake of polygenic risk information among women at increased risk of breast cancer. Clinical Genetics. 2020 mar;97(3):492–501. Available from: https://doi.org/10.1111/cge.13687.

[245] Mina LA, Arun B. Polygenic Risk Scores in Breast Cancer. Current Breast Cancer Reports. 2019;11(3):117–122. Available from: https://doi.org/10.1007/s12609-019-00320-8.

[246] Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. Evolution, Medicine, and Public Health. 2018 dec;2019(1):26–34. Available from: https://doi.org/10.1093/emph/eoy036.

[247] Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nature reviews Genetics. 2018 sep;19(9):581–590.

[248] Läll K, Lepamets M, Palover M, Esko T, Metspalu A, Tõnisson N, et al. Polygenic prediction of breast cancer: comparison of genetic predictors and implications for risk stratification. BMC Cancer. 2019;19(1):557. Available from: https://doi.org/10.1186/s12885-019-5783-1.

[249] Shi M, O'Brien KM, Weinberg CR. Interactions between a Polygenic Risk Score and Nongenetic Risk Factors in Young-Onset Breast Cancer. Scientific Reports. 2020;10(1):3242. Available from: https://doi.org/10.1038/s41598-020-60032-3.

[250] Konuma T, Okada Y. Statistical genetics and polygenic risk score for precision medicine. Inflammation and Regeneration. 2021;41(1):18. Available from: https://doi.org/10.1186/s41232-021-00172-9.

[251] Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nature Protocols. 2020;15(9):2759–2772. Available from: https://doi.org/10.1038/s41596-020-0353-1.

[252] Cecile A, Janssens JW, Joyner MJ. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? Clinical Chemistry. 2019 05;65(5):609–611. Available from: https://doi.org/10.1373/clinchem.2018.296103.

[253] Wu J, Pfeiffer RM, Gail MH. Strategies for Developing Prediction Models From Genome-Wide Association Studies. Genetic Epidemiology. 2013;37(8):768–777. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21762.

[254] Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. The American Journal of Human Genetics. 2015 2021/08/31;97(4):576–592. Available from: https://doi.org/10.1016/j.ajhg.2015.09.001.

[255] Bogaert KA, Manoharan-Basil SS, Perez E, Levine RD, Remacle F, Remacle C. Surprisal analysis of genome-wide transcript profiling identifies differentially expressed genes and pathways associated with four growth conditions in the microalga Chlamydomonas. PloS one. 2018 apr;13(4):e0195142–e0195142. Available from: https://pubmed.ncbi.nlm.nih.gov/29664904https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5903653/.

[256] Eccles D, Gerty S, Simmonds P, Hammond V, Ennis S, Altman DG. Prospective study of Outcomes in Sporadic versus Hereditary breast cancer (POSH): study protocol. BMC cancer. 2007 aug;7:160.

[257] Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–678. Available from: https://doi.org/10.1038/nature05911.

[258] Kadalayil L, Khan S, Nevanlinna H, Fasching PA, Couch FJ, Hopper JL, et al. Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. Nature communications. 2017 nov;8(1):1632.

[259] Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nature genetics. 2010 jun;42(6):504–507.

[260] Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–678. Available from: https://doi.org/10.1038/nature05911.

[261] Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Current protocols in human genetics. 2011 jan;Chapter 1:Unit1.19–Unit1.19. Available from: https://pubmed.ncbi.nlm.nih.gov/21234875https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066182/.

[262] Wellcome Centre Human Genetics;. Available from: http://www.well.ox.ac.uk/{~}wrayner/tools/.

[263] Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics. 2012;3.

[264] Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The International HapMap Project. Nature. 2003;426(6968):789–796. Available from: https://doi.org/10.1038/nature02168.

[265] Mak TSH, Kwan JSH, Campbell DD, Sham PC. Local True Discovery Rate Weighted Polygenic Scores Using GWAS Summary Data. Behavior Genetics. 2016;46(4):573–582. Available from: https://doi.org/10.1007/s10519-015-9770-2.

[266] Cover TM, Thomas JA. Elements of information theory. New York: Wiley; 1991.

[267] Levy R. Expectation-based syntactic comprehension. Cognition. 2008;106(3):1126–1177. Available from: https://www.sciencedirect.com/science/article/pii/S0010027707001436.

[268] Gaudet MM, Kuchenbaecker KB, Vijai J, Klein RJ, Kirchhoff T, McGuffog L, et al. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. PLoS genetics. 2013;9(3):e1003173.

[269] Milne RL, Antoniou AC. Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. Annals of oncology : official journal of the European Society for Medical Oncology. 2011 jan;22 Suppl 1:i11–7.

[270] Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, et al. Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. JNCI: Journal of the National Cancer Institute. 2017 mar;109(7). Available from: https://doi.org/10.1093/jnci/djw302.

[271] Sieh W, Rothstein JH, McGuire V, Whittemore AS. The role of genome sequencing in personalized breast cancer prevention. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2014 nov;23(11):2322–2327.

[272] Jia G, Lu Y, Wen W, Long J, Liu Y, Tao R, et al. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. JNCI Cancer Spectrum. 2020 mar;4(3). Available from: https://doi.org/10.1093/jncics/pkaa021.

[273] Godet I, Gilkes DM. BRCA1 and BRCA2 mutations and treatment strategies for breast cancer. Integrative cancer science and therapeutics. 2017 feb;4(1):10.15761/ICST.1000228. Available from: https://pubmed.ncbi.nlm.nih.gov/28706734https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5505673/.

[274] Antoniou AC, Pharoah PDP, McMullan G, Day NE, Stratton MR, Peto J, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. British journal of cancer. 2002 jan;86(1):76–83.

[275] Larsen MJ, Thomassen M, Gerdes AM, Kruse TA. Hereditary breast cancer: clinical, pathological and molecular characteristics. Breast cancer : basic and clinical research. 2014 oct;8:145–155. Available from: https://pubmed.ncbi.nlm.nih.gov/25368521https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4213954/.

[276] Lei X, Zhang Y. Predicting disease-genes based on network information loss and protein complexes in heterogeneous network. Information Sciences. 2019;479:386–400. Available from: http://www.sciencedirect.com/science/article/pii/S0020025518309514.

[277] Kiryo R, Niu G, Plessis M, Sugiyama M. Positive-Unlabeled Learning with Non-Negative Risk Estimator. 2017 mar.

[278] Hou M, Zhao Q, Li C, Chaib-draa B. A generative adversarial framework for positive-unlabeled classification. ArXiv. 2017;abs/1711.08054.

[279] Fullerton JM, Nurnberger JI. Polygenic risk scores in psychiatry: Will they be useful for clinicians? F1000Research. 2019 jul;8:F1000 Faculty Rev–1293. Available from: https://pubmed.ncbi.nlm.nih.gov/31448085https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6676506/.

[280] Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science (New York, NY). 2018 dec;362(6420).

[281] Fragomeni SM, Sciallis A, Jeruss JS. Molecular Subtypes and Local-Regional Control of Breast Cancer. Surgical oncology clinics of North America. 2018 jan;27(1):95–120. Available from: https://pubmed.ncbi.nlm.nih.gov/29132568https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5715810/.

[282] Griffis JC, Allendorfer JB, Szaflarski JP. Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. Journal of neuroscience methods. 2016 jan;257:97–108. Available from: https://pubmed.ncbi.nlm.nih.gov/26432931https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4662880/.

[283] Ben-Hur Asa, Horm David, Siegelman Hava VV. Support Vector Clustering. Journal of Machine Learning Research. 2001;2:125–137.

[284] Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 2004;88(2):365–411. Available from: http://www.sciencedirect.com/science/article/pii/S0047259X03000964.

[285] Minka T. A comparison of numerical optimizers for logistic regression. CMU Technical Report. 2003 jan;2003.

[286] Amato F, Mazzocca N, Moscato F, Vivenzio E. Multilayer Perceptron: An Intelligent Model for Classification and Intrusion Detection. In: 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA); 2017. p. 686–691.

[287] Breiman L. Bagging Predictors. Machine Learning. 1996;24(2):123–140. Available from: https://doi.org/10.1023/A:1018054314350.