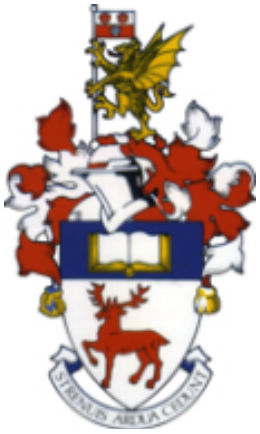
UNIVERSITY OF Southampton

University of Southampton Research Repository
Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.
When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.
Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.
Data: Author (Year) Title. URI [dataset]



UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

Genetic Epidemiology and Bioinformatics Research Group

Development and application of methods for resolving molecular diagnoses from patient sequence data for monogenic diseases

by

Dareen Mohammed S. Alyousfi

Thesis submitted towards a Ph.D.

July 2021

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE

Genetic Epidemiology and Bioinformatics Research Group

Thesis submitted towards a Ph.D.

Development and application of methods for resolving molecular diagnoses from patient sequence data for monogenic diseases

Dareen Mohammed S. Alyousfi

Identifying molecular causes of disease from sequenced genomes can be extremely challenging, and usually requires tiered filtering with the possibility of causal variant(s) being missed. **The first stage** of this study was focused on understanding the specific properties and features of genes including essentiality, haploinsufficiency, and selection and therefore, linking these properties to facilitate the prediction of disease causal genes. Gene essentiality refers to genes that is required for the survival of the cells. This study found 20 gene-specific scores in the literature, each of which measures various genetic features. It then showed that until now, no reliable single score has been predictive of genic deleteriousness. This systematic review helped in identifying the gaps and challenges in the prediction of disease genes that might have an impact on the diagnosis of monogenic diseases. This information on genes rather than variants broadens the scope of thinking to better assess gene pathogenicity. **The second stage** gathered all this information to build a model to filter the clinical sequence data and decrease the number of potential disease-causing genes to follow-up. Further, essentiality specific pathogenicity prioritisation (ESPP) was constructed to prioritise disease causing genes and showed improved performance in identifying disease genes that score high—helping to exclude non-disease genes that score low—as compared to any single score. **The third stage** evaluated the proposed gene-level score to guide prioritization of disease genes by testing the score using multiple databases and integration with alternative scores. This contributes significantly to improving data interpretation. The results were encouraging as two genes, named *CNOT1* and *RYR3*, that were prioritised by ESPP as strong candidates for Mendelian diseases, were subsequently confirmed to be causal. Another finding from the sum of ranks of alternative scores (ESPP, LOEUF and CoNeS) found four genes (SETD1A, SMARCC2, KDM3B, MED12L) that were ranked highly

and are now known to contain disease variations. Ultimately, applying such models to monogenic disease patient sequence data will help identify molecular causes for these conditions.

Table of Contents

List of tables

List of Figurs

Academic Thesis: Declaration of Authorship

I, ... [Dareen Mohammed S. Alyousfi]

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

[Development and application of methods for resolving molecular diagnoses from patient sequence data]

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
    - 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*. 2019 Jan;20(1):267-73.
    - 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*. 2019 Jan;18(1):23-9.
    - Alyousfi D, Baralle D, Collins A. Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data. *Briefings in Bioinformatics*. 2020 Mar 18.

Signed:

……Dareen……

……Date: 07/06/2021

Acknowledgements

First, I would like to acknowledge my country and King Abdulaziz University for providing me with scholarship funding and facilitating my entire Ph.D. journey with their full support. I thank them for the incredible professional and personal development experiences I consequently experienced.

Second, I would like to express my gratitude towards my parents and family and my deep appreciation to Prof. Andrew Collins, my first supervisor, Prof. Diana Baralle, my second supervisor as well as Dr. Fatima Asiri, Dr. Jenny Lord, Dr. Zoe Walters, and Dr. Gabrielle Wheway along with my colleagues, Mr. Reza Jabalameli, M. and Miss. Clare Horscroft for their endless support, inspiration, and understanding during my thesis.

Once again, I thank you all.

Definitions and Abbreviations

ACHG: American College of Human Genetics

ACMG: American College of Medical Genetics

BioGRID : Biological General Repository for Interaction Datasets

BSB: backward snowballing

CADD: Combined annotation dependent depletion

CEPH: Center d'Etude du Polymorphisme Humain

CFV: common functional variation

CM: Complex Mendelian

CMG: Centre for Mendelian Genomics

CML: chronic myelogenous leukaemia

CNM: Complex non-Mendelian

CRISPR: clustered regularly interspaced short palindromic repeats

DDG2P: Developmental Disorder Genotype-Phenotype Database

ddNTPs: dideoxynucleotide triphosphates

DIP: Database of Interacting Proteins

DL: Developmental lethal

DNA: Deoxyribonucleic acid

dN/dS: non-synonymous to synonymous substitution rates

Ebp: Exa-basepairs

END: Essential non-disease

ExAC: Exome Aggregation Consortium

FSB: forward snowballing

FUSIL: FUll spectrum of intolerance to loss-of-function variation

GS: Google Scholar

GWAS: genome wide association study

HGMD: human genome mutation database

hOMIM: hand-curated Online Mendelian Inheritance In Man

IntAct: an open source molecular interaction database

JAMA: Journal of the American Medical Association

Ka/Ks: non-synonymous to synonymous ratios

Kruskalmc: Kruskal Wallis multiple comparison

LD: Linkage disequilibrium

LOD: logarithm-of-odds

LoF: loss of function

LOFTEE: loss-of-function transcript effect estimator

miRNA: microRNA

MDR: multifactor dimensionality reduction

MINT: Molecular INTeraction database

MIPS: Munich Information Center for Protein Sequences

MNC: Mendelian non-complex

mRNA: messenger ribonucleic acid

NDNE: Non-disease non-essential

NGS: Next generation sequencing

NHS: National Health Service

OMIM: Online Mendelian inheritance in man

PAG: polyacrylamide gel

PCA: principal component analysis

Peta-basepairs (Pbp)

pLoF: predicted loss of function

PPI: protein–protein interactions

RNAi: RNA interference

RSPP: recessive-mode specific pathogenicity prioritization

SB: snowball strategy

SBS: sequence by synthesis

SHGP: Saudi Human Genome Programme

SIFT: sorting intolerant from tolerant

siRNA: small interfering RNA

SLR: systematic literature review

SNPs: single nucleotide polymorphisms

SNVS: single nucleotide variants

Tbp: Tera-basepairs

VIRs: variant intolerant regions

WGS: whole genome sequence

Zbps: Zetta-basepairs

List of Scores and software packages abbreviations

DNE: Gene constraint score-*de novo* excess

DOMINO: machine learning to predict genes associated with dominant disorders

EvoTol: evolutionary intolerance

GeVIR: Gene-level variation intolerance metric

GHIS: Genome-wide haploinsufficiency score

GIMS: gene-level integrated metric of negative selection

HI: Haploinsufficiency

HIPred: haploinsufficiency predictor

LOEUF: Loss-of-function observed/expected upper-bound fraction

mirDNMR*: de novo* mutation rate

NET: Gene position in networks indispensability score

pLI: Loss intolerance probability

REC: Recessive score

RVIS: residual variation intolerance score

SIS: substitution intolerance score

SnIPRE: Selection Inference Using a Poisson Random Effects Model

subRVIS: sub-regions residual variation intolerance score

XL: X-linked

Chapter 1 Introduction

In this chapter, the key concepts are explained to ease the understanding of this thesis. The first section introduces the human genome and some of its relative genetic properties. The second section presents a brief background on common and rare genetic diseases. The third section introduces next generation sequencing (NGS) and the gap between sequence data production and interpretation. The fourth section is on essential gene prediction and available methods to predict gene pathogenicity. The fifth section describes several gene/variant specific scores and their usage to predict genetic deleteriousness. The sixth section introduces the research question framework for the thesis, and the final section outlines the aims and contribution.

The purpose of the introduction chapter is to supply the reader with a brief review on the basic concepts of understanding the human genome, as well as NGS analysis and how it contributes to the prediction of disease causal genes. This gives context and helps in the understanding of what will be explored in subsequent chapters.

Subsequently, a broader coverage of the literature related to gene-specific metrics and assessments of gene essentiality are provided. This section is essential to outlining the available information on gene-level scores and how these may be utilised to improve the prediction of disease-causing genes.

1.1 The human genome

The human genome consists of a long sequence of bases called (nucleotides): adenine (A), guanine (G), thymine (T) and cytosine (C). These four bases are packed tightly into 23 pairs of chromosomes with a total of approximately three billion DNA base pairs, with specific regions representing genes with different biological significances. There are an estimated 20–25 thousand genes in the human genome; they are divided into exons (coding) and introns/intragenic (non-coding). The sequences located between genes, called intergenic regions, differ from introns in that they are located outside genes. More specifically, exons account for only 1.1% of the human genome, with 24% being introns, and the remaining approximately 75% being intergenic regions (1,2).

When a trigger stimulates a cell to produce a specific protein, the respective gene is transcribed to synthesise a messenger ribonucleic acid (mRNA) sequence, which will directly produce the specific protein. Each protein consists of a specific amino acid (a.a) sequence, the alteration of which may affect protein function. Diploid organisms such as humans carry two copies of each gene. If a sequence variant is observed in a single copy of the gene, the genotype is described as heterozygous, while if the mutation is observed in both copies, the genotype is homozygous.

However, genes are not independent, and several interactions between genes might affect the phenotype produced—this phenomenon is known as epistasis. For many years, identifying the causal genes for complex and rare disease was challenging, with epistasis being a particular cause for concern especially in complex diseases (3). Further, a locus might be enhanced or concealed by the impact of another locus, causing difficulty in the detection of the genuine variant. Moreover if more than two loci are involved the prediction is more complicated (3). Historically, an important aim of genetics was the correlation of a phenotype with its specific genotype (4), and although this has improved, much work remains to be done. Understanding genetic properties such as selection, essentiality, and haploinsufficiency (HI) along with gene interactions is of great importance in improving our understanding of how candidate genes are to be prioritised and subsequently, the prediction of disease-causality is to be improved.

The following section will describe how linkage and physical maps can be utilised to identify the location of certain genes, which might be the cause for a disease.

1.1.1 Linkage mapping

Mapping is used to determine the location of genetic elements using identifiable landmarks, which can be functional segments of the DNA; e.g., genes or non-functional sequences (5). In this context, linkage mapping or genetic linkage (ASA genetic map) shows the position of genetic markers in relation to each other with regard to the recombination frequency rather than the physical distance along each chromosome. In other words, using linkage mapping, the location of a gene can be determined corresponding to other gene; however, the exact location cannot be determined. Thus, the analysis of these maps is a powerful tool to detect the location of disease-genes along the chromosome (6). Here, linkage mapping of disease genes can be constructed using polymorphic markers (Figure 1-1) or family pedigrees.

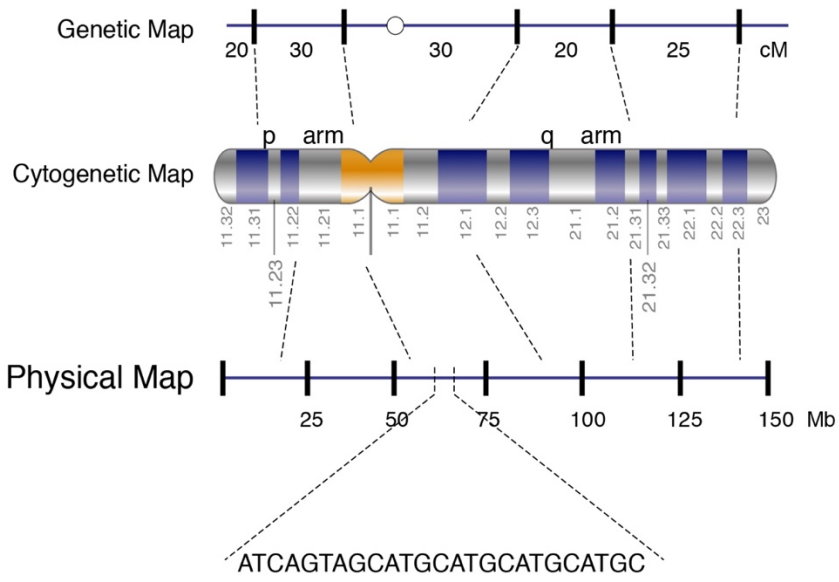


Figure 1-1 A genetic map of a chromosome measured by centimorgan (cM), a cytogenetic map, and a physical map measured by megabases (Mb) (5).

Meanwhile, physical maps describe the relative order of markers across a chromosome. Further, high quality maps can show the exact distances between adjacent markers. At the present time, these have become more accurate as they are built based on the chromosome sequence, so the exact base pair distances between markers can be identified (Figure 1-1) (5).

The first-generation combined linkage–physical map (7) was constructed in 2004 and contained approximately 14,800 markers. The second-generation such map of the human genome was constructed using the previous data of the first generation and around 13,700 SNPs genotyped in CEPH (Center d'Etude du Polymorphisme Humain) (8).

This map also referenced pedigrees at the following companies: Applied Illumina, Affymetrix, and Biosystems (7). Additionally, this linkage map has approximately double the number of markers as the previous one, thus providing a useful map for genetic analysis. Every single marker on this map is supported by recombination-based and physical data. Moreover, there are two ways to use the confidence intervals (CI) generated from the second-generation combined linkage–physical map of the human genome: First, it can be used to quantify the impact of map uncertainty on a genetic analysis and second, to integrate the information in this map with the independent map estimates obtained from individual studies (7).

In the early 80s, and with the introduction of genome-wide linkage analysis using anonymous DNA polymorphisms, connecting phenotypes to genes was made possible using human genetic linkage mapping (4). Previously, approximately 1,200 disease causing genes had been discovered using linkage-based 'positional cloning'. This is a laboratory method that localises the chromosomal position of the disease-causing gene of Mendelian phenotypes, using the knowledge of the inheritance pattern of the phenotype. In this approach, no prior knowledge about the biological process of disease is required, apart from that required to evaluate the phenotype. Identifying causal genes for hemochromatosis, cystic fibrosis, and lactose intolerance are successful examples of positional cloning (4). More specifically, this cloning starts with linkage analysis; here, a particularly important example of the use of linkage mapping to identify disease genes is the identification of the cystic fibrosis causing gene (4). Advances in this area led to the development of polymorphic markers and then, simple sequence repeats, with several studies ultimately benefiting from the hundreds and thousands of single nucleotide polymorphisms (SNPs) obtained through the sequencing of human genomes (4).

- **Linkage disequilibrium mapping**

Linkage disequilibrium (LD) is defined as 'the non-random association of alleles at different loci' (9)**.** In evolutionary biology, LD reflects the history of natural selection, mutation, and other factors that might affect gene-frequency evolution. It also reflects the population history by providing information on past events and constraints to the potential response to natural selection. If two loci are very closely linked, recombination will be reduced, and LD might be stronger to a good approximation (9). In other words, when two alleles are located close to each other in a chromosome, the chance of recombination events are reduced, and they are likely to be inherited together.
The theory behind LD is well-established and has improved understanding of evolutionary history as well as provides tools and insights for gene mapping in humans and other species. Further, in humans, LD has allowed for the fine-scale gene mapping of alleles and specific traits (9).

As the focus of this thesis is the prediction of monogenic disease genes, parametric linkage analysis will be used as an effective method to map genes in single-gene disorders. This approach is based on the information related to the penetrance of the disease, mode of

inheritance, and disease gene frequency. Conversely, due to a limited understanding of genotypic penetrance and mode of inheritance in complex diseases, parametric linkage analysis has limited power in this group of diseases (Figure 1-2) (4).

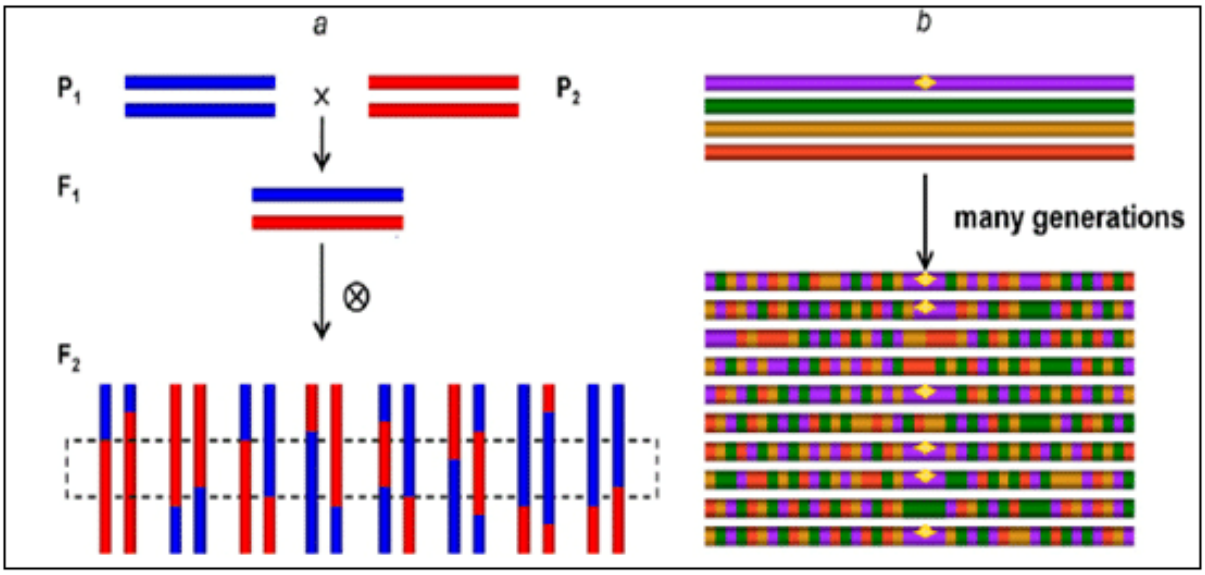


Figure 1-2 Linkage analysis and association mapping: a. linkage mapping relies on the co-segregation of a phenotype and gene through the generations; b. high-resolution mapping shown only in haplotypes. For genes within a short distance, the LD between a functional allele (yellow diamond for mutation) shows that the marker is low (10)

After exploring linkage and physical mapping and how these maps can help in detecting the position of a gene, it is important to understand how disease genes can be identified among the whole genome by recognising some genetic properties of those genes, which might improve their detection.

1.1.2 Loss of function variation (LoF)

Genetic variation can be classified by its effect on function. This includes loss of function (LoF) and gain of function mutations. The former severely disrupt the function of a protein (11) and are often nonsense sequence changes. However, they have also been identified in the genome of healthy individuals, thus distinguishing those that are pathogenic can be difficult (12,13). In this context, recent studies suggested that the healthy human genome contains approximately 200 to 800 anticipated LoF variants, which have an impact on the clinical interpretation of sequence data (12). This raises the question of whether the unit of the whole gene is the correct one to be used when assessing patterns of intolerance, which will be addressed later in the thesis. Moreover, large sequencing and genotyping projects are

important for the discovery of LoF variants and their impact on the risk of human disease (12).

1.1.3 Gene interactions

The phenomena of locus heterogeneity and phenocopy contribute to the challenges of mapping genotypes to a specific phenotype. Gene–gene interaction, sometimes called epistasis, is one of the factors that complicates identification of disease susceptibility genes (14). A common definition of epistasis is 'one gene masking the effects of another gene' (12,13). Further, Moore et al. (15) built a hypothesis that epistasis is a present component of the genetic structure of complex diseases and that complex interactions are highly substantial as compared to the independent main impact of any single susceptible gene. Here, the key question is 'why are gene–gene interactions likely to be common?' (14). First, the ubiquitous component of the genetic architecture of common human diseases hypothesis as explained by Moore (15) state that interactions between genes are important factors that lead to deviation from Mendelian ratios (14) and have been known for nearly a hundred years. Second, the biomolecular interactions, genetic regulatory functions, and metabolic systems suggest that gene–gene interactions are involved in the connection between DNA sequence variations and clinical results (14). For instance, any specific gene can be regulated by nearly a hundred or more proteins that might affect the gene through protein–DNA or protein–protein interactions (PPI). These interactions are most likely mediated by variations in the DNA sequence of genes that encode for proteins. There are several traditional and new statistical methods for identifying gene–gene interactions in association studies; examples of these methods include logistic regression and multifactor dimensionality reduction (MDR) (14). The advantage of logistic regression is that the statistical concept is very well described. However, the main disadvantage when having several independent variables is that large sample sizes are required to accurately estimate the parameters in the model. Nevertheless, MDR is designed to identify interactions in the absence of detectable main effects (14).

- **Protein–protein interaction networks (PINs)**

Generally, proteins do not function separately, but interact with each other to ensure the internal function of the cell. Studying PINs show that the network centrality of certain proteins in PINs is closely related with the essentiality of that protein (16). There are several publicly available protein–protein interaction databases, and each has its own characteristics

and different levels of annotation. Examples include the Database of Interacting Proteins (DIP) (17), STRING (a web source of predicted PPIs) (18), Biological General Repository for Interaction Datasets (BioGRID) (19), The Munich Information Center for Protein Sequences (MIPS) (20), IntAct (an open source molecular interaction database) (21), and MINT (Molecular INTeraction database) (22). More specifically, some PIN databases, such as MINT, STRING, and IntAct, provide scores with reliability of interactions acquired from different sources. Interactions with low scores can be filtered out by setting thresholds to achieve PPI data with high reliability (16).

1.2 Human genetic disease

It has been shown that functionally related genes might produce similar phenotypes. Examples include Fanconi anaemia or Usher syndrome, which are genetically heterogenous disorders, where multiple genes play a role in a single biological system (23). Moreover, several hypotheses have evolved regarding human phenotypes. First, bioinformatic analysis show that genetic diseases can be grouped depending on their functional similarities, and this represents the true biological connections of the genes involved. Second, specific phenotypic similarity can be used to identify the influence of unrelated genes on the same functional system. An example of this was testing yeast in two-hybrid screens of all the recognised genes for inherited ataxias, which showed that they all work on a single PPI network (23). Third, bioinformatics tools can be used to predict new genes for diseases that are part of the same phenotype group. This is possible by studying the known disease genes and then searching for other genes that share the same function (23).
To date, there are 6621 known genetic diseases, in which 3865 genes cause monogenic disorders, and 502 are associated with complex diseases (24).

1.2.1 Common and rare genetic diseases

Common genetic diseases, such as asthma and ischemic heart diseases, are usually caused by multiple genetic variants that work in conjunction with environmental factors (25). Although patients may have similar clinical presentations, the causative genetic variants might differ. Traditionally, genome-wide association studies (GWAS) were used to link genetics to the molecular bases of complex diseases before the introduction of Next Generation Sequencing (NGS) analysis. Eventually, more than a million common single nucleotide variants (SNVs) have been identified using GWAS. However, since they require

huge cohorts of patients and controls, they were generally uninformative on an individual patient basis, and consequently, the application of GWAS outcomes to individual patients could not be transferred into clinical practice.

Moreover, rare disease is defined as a health condition that affects a small number of individuals compared with other widespread diseases in the general population; this is commonly said to be < 1:2000. Currently, approximately between 5000 to 8000 distinct rare diseases have been identified (26). Although there is no uniform definition of rare disease, the terms ‘rare disease(s)’ and ‘orphan drug(s)’ are the most widely used among different organizations—by approximately 38% and 27%, respectively—as compared to other terminologies such as ‘neglected disease’, ‘rare and neglected disease’, ‘syndrome without a name’, ‘ultra orphan disease’ or ‘undiagnosed disease’ (26). Studies showed that the average prevalence threshold of rare diseases ranged from five to 76 cases/100,000 individuals. This signifies a fifteen-fold relative difference in the average prevalence thresholds used to label rare diseases. The worldwide average prevalence threshold among all organizations was 40 cases/100,000 individual (26). Despite being described as rare, they are collectively common as approximately three million people in the United Kingdom have been diagnosed with rare diseases; furthermore, 1:17 will have a rare disease at some point in their lives (27,28). Here, it is worth noting that rare diseases can be due to genetic and non-genetic causes and accounts for 80% of cases. Meanwhile, genetic disorders can be classified into hereditary or non-hereditary and the former are further classified into:

1. Single gene inheritance/Mendelian diseases
2. Multifactorial inheritance/Complex diseases
3. Chromosome abnormalities
4. Mitochondrial inheritance

Ultimately, since rare diseases are collectively common, and many patients with rare genetic diseases have not yet been diagnosed, studying this group of diseases has the potential to increase the proportion of cases with a molecular diagnosis, and thus, have a significant clinical impact.

1.3 DNA Sequencing

1.3.1 Sanger sequencing

Sequencing technology has vividly transformed biology as it allows researchers to combine DNA sequenced data with clinical phenotypes. In 1977, classical Sanger sequencing was introduced; it is based on base-specific chain terminations in four separate reactions (A, G, C, and T) (Figure 1-3, a) corresponding to the four different nucleotides in the DNA structure (29). The usage of dideoxynucleotide triphosphates (ddNTPs) in Sanger sequencing was a novel technique at the time, in which a specific 2',3'-ddNTP is added to every reaction in the presence of all four 2'-deoxynucleotide triphosphates (dNTPs). In total, four reactions are undertaken with each terminating at a different base. However, due to a small amount of ddNTP (~1%), the termination only occurs occasionally, resulting in strands of all lengths (Figure 1-3, a and b). This technique produces better results when compared to the prototype approach developed earlier in 1975 by the same group called 'plus and minus' (29). Further, when the corresponding ddNTP is incorporated, the newly synthesised DNA strand extension will terminate (chain-termination method).

Another novel aspect of the Sanger approach is the use of fluorescent labelling incorporated into the new synthesised DNA strand by a labelled precursor (the sequencing primer or dNTP) to make it detectable by radiography. The last step is to separate and detect the complex radioactive DNA molecules produced from each extension reaction, which can be done using polyacrylamide gel (PAG) that allows for the specific sizing of termination products by electrophoresis followed by in situ autoradiography. Then, by taking advantage of a physically compact DNA separation device combined with laser-based fragment detection, ultimately capillary electrophoresis became compatible with 96- and 384-well DNA plate formats, thus producing high parallel automation (29).

Although next generation sequencing has substituted Sanger sequencing as a gold standard test in the diagnostic field, there is a debate within the scientific community on the importance of confirming NGS variants using this test to maintain high sensitivity (30).

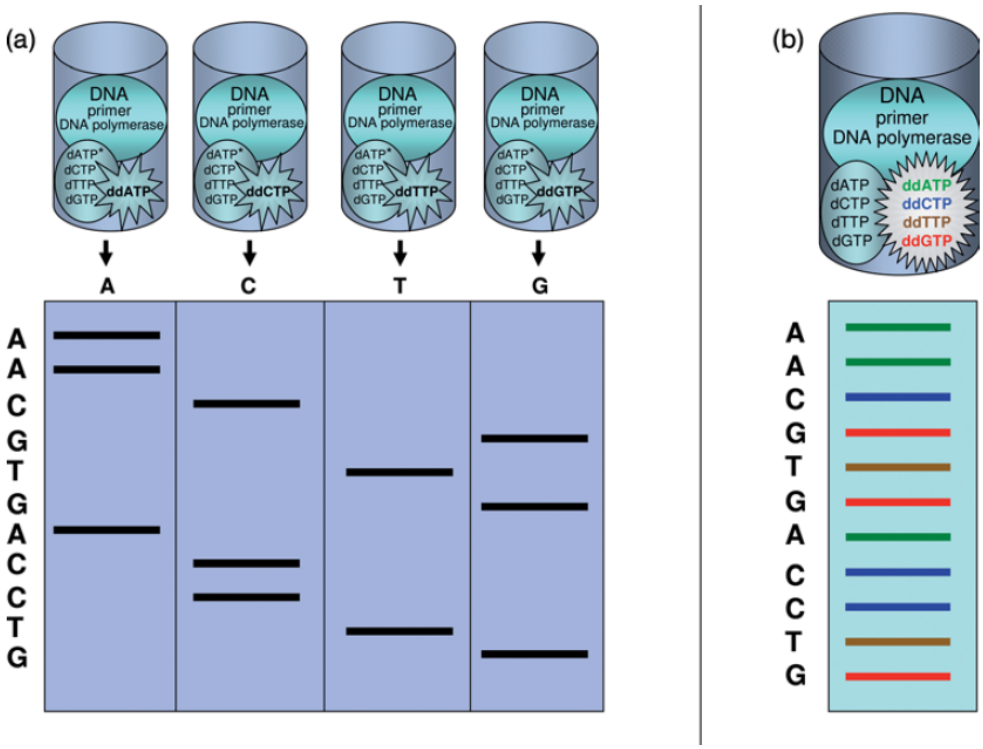


Figure 1-3 The principles of Sanger sequencing (Adopted from Men et al.) (29)

(a) Each of the four separate DNA extension reactions contain a single-stranded DNA template, primer, DNA polymerase, and all four dNTPs to create new DNA strands. Each one of the dideoxynucleotide triphosphates (ddATP, ddCTP, ddTTP, or ddGTP) is spiked with a corresponding reaction. In the presence of dNTPs the newly produced DNA strand will extend until the ddNTP is incorporated to stop further extension. The radioactive products are then detached using polyacrylamide gel in four lanes and scored according to their molecular sizes. On the left, the inferred DNA sequence is shown (29).
(b) As an alternative to adding radioactive dATP, all four ddNTPs are tagged with different fluorescent dyes. The extension products are separated using electrophoresis in a single glass capillary fully occupied with a polymer. Depending on their molecular sizes, the DNA bands pass inside the capillary. Fluorophores are excited by the laser at the end of the capillary. Ultimately, the DNA sequence can be interpreted by the colour that corresponds to a particular nucleotide (29).

1.3.2 Next-generation sequencing (NGS)

The advancement of 454 pyrosequencing and first generation Sanger DNA sequencing is the basis of the development of the second/next generation sequencing and the ongoing and increasing understanding of the genetic code (31,32). Following this, there were

advancements in NimbleGene technology, through which the sequencing capture allowed the enrichment of pre specified fragments of the genome in microarray applications. This technique led to markdown through the concentration of genotyping effort on the targeted region (33). Further, the development of this technology led to the invention of whole exome sequencing (WES), where only the selected genome coding regions are sequenced. This transformation of gene sequencing revolutionised gene hunting for monogenic diseases (34).

The field of genomics is an extensive data science topic that has grown substantially post the introduction of next generation sequencing. Following the completion of the first human genome project in 2003 (35), the post genomic era evolved, characterised by a paradigm shift in genomic research and understanding of the human genome. Projects included multiple sequencing programmes such as the 1000 genomes project (36), the 100K genome project (37) that involved 10 countries, Exome Aggregation Consortium (ExAC) (38), etc. Thus, the volume of genomic data continues to increase, which serves clinical and research communities.

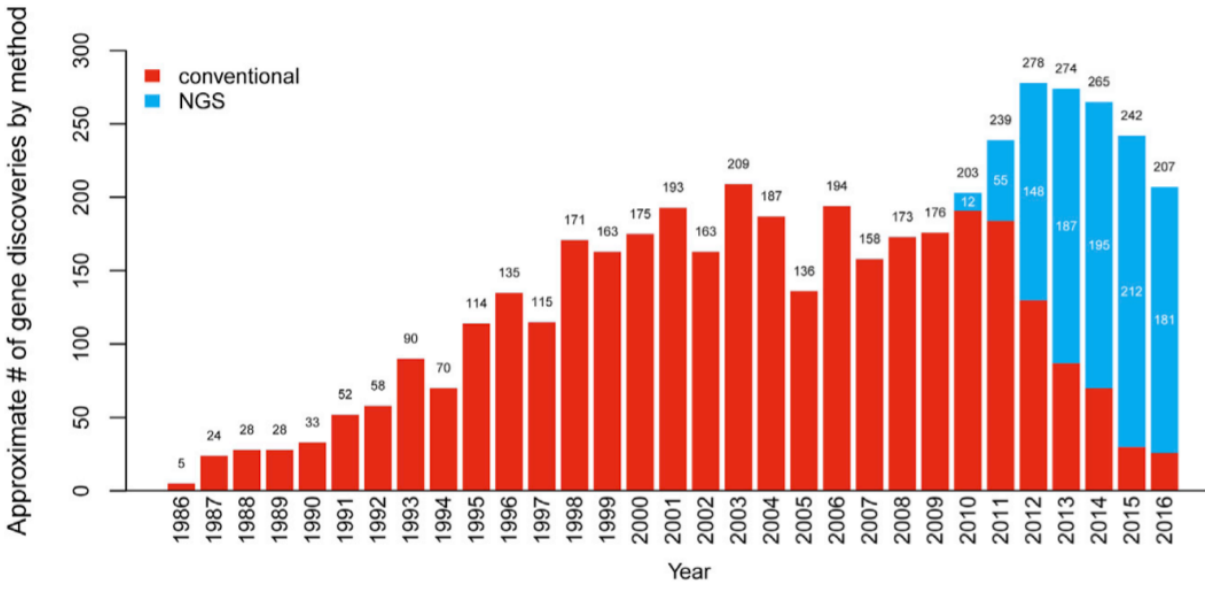


Figure 1-4 Number of genes discovered by WES and whole genome sequence (WGS) versus conventional methods since 2010 according to OMIM data (Adopted from Chong et al.) (27, 39).

Figure 1-4 shows the increase of discovered rare disease genes (WGS and WES in blue; conventional method in red) from 2010 until 2016. Two years after the introduction of the

NGS analysis, since 2010, approximately three times as many genes were discovered in comparison to conventional methods (27). This included rare disease genes (Figure 1-4). Further, significant improvements in identification of disease genes for monogenic disorders was achieved after the introduction of WES.

Moreover, the discovery of the causal variant in Miller syndrome using exome sequencing (40) made WES the state-of-art method used for discovering the causal genes of Mendelian diseases. The ability to sequence millions of DNA fragments from different samples simultaneously is a common feature of NGS platforms. The NGS approaches depend on alignment or *de novo* assembly of abundant short overlapping reads produced from fragmented genomic DNA (gDNA) (41).

- **Ilumina sequencing**

The Illumina platform employs a sequence by synthesis (SBS) approach, which has the ability to sequence the ends of billions of DNA fragments in parallel and perform read assembly for analysis. The standard sequencing procedure of illumina sequencing includes four steps: sample library preparation, cluster generation, SBS, and data analysis. The library samples are composed of double-stranded DNA flanked by known adapter sequences and have the ability to hybridise to the oligonucleotides on the surface of a flow cell (42).

The flow cell is a key element of this technology, and comprises of a thick glass slide with single or multiple channels (lanes) coated with a lawn of two designed oligonucleotides. The advantage of using a specific pattern flow cell is the production of a higher data output—more sequencing reads with a faster run time. In this context, clustering is a process where each fragment DNA molecule is amplified iso thermally. The cluster generation will start when denatured DNA libraries are permitted to randomly hybridise to the oligonucleotide lawn in the flow cell channels by their adapter ends (Figure 1-5) (42). Here, a polymerase creates a complement of the hybridised DNA fragment. Then, the double stranded DNA is denatured, and the original strand washed away. The cloned strand starts amplifying through bridge amplification, which is the process when the cloned strand bends over and the adapter hybridises to an adjacent and complementary oligonucleotide on the flow cell. Subsequently, a complimentary strand allows polymerases to end up with a double stranded bridge, after which the bridge denatures, creating two covalently bound complementary copies of the original DNA fragment. The same process is repeated 24 times until clonal amplification is achieved for all fragments. In the final step of bridge

amplification, the reverse strands are washed off, leaving the forward strand ready for sequencing (42).

Sequencing by synthesis begins with extension of the first sequencing primer to produce the first read. Fluorescently labelled nucleotides compete for addition to the producing chain. Only one nucleotide is incorporated based on the template sequence. After the insertion of each nucleotide, the clusters are imaged by laser excitation, and a distinctive fluorescent signal is omitted to permit for the incorporation of the next base. This process is called SBS. Here, it is worth noting that the length of the read is determined by the number of cycles, and the base call is determined by wave length and signal intensity (42).
Moreover, for a given cluster, all matching strands are read at the same time. When the first read is completed, the read product is washed away. In this step, the index one read primer is introduced and hybridised to the template (42). Then, another read is generated in the same way as the first read. When the index read produced, the read product is washed off and the three prime ends of the template are deprotected. Subsequently, another round of bridge amplification takes place. The new index is read in the same way as the original one. Polymerases extend the second flow cell oligo, thus producing a dual bridge. This double stranded DNA is then cleaved, and the three prime ends are blocked. The original forward strand is washed away, leaving the reverse strand. Next, read two starts with the insertion of the read two primer. Likewise, the sequencing process is repeated until the anticipated read length is reached (42).

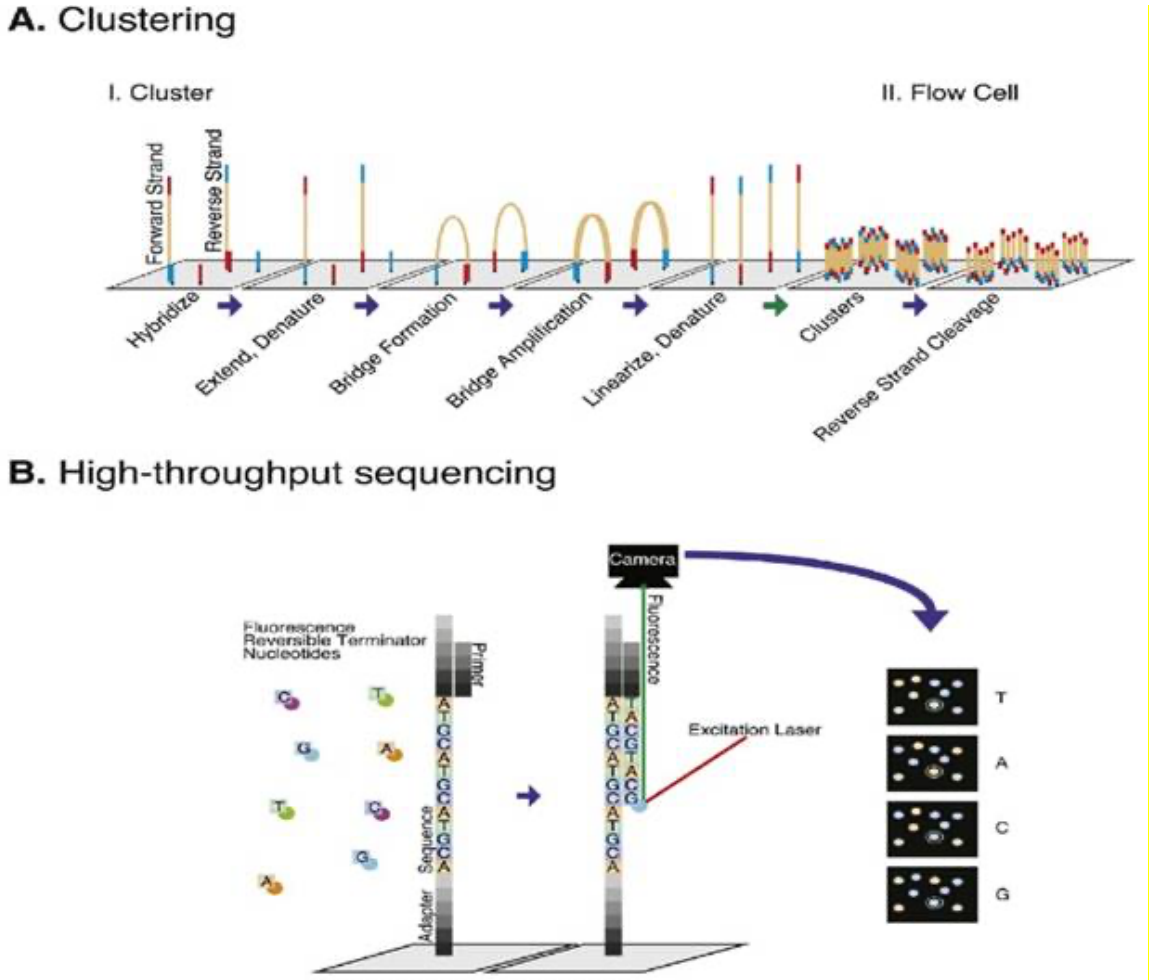


Figure 1-5 Illumina sequencing by synthesis (SBS). Adopted from (Chaitankar et al.) (42).

After completion of the library preparation and the fragmented DNA flanked with adaptor sequences (contain primer sequences), the following steps are undertaken.

A. Bridge amplification begins by binding single strands to sequences attached to a solid surface: The free ends of the strands bend and bind to adjacent complimentary sequences, forming a bridge. Then nucleotides are added to produce a double strand, and the original strand is detached and washed away. This process is repeated to produce local clusters of copies of the same sequence.

B. Adding four modified dNTPs and correct base incorporations: Numerous bases cannot be combined due to the blocking group. Thus, lasers are used to excite the fluorescent dye to distinguish the bases. Next, the blocking group and the fluorescent dye are denatured, and the cycle is then repeated.

- **Paired end sequencing**

The Illumina sequencing method allows for single- or paired-end sequencing for the multiple loading of libraries. In paired-end sequencing, both ends of the DNA fragment (Sequence (~100bp) of a larger fragment of the sequence (~500bp)) are sequenced, producing two reads. Bridge amplification is performed to produce a second read, and consequently, the forward strand is washed out to start the next round of SBS (42). The two reads are processed simultaneously. To identify the libraries from which reads they were originated, each library adapter contains a specific index sequence. When using single indexing, 24 libraries can be combined in each lane of the flow cell, while dual indexing allows up to 96 libraries to be pooled together (42).

- **Whole Exome Sequencing**

WES is a genomics approach, in which the protein-coding regions of the genome are sequenced, producing a powerful high throughput of exome data. The purpose of this technique is to identify genetic variants that affect the final protein product. This approach has been applied clinically in research and academia. Multiple steps are required in WES, which include:

1. Library preparation
2. Amplification and enrichment
3. Sequencing
4. Analysis of the data

In the first step of WES, library preparation, the following steps are required:

DNA fragmentation and target selection. This step typically aims to produce DNA fragments and ligate specific adapters to both ends of the fragment. DNA fragmentation can be undertaken using physical or enzymatic methods. These libraries are known as fragment libraries (43). On the other hand, if the sequence of the DNA target is known, PCR amplification might be used to produce DNA amplicons, known as the amplicons library. Specific DNA adaptor sequences are ligated near 3’ or 5’ of the fragmented or amplicon DNA. The DNA adapters are usually 20–40 bp of known sequence. After preparing DNA fragments with known adapters, the library fragment sizes need to be selected. This step can be achieved either by using gel electrophoresis (i.e., separating the fragments by size using

gel-based approach) or the bead-based size selection method (i.e., using magnetic beads to isolate fragments size of interest). This is a key step to achieve high quality DNA sequencing (43).

The second step is library quantification and quality control. This step can be done using the BioanalyzerTM system, which provides fragment size information and library concentration. Otherwise, this can be done using qPCR, which provides library quantification information with high accuracy; however, it lacks library size information (43).

Further, the amplification step uses PCR to increase the number of fragmented segments for subsequent sequencing using one of the major sequencing methods such as sequencing by synthesis or ligation or any other such method. In DNA SBS, fragments are read by producing a complementary fragment with polymerase enzyme and florescent nucleotide. Each colour represents a nucleotide sequence, so it can be determined by differences in florescent colours (44). Moreover, SBS uses ligase enzyme to determine the underlying sequence of the target DNA. This method depends on the sensitivity of DNA ligase enzyme for base pair mismatches (45). Post sequencing, a data file is generated containing the observed bases and the relative per base quality. This file can then be analysed using bioinformatics tools, usually in FASTQ format, although sometimes the file generated can be of a different format. A FASTQ file typically uses four lines per sequence as the following:

- Line 1 is a sequence identifier, and begins with an '@' character.
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally proceeded by the same sequence ID (or any description) again.
- Line 4 encodes the quality values for the sequence presented in Line 2, and must be comprised of the same number of symbols as letters in the sequence (46).

An example of a FASTQ file format is shown below (46):

@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTGAATTTCA
AACCCTTTCGGTTTCCAACCTTCCAAAGCAATGCCAATA
+

SRR014849.1EIXKN4201CFU84length=933+&$#"""""""""""""""7F@71,'";C?,B;?6B;:EA1E
A1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@/=<?7=9<2A8==
Furthermore, the purpose of exome sequencing is to target regions that code for protein. In comparison to whole genome sequencing, this reduces output data and is more cost-effective. There are several methods to capture only the coding parts of the genome (47). One of the most common approaches is in-solution capture and uses a pool of custom oligonucleotides (probes) synthesised and hybridised in solution to a fragmented DNA sample. The probes are labelled with beads mixed to the genomic regions of interest, so that this segment of interest can be retained, and non-coding regions are washed out. The beads are then removed, and the genomic fragments of interest can be sequenced (47).

- **Challenges of NGS technologies**

Although NGS technology has been widely used and has been helpful in research as well diagnostic settings, NGS still has challenges. One of the major limitations is the error rate associated with the base calling, which differs for each NGS platform, ranging from 0.1 to 15%. This error rate is higher than in Sanger sequencing (41). The problem with a high error rate is that it leads to false positive or negative results (48). Recently, machine learning methods have been proposed to solve this issue (49).

A further limitation of NGS platforms is the short-read length, increasing the difficulty of variant calling in repetitive and low complexity regions of the genome. Initially, this was just ~35bp of sequence, but currently, the read length of most of the conventional NGS platforms is up to 300bp. Presently, mapping-based variant calling where short-reads are uniquely mapped to the reference genome are considered the best available practice. Additionally, this problem was also addressed by developing platforms with long-read sequencing competency (50); however, read lengths and price were proportionally correlated, limiting the extension of this platform into clinical and research settings (41).

1.3.3 Growth of DNA sequencing

Prior to 2015, the rate of growth of genomic data almost doubled every seven months. Of all the sequencing equipment in the world, the Omics Maps Catalogue recorded approximately 2,500 high-throughput instruments in more than 55 countries (51). In the future, the sequencing capacity is expected to further increase dramatically due to rapid developments in this field. Furthermore, if the growth continues at the rate reached in 2015 (by increasing

two-folds every seven months), approximately one exabase of sequence/year in five years and one zettabase of sequence/year by the year 2025 might be reached (51) (Figure 1-6). However, currently, it is quite difficult to analyse and interpret this huge amount of data, unless a way is found to prioritise disease genes to reach a diagnosis. One way of doing this is to create filters or scores that prioritise genes and score the genes based on their potential for causing disease.

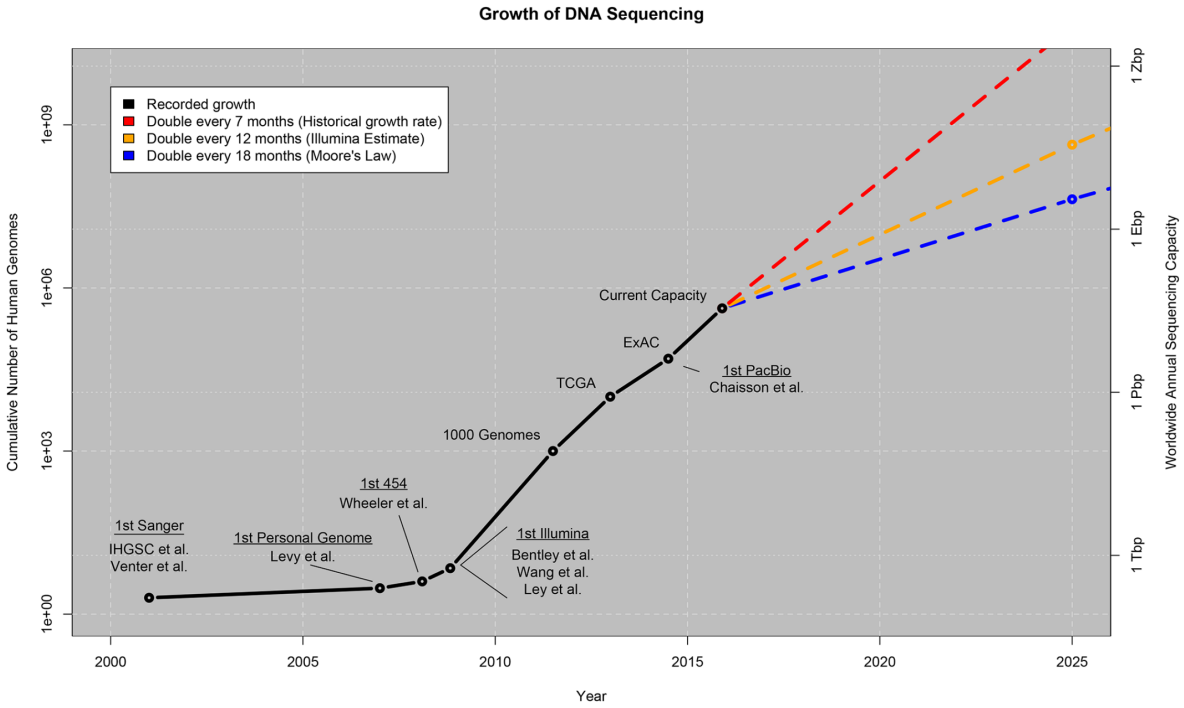


Figure 1-6 Growth of DNA sequencing timeline, adapted from Stephens et al. (51).

This timeline figure demonstrates the growth of DNA sequencing both in the total number of human genomes sequenced represented in the left axis along with the worldwide annual sequencing capacity represented in the right axis with abbreviations describing the length of D/RNA molecule: Tera-basepairs (Tbp) = 1,000,000,000,000 bp, Peta-basepairs (Pbp), Exa-basepairs (Ebp), and Zetta-basepairs (Zbps).

Sequencing the human genome has improved our understanding of genetic variations, gene properties, gene interactions, and how changes in nucleic acid (D/RNA) might manifest a phenotype.

1.3.4 The gap between sequence data production and interpretation

The development of DNA sequencing methods has contributed to the production of a considerable throughput of data. This has been a quantum leap in the field of genomics (Figure 1-1). However, interpretation of these sets of sequenced data is now a major challenge, due to the fact that data interpretation techniques have not achieved the same growth as sequencing of data, leading to a considerable gap between available genomic data and our utilization of it. To facilitate massive genomic data interpretation, data filtering has become a necessity and has gained significant popularity to save time and resources. In the past, prior studies focused on predicting variant pathogenicity rather than studying how likely a gene might contain disease variations. Recently, and by integration with variant level scores, a sustained growth of evidence has shown how gene-specific characteristics might impact predicting gene pathogenicity.

1.4 Minimal genome

The term 'minimal genome' refers to the minimal genome content indispensable to creature survival referred to "essential genome" in human (52). This concept arose from observations that several genes do not seem to be required for an organism's survival. The very small genome size of *Mycoplasma genitalium* makes this bacterium a good model for a minimal genome. Most genes used by it are commonly considered vital for cell survival; thus, a set of around 250 genes has been proposed as essential genes based on this concept (53). At that time, it was thought that essential genes could be inferred from minimal genomes, which apparently comprise only essential genes. In 2009, this concept was challenged when McCutcheon et al. (54) showed that the smallest genomes belong to a parasitic species, which have the ability to survive with only a few genes obtained from their hosts. *Hodgkinia cicadicola* is a good example as it has one of the smallest genomes containing approximately 180 genes. Further, similar to other parasites, it obtains its nutrients from its host, so its genes do not need to be essential.

Moreover, essential genes and proteins play a major role in cell differentiation, metabolism, and core biological functions responsible for cell survival. Alteration of these genes might be lethal for an organism (16). Studies have identified that there is a close relationship between essential and disease genes as the latter fall towards the essential end of the gene spectrum; therefore, identification of the essential genes is of great importance to the

detection of disease genes (16). This was further investigated by Park et al. by separating mutant orthologs of the mouse genome into two groups as causing a ‘phenotype’ and ‘no phenotype’. Phenotype genes tend to be disease genes more than genes in a no phenotype group. The phenotype genes were further classified into ‘lethal phenotype’ and ‘non-lethal phenotype’ subgroups. Based on the results of the logistic regression, the frequency of disease genes in the lethal phenotype genes was 38% (odds ratio = 3.02) and 34% in the non-lethal phenotype (odds ratio = 2.47), while the frequency of disease genes in no phenotype was only 17% with odds ratio of 1.16. This suggests that lethal genes in mice can have a disease-causing orthologue in humans (55).

1.4.1 Essential gene prediction

In the field of genomics, identifying essential genes relies mainly on laboratory methods like gene knockouts (56), RNA interference (RNAi) (57), CRISPR/Cas9 (58,59), transposon mutagenesis (60), and antisense RNA (61,16).

Gene knockouts is a way of studying gene function by investigating gene loss. To this end, a gene is knocked out in a model organism and the phenotype is observed (56). Another way of testing gene function is RNA interference—a process of inhibiting gene expression or translation by using either microRNA (miRNA) or small interfering RNA (siRNA). The mechanism of RNAi is through an enzyme complex degraded DNA methylation at genomic sites corresponding to siRNA or miRNA. This technique can be used in model organisms to determine gene function, discover new drugs, and study cellular processes to predict gene essentiality (57).

Meanwhile, CRISPR/Cas9 is a method that can be used for gene editing. Here, the enzyme Cas9 is one of the enzymes produced by the clustered regularly interspaced short palindromic repeats (CRISPR) system. This enzyme has the ability to cut the DNA and turnoff the targeted gene, which facilitates the studying of gene functions (59). Later, a comparison was made comparing the roles of RNAi screens and CRISPR/Cas9 in identifying essential genes in human chronic myelogenous leukaemia (CML). The results demonstrated that the accuracy of both libraries in identifying essential genes was similar, suggesting the benefits of combining data from both screens (58).

A further method of identifying nonessential genes is to use global transposon mutagenesis, which is a method used to study whether the naturally occurring gene complement is a true minimal genome under laboratory growth settings. This is performed by applying

transposon mutagenesis to completely sequenced genomes, allowing for the accurate localization of insertion sites with respect to each of the coding sequences (53). However, not every transposon insertion within a gene will lead to gene function disruption. Nevertheless, if it is done near the 3′ end of a gene, it may not destroy gene function. Likewise, an insertion close to the 5′ end of a gene does not always affect gene function (53).

Moreover, antisense is a technique used to inhibit gene expression in a diverse organism. It has been used in combination with regulated expression for rapid identification and characterization of essential genes from the human pathogen, *Staphylococcus aureus* (61). Further, the authors created many defined strains exhibiting conditional growth phenotypes. These bacteria, in which certain genes are controlled by antisense RNA, were used to examine the impact of staphylococcal gene products on growth in bacterial cultures and animal models of disease (61).

Due to the fact that the number of genes in the genome is massive, these experimental strategies are certainly sophisticated. However, the complexity of the human genome makes these techniques difficult to apply in humans. For this reason, researchers have focussed on developing computational techniques to predict essential genes and proteins (16).

1.4.2 Essentiality hypothetical model

Recently, a hypothetical model that outlines potential relationships between gene essentiality, recombination, and selection, and how these relate to gene categories including disease genes, essential and non-essential genes (see Figure1-7) was described. In this project, the strength of LD for candidate genes are considered along with other genic scores using the same gene groups used in Pengelly et al. (62).

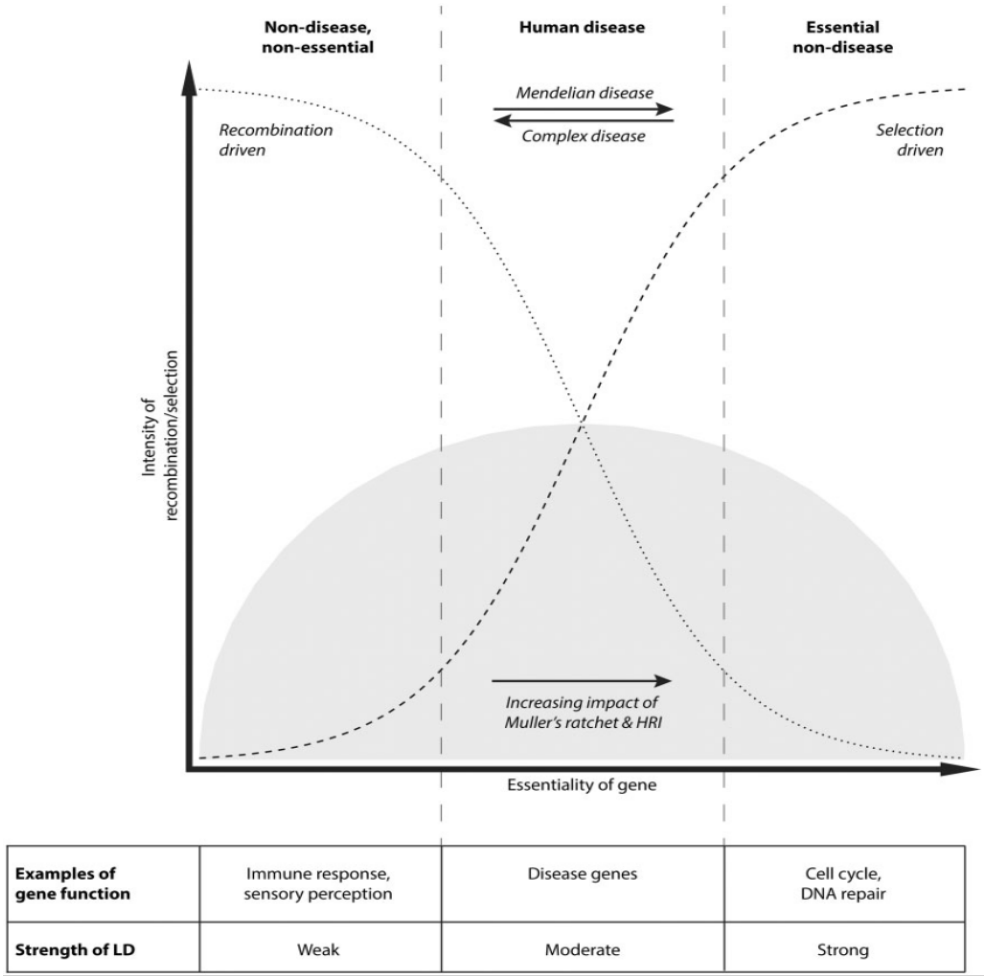


Figure 1-7 Hypothetical relationship between gene essentiality, recombination, and selection and different gene groups including non-disease, non-essential, human disease and essential non-disease. Genes towards the high essentiality end of the spectrum tend to be highly intolerant to LOF variants, have a low recombination rate, and are strongly impacted by negative selection. While genes in the non-essential end tend to be less intolerant to LOF variants, with high recombination rates and less impacted by negative selection. Adapted from Pengelly et al. (62).

As demonstrated in Figure1-7, an increasing gene essentiality measure was assumed for genes involved in core biological functions and vital for survival; any disruption in this group of genes is fatal at the pre-natal stage. Highly essential genes tend to have strong LD with low haplotype diversity. Since these genes are mutation intolerant, they tend to have a low recombination rate (62). At the other end of the spectrum, less essential genes, which are involved in sensory perception, for example, are more tolerant to mutation and may have high recombination rates. This group of genes tend to have high haplotype diversity, weak LD, and are weakly impacted by selection (62). According to this model, monogenic genes

occupy the intermediate position that is shifted towards the essential end, and complex disease genes are closer to the non-essential end.

1.5 Methods to predict gene pathogenicity

1.5.1 Variant-level predictors

Several tools have been produced to predict the potential impact of genetic variants on the function of gene or proteins. These tools use a variety of algorithms that can measure one or more biological feature(s) to predict variant deleterious impact (63). The following scores predict gene conservation: phastCons (64), GERP++ (65), and phyloP (66); variants where the homologous position in other species has persisted as constrained over evolutionary history and are scored as high deleterious variants (63). Another group of variant-level scores focus on prediction of the effect of protein function through disruption of the amino acid sequence such as FATHMM (67), SIFT (68), fathmm-MKL (69), and PolyPhen2 (70) (63). Numerous scores are described in the literature, which predict the possibility of a gene having a LoF mutation. However, until now, no single measure is entirely reliable in predicting gene pathogenicity. Here, the integration of gene-level scores may help in identifying disease causal genes for monogenic diseases.

SIFT

Genetic mutation studies have identified substitutions of a.a (amino acids) in protein-coding regions. These might affect protein function and produce the disease phenotype. In this context, the sorting intolerant from tolerant (SIFT) score is a variant- level predictor that predicts whether an a.a substitution disturbs protein function, so that these substitutions can be studied further. SIFT score differentiates between the functionally neutral and damaging alterations in a.a sequences (68). The prediction of a SIFT score is based on the assumption that important a.as will be conserved in the protein family, and any alteration at well-conserved positions are predicted as damaging. This can be achieved by choosing related proteins in a given protein sequence and obtaining alignments of this protein. Based on the amino acids at each position in the alignment, SIFT estimates the probability that an a.a is tolerated at a given position (68).

PolyPhen-2

This is a tool that predicts the deleterious effects of missense variations. There are three main differences between PolyPhen-2 and the previous PolyPhen tool: the set of predictive features, alignment pipeline, and method of classification (70). PolyPhen-2 uses three structure-based predictive and eight sequence-based features. These features distinguish the wild (ancestral) type from the mutant (derived) allele. Further, Polyphen-2 was tested against PolyPhen and found to be consistently superior (70). More specifically, two pairs of datasets were used to test and train PolyPhen-2. The first was HumDiv from approximately 3,100 damaging alleles annotated in the UniProt database as causing human Mendelian diseases and affecting protein function or stability, along with approximately 6300 differences between human proteins and their closely related mammalian homologs, presumed to be non-deleterious (70). The second dataset was HumVar, in which there are around 13,000 human disease-causing variations from UniProt (71) and approximately 8,900 human nonsynonymous variants (single-nucleotide polymorphisms) without annotated involvement in diseases that are considered as non-damaging. The false positive rate was 20%, in which the HumDiv achieved true positive prediction rates of 92% compared with 73% achieved by HumVar. The lower accuracy of the HumVar prediction could be due to the fact that the nsSNPs presumed to be nondamaging in the HumVar dataset included a sizable fraction of mildly damaging alleles. On the other hand, the majority of amino acids replacements presumed non-deleterious in HumDiv dataset must be most likely close to selective neutrality (70).

Moreover, PolyPhen-2 calculates the naive Bayes posterior probability that a certain mutation is deleterious and produces estimates of false and true positive rates. The former is the possibility that the variation is categorised as deleterious when it is in fact, non-deleterious, while the latter is that the variation is prioritised as deleterious when it is indeed a damaging mutation. Thus, HumVar-trained PolyPhen-2 is recommended for predicting Mendelian disease variations (70).

CADD

Combined annotation dependent depletion (CADD) integrates a number of different annotations and produces an individual quantitative score per variant. This is implemented using machine learning (support vector machine), trained to distinguish approximately fifteen million high-frequency human-derived alleles from fifteen million simulated variants (72). In CADD, ‘C-scores’ are computed for around 8.5 billion potential human SNVs and

the short insertions and deletions are then allowed to be scored. C-scores are associated with allelic assortment, functionality annotations, deleteriousness, severity of disease, complex trait associations, and highly ranked known deleterious variants within individual genomes (72). CADD scores integrate various genome annotations and give scores to candidate SNVs or small insertions/deletions (indel).

MutationTaster

This is a free, web-based tool developed by Schwarz et al. (73) for the quick assessment of the disease-causing potential of DNA sequence variations. This tool integrates various bits of information from different biomedical databases and uses well-established analysis applications. The analysis results of MutationTaster are reportedly completed within 0.3 seconds, providing information for the analysis of splice-site changes, evolutionary conservation, loss of protein features, and changes that might affect the amount of mRNA. The test results are assessed by a naive Bayes classifier that prioritises disease-causation (73). The outcome of a MutationTaster prediction could be silent synonymous, intronic variations or variations affecting a single a.a or causing complex alterations in the a.a sequence. However, this tool has some limitations: first, its inability to analyse insertion/deletions greater than 12 base pairs (bp); second, the analysis of non-exonic alterations is restricted to the Kozak consensus sequence, splice sites, and poly (A) signals (73). To overcome these limitations, **MutationTaster2** was developed to predict the functional consequences of not only amino acid substitutions but also intronic and synonymous alterations, short insertion/deletion (indel) mutations, and variants spanning intron-exon borders (74). It includes all freely available (SNPs) and indels from the 1000 Genomes Project along with known disease alterations from human genome mutation database (HGMD) Public and ClinVar. Any variation found more than four times in the homozygous state in HapMap or 1000G is considered neutral. Thus, the automatic prediction of a variant to be disease-causing if the variant was prioritised automatically as pathogenic in ClinVar and the identification of the presence of the disease phenotype was possible (74).

REVEL

This is an Ensemble method to prioritise pathogenic rare missense variants. It incorporates 18 pathogenicity prioritisation scores (features) from 13 different tools as predictive features

as follows: phyloP (primate) (66), phastCons (primate) (64), phastCons (placental), SiPhy (75), GERP++ RS (65), phyloP (placental), MutationTaster (73), LRT, phyloP (vertebrate), phastCons (vertebrate), FATHMM (67), MutPred (76), SIFT (68), MutationAssessor (77), PROVEAN (78), VEST (79), Polyphen2 HDIV, and Polyphen2 HVAR (70)). The MutPredictor was constructed for the purpose of this study with the UniProt canonical protein sequence when available and the Ensembl canonical transcript otherwise. FATHMM (67), VEST (79), Mutation-Assessor (77), MutPred (76), and PolyPhen-2 HVAR (70) were the most important features in REVEL. Further, the important measure of each of these five features reflects correlations with other features in addition to this method's intrinsic prioritisation ability, as significance can be mutual among correlated features. The performance of REVEL (80) was compared to other ensemble methods—such as (MetaLR (81), MetaSVM (81), KGGSeq (82,83), CONDEL (84),CADD (72), DANN (85),and Eigen (86)—for distinguishing HGMD disease variations (rare variants) from neutral exome sequencing variants with allele frequency (AF) ranging from very rare (0.1–0.3%) to common (> 5%). It was found that all of the ensemble methods performed poorly as compared to REVEL, which had the best scoring in predicting disease variants from uncommon neutral missense variants with an AF below 3% (80). Thus, this method has the following strengths: first, REVEL was trained and tested on recently identified pathogenic and neutral variants that resemble novel variants identified by future NGS studies, which are likely to comprise variants with lower allele frequencies and more modest effects than previously identified variants; second, to improve performance when interpreting rare variants, the AF of REVEL neutral variants were restricted exactly between 0.1% and 1%; third, a larger number of individual predictors were incorporated by REVEL than any method prior to ensembling (80).

1.5.2 Gene-level predictors

From 2005 onwards, the number of gene-specific pathogenicity predictors has increased dramatically. Various measures considered in this project include the following: residual variation intolerance score (RVIS), which is an essentiality predictor measuring gene intolerance to rare and common functional variation (CFV); pLI, which is another essentiality predictor that measures the probabilities of a gene carrying LoF variation; and substitution intolerance score (SIS), measuring the probability that a gene is intolerant to

functional variation using data from the 1K genomes project; and other scores that predict essentiality, HI, and selection (87,88,38).

1.6 Research question framework

This project aims to assess how gene-specific features predict the susceptibility of genes to monogenic disease variation. To this end, a gene-based score was developed and evaluated to allocate and prioritise disease causal genes. Additionally, this score was applied to monogenic disease data to aid in identifying molecular causes for those conditions. This aim can be achieved by answering the research question presented in Table 1-1, which was formulated using the PICO (population, intervention, comparison, and outcome) format. This format needs to be directly relevant to a problem or a patient and is widely used in evidence-based practice. It was designed to facilitate answering the research question by dividing the question into four parts to digest the main idea of the research (89). Thus, this project aims to answer the following research question: 'Can the use of gene-specific metrics facilitate the identification of disease genes in patients with monogenic diseases?'

Table 1-1 Research question framework

Population	Intervention	Comparison	Outcome
Patients with monogenic diseases	Using gene-specific metrics	None	Facilitated identification of disease genes

1.7 Thesis outline, aims, and contribution

The role of gene-specific metrics in the identification of Mendelian disease genes is not yet established. Thus, the aim of this thesis was to apply gene-level approaches to predict disease genes for monogenic diseases. Utilising the available range of gene specific scores in addition to the integration of those metrics might potentially improve the prioritisation of disease genes.

Aim 1—Systematic review summarising gene-specific scores using the snowballing technique.

The aim of chapter 2 was identifying gene-specific scores described in the literature and utilising them to help in prioritising disease genes, and the specific objectives are as follows: 1. Identifying an effective study design to cover the whole literature in predicting disease genes at the gene level; 2. Classifying the identified scores in a way that can correlate each group together.

For this study, 20 gene-level predictors were found, and each score predicts certain genetic features. These scores were then classified into three main classes that predict essentiality, HI, and selection. This classification was based on score properties and what they are intended to measure. The aim was to identify gene-level scores that provide scores per gene to create a base for the next project, which is constructing a composite score by combining these predictors together. Among the twenty gene-level scores identified in the systematic review, 10 scores provided scores per gene. Combining these metrics into a single composite score established the foundation for building the Essentiality specific pathogenicity prioritisation (ESPP) score.

My contribution was in choosing an appropriate method (study design) and using it to achieve broad coverage of the literature by gathering information about each predictor and allocating specific metrics that might have potential when integrated together to produce a composite score to predict disease genes.

Aim 2—Devising methods to facilitate the construction of an essentiality predictor.

The aim of chapter 3 was preparation to build an essentiality score that predicts disease genes. The objectives were to find the highest scores (from the scores identified earlier from the systematic review) representing the data and following the same direction of the essentiality hypothetical model proposed by Pengelly et al. (62) to evaluate and update the Spataro et al. classification (90) to improve the prioritisation of disease genes.

My contribution was gathering information of all scores that can be used in the composite score, evaluating the relationship between those scores, correlating them with Spataro groups and undertaking all the statistical analysis.

Aim 3—Constructing a composite gene score: ESPP.

Chapter 4 shows the development of a gene essentiality predictor that will help in predicting disease causal genes. The aim was to integrate gene-level scores that align with gene properties that can be placed on a broad ‘essentiality’ scale corresponding to that described

in the Pengelly et al. (62) hypothetical model, to create an essentiality score that predicts disease genes and distinguishes them from the essential genes. In order to achieve this, the gene classification proposed by Spataro et al. (90) was used. This classifies genes into five groups from the least to most essential. Using principal component analysis (PCA), 10 scores identified from the systematic review were integrated with another additional score measuring LD per gene obtained from Vergara-Lope et al. (91) to produce the ESPP score. To improve the results, several steps were undertaken to refine the model prediction and provide better separation between disease and essential genes.
The work conducted in this chapter was carried out by this author under the supervision of Prof. Andy Collins and Prof. Diana Baralle. My contribution consisted of every step outlined in this chapter including the curation of the research database, identification of possible gene classification strategies, performing of statistical tests, interpretation of results as well as pipeline and score development.

Aim 4—Using/integrating scores to predict new Mendelian disease candidates.

Chapter 5 tests the performance of the ESPP score across several data sets. The aim was to identify unknown candidate genes that score high based on this study's ESPP score; these genes are expected to be candidate genes for monogenic diseases. These will be tested using the ESPP score across available data. The first objective was to test the ESPP score in predicting dominant and recessive genes. The second objective was to compare ESPP score with the most recent gene-level predictors.
The work conducted in this chapter was carried out mostly by this author under the supervision of Prof. Andy Collins and Prof. Diana Baralle. My contribution consisted of every step outlined in this chapter with help in accessing some of the data, I acknowledge the help of Dr Jenny Lord in accessing DECIPHER and Dr Gabrielle Wheway in accessing GEL data. I tested the performance of ESPP using the previous mentioned data and the Saudi data in addition to the comparison of ESPP with LOEUF and CoNeS.

Aim 5—Constructing a recessive gene score: RSPP (future work).

As the distribution of recessive disease genes by ESPP were not as clear as the dominant disease genes, I decided to build another classifier to improve prediction of recessive disease genes (Figure 1-5). Chapter 6 demonstrates the future plans to construct recessive-

mode specific pathogenicity prioritization (RSPP) score and test the score using Genomics England (GEL) and Saudi Human Genome Programme (SHGP) data.

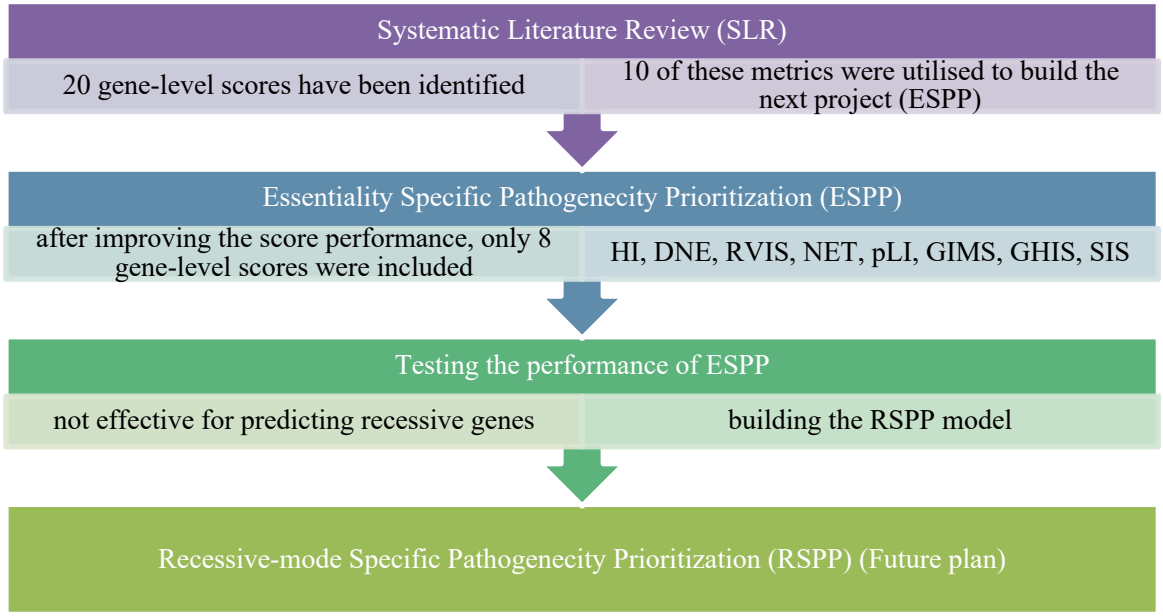


Figure 1-8 Thesis pipeline and future extension

Chapter 2 Literature Review of Gene-Specific Pathogenicity Prediction Scores

2.1 Introduction

The advancement of NGS technologies has been an exceptional contribution to the detection of deleterious variants in previously undiagnosed diseases. Nevertheless, in genome sequencing studies, because of the large number of variants seen (on average around 10,000 variants/genome that cause alteration in the amino acid sequence (92)), a major challenge is distinguishing which of these sequence variants are truly deleterious.

Several variant prediction tools have been constructed to improve detection of disease-causal sites. However, as many variants scored as deleterious by these tools are often in fact tolerated, the results might be equivocal. Further, there is a lack of consistency in the prediction results among the various variant-level prioritization tools (93). Latterly, researchers believe that understanding gene properties, such as selection, mutation, recombination, and HI, may help in the prediction of genes that might be disease causing and that utilizing this information will refine molecular diagnostics (93). Thus, the motivation behind this review is to assess how understanding gene-specific features may improve filtering strategies in clinical sequence data to allocate possible disease variants. Additionally, the ability to recognise the 'disease genome,' which comprises coding, non-coding, and regulatory variation, might lend a hand in resolving undiagnosed cases. This review delivers an extensive assessment of existing gene prioritizing methods, the association between measures of gene-deleteriousness, and how utilizing these prediction tools can be improved for molecular diagnostics.

2.2 Systematic literature review

2.2.1 Introduction

NGS and whole genome sequencing, in particular, produce enormous datasets that generate substantial analytical challenges for the prediction of disease-causal genes. Within the scope of knowledge, a subgroup of human genes contains, or are related to, rare and/or common

variations, which are important in disease formation (the 'disease genome'). Nevertheless, the identification of causal variants amongst many thousands of mostly neutral variants is not entirely established and a pressing challenge. For instance, Chong et al. (39) state that the genes underlying 50% of all Mendelian diseases are still unknown, and many Mendelian conditions are yet to be described (39). Further, according to OMIM (24), currently, this percentage has decreased by almost 20% (39). In conjunction with methods for anticipating the potential deleteriousness of individual DNA variants, a sum of gene-specific scores was developed, which may help facilitate the recognition of disease-causing variations. Recognising the disease genome properties and combining existing models that score genes may help in categorizing genes based on their specific characteristics to improve molecular diagnosis (93). However, there is conflicting evidence about the reliability of pathogenicity scores for individual DNA variants due to the inconsistencies in the results of the different methods used (94). Owing to redundancy in the genome, predicting the potential causal variation can be challenging. Thus, this study proposes an integrated approach that appraises evidence at a variant level and with a broader scope (gene level). To this end, it was identified that variant prioritization metrics alone are presently not accurate and evidence at the gene-level has the potential to improve the prediction of variant pathogenicity (93). This systematic review brought together the literature related to gene-level approaches and their relevance in enhancing filtering of genome sequence data. The goal is to generate a satisfactory answer to the following research question: 'Can the use of gene-specific metrics facilitate the identification of disease genes in patient genomes?' It is presented in Table 2-1 below.

Table 2-1 Systematic literature review question framework

Population	Intervention	Comparison	Outcome
Patient genomes	Using gene-specific metrics	None	Facilitated identification of disease genes

2.2.2 Methodology

The research question framework was structured using the PICO framework presented in Table 2-1. This framework was particularly chosen to enhance specificity and eliminate the conceptual uncertainty of clinical issues (89).

Moreover, a systematic literature review (SLR) uses a systematic method to examine the literature related to a particular topic by collecting and criticizing secondary data to answer a specific research question. The top of the hierarchical pyramid of scientific evidence is occupied by SLR and meta-analysis, considered the highest level of evidence (95). The difference between a regular review and the SLR is that the latter is formally designed to cover the entire literature related to a well-defined research question and is conducted in a systematic manner (96). For the sake of achieving a high level of validity, SLR cautiously selects the available literature in the area of interest and combines scattered evidence into a single article (95,96).

The purpose of this review is to clearly identify articles making a central contribution in the understanding of the disease genome. The double-sided systematic snowball strategy (SB) was preferred over a keyword-based reviewing method due to the difficulties in selecting appropriate keywords for the search strategy as a consequence of the paucity of available gene-specific scores in the literature. Additionally, the SB technique is a systematic way to screen the literature by collecting papers that are closely related to the area being studied, called the start set. This is followed by screening of the bibliographies of these papers for any related articles, and this is repeated for every new article identified until there are no more related articles to be found—this is called backward snowballing (BSB). On the other hand, forward snowballing (FSB) also uses a start set of papers, scans all citation papers and does the same for any new citation identified until no new article can be found. The term snowballing is used in this context as the search strategy can be started with small number of papers that increases gradually until numerous papers that cover the topic of interest are found.

Moreover, the SB technique has been found to be more competitive compared to a database search when using general terms and reducing the amount of noise (97). In addition, it is reported to be more efficient at finding relevant articles: it identified 85% relevant papers compared to a database search that identified 45.9%. Ultimately, the SB technique shows higher reliability when the starting set of papers is chosen as compared to a database search (98).

An extensive scoping search was undertaken to make sure that there was no systematic review addressing the same area of study. The scoping search was conducted through four major databases: Cochrane Library, Prospero, The Trip and Evidence search (NICE) along with two major platforms (EBSCOhost and OvidSP) to ensure the originality of the topic and guarantee a comprehensive screen of the literature. The EBSCOhost platform comprises three mega databases, which are MEDLINE, CINAHL, and SportDiscus, while OvidSP hosts Medline, Embase, and EBM reviews. This was followed by a screening of grey literature through Open Grey and MedNar databases, but no relevant SLR was identified. More details on the scoping search approach are shown in Appendix A.

Four search key phrases (KP1-4) were set to identify the start set of ('seed') papers:

KP1: 'Gene-level approach.'

KP2: 'Protein coding variants.'

KP3: 'Pathogenicity prediction at gene-level.'

KP4: 'Pathogenicity prioritization at gene-level.'

The four search key phrases were selected based on the scoping search that was initially performed.

To conduct a high-quality systematic review and among several search databases, Google Scholar (GS) was selected for the snowballing search. A comparison was conducted between GS and other trusted sources of information, such as Cochrane Database Systematic Review and The Journal of the American Medical Association (JAMA), to assess the coverage of GS, especially in identifying systematic reviews (99). The authors selected original articles that were involved in around 29 systematic reviews identified in aforementioned information sources in 2009. Then, they searched for all these papers in GS to predict the adequacy of coverage if it is used as the only source. They concluded that GS provides broad coverage for studies included in the systematic review, reaching 100% coverage (99). For this reason, GS was chosen as the leading search engine for this study. In addition to the high sensitivity, it had already been approved when used alone for systematic reviews (99). Further, besides the broad variety of articles and resources within its database, the usage of GS within the research environment has increased (100,101). A further advantage is that it is not restricted to particular publishers (98).

Initially, five original papers were identified, and these were the seed set of papers chosen to start the SB technique. As any starting sets have limitations, it was assured that using the double-sided SB technique instead of using the single SB strategy (either backword or forward SB) would widely cover the literature.

Conducting backward snowballing

For the retrospective systematic snowball technique, the bibliographies of the initial (n = 5) seed articles were used as a starting point. The references in those five papers were analysed in three phases (Figure 2-1). The following criteria were used to review the bibliographies of each paper. The inclusion criteria were as follows: (i) 'Papers published during a 2005–2018 timeframe'; (ii) 'peer-reviewed papers'; (iii) 'English language;' The exclusion criteria were as follows: papers that did not offer a score per gene or that described a method to predict gene-specific properties; those published prior to 2005 because of limited genome sequencing and weak understanding of gene-specific features before then (102). As a result of the BSB, n = 6 of new relevant papers were identified based on the inclusion and exclusion criteria. The BSB process is illustrated in Figure 2-1. The 11 identified papers (n = 5 (seeds) + n = 6 (BSB)) were made the starting set of papers to conduct the FSB.

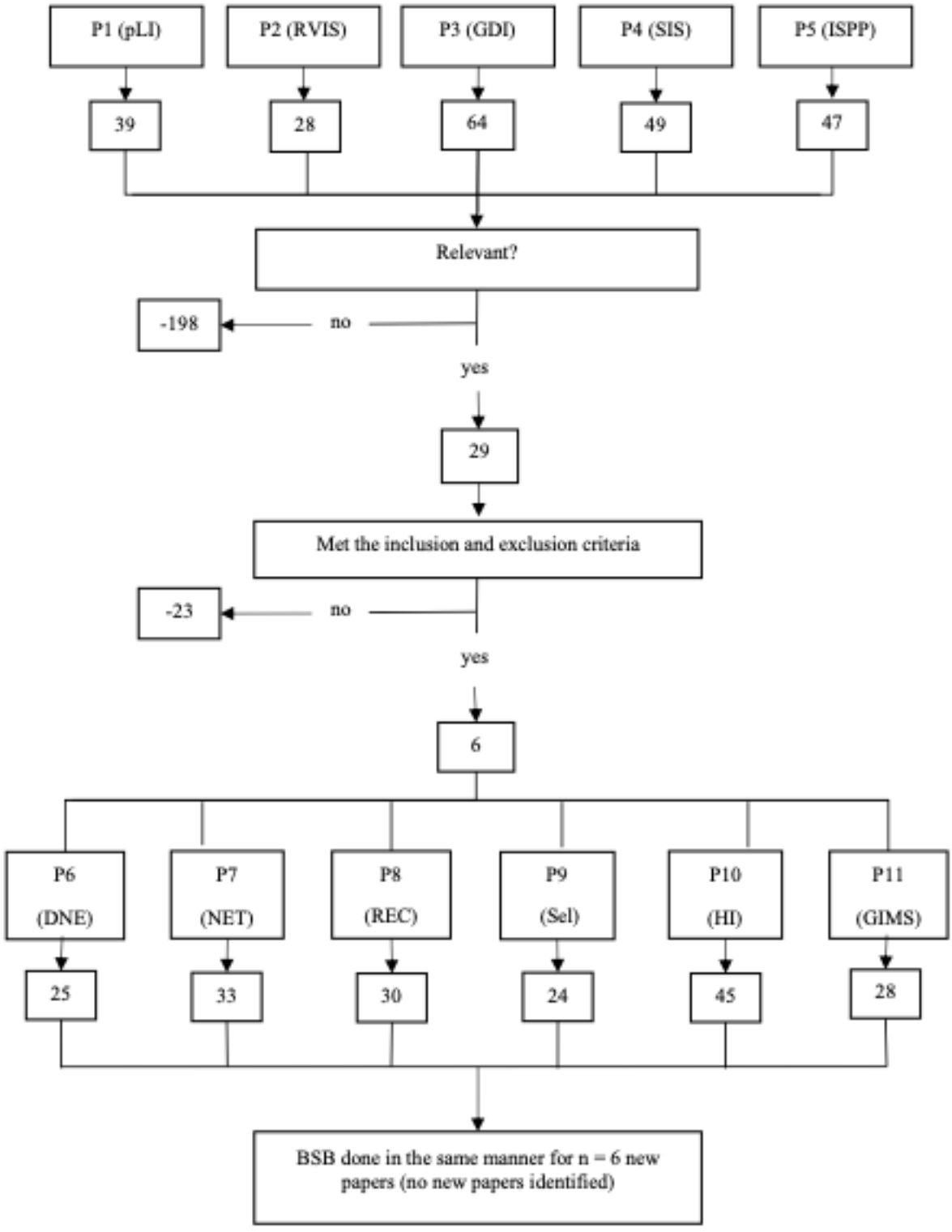


Figure 2-1 Backward snowballing search process: (i) The initial phase—scanning of the bibliographies of the initial set of papers was demonstrated including Loss Intolerance

probability (pLI), RVIS, gene damage index (GDI), SIS, and inheritance-mode specific pathogenicity prioritization (ISPP). The relevant papers were identified based on title, abstract, introduction, and results when needed, and sometimes, more sections were reviewed. In this phase, the review of the whole text was not done; (ii) The second phase—application of the inclusion and exclusion criteria; (iii) The third phase—in-depth analysis of n = 6 new papers was conducted to confirm relevance including gene constraint score-*de novo* excess (DNE), gene position in networks (NET), indispensability score, recessive (REC) score, negative selection (Sel), deletion-based HI score, gene-level integrated metric of negative selection (GIMS). Then, it was run through BSB again in the same manner as for the initial set of papers (three phases of analysis) until no further papers were found.

Forward snowballing

The following steps were undertaken for each of the 11 articles:

- Identifying the article in Google Scholar;
- By using the citation link, examining all citing papers;

The frequency of citations for each of the n = 11 papers were as follows: P1—1710; P2—405; P3—51; P4—29; P5—4; P6—325; P7—92; P8—731; P9—723; P10—311, and P11—6.

For all papers P1–P11, 4387 citations were retrieved using FSB. At the start, the relevant papers were identified by using a three-phase analysis (as explained in Figure 2-2 legend) in the same way as BSB (Figure 2-1). The analysis of the citations and the bibliographies was undertaken in February 2018. A number of (n = 9) papers were identified that met the previously defined inclusion and exclusion criteria for prospective snowballing; these nine papers were provisionally added to the n = 11 primary papers; accordingly, a total of 20 papers were included in this systematic review. A critical analysis for these twenty papers was obtained to try to understand how a gene's pathogenicity was determined and to identify how these gene-specific scores might improve detection of genes that are disease-related, while also highlighting that the improvement of the filtering of sequence data was the primary focus.

Before extracting any data from the reviewed articles, quality assessment of the papers identified was undertaken. Since all included articles were published in peer-reviewed and

reputable journals (such as Nature Genetics, Nature, PLoS Genetics, Science, PNAS, Oxford Academic), no further quality assessment was undertaken.

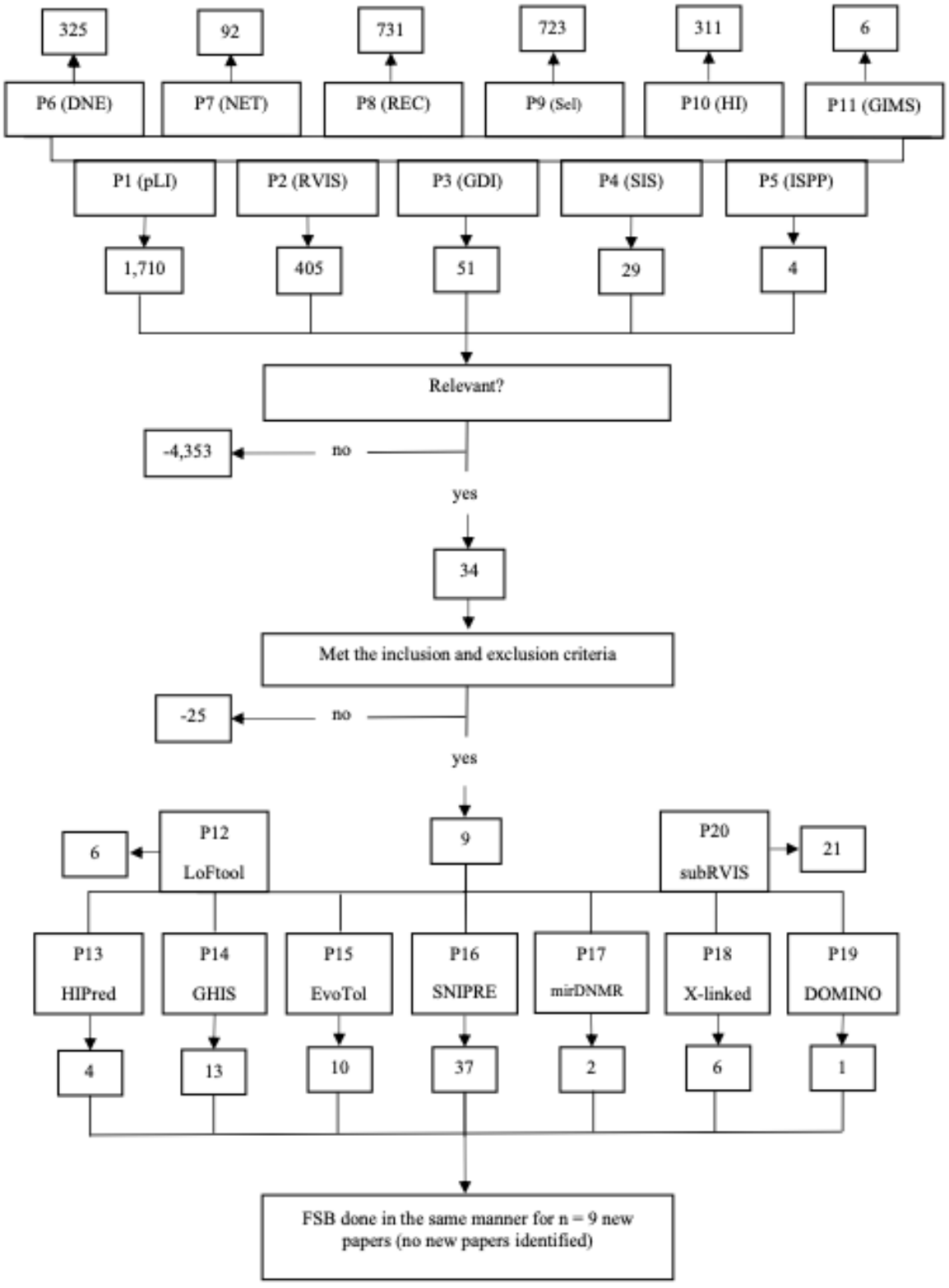


Figure 2-2 FSB search process. (i) The initial phase: scanning of the citations of n = 11 papers including DNE, NET indispensability score, REC score, negative selection (Sel), deletion-based HI score, GIMS, pLI, RVIS, GDI, SIS, and ISPP, which were identified

through BSB. The relevant articles were provisionally included based on title, abstract, introduction, and results when needed; sometimes, more sections were reviewed as well, In this phase, review of the whole text was not done; (ii) The second phase: application of the inclusion and exclusion criteria (iii) The third phase: analysis of n = 9 new papers was made in depth to confirm relevance including gene intolerance score based on loss-of-function variants (LoFtool), haploinsufficiency predictor (HIPred), genome-wide haploinsufficiency score (GHIS), evolutionary intolerance (EvoTol), Selection Inference Using a Poisson Random Effects Model (SnIPRE), *de novo* mutation rate (mirDNMR), X-linked (XL) score, machine learning to predict genes associated with dominant disorders (DOMINO),and sub-regions residual variation intolerance score (subRVIS). FSB was undertaken again in the same manner as for the initial set of papers (three phases of analysis) until no further papers were identified.

2.2.3 Results

Findings: Key models

A total of 20 models retrieved by the systematic review method were classified into three groups determined by the concept of each approach and the corresponding scores: (i) Essentiality (conservation), (ii) HI, and (iii) Selection.

Characteristics of essential (conserved) genes

Essential (conserved) genes are considered genes that are required for existence; they encode proteins that have vital biological functions, playing a crucial role in an organism's viability. However, each gene might be considered to have a different degree of essentiality, and there are numerous measurable scores estimating gene essentiality. To assess the level of essentiality for a given gene, several methods can be used starting from the estimation of the expected rate of *de novo* mutations for that gene to the prediction of whether that gene is tolerant or intolerant to a LoF mutation (62). Table 2-2 summaries the critical approaches in this category.

The RVIS prioritises genes by the likelihood of carrying more, or less, functional genetic variation than expected, thus highlighting genes intolerant to CFV (87). This uses the excess of rare versus common missense variation within the human genome and highlights genes intolerant to CFV. Therefore, the number of observed variants in a given gene can be

compared to the observed CFV. Moreover, genes involved in monogenic diseases were found to have lower RVIS scores than other genes. Positive scored genes indicate that they are prone to have more CFV, whereas genes that scored negatively are less tolerant and have less CFV.

Meanwhile, Rackham et al. constructed the evolutionary in tolerance score (EvoTol) based on the evolutionary conservation of protein sequences to predict genes that are intolerant to mutation (103,104). Due to the fact that only a small part of a gene may be intolerant—for instance, the protein-coding region—these sub-regions, particularly, might be considered essential (104). Further, EvoTol makes the prioritization of intolerant protein sub-domains in conjunction with the prediction of intolerant genes more generally.

Moreover, the advancement of NGS technologies facilitates the detection of newly arising (*de novo*) mutations (DNMs) and their possible effect in causing monogenic disease. This type of mutation is not considered to have a role in the development of complex diseases (105). Further, the meticulous evaluation of gene mutability is needed to predict the expected rate of *de novo* mutation in a given gene. Two essential factors underlie mutation rate differences: gene length and local sequence context (11). Based on the observed counts of rare missense variants in the Exome Sequencing Project (ESP) data set, Samocha et al. calculated the possibility of a gene being mutated. Taking into consideration the depth of coverage, which is defined as how many sequence reads were present on average per base and the regional divergence in genes between humans and macaques (105). Here, they were able to expand a model that evaluated *de novo* mutations in epileptic encephalopathy patients (Epi4K consortium). Several genes with missense variant deficits were observed and compared to expectations from expected mutation rates. The deficit indicates a strong evolutionary constraint with damaging variation that was removed from the population by negative selection (105,106). The Samocha et al. score utilises whole exome sequence (WES) data to calculate the DNM rate (DNMR) for a single gene, and on a gene set basis (105); this model is referred to as '*de novo* excess'. Additionally, this model can predict the selective constraint in the human genome, and recognises approximately 1000 constrained genes that are known to cause severe conditions (105). Here, it is worth noting that constrained genes are prone to a higher *de novo* LoF mutation rate than expected by chance (105).

The LoFtool calculates the ratio of LoF mutations to synonymous mutations for every single gene. The performance of RVIS, DNE Z-score, and EvoTol were compared with the

performance of the LoFtool, which proved to have a better performance in predicting *de novo* haploinsufficient disease-related genes. The values of LoFtool are represented as intolerance percentiles: a low LoFtool percentile indicates a less tolerant gene to LoF mutation (103). All the genic intolerance predictors outlined so far were described by Bartha et al. as essentiality scores (11).

In 2016, Aggarwala et al., proposed the 'substitution intolerance score', a gene-specific predictor of essentiality calculated using data from the 1000 Genomes Project. High functionally constrained genes scored high based on this scoring system, whereas genes that are more tolerant to the functional mutation in the protein scored low; this might have arisen through mutations in the DNA sequence (88).

Meanwhile, SubRVIS is another gene scoring model produced by Gussow et al. that predicts genic intolerance in sub-regions to identify where deleterious mutations are likely to fall within genes, suggesting that regions that are highly conserved are more likely to contain more deleterious variants (107). The ranking of these regions was based on RVIS with a combination of data on conservation. Regions intolerant to functional variation were shown to have low scores by the subRVIS prediction model. Here, the GERP++ score was calculated to estimate evolutionary constraint for bases in each sub-region (107).

By quantifying the probability that a gene is intolerant to a mutation that produces LoF in the protein product, Lek et al. produced the pLI score (38). The score is derived from an extensive catalogue of human genetic diversity called the ExAC database. This database provides a substantial filter for analysis of candidate pathogenic variants in severe Mendelian diseases; this is due to its ability to detect one variant for every eight bases on average in the exome, which provides evidence for the existence of extensive variant recurrence (38). The pLI score estimates the AF for genetic variants in protein-coding regions. Genes intolerant to LoF mutation are those scored high by the pLI score (pLI > = 0.9), indicating the most evolutionarily constrained genes. Low pLI scores (pLI < = 0.1) are usually associated with LoF-tolerant genes, typically involved in the least constrained biological pathways, like sensory perception, wherein high haplotype diversity is probably beneficial (38).

However, difficulties remain in the assessment of the relationship between the DNMR and genes involved in diseases. Recently, Jiang et al. (108) applied existing DNM data to correct for the background mutation rate, which was considered one of the main limitations of the Samocha et al. model (105) The issue was detected by sequencing more individuals, as more DNMs will inevitably be observed in the same gene by chance. Thus, for diseases

associated with *de novo* mutations, it is expected that disease-genes might contain more *de novo* mutations than expected from background rates. In this context, Samocha et al. describe the development of a database that defines the background DNMR acquired from population variation data (108).

Details on gene essentiality predictors with their specific properties are provided in the table below. Additionally, the approaches used to calculate each score along with access to information on each score when feasible are included.

Table 2-2 Gene essentiality and conservation metrics.

Essentiality measures	Non-essential genes	Essential genes	Score characteristics	Method	Weblink/Score data	Reference
RVIS	+++	+	Applying human population genetic data to predict pathogenicity and essentiality.	Linear regression of the number of common functional variants on the total number of variants	RVIS score data available for 18,860 genes in Supplementary Table S4, S1 sheet of Hsu *et al.,* 2016.	Petrovski et al. (87) Hsu et al. (109)
EvoTol	+++	+	Integrates intra and inter-species information (as RVIS). Considers intolerance of gene sub-regions (e.g., protein domains).	Linear regression analysis of the number of common functional variants on the total number of variants.	http://www.evotol.co.uk/	Rackham et al. (104)
DNE	+	+++	Powerful recognition of constrained genes. Employs a neutral mutation model as a baseline. The background rate of DNMs is the main limitation.	Z-Score (calculating the difference between the observed and the expected missense variants) based on a mutation model.	DNE score data available for 18,860 genes in Supplementary Table S4, S1 sheet of Hsu *et al*., 2016. This score referred as CONS score in Hsu *et al*. paper.	Samocha et al. (105) Hsu et al. (109)
LoFtool	+++	+	Non-parametric combination of functional prediction scores and mutation rates. Prediction of *de novo* haploinsufficient disease-causing genes.	Heuristic model.	None	Fadista et al. (103)
SIS	+	+++	Essentiality score depends on the following factors: sample ascertainment, population history, selection, and local context features that influence the rate of mutation.	Posterior substitution probabilities.	SIS score data available for 16,387 genes in Supplementary 3, table 15 of Aggarwala *et al*., 2016 [18]	Aggarwala et al. (88)

Sub RVIS	+++	+	Predicts genic intolerance to functional variation in gene sub-regions.	Logistic regression model.	None	Gussow et al. (107)
Sub GERP	+	+++	Predicts genic intolerance to functional variation in gene sub-regions. Utilises variant-level scores to predict genic intolerance.	Logistic regression model.	None	Gussow et al. (107)
pLI	+	+++	The probability that a gene is intolerant to a loss of function (LoF) mutation.	Posterior probabilities from a Poisson mixture model.	pLI score data available for 18,226 genes in Supplementary table 13, of Lek *et al*., 2016 paper.	Lek et al. (38)
mirDNMR (*de novo* mutation rate)	+	+++	Database predicts the background DNMRs by four methods based on: GC content (DNMR-GC), multiple factors (DNMRMF), sequence context (DNMR-SC), and local DNA methylation level (DNMR-DM).	Three statistical methods were used (TADA, Binomial and Poisson test).	https://www.mirdnmr/index.php	Jiang et al. (108)

Note: +, +++ is relative magnitude of score value.

Characteristics of haploinsufficient genes

HI is a mechanism where a diploid species has one copy of a gene missing and remains with a single functional copy that is insufficient to maintain normal function (93). LoF mutations typically cause HI leading to dominant diseases. Recognizing haploinsufficient genes may facilitate filtering of disease genome data, where the disease manifestation is expected to have arisen through lower levels of the gene product.

In 2010, a deletion-based HI model was proposed by Huang et al., and recognised differences between HI and haplosufficient genes to achieve better discrimination between deleterious and benign deletions, which, in turn, led to better variant prioritization (93). The analysis produced a logarithm-of-odds (LOD) score to better predict the likelihood of a deletion leading to a HI phenotype. Deleterious deletions are prone to high LOD scores through HI and subsequently, might produce dominant traits. However, no assumption of statistical interactions between the genes was considered in this model (93). To assess the deleteriousness of a deletion, the length of a deletion or the number of genes deleted were the main factors considered by clinicians.

The Huang et al. score provides a rational basis to classify pathogenic deletions by comparing deletions seen in patients with those in controls and calculating the fraction of controls with a deletion at least as deleterious as that seen in the patient (93).

Moreover, in the interest of achieving accurate molecular diagnoses, it is crucial to differentiate false-positive disease mutations from true causal mutations. In this context, MacArthur et al. applied their REC score for differentiating genes involved in recessive diseases from those that are LoF-variation tolerant (12). The non-disease 'healthy' genome contains approximately 100 genuine LoF variants, mostly in the heterozygous state. Further, in a single copy of the human genome, it was reported that there are on average five recessive lethal alleles. Consequently, a greater portion of LoF variants are considered as common variants, which may still have a phenotypic effect (12). Additionally, the alterations in functional and evolutionary properties between recessive disease and LoF-tolerant genes was also demonstrated by MacArthur *et al*. This is of great use in the development of a prioritization approach to predict recessive disease variants (12).

In order to investigate the relationships between the degree of network centrality of a gene and selection within biological networks, Khurana et al. considered a range of biological networks (i.e., phosphorylation, signalling, PPI, regulatory and genetic networks) and developed a score called 'gene position in NETworks' (NET) indispensability score (110). The degree of centrality of a gene in any network represents the number of its interacting partners in that specific network. Network centrality measures highlight nodes (each node represent a gene) based on their significance to the network topology to be able to identify critical genes and proteins in biological networks (111). Genes with significant functions are the ones with a high connection to several biological networks. Consequently, alterations in those genes might lead to severe conditions (110). Nevertheless, genes associated with metabolic networks had higher numbers of duplicated copies through a high number of paralogs with additional LoF mutations (110).

Additionally, the Khurana et al. metric was included as a HIPred score in the paper by Hsu et al. (109, 110). Further, the ratio of non-synonymous to synonymous substitution rates (dN/dS) for X-chromosome genes was used by Ge et al. to build their gene-level pathogenicity score (112). A low ratio indicates that the gene is intolerant to non-synonymous variation, signifying these are liable to disease-causal variation. This approach shows a correlation between genomic regions depleted due to missense mutation with disease-related variants (112).

Meanwhile, a study undertaken by Steinberg et al. proposed that current biases in many biological networks may interfere with the ability of the existing HI predictors to prioritise the true haploinsufficient genes. For the sake of eliminating study bias effects, they built a novel, unbiased HI score—called the 'genome-wide haploinsufficiency score' (GHIS)—by substituting biological networks with co-expression networks (113,114). The GHIS model was compared with the three pre-existing methods (i.e., HI (93), NET (110) and RVIS (87)), and it was suggested that GHIS-scored genes that were not scored by former approaches (113) performed better in categorizing less-well-studied genes (113).

To identify Mendelian genes with different inheritance modes, Hsu et al. developed a score by considering Mendelian disease gene properties based on their mode of inheritance. One of the essential properties of such genes, with an autosomal dominant (AD) mode of inheritance, is HI; this specific group of genes is sensitive to *de novo* mutations (109). On the other hand, disease genes with the autosomal recessive (AR) inheritance mode were recognised to have more non-synonymous variants and regulatory transcript isoforms (109). Conversely, the XL inheritance pattern is mostly related to less non-synonymous and synonymous variants (109). Hsu et al. utilised this information to construct a score predicting Mendelian disease genes, based on their mode of inheritance (AD, AR, and XL), called 'inheritance-mode specific pathogenicity prioritization' (109). The new score combines six pre-existing gene-level prediction approaches—i.e., HI (23), NET (25), REC (24), RVIS (13), GDI, DNE (16,35)—along with various genic characteristics including global expression from RNA-Seq data, the noncoding (intronic region) mutation rate, and DNA replication time (109). However, challenges remain in the prioritization of dominant mutations for monogenic disorders; due to the abundance of non-deleterious heterozygous variants in the human genome. Using machine learning, Quinodoz et al. produced DOMINO, a method to predict whether a given gene is liable to carry dominant changes (116). Nevertheless, this method does not provide a score per-gene.

Compared to less-studied genes, well-studied genes are unsurprisingly over-represented in most biological networks used to create metrics that estimate HI. Therefore, the study bias likely affects the majority of these networks. (114). By eliminating the effect of study bias and combining functional annotations with genomic and evolutionary characteristics to prioritise HI genes, Shibab et al. (114) created an integrated approach called (HIPred) using machine learning and data from NIH Roadmap Epigenomics (117) and the ENCODE (118) project. This approach exceeds the performance of the six pre-existing HI predictors (114). The fundamental methods in this category are outlined in Table 2-3.

The table below provides details about each HI gene score along with their specific characteristics as well as the methods used to calculate the score with access to each score's information when possible.

Table 2-3 HI gene metrics.

HI measures	HS	HI	Score Characteristics	Method	Weblink/data provided	Reference
Deletion-based HI score	+	+++	Combines a list of biological properties (genomic, evolutionary, functional and network) by examining copy number variations (CNV) among many healthy individuals.	Linear discriminant analysis (LDA)	HI score data available for 18,860 genes in Supplementary Table S4, S1 sheet of Hsu et al. (109).	Huang et al. (93)
REC score	+++	+	Based on human-macaque conservation and adjacency to recessive disease genes in a protein interaction network to categorise genes into recessive disease and LoF-tolerant classes.	Linear discriminant model	REC score data available for 18,860 genes in Supplementary Table S4, S1 sheet of Hsu et al. (109).	MacArthur et al. (12) Hsu et al. 2016 (109)
NET indispensability score	+	+++	Calculates gene centrality and indispensability in various protein–protein interactions (PPI) and regulatory networks to assess the gene importance.	Logistic regression model	NET score data available for 18,860 genes in Supplementary Table S4, S1 sheet of Hsu et al. (109).	Khurana et al. (110) Hsu et al., 2016 (109)
XL	+++	+	Based on the ratio of non-synonymous to synonymous substitution rates (dN/dS) on X-chromosome genes.	Logistic regression model	None	Ge et al. (112)
GHIS	+	+++	Utilises gene features that eliminate study bias for the predictions, called the co-expression with known haploinsufficient genes in the COEXPRESdb and GTEx co-expression networks.	Eliminating study bias using R. Constructing GHIS using support vector machine (SVM) method	Data on 19,701 genes in Supplementary Table S3 of Steinberg et al. (113)	Steinberg et al. (113)
ISPP	+++	+	Combines six gene-level metrics (RVIS, NET, DNE, GDI, HI and REC) with many gene features to predict the pathogenicity of a gene in different inheritance modes.	Machine learning method: random forest algorithm.	None	Hsu et al. (109)
DOMINO	+++	+	Considers genomic data, interspecies conservation,	Machine learning method. based on	None	Quinodoz et al. (116)

			gene expression, PPI, and protein structure to evaluate the probability of a gene to harbour dominant changes. This method does not provide a score per gene.	linear discriminant analysis (LDA)		
HIPred	+	+++	Integrates genomic and evolutionary features (number of transcripts, length of the gene, and the average number of predicted protein domains across transcripts) with functional annotations from ENCODE and NIH Roadmap Epigenomics to predict HI.	Machine learning method	None	Shihab et al. (114)

Note: +, +++ is the relative magnitude of score value.

Characteristics of genes under selection

Advantageous genetic variants will increase in frequency if they are subjected to positive selection. This is contrary to negative selection, which acts to eliminate damaging alleles. Measuring the intensity of negative selection acting on genes gives valuable insights into which genes are liable to a mutation that might lead to serious conditions. Since some essential genes are not recognised to have any disease-causal variation and are possibly subject to purifying selection at high intensity, the pattern is quite complex (90). By measuring the extent and directionality of selection applied on a particular gene, Bustamante et al. developed a score referred to here as 'Sel'. Initially, they compared constant sequence alterations across humans and Chimpanzees over 11.81 Mb region of aligned coding DNA, using mutually synonymous and non-synonymous variants. This study showed that the ratio of non-synonymous to synonymous differences (divergence) is smaller than the ratio of non-synonymous to synonymous polymorphisms (23.76%, 38.42%, respectively). This indicates a substantial excess of amino acid variation, proportional to divergence, supporting previous work revealing that a large proportion of amino acid variation in the human genome is damaging to some extent (119).
Another score was constructed by Eilertson et al. to recognise genes under natural selection and robust demography using a non-parametric approach (i.e., no assumption of a specific population genetic model) (120). This method, called 'selection inference using Poisson

random effects' (SnIPRE), takes advantage of polymorphism and divergence data and from non-synonymous to synonymous ratios (Ka/Ks) within genes (120).

Meanwhile, by integrating two meta-analyses, Sampson et al. created a score called the 'gene-level integrated metric of negative selection (GIMS). The former meta-analysis unified comparative genomic metrics (GERP++) and functional genomic metrics (Poly-phen2), and the subsequent meta-analysis integrated mutation rates (as SNPs/kb) and allele frequencies (as % rare) from the 1K Genomes Project. By combining these two metrics, a meta-analysis was achieved providing GIMS scores for 20,079 human genes (121).

Owing to the fact that the majority of genes are under the effect of negative selection, the target was to measure the degree of purifying selection applied to genes. 'Functional genomic metrics' is a combination of conservation and functional approaches. These were combined with 'population genetic metrics', which is an integration of mutation rates and a fraction of rare variants. The Sampson et al. score integrates these two metrics and produces a combined score per-gene. Thus, GIMS provides a range of probabilities in quantiles; genes scored low are those under negative selection. This study reinforced that under stronger evolutionary constraints, a single pathogenic variant is enough to manifest disease than variants that require two alleles to produce the same. Therefore, the GIMS score is robust in detecting variants under negative selection in dominant diseases (121).

Considering the influences of selection and genetic drift, Yuval et al. created the GDI, a gene-level predictor that estimates if a human protein-coding gene is likely to contain disease-causal variants. For GDI, they used the variant-specific damage prediction score, which is the CADD score. This specific score was chosen because of its efficiency at differentiating between benign and pathogenic variants and its intense reliance on evolutionary conservation (115). Other scores like Poly-Phen-2 and SIFT, are only used to assess missense variants, whereas CADD scores can predict the majority of variants types. Moreover, the GDI score was created by considering the cumulative predicted damage in exonic regions of the gene using the CADD score for each allele compared to the expected score for variants with similar allele frequencies. Yuval et al. calculated the homogenised Phred I-score for every single metric to identify the position of the targeted gene relative to the rest of the genes. Here, human genes with a low GDI scored low as per the Phred score. However, a high Phred score reflects high susceptibility of a gene to contain deleterious variation.

The GDI score is interpreted as follows: High GDI represents genes enriched in sensory perception (for example, the olfactory receptor superfamily) reflecting redundancy, positive

selection constraint, tendency to be under less purifying selective pressure, more significant numbers of paralog, low D—complexity of protein amino acid composition (i.e., relatively unbiased a.a composition), and long coding DNA seq (CDS). GDI of genes containing FPs in patients was much higher than the GDI of disease-causing genes. Low GDI score represents highly conserved genes (enriched in ribosome, chemokine signalling proteasome, spliceosome) reflecting indispensability, and tend to be under purifying selection stronger than the median selective pressure acting on human genes, smaller numbers of paralogs, high D (i.e., low complexity, biased amino acid composition with respect to the median composition of human proteins), and short coding DNA seq (CDS) (115). Key approaches in this category are outlined in Table 2-4.

The table below illustrates each score measuring selection in details with their specific features as well as the methods used to build the score, providing access to information on each metric when possible.

Table 2-4 Interpretation of scores measuring selection.

Selection measures	+ve selection Less-essential genes	-ve Selection More essential genes	Score Characteristics	Method	Weblink/ Data	Reference
Sel	+++	+	Compares polymorphism versus divergence at synonymous and non-synonymous sites to quantify the extent of selection on a given gene.	Logistic regression analysis	None	Bustamante et al. (119)
SNIPRE	+++	+	Non parametric approach, which is robust to demography.	Generalised linear mixed model	None	Eilertson et al. (120)
GIMS	+++	+	GIMS integrates GERP++ scores as comparative genomic metric, Polyphen2 as functional genomic metrics, and population genetic metrics (SNPs/kb and %RARE). GIMS measures the strength of negative selection.	SVM-based learning approach.	GIMS score data is available for 20,080 genes in Table S1, of Sampson et al., 2013 paper.	Sampson et al. (121)
GDI	+++	+	Filters out false positive variants in genes that are susceptible to damaging variation in the general population.	Analysis of outliers based on modified Z-score.	GDI score data available for 18,860 genes in Supplementary Table S4, S1	Itan et al., (115) Hsu et al. (109)

sheet of Hsu et al. (109).

Note: +, +++ is relative magnitude of score value

2.2.4 Discussion

Benefits of combining DNE and RVIS

The DNE score offers some benefits that might be advantageous when scoring genes based on their essentiality and conservation properties. However, the specific validity of DNE score for interpretation of *de novo* mutations only considered the main limitation (109). More specifically, compared to other methods like RVIS and Sel, DNE considers more variables related to mutation rate beyond sequence context. Further, the sequence depth of coverage and regional divergence in genes between humans and macaques are independently additional variables, which both enhance the predictive value of this score (105). When comparing the RVIS and negative selection score Sel to the DNE model, the results showed similar effectiveness of DNE and RVIS; therefore, combining the two metrics will be of great benefit (105,122).

Evaluation and recommendation of DNE model

The Samocha et al. model was improved by incorporation of regional divergence in genes between humans and macaques independently, and the depth of coverage in the latter reflects the number of reads that were present on average per base. These factors play a major role in the development of their predictive score. It appears that there is high correlation between the number of rare synonymous variants in the ESP and the probability of a synonymous mutation determined by their approach. As rare variant allele frequencies are impacted by sample size, it would be worth evaluating this in bigger databases such as ExAC. (122).

Comparison of EvoTol and RVIS

EvoTol has a higher performance at categorizing intolerant genes compared to RVIS. Moreover, it was highly sensitive and more potent in distinguishing genes with high pathogenicity (18). RVIS and EvoTol scores are not shown to be highly correlated, although application of both models simultaneously is likely to be beneficial (104).

HIPred and REC outperform different scores

Considering the scoring of genes for potential involvement in HI phenotypes, Shihab et al. found that the HIPred score outperforms all other scores in predicting HI genes when evaluated against five predictive models (HI approach, EvoTol, RVIS, GHIS and NET, Tables 2-2 and 2-3) (114). Having different views on the 26 disease-associated gene lists and evaluating the potency of several models that prioritise gene pathogenicity, Hsu et al. displayed a more positive correlation between HI and REC, whereas the six models have a moderate correlation with each other (correlation r = 0.77, r = 0.46) respectively (109). The REC score has been shown to outperform five gene-level predictors, which are RVIS, HI, DNE, NET, and GDI, in predicting disease-related genes (109,122).

Best method to filter out false positive variants

The ISPP score shows high performance in prioritizing AR and XL disease-related genes reflecting selective pressure, whereas DNE was constructed dependent on mirDNMR estimates. Both ISPP and DNE approaches do not quantitatively predict the mutational load for a gene in a healthy human population. Therefore, these two methods are not robust for filtering genes that are highly mutated and consequently, many residual false positives might be expected. One of the most robust models to predict genes known to have deleterious variation is GDI, which proves the efficiency to filter out false positive mutations in genes known to contain damaging variations (115,122).

Features of the dN/dS ratio

The dN/dS ratio can be measured for any protein-coding gene, and the XL scoring system is not limited by former gene annotation. The XL approach can be applied to all X-chromosome protein-coding genes and can evaluate genes for various disease phenotypes (112). The intra-human dN/dS ratio is not specific to the X-chromosome. It is recommended that the dN/dS ratio be used to analyse more genomic data for future studies to prioritise genes containing disease variation (112,122).

Aim of the systematic review

This project aims to place emerging evidence on gene features and variant scoring models that might have a significant role in filtering disease sequence data in proximity (12,113,114,(12)(12)(12)(12)(12)(12)(12)(12)(12)(12)(12)(12)(12)(12)(12)123). However, distinguishing the LoF variants causing disease phenotypes from others that do not cause

any functional disruption remains a challenge (113). Based on the 1000 Genomes Project data, it has been demonstrated that on average, a healthy person might carry 250 to 300 LoF SNVs (1000 Genomes Project Consortium et al., 2010 (124); The 1000 Genomes Project Consortium, 2012 (125)) (109).

Recommendation to use wide-coverage databases (ExAC)

Due to the accumulation of sequence data in publicly available databases, understanding of human genomes becomes promising. The ExAC resource provides a strong filter available publicly to serve the wider research community and aid identification of deleterious variants in severe Mendelian diseases. Using ExAC for filtering and elimination of false positive variants, reduces the number of candidate variants that affect protein function by seven-fold compared to the ESP database, which has fewer exome sequences and lacks the power to filter at 0.1% AF without excluding many true rare variants (38).

Gene-level scores versus variant-level scores

Based on previous evidence, the missense Z score, which represents genes rather than variants, provides more data than variant-level Poly-phen2 and CADD (72) scores, emphasizing that gene-specific metrics of constraints add more detailed information to variant-specific metrics in predicting deleteriousness (38). Moreover, limitations of variant level scores (for instance, SIFT and POLYPHEN) were considered as not providing information on whether purifying selection at a particular site is acting in a recessive, additive, or dominant mode based on cross-species alignments by analyse Huang et al. (93).

Is the whole gene the best unit to use to test patterns of intolerance?

Based on dividing genes into sub-regions, Gussow et al. proposed a model called subRVIS to precisely predict where deleterious mutations are likely to present (107). This paper brought new insight into which unit is better used to judge genic intolerance, raising an important question: is the unit of the whole gene the correct one to assess patterns of intolerance? Future studies could consider modified gene-specific approaches and consider them within-gene district patterns of intolerance for further examination.

The challenge of non-unified nomenclature

The challenges in interpretation of benign LoF variants due to non-unified nomenclature

(using different terminologies to refer to LoF variant) is a major debateable issue worldwide. It is essential to be aware that in healthy individuals, there are overlaps in the interpretation of LoF variants. The following are some of the terminologies that represent LoF variants in a healthy person: 'non-deleterious or less-deleterious variants that have an impact on risk of phenotype or disease', 'true variants that do not seriously disrupt gene function', and 'benign LoF variations in redundant genes' (12).

Limitations of individual score

Despite the fact that each genic scoring approach only considers genetic architecture from a specific angle, each score has limitations as follows: (i) Dominant disease-predisposing genes are not considered by the REC score; (ii) Lack of non-CNV genetic variants in the HI prediction approach; (iii) NET score does not include the systematic comparison of diverse known disease-causal genes; (iv) RVIS score did not consider the differences in the allele frequencies among several populations; (v) Limited applicability of DNE approach for evaluating *de novo* mutations; (vi) 'The GDI score only considers mutation profiles' (109); (vii) the GHIS does not consider the genetic background in individuals, which is a significant limitation since genetic variants do not act independently, and disruption of individual genes within a particular biological pathway may affect disease risk (113). Here, it is worth noting that providing an inclusive review of the individual prediction score is a step forward in launching new paths for prioritizing disease-related variants.

2.2.5 Conclusion

In the final analysis, a range of well-studied gene-specific predictors were explored and investigated with various independent genetic features. Handling the limitations of each score or utilizing the established predictors of pathogenicity and merging these approaches in an integrated score may enhance prediction of disease-related genes as at present, no single scheme has high reliability in prioritizing genes based on pathogenicity.

Several methods were established to evaluate whether a gene is tolerant or intolerant to CFV. Initially, metrics were developed per gene. Subsequently, advanced studies revealed that dividing the gene into sub-regions is more potent in pointing out the location of the mutation precisely. At that point, prior disease knowledge was required by all scores that predict genic intolerance. An important step to overcome this limitation and to improve prediction of gene intolerance is the creation of a tool with no prior disease knowledge.

The American College of Medical Genetics (ACMG) guidelines evaluate several *in silico* variant predictors of whether a variant is involved in disease. However, the guidelines do not determine which or how many variant prediction tools to use. Therefore, it was recommended that these tools be used only as 'supporting' evidence for variant interpretation. Several challenges persisted with respect to validation of these tools, with a relatively elevated error rate and many deleterious variants being evaluated as benign by these tools and *vice versa* (94,122). Furthermore, currently, the ACMG guidelines do not consider gene-level scores, which are the focus of this systematic review. However, it may be possible to use this to establish supporting evidence alongside stronger independent evidence signifying the role in development of disease. Overall, it is optimal to undertake functional validation, but this can be difficult (94,122).

The research question of this review was 'Can the use of gene-specific metrics facilitate the identification of disease genes in patient genomes?' To answer this, various available gene-level scores were reviewed here with different independent genetic features. Recognizing the limitations of each predictor and possibly using these gene-specific predictors in conjunction with variant-specific predictors could achieve better prediction, particularly as there is currently no particular score predictive of gene pathogenicity with high reliability. Thus, this review is intended to outline existing information available to identify and explain different gene-level pathogenicity scores, as well as determine the gaps in disease-gene prioritization and annotation issues to form a solid base for new scores and better prediction of disease-related genes.

Chapter 3 Devising a Method to Reduce the Number of Candidate Genes to Follow up

3.1 Introduction

The aim of the 100,000 Genomes Project (37) was extended to broaden the plan to sequence the genomes of five million patients diagnosed with rare disease and cancers, over the coming five years. The plan of the 100K genome project was to improve the National Health Service (NHS) infrastructure, data security, and clinical training (126).

This is an international endeavour, and since 2013, fourteen countries have invested more than four billion US dollars in establishing national genomic programmes to address the challenges and transition testing from centres of excellence to medical practice. In countries like Australia, United Kingdom, France, Saudi Arabia, and Turkey, the development of manpower and infrastructure has been coupled with testing numerous patients who are known to have rare diseases or cancer. Further, US, Japan, and Qatar invested in population-based sequencing with the involvement of participants and return of results to patients (126).

In order to achieve the goals of the 100K genome project and the progression of similar initiatives, it is vital to develop plans that advance the disease genome data interpretation, so that the true causal variant can be easily differentiated from the plausibly pathogenic, but in fact neutral, variant.

In this context, WES has detected variants associated with monogenic diseases and complex diseases. However, distinguishing which variants might be causal needs careful variant filters and robust tools to predict variant pathogenicity (103). Currently, numerous tools exist to predict variant/gene's deleteriousness like SIFT (68), PolyPhen2 (70), pLI, and RVIS. Nevertheless, new tools are needed to efficiently predict pathogenicity.

Due to the effects of natural selection, pathogenic variants are expected to have low AF compared to those that are non-pathogenic, which has been demonstrated in human population sequencing data (38). Examples of scores that have been produced measuring the impact of selection on genes include Sel (119), SNIPRE (120), and GIMS (121).

Several methods have been used that attempt to predict variations in terms of their previous probabilities of conferring risk of disease, particularly population AF and conservation measures at two levels—the phylogenetic level or amino acid properties (87). Here, this study seeks to investigate certain gene-level scores and clarify how they can be used to develop a

gene-specific metric that can predict which genes have high likelihoods to influence diseases (87).

Meanwhile, Spataro et al. (90) classified genes into five discrete groups by degree of gene essentiality as the following: NDNE, CNM,CM,MNC, and END, which shape the foundation for the model proposed by Pengelly et al. (62). Essential genes are those required for cell survival and are responsible for key biological functions the creatures (127). In the Pengelly et al. hypothetical model, candidate disease genes fall in an intermediate position on the essentiality spectrum between two other groups of genes: a large group of genes considered tolerant to functional variation which are non-disease, non-essential genes (NDNE) and essential non-disease genes (END), which are highly intolerant of functional variations. The essential gene group includes lists of genes defined as essential through mouse knock-out experiments, excluding human disease genes that are listed in OMIM (24) and common disease genes that were identified by genome-wide association studies (90). Metrics related to gene essentiality for individual genes include measures such as the RVIS (87) and pLI (38), both of which quantify tolerance of LoF variations. Other scores focus on the degree of conservation: the REC score (12); the position of genes in regulatory and other networks (for instance, the NET score (110)); or scores that consider local sequence context like SIS (88). Further, LD is another factor associated with gene essentiality; high essential genes tend to have low haplotype diversity and therefore, strong LD (62). Moreover, groups of genes already identified to contribute in disease process (90) comprise of those that might cause only complex diseases (complex non-Mendelian: CNM genes), those which contribute to both monogenic and complex variations (complex Mendelian: CM), and those associated only with monogenic variations (Mendelian non-complex: MNC). Through investigation of different gene-level metrics, each of which is generally linked to gene essentiality, the understanding of the impact of each score and how it can help in identification of genes most likely to include monogenic disease variation is facilitated.

The aim of this study was to investigate how to utilise gene-specific scores identified by the systematic review (122) to later calculate a new composite gene-level score that predicts essentiality of individual genes and possibly determine candidate genes that might cause Mendelian diseases. This may help to prioritise genes based on their essentiality and contribution to diseases. This can be achieved through selecting gene-level predictors from the systematic review that provide scores per gene, and then testing the power of each score in representing a clear direction of the data in consideration with their biological meaning (Table 3-1) by performing principal component analysis, thus identifying the most useful

scores that represent the data and might be helpful to improve the prediction of disease genes.

Table 3-1 Biological interpretation of the scores used to build ESPP

Score	Biological interpretation
DNE	the rate of de novo mutation per gene (105)
GDI	Genes with High GDI are the ones functionally related and strongly enriched in sensory perception while genes with low GDI are the onse enriched in ribosome,proteasome, and spliceosome genes (115)
GHIS	Using features of co-expression networks to eliminate study bias(113)
GIMS	GIMS has been created to give a new insight on glomerular biology in terms of evolutionary selection. They test the enrichment of negative selection in high quality gene sets contain genes enriched for expression in the renal tubular compartments (121)
HI	Huang et al. found that functional interaction with known HI genes was the most predictive property of HI genes might impacts the modularity of the interaction network, suggesting certain pathways or biological processes, like early development morphogenesis (93)
NET	Mesure gene centrality and indispensability in various protein–protein interactions (PPI) (110)
pLI	High pLI genes play a role in core biological process like spliceosome, ribosome, and proteasome components, while olfactory receptors are among the least constraint with low pLI (38)
RVIS	RVIS applying human population and genetic data to prioritise essential genes (87)
SIS	Using multiple gene features that affect the rate of mutation of certain gene (88)
REC	Prioritise genes that might play a role in recessive diseases (12)

3.2 Materials and methods

This chapter was in preparation of developing an essentiality score to prioritise disease genes. To this end, current data of the available gene-level metrics to be analysed were gathered in terms of which score shows a clear direction in predicting disease genes using updated gene classification.

3.2.1 *Gene-specific scores*

The database was generated using 10 scores identified by Alyousfi et al. (122). A pLI score that measures gene essentiality by predicting the tolerance of a gene of carrying LoF variation was used; genes that have high pLI scores represent the most constrained genes. In the end, pLI scores for 18,226 genes were obtained from Supplementary table 13 from Lek et al. (38). The SIS score is another gene essentiality predictor, in which high scores represent highly constrained genes; this score was obtained for 16,387 genes from Supplementary 3, table 15 from Aggarwala et al. (88). Further, deletion-based HI score, REC, NET indispensability, GDI, and RVIS scores were obtained for 18,860 genes from Supplementary Table S4, S1 sheet from Hsu et al. (109).

More specifically, the HI score is used to assess whether a deletion is benign or pathogenic using a LOD score. The LOD score is a statistical test used in genetic linkage analysis to observe the linkage of two loci and the probabilities of them being inherited together (128). Here, a high LOD score is a feature of high linkage (128), and in this model, represents pathogenic deletions and dominant traits. Meanwhile, MacArthur et al. produced the REC score to prioritise recessive disease variants using human–macaque conservation data (12), and Khurana et al. produced their NET scores based on several biological networks to predict gene HI (110). In this context, genes with high NET score are considered haploinsufficient. Furthermore, using linear regression analysis, the RVIS score predicts essential genes by measuring the intolerance of a gene to CFV, with low scored genes considered essential by RVIS predictors. Moreover, the GDI is one of the scores that predict whether a gene is impacted by selection; genes that score low by GDI are the ones highly impacted by negative selection. More information on each score is detailed in Chapter 2. Additionally, the DNE score can predict essential genes using a Z score by estimating the difference between the observed and the expected missense variants. This score—referred to as the CONS score—was obtained for 18,860 genes from Supplementary Table S4, S1 sheet

from Hsu et al. (109). The GIMS measures the impact of selection on a gene using machine learning methods. Data using this score was obtained for 20,080 genes from Table S1 from Sampson et al. (121). The genome-wide haploinsufficiency (GHIS) score predicts HI using the machine learning method; genes with high GHIS score are considered haploinsufficient. Data on 19,702 genes with GHIS scores were obtained from Supplementary table 3 from Steinberg et al. (113). Further, a pre-existing gene-level LD score for 18,269 genes was obtained from LD maps in LD units (LDUs) (91). These maps were built using publicly available 1000 Genomes Project data that were derived from the Wellderly study (129). Over 400 WGS samples were used to construct the LD maps. The LDU lengths of genes were corrected for physical gene length by regression for the 'LDU_res-fit' scores used as the 11th score in the analysis. Further information about LD maps can be found in Vergara-Lope et al. (91). A very high-resolution understanding of patterns of LD in the genome was provided by these maps (91).

Other scores that were described in Alyousfi et al. (EvoTol, LoFtool, subRVIS, mirDNMR, XL, DOMINO, Sel, and SNIPRE) do not provide score per gene, therefore these were excluded from the study. ISPP and HIPred were excluded as well due to the unavailability of these scores.

For GIMS (121), SIS (88), and GHIS (113) scores, ensembl IDs were transformed to the corresponding gene name to allow for data to be aligned correctly. Here, the original scores' names were maintained for all metrics except for SIS and DNE scores that were abbreviated for ease of usage in the tables—Substitution intolerance score was shortened to 'SIS' and *de novo* excess was shortened to 'DNE' for better recognition of the score property as it was referred to as the CONS score in Hsu et al.(109).

3.2.2 *Gene classification*

In this study, the classification of gene groups by Spataro et al. (90) was used as were the 17982 genes in their supplementary table S2. Here, it is worth noting that they classified genes into two major groups: disease-genes (which are genes presented in both the hand-curated Online Mendelian Inheritance In Man (hOMIM) and GWAS catalogues); and non-disease-genes. They further classified the disease gene group into three subgroups: CNM, CM and MNC. CNM are genes represented in GWAS but not in hOMIM, and the opposite is true for MNC. However, CM are genes were represented in both the hOMIM and GWAS catalogues. Moreover, Spataro et al. described the biological property differences between

complex disease genes in the two groups, CM and CNM. Compared to the CNM group, CM genes tend to have high expression levels in the protein network, and they are enriched in certain relevant protein function categories. CM genes also have higher Odds Ratios (ORs) than CNM which suggests that variants around CM genes have a stronger impact on the complex phenotypes (90).
Meanwhile, the non-disease genes were classified into two groups: END and NDNE (90). Essential genes have been identified through knock-out experiments in mice, so any essential gene in mice is putative essential in humans. Then, all genes in that list that were found to be disease causal were removed. The number of genes in each group were as follows: NDNE (13135 genes), CNM (2388 genes), CM (203 genes), MNC (684 genes), and END (1572 genes). After aligning the Spataro list with the other score data, 1273 (NA) genes could not be categorised. These are of interest with regards to the testing of their essentiality and might include candidate genes for rare monogenic diseases based on this study's future model.

The results of the Mann Whitney U test revealed no significant statistical differences between the two groups (CM and MNC), reflecting the overlap between these two groups. As the focus of this project is facilitating recognition of genes implicated in monogenic disease, merging the two groups into a single Mendelian disease gene group (MDG) was considered to gain full coverage of the gene group of interest—the Mendelian disease genes (Figure 4-1).
All previous tests including the Mann Whitney U, Kruskal Wallis, and Kruskal multiple comparison tests were repeated for the new classification after merging Mendelian disease gene groups to test whether the previous approach of merging the CM and MNC will improve group separation and ensure better results.

A list of all dominant genes and all recessive genes were obtained from the union of the Berg and Blekhman lists of dominant and recessive genes (128,129). These lists were cross-referenced with the findings from Spataro et al. on Mendelian genes to improve the model's performance by covering most of the discovered Mendelian disease genes and improve separation between groups (90). However, for better coverage, the aforementioned study's list was updated using the OMIM database (Figure 4-1)(24), which is the current most comprehensive source of Mendelian disease genes. By November 2019, OMIM reported 3,799 genes and 5,483 phenotypes related to single gene disorders (24).

3.2.3 *Analysis*

The preliminary analysis was exploratory. Box plots (Figure 3-1) were produced to compare medians of these metrics for the five group of genes that were defined by Spataro et al. (90). A number of non-parametric tests were performed as the data were not normally distributed. Further, since the data are ordinal, a Mann-Whitney U test was performed to explore the distribution of gene groups by comparing the means of each of the two groups. This test was undertaken for dual comparisons of each two groups as it allowed only paired comparisons. For instance, NDNE and CNM, NDNE and CM, NDNE and MNC, NDNE and END, and so on.

The second non-parametric test performed was the Kruskal-Wallis Test, which was conducted to compare the mean rank of each variable in the five groups of genes to determine whether there are statistically significant differences between the five groups of genes defined by Spataro et al. (90). Here, as the outcome of the p-value was less than the significance level 0.05 for all the variables, the mean rank was produced and transformed to a percentage to simplify interpretation. Table 3-2 in the results section shows the results of the Kruskal-Wallis test in combination with that of the Mann-Whitney U test. The third non-parametric test performed was the Kruskal Wallis multiple comparison test and was integrated with the results of Mann Whitney U test as shown in Table 3-1 in the results section.

Following this, after data standardization to mean zero and SD=1, the first PCA was performed to decrease the dimensionality of the variables. PCA is a dimensionality reduction algorithm that helps in compressing a dataset with several dimensions and flattens it into two or three dimensions in a way that captures the essence of the data and provides better visualisation. In other words, PCA is an analytical method to identify a meaningful way to look at the data by concentrating on the differences between variables (132). The first PCA was undertaken for 11 gene-level scores including LDU_res-fit, HI, REC, NET, DNE, GDI, pLI, RVIS, SIS, GIMS, and GHIS. Subsequently, a second PCA was done using only the most informative scores from the first PCA. The results of the first and second PCAs are presented in Table 3-3 in the results section.

3.2.4 *Evaluation of the relationship between measures of essentiality*

To test the relationship between measures of essentiality, Spearman's correlation was performed for eight gene specific scores, and the results are presented in Table 3-6 of the results section.

3.3 Results

Ten gene-specific scores were derived from the systematic review in this study (122) as the benchmark dataset to prepare for the essential gene prediction model. Further, the LDU_ res-fit from pre-existing whole-genome LD maps was included as a linkage-disequilibrium predictor (91). Next, the aforementioned 11 scores were evaluated among gene groups described by Spataro et al. (90).

3.3.1 *Relationship of gene-specific metrics in gene groups*

- ***Results based on Spataro et al. gene groups for 11 scores***

Boxplots were produced for all the variables to represent the influence of each score amongst different gene classes. The boxplots provide only narrow median differences among the three disease-gene groups (CNM, CM, MNC) for all scores. All the predictors display non-significant differences between the two ends (NDNE and END) apart from NET and pLI (which show significant differences), in addition to GIMS and HI (which reveal less significant discrimination among the two groups) as demonstrated in Figure 3-1.

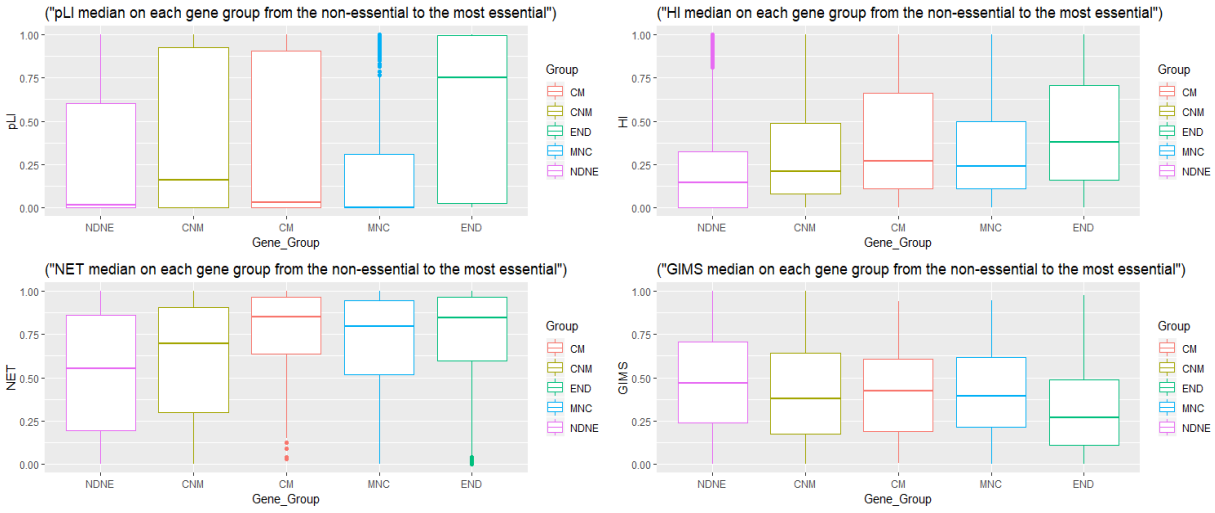


Figure 3-1 Boxplots representing the medians of pLI, HI, NET, and GIMS scores among five gene groups.

The above boxplots represent the medians of pLI, deletion-based HI, NET indispensability, and GIMS scores among the five gene groups, which are NDNE, CNM, CM, MNC, and END. The boxplots provide limited discrimination among the disease gene groups for the four scores. pLI shows the most significant difference between the essential and non-essential groups, while the HI, NET and GIMS show less significant differences between essential and non-essential groups.

Table 3-2 The significance of Kruskal Wallis multiple comparison (Kruskalmc) and Mann Whitney U tests among the Spataro et al. gene groups (90), which are NDNE, END, CNM, CM, and MNC.

Variable 1	NDNE-CNM[2]	NDNE-CM	NDNE-MNC	NDNE-END	CNM-CM	CNM-MNC	CNM-END	CM-MNC	CM-END	MNC-END
DNE	< 0.0001	< 0.0001	< 0.0001	< 0.0001	**0.169**	**0.1884**	<0.0000	0.0394	<0.0001	<0.0001
GDI	**0.5336**	< 0.0001	< 0.0001	< 0.0001	<0.0001	<0.0001	<0.0000	**0.0906**	<0.0001	<0.0001
GHIS	<0.0001	0.0145	**0.3089**	< 0.0001	<0.0001	0.002309	<0.0000	**0.0676**	<0.0001	<0.0001
GIMS	<0.0001	<0.0001	<0.0001	< 0.0001	**0.8771**	**0.3849**	<0.0000	**0.4897**	<0.0001	<0.0001
HI	<0.0001	<0.0001	<0.0001	< 0.0001	0.0061	<0.0001	<0.0000	**0.5871**	0.0018	<0.0001
LDU_res-fit	0.0459	**0.0983**	**0.8616**	<0.0001	**0.0611**	**0.2625**	<0.0000	**0.1813**	<0.0001	<0.0001
NET	<0.0001	<0.0001	<0.0001	< 0.0001	<0.0001	<0.0001	<0.0000	0.0546	**0.6000**	<0.0001
pLI	<0.0001	**0.3119**	<0.0001	< 0.0001	0.0518	<0.0001	<0.0000	0.0192	<0.0001	<0.0001
REC	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0000	0.0546	<0.0001	<0.0001
RVIS	<0.0001	<0.0001	<0.0001	<0.0001	**0.389**	0.0464	<0.0000	**0.8312**	0.0032	<0.0001
SIS	<0.0001	**0.1075**	<0.0001	<0.0001	**0.2659**	0.0585	<0.0000	**0.7711**	<0.0001	<0.0001

[1] Gene constraint score- *de novo* excess (DNE), Gene damaged index (GDI), genome-wide haploinsufficiency score (GHIS), **G**ene-level **I**ntegrated **M**etric of negative **S**election (GIMS), Deletion-based haploinsufficiency score (HI), residuals generated after linear regression analysis to correct the LDU length of a gene for physical gene size (LDU_Res-fit), Gene position in networks (NET) indispensability score, Loss Intolerance probability (pLI), Recessive (REC) score, Residual Variation Intolerance Score (RVIS), Substitution Intolerance Score (SIS).

[2] non-disease, non-essential (NDNE), essential non-disease genes (END), complex non-Mendelian (CNM), complex Mendelian (CM), and Mendelian non-complex (MNC)

Underlined = No significant difference between the two groups as per Kruskalmc (multiple comparison)
Bolded = no significant difference as per the Mann Whitney U test.

The results of the Mann Whitney U and Kruskal multiple comparison (Kruskalmc) tests are illustrated in Table 3-1. More specifically, the results of the former showed the highest statistical significant difference between the END and NDNE and CNM and END for all the scores. The differences between the CM and MNC groups were the least significant. These results were supported by the Kruskalmc. Further, the overall results showed a significant difference between the essential and non-essential genes for all the scores. Within the rest of the disease-gene group, the direction is less clear, which makes the explanation for this group difficult. There was no statistical significance difference between the CM–MNC by all the variables and CM–CNM by most of the variables. The LDU_res-fit, which characterises the effect of LD, was the least good in distinguishing the groups. Whereas, the REC score distinguished all the groups based on the Mann-Whitney U test.

Table 3-3 Mean rank scores in the Kruskal-Wallis Test as percentages of the highest mean rank amongst the Spataro et al. five gene classes (90).

Variable[1]	**NDNE**	**CNM**	**CM**	**MNC**	**END**	**N genes**
DNE	70.054	79.200***	82.981***	76.969***	100***	16840
GDI	87.004	85.269	100***	94.061***	75.981***	16840
GHIS	90.237	88.146***	74.506*	78.886	100***	14914
GIMS	100	84.296***	83.697***	80.880***	64.111***	16485
HI	69.054	82.275***	90.612***	89.86***	100***	16840
LDU_res-fit	93.861	91.397*	100	94.203	82.497***	16995
NET	68.481	81.626***	99.422***	95.134***	100***	16840
pLI	73.757	81.789***	74.536	64.29***	100***	16161
REC	60.640	74.222***	99.777***	100***	88.734***	16840

RVIS	100	86.471***	83.412***	81.492***	68.928***	16840
SIS	76.395	81.55***	77.851	81.406***	100***	14502

*Significantly different from NDNE as per Mann-Whitney Test; * = P < 0.05, ** = P < 0.01, *** = P < 0.001.

The Kruskal-Wallis and Mann-Whitney U test results are presented in Table 3-2. The former suggests that the difference between NDNE and END is statistically significant, representing a high P value for all the scores. The same was observed for the MNC group, except for the GHIS and LDU_res-fit, which showed non-significant P values. Meanwhile, CM and CNM show less significant differences from the NDNE gene group. Further, essential genes show a statistically significant difference from NDNE among all variables. For the rest of the groups, there were less significant differences. There were high scores for DNE, GHIS, HI, NET, pLI, GHIS and SIS, showing more intolerance of variation; conversely, low scores of GDI, GIMS, REC, and RVIS showed more intolerance of variation.

As there was no significant statistical difference between the two groups (CM and MNC) based on the results of the Mann Whitney U and Kruskalmc tests presented in Table 3-1, the Mendelian genes (CM and MNC) groups were merged into a single group called MDG.

The first principal component analysis (PC1) (Table 3-3, Figure 3-2 and 3-3) showed that some scores are less informative than others. For instance, three variables—LDU_res-fit, GDI and REC—explained less than 20% of the variance in PC1; thus, these scores were eliminated from further analysis. Furthermore, it was expected that REC and GDI would have negative scores under the model due to increasing essentiality, but this was not observed.

Table 3-4 Principal components for essentiality scores.

Essentiality measure	11 scores: PC1	8 scores: PC1
DNE	0.3473	0.3574

GDI	0.0129	-
GHIS	0.3494	0.3607
GIMS	-0.3878	-0.3972
HI	0.3029	0.2908
LDU	-0.0549	-
NET	0.2784	0.2729
pLI	0.3442	0.3527
REC	0.1866	-
RVIS	-0.3374	-0.3494
SIS	0.4090	0.4230
Total variance explained	0.3615	0.4825

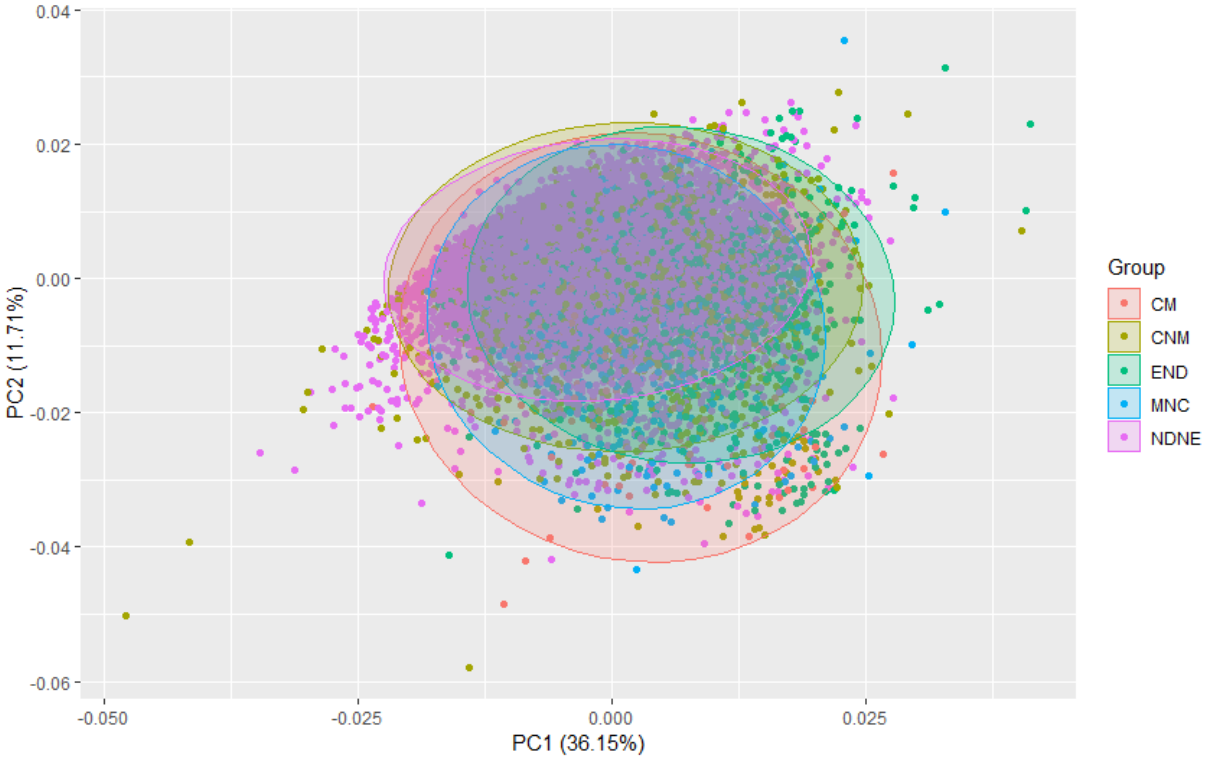


Figure 3-2 plot of the first principal component analysis based on the Spataro et al. gene classification (90). This plot shows PC1 against PC2 of the first principal component analysis. There is substantial overlap between the five groups

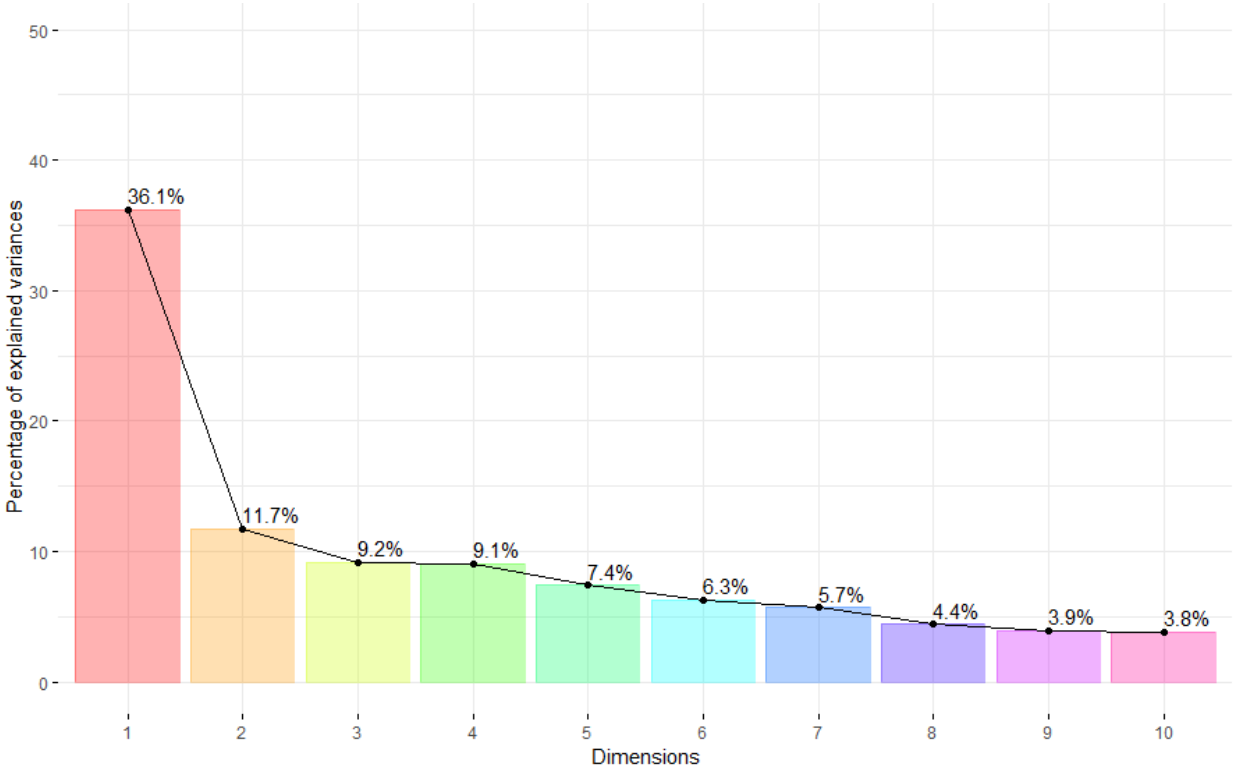


Figure 3-3 Scree plot of the first principal component analysis. The scores used in the first principal component analysis represent 36.1% of the data in PC1.

The second PC analysis for the eight variables, after discarding the uninformative scores from PC1, revealed enhanced representation of all the variables as shown in Table 3-3. The percentage of variance explained went from 36% in the first PCA for PC1 to 48% in the second PCA for PC1 (Table 3-3, Figure 3-4 and 3-5).

Since PCAs provide an orthographic transformation of variables that may be originally connected and produce linearly uncorrelated variables with close correlations between sub-sets of scores (Table 3-6), thus decreasing the independent contribution of a number of scores and justifying the dimensionality reduction. Hence, a PCA using the remaining eight variables was undertaken (Table 3-3); this model describes a high proportion of the variance increasing from 36% till 48%.

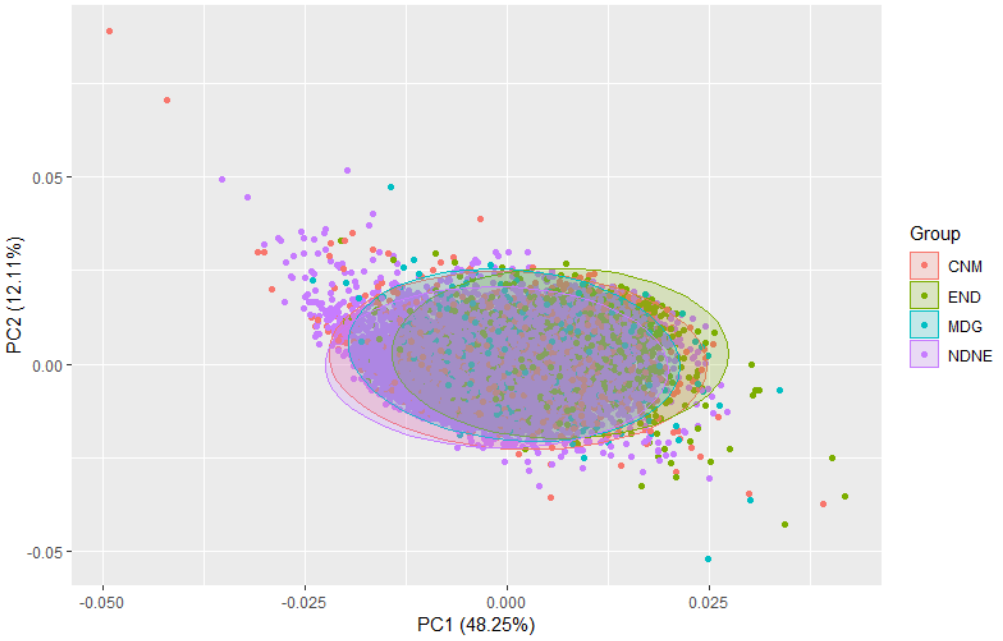


Figure 3-4 plot of the second PCA against four gene groups.

This plot presents the PC1 against PC2 of the second PCA. It became more noticeable that the END genes are right shifted and NDNE left shifted, which gives better representation than the first principal component analysis. However, there is substantial overlap between the disease groups, which makes it hard to understand the direction of these two groups.

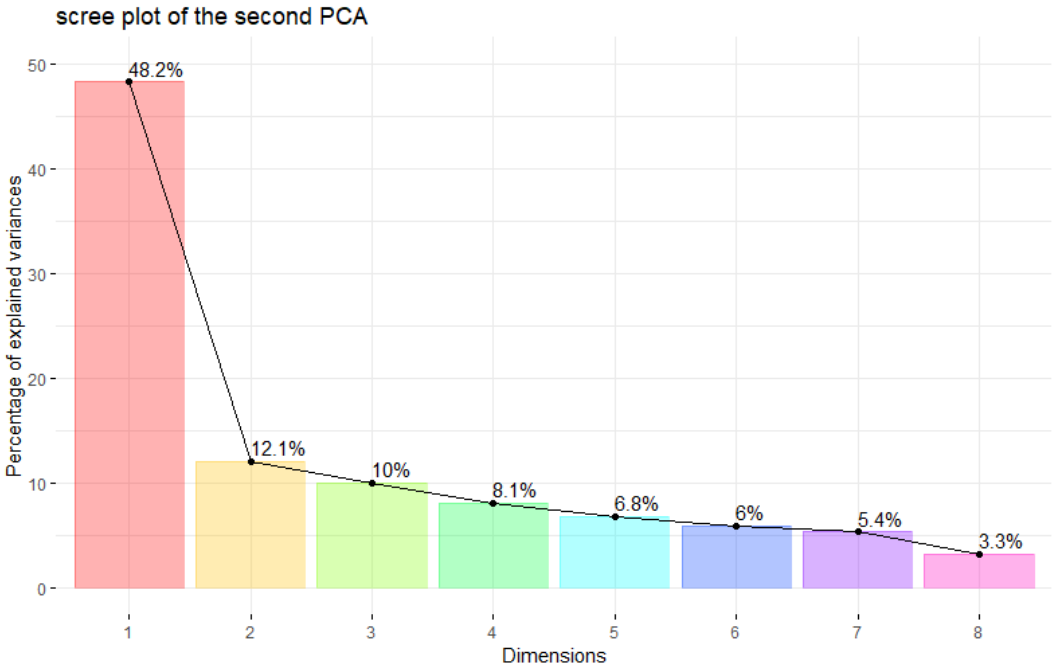


Figure 3-5 Scree plot of the second PCA, which represents 48.2% of the data in PC1.

- ***Results after merging (CM and MNC) into MDG for eight variables.***

The table below represents the result of eight scores using the combined MDG group, which showed improved results when compared to the Spataro et al. gene classification (90). Most of the binary comparisons show high statistically significant differences, except for CNM-MDG, which show less significant differences.

Table 3-5 The significance of Kruskal Wallis multiple comparison (Kruskalmc) and Mann Whitney U tests for eight variables using MDG.

Variable	NDNE-CNM	NDNE-MDG	NDNE-END	CNM-MDG	CNM-END	MDG-END
DNE	<0.0001	<0.0001	<0.0001	**0.5845**	<0.0001	<0.0001
GHIS	<0.0001	0.0448	<0.0001	<0.0001	<0.0001	<0.0001
GIMS	<0.0001	<0.0001	<0.0001	**0.4985**	<0.0001	<0.0001
HI	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
NET	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0016
pLI	<0.0001	0.0050	<0.0001	<0.0001	<0.0001	<0.0001
RVIS	<0.0001	<0.0001	<0.0001	0.0393	<0.0001	<0.0001
SIS	<0.0001	<0.0001	<0.0001	0.0373	<0.0001	<0.0001

Underlined = No significant difference between the two groups as per the Kruskalmc (multiple comparison) test

Bolded = no significant difference as per the Mann Whitney U test

Table 3-6 Mean rank scores of the Kruskal-Wallis test as percentages of the highest mean rank amongst four gene classes after combining MDGs.

Variable	NDNE	CNM	MDG	END	N genes
DNE	70.054	79.200***	78.312***	100***	16840
GHIS	90.237	88.146***	77.908*	100***	14914
GIMS	100	84.296***	81.509***	64.111***	16485
HI	69.054	82.275***	90.028***	100***	16840

NET	68.481	81.626***	96.091***	100***	16840
pLI	73.757	81.789***	66.579**	100***	16161
RVIS	100	86.471***	81.921***	68.928***	16840
SIS	76.395	81.55***	80.612***	100***	14502

*Significantly different from NDNE as per Mann-Whitney Test; * = P < 0.05, ** = P < 0.01, *** = P < 0.001

There was much more consistency with the hypothesised essentiality model, considering the results of the Mann Whitney U and Kruskal Wallis tests after combining the Mendelian gene groups as demonstrated in Table 3-4 and 3-5. The Mann Whitney U test suggests that the eight scores demonstrated a significant statistical difference between all the groups (pairwise), apart from the group CNM-MDG, where the two variables did not reach a statistically significant level. Almost half of the scores showed insignificant difference in the CNM-MDG group as per the Kruskalmc (multiple comparison) test.

There was improved discrimination between the variables in the CNM-MDG group by more than 50% of the score (Table 3-4), when compared to the differences between the complex gene groups (CNM-CM) and the Mendelian gene groups (CM-MNC) before the combining of the Mendelian genes (Table 3-1), signifying that merging the Mendelian gene groups may enhance results.

Following omission of the least useful scores (LDU_res-fit, GDI and REC), the results of the Kruskal-Wallis test (Table 3-5) are consistent with the improved modelling of the data. To summarise, merging the Mendelian gene group enhances the discrimination between the groups and delivers a better understanding of the data. The high value of all the predictors indicates high-essential genes, apart from two variables (GIMS and RVIS), in which increasing their values results in them going towards the non-essential spectrum.

3.3.2 *Relationship between measures of essentiality.*

- ***Spearman's correlation***

In order to test the power and direction of a monotonic relationship between the scores, Spearman's correlation was performed to produce a correlation coefficient. Table 3-6 provides the results of this analysis for the eight scores. The correlations are relatively high throughout.

Table 3-7 Spearman's correlation coefficients for the eight scores.

Essentiality measure[1]	GHIS	GIMS	HI	NET	pLI	RVIS	SIS
DNE	0.3107	_0.4437	0.3043	0.3082	0.5315	_0.3240	0.6117
GHIS	-	_0.5224	0.3297	0.3387	0.4059	_0.5587	0.5466
GIMS	-	-	_0.3901	_0.3305	_0.4613	0.5364	_0.6441
HI	-	-	-	0.3714	0.3416	_0.2740	0.3273
NET	-	-	-	-	0.3062	_0.2472	0.3236
pLI	-	-	-	-	-	_0.3443	0.5584
RVIS	-	-	-	-	-	-	_0.6111
SIS	-	-	-	-	-	-	-

The values in Table 3-6 represent Spearman correlation coefficient (r_s), which has values from positive 1 to negative 1. A r_s of +1 indicates a perfect association of ranks, while a r_s of 0 indicates no association between variables, and a r_s of -1 indicates a perfect negative association of ranks. Thus, the closer r_s values are to zero, the weaker the association between the ranks, either in positive or negative directions.

3.4 Discussion

Despite the development of sequencing technology, the difficulty of making a solid molecular diagnosis from genome sequences remains a challenge in many cases. Since the understanding of several features of gene function are not well recognised and genes might have overlapping functions and a high degree of redundancy, the challenges persist even for highly penetrant monogenic diseases.

While methods that are intended to estimate the deleteriousness of individual DNA variants are widely used to assist in interpreting genome variation, gene-specific scores are less frequently considered. Here, 11 quantitative measures—including LDU_res-fit, HI, REC, NET, DNE, GDI, pLI, RVIS, SIS, GIMS, and GHIS—that predict gene essentiality are evaluated. The metrics have diverse characteristics and gene properties, such as degree of intolerance of genes to functional variation, the local sequence context of a gene, and the position of genes in gene interaction networks like phosphorylation, signalling, metabolic, and physical PPIs.

Since the results of the principal component analysis showed that LDU_res-fit, GDI, and REC consistently explain less of the variance in the data than other scores, these scores were eliminated; here, it is worth noting that GDI and REC were the only two scores that had opposite directions from what was expected. LD patterns have been found to have some relationship with disease/essentiality, but the effect might be small.

Moreover, the reason why the REC score might have had little effect is because the aim of the score is to predict recessive genes particularly and differentiate them from the LoF tolerant genes. Meanwhile, in the case of GDI, it might have been because it is based only on CADD score, which might be not as powerful as the rest of the scores in predicting gene pathogenicity.

While the prediction of Monogenic disease genes seems to be easier than that of genes associated with complex diseases, this study is only looking for single genes that might cause the disease. In fact, gene interactions and the influence of certain genes on others is not yet well understood, making it complex to predict Monogenic diseases.

There were few limitations in this study. The first limitation was the number of genes available for each individual score; for instance, the SIS score is available for 16,387 genes, and by matching this score with the rest of the data, the number of genes drops to 14,503.

Another issue was the non-unified nomenclature of genes, so for some scores, there were genes that did not exist in the list of genes for another score; therefore, these genes are lost. Further, it seems that the available gene classes are not ideal as there are still some groups not differentiated well from the rest of the groups; the reason might be because of the following limitations of the Spataro et al. gene categorization (90). The major is the precision of the genetic information presently available for human diseases, in addition to insufficient up-to-date knowledge regarding the true susceptibility of genes/variants to cause diseases. For example, the genetic bases of 50% of all known Monogenic diseases are not well understood, and most complex diseases remain unsolved; the real elements producing a lot of human disease phenotypes are not yet recognised. Furthermore, as the GWAS catalogue contains false-negatives and false-positives, the list of candidate genes harbouring the causal variants is usually reported based on biologists' knowledge and experience. Therefore, a proportion of human disease genes may be mis-allocated to corresponding phenotypes (90).
In this context, defining well-characterised gene groups is a great area for future study. Further, gene essentiality patterns and how essentiality in humans can be studied might help in prioritising disease genes. Moreover, identifying a criterion for genes that are highly likely to cause disease as well as Mendelian or complex diseases, along with the potential common genes for both types can improve genetic prediction. In addition, the role of non-essential genes might have a regulatory function or their influence and interaction might differ from one gene to another. All these factors need to be studied in order to achieve improved gene classification. Thus, this study updated the Spataro et al. gene groups (90) to improve group separation and improve gene prediction. Ultimately, there is now better coverage of Mendelian disease gene groups based on the best available evidence, and this helps in better understanding the data and makes the process of creating an essentiality prediction model feasible.

3.5 Conclusion

In this chapter, 11 gene-level metrics were evaluated; the aim was to assess the available gene-specific scores that might later be utilised to improve clinical diagnosis for patients suffering from monogenic diseases. Using the Spataro et al. classification model (90), the Mann Whitney U test was applied to evaluate the statistical differences between the two groups in pairwise comparisons using P values. Here, the aim was to evaluate the differences between the groups for better gene categorisation. The results of the Mann-

Whitney U test showed no statistically significant difference between CM and MNC groups for most of the variables, suggesting the combining of the two groups into one group called ‘Mendelian disease genes’. Furthermore, the results of the Kruskal Wallis test was integrated with the Mann Whitney U results, creating a clear direction of the data and showing a statistically significant difference between the non-essential and essential gene groups. To investigate each score further, I applied the first PCA analysis to predict the weight of each metric in the data for 11 scores; three scores were eliminated as they were the least useful in representing the data, and two of them were going in the opposite direction from what was expected. Subsequently, the second PCA analysis was produced for the eight remaining variables, showing better results by 12%. Ultimately, the next step will be utilising these eight scores to produce a composite score that might help in identifying Mendelian disease genes.

Chapter 4 Essentiality-specific Pathogenicity Prioritization

4.1 Introduction

Single gene disorders include those that follow the Mendelian pattern of inheritance in relatives and conditions arising in individuals through *de novo* deleterious variations. To resolve the underlying cause of these cases at a molecular level, it is essential to understand the disease phenotype in terms of the patient's genotype. This accomplishment can help refine diagnoses and make possible routes for better clinical management available. The OMIM database (24) lists approximately 3,800 genes underlying 5,470 Mendelian phenotypes. However, although ~ 69 percent of all known Mendelian phenotypes have a determined genetic source, many more Mendelian conditions have yet to be characterised (24).

A fresh review, using data from approximately 60 NHS hospitals in the United Kingdom and around 25 hospitals in other countries, found that only a small number of patients with hereditary rare diseases receive a genetic diagnosis (133). Even in cases in which the genetic cause is recognised, the chances of making a firm diagnosis may be reduced through incomplete characterisation of the patient phenotype or incomplete genetic testing, which might be restricted to a set of candidate genes that may not include the gene at fault. For some patients, the molecular basis of the condition is recognised after as many as 16 clinic visits, following almost three misdiagnoses in a journey, which might last more than two years (133).

Essential gene candidates have also been determined through experiments using technologies like CRISPR-Cas9 (134). These genes are very important for survival as the damaging variation is intolerable and likely to be preserved only by a selection/mutation balance. These genes are responsible for core cellular regulation, and any disruption of these functions may lead to fatal illnesses (134). Within the spectrum of essential genes, Cacheiro et al. (135) recognised variations between cellular lethal (CL) genes, which demonstrate nearly complete concordance with mouse lethal genes and are vital for both cell/organism survival, and developmental lethal (DL) genes, which are not essential at a cellular level, however, LoF variation in these genes might be fatal.

Recently, in 2019, Cacheiro et al. produced a new classification of genes considering a sub-

division of essential genes through their cross-species gene categorization called 'FUll spectrum of intolerance to loss-of-function variation' (FUSIL) (135). Their analysis supports the model by Pengelly et al. (62) model, which demonstrates disease-genes having intermediate essentiality.

This information is gene specific; consequently, the focus of this study is to build a model of genes, rather than just variants, which might improve filtering of disease genes and therefore, enhance identification of disease-causal genes.

Combining the most useful scores that represent the data as a single (ESPP) score, the new composite score will be evaluated against each individual predictor. Here, it is proposed that combining several gene predictors that measure different genetic features will enhance the composite score and produce a more powerful predictor than any of the individual scores.

Consequently, relationships between essentiality measures in different gene groups including non-essential genes not involved in disease, Mendelian disease genes, and genes classed as essential were assessed. Further, the ESPP score for almost 12, 000 genes that were constructed using gene-level predictors including RVIS, pLI, HI, SIS, etc. were introduced. These genes were classified into five groups using the Spataro et al. classification, which includes NDNE, CNM, CM, MNC, and END (90).

4.2 Materials and Methods

As the purpose of this study was to create a simple predictor for disease genes, a score was created based on the current available data at the gene-level. The performance of this classifier was then assessed.

4.2.1 *Gene classification*

First, the distribution of ESPP scores within different gene groups as defined by Spataro et al. (90) were considered, and 17982 genes were listed as done in their supplementary table S2. The aforementioned study along with its updated classification that has been done previously was used including four gene groups: NDNE, CNM, MDG, and END (Figure 4-1) (refer to Chapter 3 for more details).

Meanwhile, Cacheiro et al. identified a set of genes as strong candidates for developmental disorders, which was considered in this study. This is a sub-set of 163 genes classed as likely to be developmentally lethal (Supplementary table 7 in Cacheiro et al. (135)). These comprise

genes that are 'highly intolerant to loss of function variation' (pLI > 0.9) (38) or have 'gnomAD's observed/expected LoF scores with upper boundary <0.35' (https://gnomad.broadinstitute.org/)(136) or 'haploinsufficiency score (HI) < 10 (93) and [are] not currently associated with human disease by OMIM (24), Orphanet (137), or the Developmental Disorder Genotype-Phenotype Database (DDG2P)' (135, 136). Moreover, the gene sub-set are genes known to have *de novo* variants in the 100K Genomes undiagnosed cases with intellectual disability (around 50 genes), DDD cases with variants of uncertain significance (VUS) in undiagnosed children (approximately 50 genes), and ~ 15 genes from the Centre for Mendelian Genomics (CMG). The latter Mendelian candidate genes consist of Tier 1 genes, which have variations in multiple kindreds or are located within a linkage peak or linked with a phenotype summarised in a model organism or Tier 2, which are considered strong candidates, but with mutations only known in one kindred. Accounting for overlaps, this is a set of 82 genes (135).

4.2.2 *Constructing a gene-level score*

The aim of this project is to build a composite score at the gene-level to predict disease genes based on different genetic features measured by each single score involved in the composite score (refer to Chapter 2 for more details about every single score), therefore providing a score per gene to predict the position of that gene in the essentiality spectrum.

This analysis was performed using R Studio statistics software (140), version 1.0.153, 2009–2017, RStudio Inc. First, the list of genes (~18,269 genes) provided from ensembl was aligned with several essentiality scores from various studies along with the LDU_res-fit produced from LD maps. The following scores were chosen from this study's systematic review as approximations to essentiality: HI, RVIS, pLI, SIS, NET, REC, DNE, GDI, GIMS, and GHIS (122).

However, some of the scores were less informative than others and included LDU_res-fit, GDI, and REC as they explained <20% of the variance; these scores were therefore eliminated from further analysis (chapter 3). PCA was repeated using only eight scores, and the result is shown in Table 3-3 in the results section of Chapter 3. The result of PCA was improved after excluding the least informative scores, and it was possible to then produce a formula from PC1 to calculate the new score (ESPP) using the following equation: HI x 0.290 + DNE x 0.357 + RVIS x – 0.349 + NET x 0.272 + pLI x 0.352 + GIMS x -0.397 + GHIS x 0.360 + SIS x 0.423.

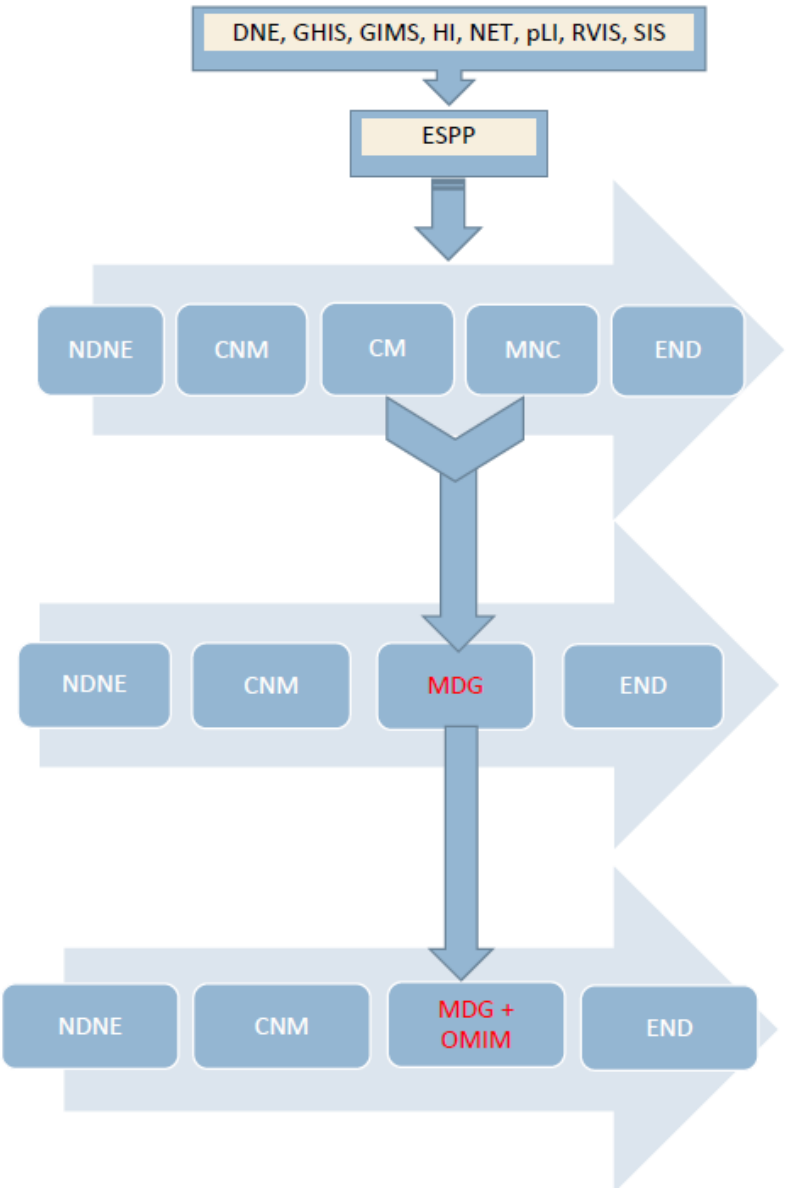


Figure 4-1 Essentiality specific pathogenicity prioritisation (ESPP) workflow.

The above simple diagram demonstrates the pipeline of constructing the ESPP score through merging eight gene-level scores and then combining them into a single score. ESPP utilised the Spataro et al. gene classification including the following five gene groups: NDNE, END, CNM, CM, and MNC (90). The arrows represent the direction of essentiality and gene groups from the least to the most essential. Next, CM and MNC groups have been merged as Mendelian disease genes (MDG). This group of genes was then updated using the OMIM updated list of genes to ensure coverage of all Mendelian disease genes (24).

4.2.3 *Evaluation of the relationships between measures of essentiality*

Spearman's correlation was performed for nine gene specific scores, including the ESPP score, to test the association of measures of essentiality with each other. The results are presented in Table 4-5 of the results section.

4.2.4 *Evaluation of ESPP performance*

Assuming that genes within the MDG group will score high with ESPP along with the essential genes, other groups were expected to score low. This study was interested in Not Available (NA) genes with high ESPP scores as these genes might be Mendelian disease candidate genes.

For this study, the updating of the Spataro et al. Mendelian disease gene list (90) by merging the MDG group with a list of all dominant and all recessive genes provided by Berg et al. (130) and Blekhman et al. (131) was decided. To this end, the most updated list of Mendelian disease genes was obtained from the OMIM database (24), which is updated daily to improve group separation. Ultimately, the 82 DL candidate genes recognised by Cacheiro et al. as strong candidates for developmental disorders were merged to evaluate the performance of ESPP scores among these genes. Genes with high ESPP scores > 4 and not assigned to MDG or END were extracted to investigate their function in OMIM (24) and to verify whether they were ever related to any Mendelian condition (Table 4-4).

4.3 Results

After evaluation of the 10 gene-specific scores that were derived from our systematic review (122), LDU_ res-fit from pre-existing whole-genome LD maps was used as a linkage-disequilibrium predictor (91) to construct the new ESPP model. The results of the first PCA showed that GDI, LDU, and REC were the least informative and were, thus, excluded from further analysis (refer to Chapter 3 for more details).

The outcomes of the PCA of 11 scores (Table 3-3) display relatively minor weightings for GDI (0.013), LDU (-0.055), and REC (0.1866). A second PCA was therefore undertaken for the eight scores that demonstrated a higher percentage of the variance.

The ESPP is derived from a linear combination of the first principal component weightings of the second PCA (Table 3-3). The combined variance explained is (0.48) with the highest weighting applied to SIS (0.42) and the lowest to the NET score (0.27).

Table 4-1 Numbers of genes with essentiality score assigned to each group (mean score in brackets)

Essentiality measure	NDNE	CNM	*MDG	END	Genes with score but no gene group	Totals of genes assigned to groups
Gene totals (Spataro et al. [8] (90) and OMIM [27] (24) classification)	10627	1732	4440	969	0	17768
DNE	10482 (0.621)	1730 (0.880)	3769 (1.025)	968 (1.651)	463 (0.564)	16949
GDI	10482 (192.264)	1730 (85.421)	3769 (124.199)	968 (189.18)	463 (2487.1)	16949
GHIS	8971 (0.522)	1557 (0.527)	3448 (0.532)	938 (0.566)	0	14914
GIMS	10177 (0.525)	1722 (0.463)	3722 (0.433)	958 (0.322)	371 (0.507)	16579
HI	10482 (0.183)	1730 (0.262)	3769 (0.304)	968 (0.411)	463 (0.118)	16949
LDU	10627 (-0.008)	1732 (0.237)	3836 (-0.034)	969 (-0.238)	1104 (0.041)	17164
NET	10482 (0.447)	1730 (0.557)	3769 (0.639)	968 (0.733)	463 (0.334)	16949
pLI	9956 (0.253)	1687 (0.360)	3674 (0.318)	941 (0.579)	356 (0.262)	16258
REC	10482 (0.098)	1730 (0.147)	3769 (0.232)	968 (0.200)	463 (0.059)	16949
RVIS	10482 (0.091)	1730 (-0.051)	3769 (-0.188)	968 (-0.389)	463 (0.160)	16949
SIS	9007 (-0.093)	1516 (0.077)	3174 (0.189)	805 (0.619)	0	14502

ESPP (from eight scores—excluding GDI, LDU, REC)	7076 (0.620)	1330 (0.884)	2914 (1.003)	760 (1.641)	0	12080

*MDG—Mendelian disease genes comprising combined CM and MNC from Spataro et al. (90) and the updated OMIM list (24). OMIM total number = 4428; Matched OMIM and Spataro = 3824

Table 4-2 ESPP score count by group and percentage of genes in brackets (eight scores)

ESPP score range	NDNE	CNM	MDG	END	Cacheiro et al. (135) DL Candidate genes (70 genes with ESPP score)
<-4	2 (0.02)	1(0.07)	1 (0.03)	0	0
-4 to -3	13 (0.18)	3 (0.23)	0	0	0
-3 to -2	65 (0.9)	10 (0.76)	16 (0.5)	0	1 (1.4)
-2 to -1	367 (5.2)	45 (3.4)	95 (3.2)	10 (1.3)	1 (1.4)
-1 to 0	1548 (220)	240 (18.4)	433 (14.8)	33 (4.6)	1 (1.4)
0 to 1	2576 (36.7)	411 (31.5)	1043 (35.8)	153 (21.3)	3 (4.3)
1 to 2	1678 (23.9)	373 (28.6)	724 (24.8)	276 (38.5)	26 (37.1)
2 to 3	668 (9.5)	188 (14.4)	434 (14.9)	185 (25.8)	32 (45.7)
3 to 4	82 (1.17)	30 (2.3)	121 (4.1)	44 (6.1)	4 (5.7)
4 to 5	8 (0.1)	3 (0.2)	36 (1.2)	11 (1.5)	1 (1.4)
5 to 6	0	0	4 (0.13)	2 (0.27)	1 (1.4)
>6	0	0	5 (0.17)	2 (0.27)	0
Totals	7007	1304	2912	716	70

% greater than 3 = NDNE = 1.3; CNM = 2.5; MDG = 5.7; END = 8.2. [63% of genes with ESPP > 3 are MDG/END; 82% of genes with ESPP > 4 are MDG/END]

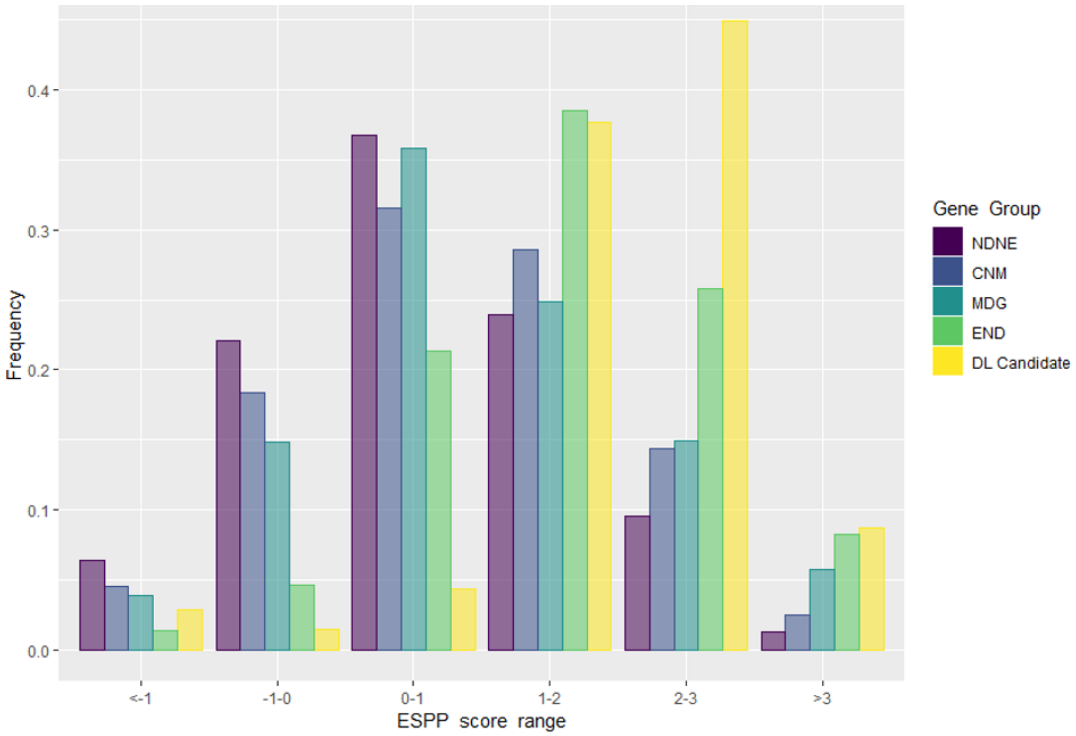


Figure 4-2 The frequency of genes with ESPP scores.

In the above figure, in each score range, it starts from the least essential (NDNE, CNM, and MDG) to the most essential (END and candidate DL genes adopted from Cacheiro et al. (135)). Further, most of the DL candidate genes fall between 1–3 ESPP score (37% fall between 1—2, and around 46% between 2—3).

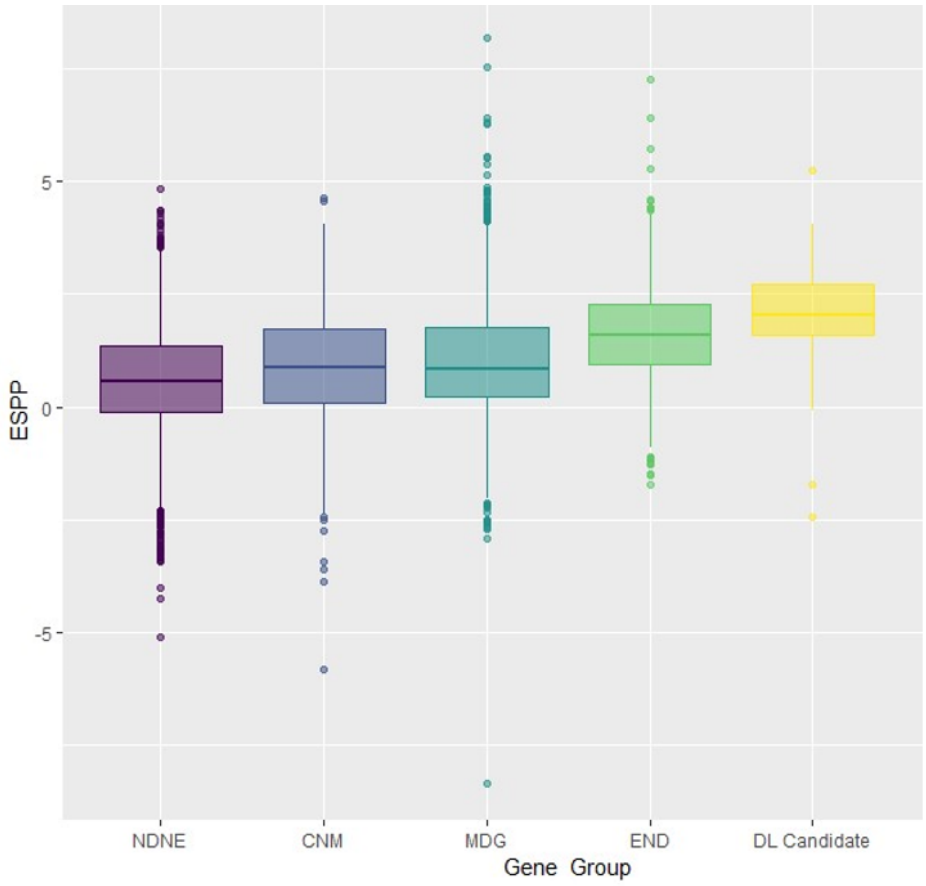


Figure 4-3 The median of the ESPP score in each gene group.

In the above figure, ESPP range starts from the least essential (NDNE, CNM, MDG) to the most essential (END and DL genes from Cacheiro et al. (135)). Here, the pattern of positive relationships between gene essentiality and the increase of ESPP values can be noted.

Figure 4-2 demonstrates the breakdown of ESPP scores within a score range with respect to each group. As shown, there is wide overlap between groups, which means that the features of genes explained by these scores cannot definitively allow genes to be placed into the groups. Nevertheless, the gene groups are classified according to current understanding; for example, unrecognised monogenic genes are mis-classified, and there is incomplete understanding of human essential genes, which might be improved by using CRISPR Cas 9 and gene trapping methods in identifying essential genes in humans (134). However, while Figure 4-2 displays the percentage of genes in each category within an ESPP score range, Figure 4-3 shows the median of ESPP in each gene group, demonstrating that the DL candidate showing 91% have ESPP > 1. Further, there is a clear separation between the peaks for NDNE and END genes. Large ESPP scores (> 2) have an excess of END genes (approximately 35% of highly essential genes scored more than 2 as per ESPP) and ESPP

scores of ≥ 3 are enriched for rare disease genes including 6% of Mendelian disease genes and 8% of essential genes as compared to 1.3% of NDNE and 2.5% of CNM gene groups. In general, 63% of genes with ESPP > 3 are essential/Mendelian disease genes, and the proportion increases to 82% percent for genes that scored more than 4 as per ESPP (Table 4-3). Thus, high ESPP scores are strongly indicative of genes at the rare disease/essential end of the spectrum.

Meanwhile, a total of 70 of the 82 genes recognised as strong candidates for developmental disorders by Cacheiro et al. (135) have values in terms of ESPP scores (Table 4-3, identified as 'DL candidates' in Figure 4-2): 10 of these genes are in the complex gene group, 25 in the essential gene group, 1 in the Mendelian disease gene group, and 34 are non-essential genes. The distribution of ESPP scores for these genes (Figure 4-2) is highly skewed towards ESPP with more than two in line with the expectation that most are strong candidate genes for monogenic diseases (139).

Table 4-3 Genes with ESPP score > 4 not assigned to MDG or END groups.

Gene	Group	ESPP score	Full name	Notes on gene function (OMIM)
ANKRD17	CNM	4.612	Ankyrin Repeat Domain 17	May mediate immune responses to bacteria and viruses
DIP2C	CNM	4.564	Disco Interacting Protein 2 Homolog C	May be involved in transcription factor binding
RYR3	CNM	4.039	Ryanodine Receptor 3	Involved in Ca(2+) signalling in neurons in the central nervous system
PLXNA1	NDNE	4.848	Plexin A1	Involved in cortico-motoneuronal connections underlying manual dexterity.
CNOT1	NDNE	4.359	CCR4-NOT Transcription Complex Subunit 1	May be involved in transcriptional regulation.

CHD5	NDNE	4.346	Cadherin 5	CDH5/beta-catenin signalling appears to control endothelial survival.
USP34	NDNE	4.281	Ubiquitin Specific Peptidase 34	May rescue ubiquitinated proteins from proteasomal degradation.
FRY	NDNE	4.200	FRY Microtubule Binding Protein	Involved in structural integrity of mitotic centrosomes and maintenance of spindle bipolarity.
SUPT5H	NDNE	4.084	SPT5 Homolog, DSIF Elongation Factor Subunit	May control key aspects of neuronal development.
PCDH17	NDNE	4.035	Protocadherin 17	May be involved in synaptic function in the central nervous system.
SUPT6H	NDNE	4.003	SPT6 Homolog, Histone Chaperone, and Transcription Elongation Factor	May regulate transcription through establishment or maintenance of chromatin structure.

- **Relationships between measures of essentiality**

Spearman's correlation

In order to test the power and direction of a monotonic relationship between the scores and the ESPP, Spearman's correlation was performed to produce a correlation coefficient. Table 4-5 gives the results of this analysis for the eight scores and combined ESPP. Correlations are relatively high throughout, and the correlation structure appears to be captured well by the combined ESPP score, which shows a higher correlation than any other scores with DNE, HI, and pLI and high correlation with other variables.

Table 4-4 Spearman's correlation coefficients for the eight scores and ESPP

Essentiality measure[1]	**GHIS**	**GIMS**	**HI**	**NET**	**pLI**	**RVIS**	**SIS**	**Combined ESPP score**

DNE	0.3107	_0.4437	0.3043	0.3082	0.5315	_0.3240	0.6117	0.8154
GHIS	-	_0.5224	0.3297	0.3387	0.4059	_0.5587	0.5466	0.4664
GIMS	-	-	_0.3901	_0.3305	_0.4613	0.5364	_0.6441	_0.5902
HI	-	-	-	0.3714	0.3416	_0.2740	0.3273	0.4894
NET	-	-	-	-	0.3062	_0.2472	0.3236	0.4908
pLI	-	-	-	-	-	_0.3443	0.5584	0.6295
RVIS	-	-	-	-	-	-	_0.6111	_0.5363
SIS	-	-	-	-	-	-	-	0.5019

4.4 **Discussion**

The combined ESPP classifier is linked to the genic essentiality hypothetical model proposed by Pengelly et al. (62), in which monogenic disease genes are located between non-essential and essential genes. Despite individual gene scores covering a variety of gene characteristics, the correlation between scores are relatively high (Table 4-5), which shows that combining these scores could be beneficial. Thus, a simple combined model was built to prioritise the recognition of monogenic disease genes by integrating the available gene-level predictors. Here, the ESPP predictor will likely enhance the modularity of various genetic properties measured by each score independently. However, this means that it is also impacted by the limitations and assumptions of every single score. Moreover, based on the PCA results of PC1, the ESPP score proposed combines all measures into a single model, explaining a higher percentage of the variance (45%, Table 4-1) than any single measure.

The NET predictor has the smallest weighting PCA of the eight scores that contribute most to the predicted variance (Table 4-1). The NET score was the first comprehensive genome-wide study that associates genetic variants at population level, in addition to disease variants with current network resources and may have been affected by the lack of biological network data at the time of the study, which was before the 1000 Genomes Project (110). On the other hand, the SIS score, which has an elevated overall influence, is more recent and utilised data from the 1000 Genomes Project. In order to improve recognition of gene properties that might be related to monogenic disease in the future, improving the understanding of genic properties—such as essentiality, selection, and mutation—as larger numbers of genomes are sequenced is critical.

Besides the effect of inconsistency in quality and completeness of every gene-specific metric, an added difficulty in the explanation of ESPP scores comes from the lack of a complete understanding of gene classification. Thus, the rationale behind this study is to recognise new genes that have not yet been assigned to the group of genes already known to be involved in monogenic disorders. Therefore, it is inevitable that genes in the existing gene group classification (Table 4-2) will be mis-classified. Moreover, to date, more than 30% of genes involved in monogenic disease have not yet been identified and accordingly, those are currently assigned to gene groups other than MDG. Additionally, recognizing essential genes is another challenge, since inactivation of an essential gene is fatal and unethical to test, so recognition of these genes in humans can only be made indirectly through homology or, lately through techniques such as CRISPR-cas9 (141).

In this context, a sub-division of essential genes was considered by Cacheiro et al. (135) through their cross-species gene classification termed FUSIL. Here, they integrated human, mouse, and CRISPR-Cas9 screening data and identified two classes of essential genes—CL and DL—as defined earlier. They also added diverse sets of sub-viable and viable genes determined from LoF mice trials (135). Their analysis is along the same lines as that proposed by the model by Pengelly et al. (62), which revealed the intermediate position of disease-genes in the essentiality spectrum. Further, their broadly characterised set of DL candidates genes show that 91% of them have ESPP > 1 (Figure 4-2), which means that the ESPP scores have a high potential in identifying disease genes.

Moreover, any gene classed as NDNE but with particularly high ESPP scores are probable monogenic disease candidates. Here, 63% of genes with a score of at least 3 are currently categorised as MDG or END. Moreover, 82% of genes with ESPP > 4 are MDG/END. Table 4-4 shows 11 genes currently assigned to these two categories, which have ESPP > 4. They consist of candidate essential genes currently categorised as NDNE (for example *SUPT6H, FRY*) and genes that are known to contain CNM variations, but have properties that suggest they are also candidate monogenic disease genes (for example *RYR3, DIP2C*). Here, it is worth noting that disease–gene relationships are complex, and the variety of gene characteristics restrict the ability of individual and combined predictors to fully distinguish certain gene classes. For instance, the score developed by MacArthur et al. was based on human-macaque conservation and proximity to known recessive genes in protein interaction networks (12). Although their recessive metric, which describes the chance of a gene having recessive variation, provides a degree of discrimination between loss-of function tolerant and recessive genes, there is a significant overlap. Thus, importance of these scores is to

supply useful information to prioritise potential candidates in a genome filtering context. Moreover, with the marked rise in the number of genomes being sequenced, a better understanding of genic properties and functions is expected to enhance identifying genes likely to contain monogenic disease variations.

Given a sequenced genome, for which there is a number of potential functional candidate variants in different genes, access to the available ESPP scores provides a basis for ranking candidates objectively. For instance, genes with ESPP scores of 2 or greater appear particularly interesting in this context. Thus, a worthwhile basis for prospective studies would be to enhance the performance of the classifier in an effort to merge additional genomic and functional gene characteristics (133,139), in conjunction with enhancing gene classification given developing knowledge.

4.5 Conclusion

In this chapter, a composite score comprising eight gene-level metrics was constructed with the aim of predicting disease genes that will eventually clinically help in the accurate diagnosis of monogenic disease patients. More specifically, PCA analysis was used to produce an ESPP score for a total of 12081 genes. The analysis was started with 18873 genes, but there were limitations with the data that interfered with the computation of ESPP scores for some genes as mentioned in Chapter 3. Ultimately, a validation of ESPP is needed to assess if the ESPP was better at predicting disease genes (scored high) than eliminating non-disease genes, which scored low.

Chapter 5 Using/integrating Scores to Predict New Mendelian Disease Candidates

5.1 Introduction

The functions of many genes in the human body remain a mystery. One way of identifying the function of a system is to introduce a variant (mutation) into a gene and explore the impact of this mutation by assaying the effect on a model organism or cell line and observing the phenotype (143).

The main obstacles preventing the large-scale engineering of LoF mutations in humans are ethical and technical restrictions. However, exome and genome sequencing technologies have revealed a high volume of natural LoF variations in humans, which can be used as natural models for human gene inactivation. These variants have facilitated the identification of disease mechanisms by studying the basis of severe Mendelian diseases. These variants have also been shown to be valuable in discovering therapeutic targets—for instance, confirmed LoF variants in the *PCSK9* gene have been proven to be associated with low density lipo-protein cholesterol levels, leading eventually to the production of *PCSK9* inhibitors that are used now to decrease the risk of cardiovascular diseases. Thus, creating a catalogue for human LoF variants and classifying genes based on their tolerance to functional variations will provide an important resource for human variation.

Additionally, *in-silico* metrics that predict the ability of a gene to tolerate LoF variation can help in the clinical interpretation of human genomes and make advancements in the discovery of human disease genes (144). In this context, the increasing size of publicly available variant/gene databases from large populations make the process of evaluation of the performance of gene metrics feasible. The aim of this chapter was to evaluate the performance of ESPP using the DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER) (138) as well as SHGP data, and integrate the most recent gene-specific scores. This work highlights a set of genes not currently known to harbour variation underlying Mendelian traits, but which are strong candidates based on their properties. Further, the recognition of these genes should facilitate the interpretation of disease genome sequence data. The new gene-specific scores include the gene-level variation intolerance metric (GeVIR) (144), the loss-of-function observed/expected upper

bound fraction (LOEUF) (143), and VIRLoF, which is a combination of the aforementioned scores, GeVIR and LOEUF (144).

5.1.1 DECIPHER

The ESPP score was evaluated using DECIPHER (138), a web-based database that enhances the clinical interpretation of a variant using a variety of bioinformatics tools and resources. The aim of this database was to improve interpretation of candidate variants from genome-wide analyses and focus on variant confirmation to place unknown variants into a known variant list (138). DECIPHER contains 439,563 sequence variants from ClinVar, 223,342,519 sequence variants from gnomAD, and another 139,452 sequence variants from HGMD (138). Genes scored high by ESPP and therefore, considered more likely to be associated with diseases were explored in DECIPHER to determine any known or predicted disease relationships,

5.1.2 GEL data

The aim of the 100K genome project, launched in 2013, was to facilitate new scientific discovery and help in the development of the UK genomics industry through sequencing 100,000 individuals (NHS patients with rare diseases, and patients with common cancers) (37). The project was focused on improving patient diagnoses in these cases. The 100,000 Genomes Project is funded by the National Institute for Health Research, The Wellcome Trust, NHS England, Cancer Research UK, and the Medical Research Council (37). GEL data has been used to check candidate genes that were found to be causal on GEL but will not be in the OMIM database until published.

5.1.3 SHGP

The SHGP is a national program funded as part of Vision 2030 for Saudi Arabia launched in 2013 in Riyadh by the King Abdulaziz City for Science and Technology. The aim of the project was to sequence more than 100,000 individuals within five years to enhance the identification of the genetic basis of rare and common genetic diseases in the Saudi population. According to their statistics approximately 40,377 individuals have been sequenced, identifying ~ 7000 disease variants (145). A portion of these data have been made available and were accessed remotely and comprise around 987 exomes in the csv

format. Ethical approval was obtained from both the King Faisal Specialist Hospital and Research Centre and the University of Southampton through the ERGO system (Submission ID: 48601). More specifically, the ESPP score will be tested against these data, and since it is rich in homozygous variants, it is a basis for future work in terms of constructing a score to predict recessive genes.

5.2 Recently developed gene-level scores for comparison and integration with ESPP

5.2.1 GeVIR

GeVIR is a continuous gene-level variation intolerance metric produced by Abramovs et al. (144). This score was built based on the length, evolutionary conservation, and number of variant intolerant regions (VIRs) in gnomAD, which is the second version of ExAC that aims to harmonise exome and genome sequencing data from different sequencing projects. The VIRs are segments lying between two protein-altering variants. To calculate the score the VIRs of each length was counted in the canonical transcript of almost 19,400 genes, and then the weights of each VIR length were produced based on frequency among all genes. To calculate the GeVIR score for each gene, the following equation was used:

$$\text{Gene score} = \sum (W \times \text{GERP}) / N \text{ regions}$$

Where ($W \times$ GERP) = high-covered VIR weights (W) adjusted by their conservation (GERP) (146), and N regions = the total number of regions in a gene including low-coverage regions (N regions), where high weights were correlated with high evolutionary conservation (144).

The aforementioned score is distinct from ESPP, which integrates a range of properties including the evolutionary conservation, length, and number of variant intolerant regions, which looks to be a fairly independent way of scoring genes and is, therefore, useful and interesting for comparison with ESPP.

5.2.2 LOEUF

Karczewski et al. placed each individual gene on a spectrum of LoF intolerance by producing a predicted loss of function (pLoF) variation measure (143). They defined the pLoF by frameshift mutations, premature stops (stop-gained), or alteration of the two-

essential splice-site nucleotides immediately to the left and right of each exon (splice) found in protein-coding transcripts. They then created the loss-of-function transcript effect estimator (LOFTEE) package to eliminate annotation artefacts in these variants. They found that this method eliminates common pLoF variants in the population, which were found to be enriched with annotation errors (143). LOFTEE discriminates annotation artefacts from high confidence pLoF variants and predicts candidate splice variants outside the essential splice site. Further, approximately 443,770 high-confidence pLoF variants in 16,694 genes were discovered using this model.

The LOFTEE model was designed to predict expected levels of variation under neutrality. Under the LOFTEE model, 'the variation in the number of synonymous variants observed is accurately captured (r = 0.979)'. Additionally, this model was used to identify depletion of pLoF variants by measuring the differences between the number of observed pLoF variants and their expectation in the gnomAD exome data from approximately 126,000 individuals. By using gnomAD data, an assessment of the degree of intolerance to pLoF variation was possible for each gene by using the continuous metric of the observed/expected (o/e) ratio, where o and e refer to the observed pLoF variants and expected number of LoF variants in the genomAD, respectively. The 90% upper boundary of the confidence interval (CI) was used to produce the LoF observed/expected upper bound fraction (LOEUF) (143). Meanwhile, the haploinsufficient genes were strongly depleted in the pLoF variations, while, in contrast, less essential genes, such as genes encoding olfactory receptors, were more tolerant of pLoF variations. Here, it is worth noting that the concept of the LOEUF score is a bit similar to ESPP, in which both scores are attempting to assess the degree of gene intolerance to LoF variants. However, the method varies as LOEUF produces predicted LoF measures and ESPP relies on multiple gene scoring systems.

LOEUF was examined for 390 genes, which are embryonically lethal upon heterozygous deletion in the mouse model system. It was found that these genes have a lower LOEUF in comparison with the rest of the genome comprising ~19,300 genes (Figure 5-1 A). Likewise, 678 genes characterised by CRISPR screens as essential were depleted in the pLoF variation (low LOEUF score) as compared to around 860 non-essential genes (Figure 5-1 B) (143).

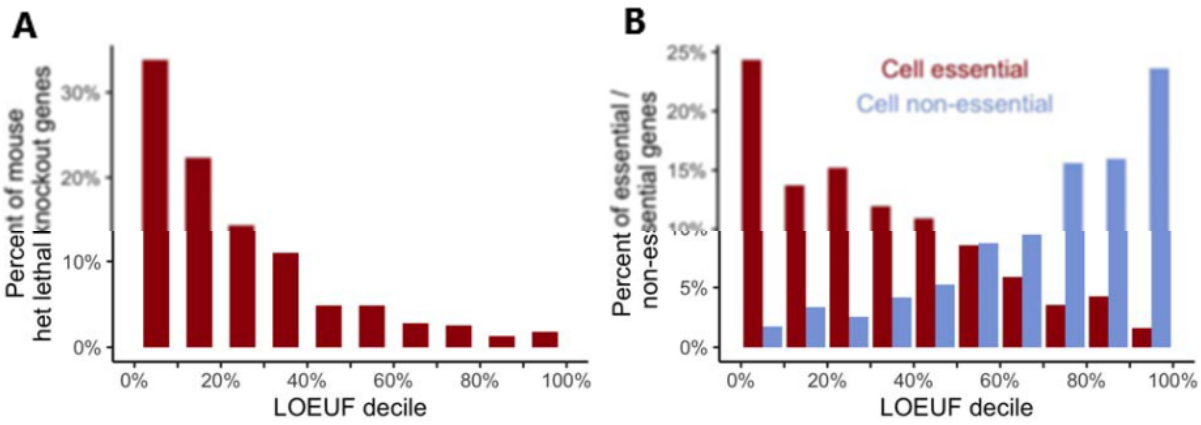


Figure 5-1 The functional distribution of LOEUF scores (adapted from Karczewski et al. (143)).

A: Highly constrained genes, when heterozygously inactivated in mice, are more likely to be fatal, and consequently, might be lethal in humans. However, unconstrained genes are shown to be more tolerant of disruptions. B: On the right side are the most unconstrained or non-essential genes, while the left side represents highly essential genes.

5.2.3 CoNeS

CoNeS is a gene-level metric that measures the strength of negative selection acting on genes (147). This score was developed by integrating information from the following scores: LOEUF, SnIPRE, LofTool, EvoTool, SIS, pLI, and RVIS. Therefore, a PCA was undertaken after data standardization. Moreover, genes with low CoNeS score are strongly impacted by negative selection. Further, this score showed a positive correlation with LofTool, EvoTool, LOEUF, and RVIS, while it showed negative correlation with pLI and SIS. Rapaport et al. used the hOMIM database and essential genes in mice and subsequently, produced the CoNeS score for 18,506 genes. It is noteworthy that some of the same component scores have been used in constructing ESPP scores, which include SIS, pLI, and RVIS; therefore, overlap is to be expected. However, the difference between the two scores is that the focus of CoNeS is on predicting genes that are impacted by negative selection, while ESPP integrated metrics in which some predict the impact of negative selection while other try to predict Mendelian disease genes.

5.2.4 Comparison of GeVIR and LOEUF

The comparison of GeVIR with LOEUF scores showed that the latter was more biased toward longer genes than the former, with Spearman's $r = -0.54$ and $r = -0.26$, respectively (144). More specifically, the first decile the median protein length of LOEUF was approximately 1.91 times longer than the expected 425 amino acids, while GeVIR was ~1.15 longer than expected. Here, VIR coverage was measured as the mean exome coverage of nucleotides, including region start and stop variants. Exome, rather than genome, coverage was assessed because the majority of samples in the gnomAD database are exomes (~ 123,136 out of 138,632), and exome coverage was less stable than genome coverage and henceforth, could better highlight potential biases in the variant load. However, as absence of a variant might be a result of low coverage, strict filters were applied to help distinguish high and low coverage VIRs (144).

Next, the GeVIR and LOEUF scores were combined by rank summation with re-sorting to develop a new score called VIRLoF to rank genes by intolerance to both missense and LOF variations. VIRLoF shows better ranking and higher performance than GeVIR and LOEUF by measuring Area Under the ROC Curve (AUC) with higher AUC and indicating better performance (144). The set of missense and LOF intolerant genes incorporate approximately 32% of the known AD disease genes and only around 2% of the known AR genes (Figure 5-1 A,B). Furthermore, a large percentage of these genes (around 70%) were not found to be associated with any OMIM phenotypes (144).

Moreover, identifying dominant genes is comparatively more challenging than identifying recessive genes due to the fact that in dominant conditions with monoallelic inheritance, there is redundancy through many actually neutral heterozygous variants that act as massive background noise (116).

Meanwhile, in Mendelian diseases, only one or two deleterious variants must be identified among a large number of variants that naturally occur in the human genome. A sequence patient human genome might carry 20,000 exonic variants, of which 400 will be good-quality, nonsynonymous, and rare DNA variations (116). If two of these variants are found in the same gene, they would most likely underlie a recessive condition, decreasing the number of candidate genes to five to 10 genome-wide. On the other hand, any heterozygous variants among these 400 variants might be associated with dominant disorders, making identification of the dominant genes more difficult (116). Accordingly, the power of NGS analysis to detect genes associated with recessive disorders is 10-fold more efficient than

the detection of dominant genes. In this context, the identification of rare alleles as a function of their deleterious potential in the heterozygous state signifies a real challenge in solving dominant cases. Several *in silico* tools have been established to prioritise the deleterious effect of DNA changes (116). So far, the majority of these methods have focused on the pathogenicity of a variant on protein function rather than differentiating dominant and recessive variants. Further, other tools were developed to predict haploinsufficient genes, which will partially help in identifying dominant genes as dominant variants produce a haploinsufficient phenotype; however, dominant variants might arise by the gain of function as well (116). Here, as this study is focused on predicting Mendelian disease genes, it is worth looking at dominant and recessive genes and exploring their patterns in the essentiality spectrum.

In order to test the performance of the ESPP score, the ESPP percentiles were compared to the three gene constraint metrics (GeVIR, LOEUF and VIRLoF) by sorting genes based on the updated Spataro et al. classification—that was used in Chapter 3 to identify the relationship between these scores and the extent to which they can predict disease genes. Further, known dominant and recessive genes in relation to the ESPP score and the Spataro et al. updated classification were also examined (please refer to chapter 3 for more details) (90).

5.3 Methods

5.3.1 Investigating candidate genes prioritised by ESPP using DECIPHER and GEL data

In this study, genes with an ESPP score > 4 that were not assigned to MDG or END groups as listed in Chapter 4 (Table 4-4) were investigated using DECIPHER and GEL data. Each gene was first examined in DECIPHER and then in GEL data to identify whether it was recently predicted to be associated with any clinical phenotypes (See results section).

5.3.2 Investigating candidate genes prioritised by ESPP using SHGP

The ESPP score was evaluated within the SHGP data (145). The data consists of 987 exomes that were accessed remotely from within UK after obtaining the ethical permission from both sides. The data was received in csv format and was accessed using a terminal

through their server. A list of 483 confirmed disease causing (mostly recessive) variants was received enabling evaluation of the performance of ESPP using confirmed cases.

To test the performance of ESPP, each confirmed variant was searched in the exome data of the corresponding patient to extract more information about this specific variant. As a result, a database of 143 variants was constructed including REVEL (80), GWAVA (148), SIFT (68), Polyphen-2 (70), MutationTaster (73), MetaSVM (81), M-CAP (149), and CADD (72) scores. Unfortunately, the majority of the accessed csv files were not annotated, and access to the data from outside Saudi Arabia became more restricted preventing wider investigations. However, considering the well-established set of 143 variants, the data were aligned with the ESPP score for each gene to create a data set of 107 confirmed variants with an ESPP value for each corresponding gene.

Details of library preparation and WES sequencing of Saudi data

The following description was received from Sateesh (personal communication, Feb, 22, 2020) at the Saudi project*:* Exome capture was performed using TruSeq Exome Enrichment kit (Illumina) following the producer's protocol. The preparation of the samples was done as an Illumina sequencing library, and in the second step, the sequencing libraries were enriched for the desired target using the protocol of Illumina Exome Enrichment. Illumina HiSeq2000 Sequencer was used to sequence the captured libraries. The reads were mapped against UCSC hg19 (http://genome.ucsc.edu/)(150) by BWA (http://bio-bwa.sourceforge.net/)(151). The SNPs and indels were detected by SAMTOOLS (http://samtools.sourceforge.net/)(152) (145).

5.3.3 Comparison of ESPP score with LOEUF

Data of the LOEUF scores were downloaded from gnomAD website (136). The LOEUF score is available for 19705 genes. This score was aligned with the ESPP score and successfully matched 12013 genes. Subsequently, the ESPP data and LOEUF were aligned with the Spataro et al. groups (90), and a comparison was made to show which of the two scores best fits the Spataro groups and whether there is a correlation between the two scores (refer to the results section).

5.3.4 Comparison of ESPP score with CoNeS

Data of CoNeS score were downloaded from this link (147). CoNeS score is available for 18506 genes. This score was aligned with the ESPP score and successfully matched 12566 genes. Data on ESPP and CoNeS were aligned with the Spataro et al. groups (90), and a comparison was made to show which of the two scores best fits the Spataro groups. Further, a regression analysis was performed to test the correlation between the two scores (refer to the results section). Afterwards, for the analysis, the data on ESPP were aligned with LOEUF and CoNeS, and a list of 11711 genes was retained.

5.3.5 Comparison of LOEUF score with CoNeS

The LOEUF score was aligned with the CoNeS score and successfully matched 11711 genes as previously mentioned. A regression analysis was done for the two scores to investigate which best fits the Spataro classification (refer to the results section).

5.3.6 Predicting dominant and recessive genes using ESPP, LOEUF, and CoNeS

Known recessive and dominant genes were categorised by ESPP, LOEUF, and CoNeS scores. A list of 985 recessive and dominant genes was obtained from Quinodoz et al. (116). Of these, 762 Dom (Dominant) and Rec (Recessive) genes matched ESPP scores, of which 232 genes were dominant and 530 were recessive. When the complete data was matched including ESPP, LOEUF, and CoNeS, 617 recessive and dominant genes were retrieved, of which 510 were recessive and the rest were dominant.

Further, the updated Spataro et al. classification of genes was used to classify genes into four categories: NDNE,CNM, MDG, and END (refer to list of abbreviations) (90). The MDG group was further classified into dominant and recessive based on Quinodoz et al. (116) (refer to the results section).

5.4 Results

5.4.1 Evaluation of candidate genes prioritised by ESPP within the DECIPHER and GEL data

The result of investigating genes with an ESPP score > 4 that were not assigned to MDG or END groups using DECIPHER and GEL data is listed in Table 4-4

The *CNOT1* gene that was prioritised as a potential candidate disease gene by ESPP (with a 4.359 ESPP value) (Table 5-1) has been recently recognised in two studies as a cause of holoprosencephaly (147,148). The first report was by De Franco et al. who investigated an international cohort of 107 individuals diagnosed with pancreatic agenesis— 'defined by requiring endocrine (insulin) and exocrine (pancreatic enzymes) replacement therapy within the first 6 months of life'; here, mutations in known genes were recognised in 98 of these individuals. In order to identify *de novo* mutations in the remaining nine individuals, exome sequencing was done as trios (the probands and unaffected parents), which was subject to availability (n = 7) (153). A heterozygous missense mutation in *CNOT1* was then identified (NM_016284.4; c.1603C>T [p.Arg535Cys]) in three individuals with pancreatic agenesis. However, the p. Arg535Cys variant was absent in dbSNP138, DECIPHER, and GnomAD. This variant affects a highly conserved residue across species. Meanwhile, the in silico prediction tools (AlignGVGD, PolyPhen2, and SIFT) predicted that the variant will have a pathogenic effect on protein function.

This variant resulted in a syndrome of pancreatic agenesis and abnormal forebrain development in three individuals with a similar phenotype to that in mice. Here, *CNOT1* was found to be an important gene for maintaining embryonic stem cells that differentiate other types of cells. Thus, these results suggest that *CNOT1* has a crucial role in the formation of pancreatic tissues (153).

At the same time, another group studied two unrelated patients with semi lobar holoprosencephaly. Exome sequencing was performed for both patients and an identical *de novo* missense variant was identified in the *CNOT1* gene. The variant (c.1603C>T [p.Arg535Cys]) is predicted to be deleterious and is not present in public databases (154).

Furthermore, considering the GEL data, it was found that the *RYR3* gene, which was prioritised by ESPP (with a 4.039 ESPP value) (Table 5-1), is related to an intellectual disability case with a *de novo* splice donor site mutation (155). This is, therefore, a strong candidate disease gene.

Table 5-1 Updates on genes with ESPP score > 4 not assigned to the MDG or END groups

Gene	Group	ESPP score	Full name	Notes on gene function (OMIM)	Update
RYR3	CNM	4.039	Ryanodine Receptor 3	Involved in Ca (2+) signalling in neurons in the central nervous system	*RYR3* is related to an intellectu disability (155).
CNOT1	NDNE	4.359	CCR4-NOT Transcription Complex Subunit 1	May be involved in transcriptional regulation.	Causes a novel genetic syndrome of pancreatic agenesis and holoprosencephaly, De Franco et al. (153). *CNOT1* is associated with holoprosencephaly; Kruszka e al. (154).

5.4.2 Results of testing candidate genes prioritised by ESPP using SHGP data

The ESPP score was examined in relation to 107 confirmed cases—for which the disease causal variant was known—from the SHGP data, in which the molecular diagnoses were made based on known MDG genes that they recognised in their patients. The data show that 82.2% out of the 107 variants are in genes classified as MDG/END by ESPP as compared to 2.8% classified as CNM and 14.9 % as NDNE. Genes that were classified as CNM and NDNE were described as novel candidate genes discovered by autozygosity mapping on 143 multiplex consanguineous Saudi families using WES in 2015 (156). This homozygosity scan, which showed genomic regions that are identical by descent, helps in identifying recessive diseases with atypical phenotypes such as those in neurogenetic diseases. Further, the autozygome in multiplex consanguineous families can be utilised after excluding known disease genes in discovering new candidate genes. These genes were assigned as causal, and

the next step was to confirm those genes using GenCC (157), which can validate causality for newly discovered genes based on various criteria and several databases (e.g., ClinGen, DECIPHER, OMIM, GEL PanelApp etc.).

Here, it is noteworthy that 100% of the genes that have ESPP > 3 were classified as MDG/END, 90% of genes that have ESPP > 2 were classified as MDG/END, and ~82% of genes that have ESPP > 1 were classified as MDG/END.

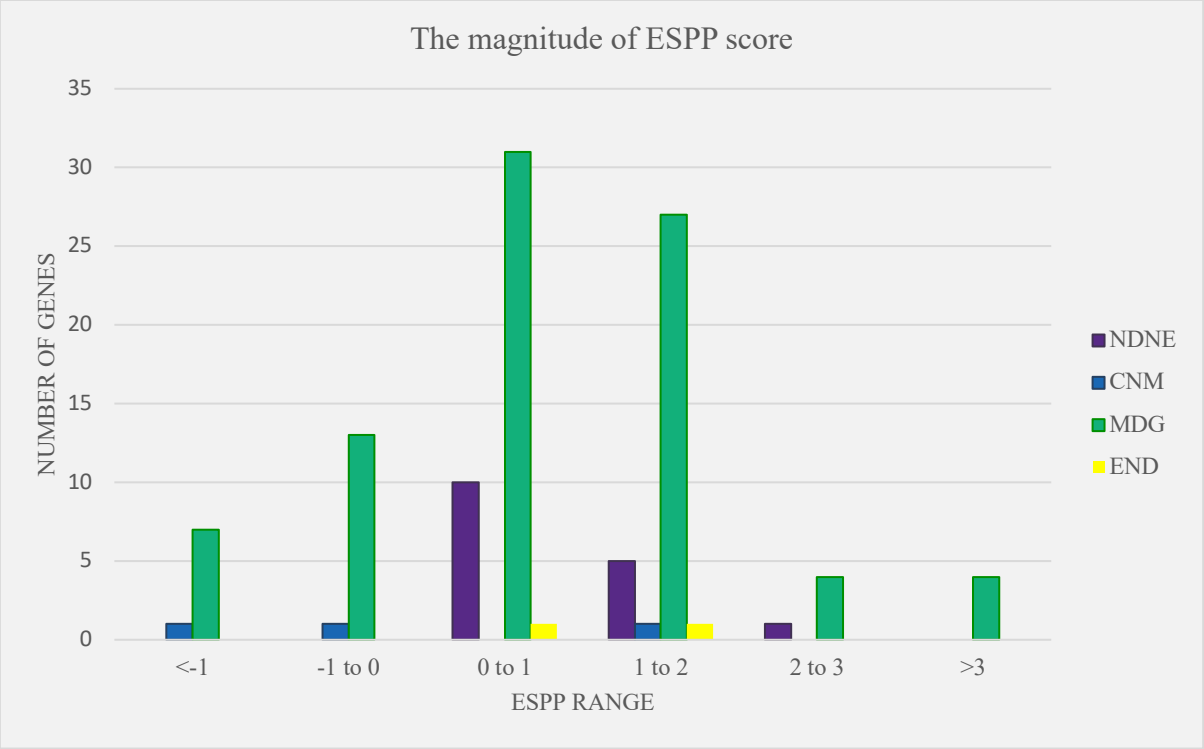


Figure 5-2 The magnitude of ESPP score for 107 confirmed cases from Saudi data in each gene group.

Figure 5-2 demonstrates that almost 78% of genes scored > 0 by ESPP were MDG, 2% were END, around 1% were CNM, and almost 19% were NDNE. Further, 81.4% of genes that scored > 1 were MDG, 2% were END, 2% were CNM, and almost 14% were NDNE. Moreover, four out of the 19 genes that were assigned as CNM/NDNE have been recently found to be causal in OMIM: *YIF1B* (158), *NEMF* (159), *METTL5* (160), and *PTPN23* (24,155). The rest of the genes have been checked using the sumRank of ESPP, LOEUF, and CoNeS. The sumRank helps in identifying genes that might be essential or disease genes which score low by sumRank. Genes that have high sumRank goes toward the non-

essential end. The scores for those genes are presented in Table 5-2. Most of those genes got very high sumRanks, and all of them had scores > 0 as per ESPP.

Table 5-2 The sumRanks of genes found to be causal in Saudi data and classified as non MDG.

Name	Group	ESPP	LOEUF	CoNeS	ESPP Rank	LOEUF Rank	CoNeS Rank	Sum Ranks
ST7	NDNE	1.988	0.264	-1.247	1921	1384	1658	4963
ZNF219	NDNE	2.547	0.124	-1.532	836	272	865	1973
NID1	NDNE	0.048	0.65	0.043	13	5	12	30
STXBP3	END	1.369	0.593	-0.293	5	4	6	15
NUDT2	NDNE	0.525	1.45	-0.010	7038	10355	6294	23687
AKR1E2	CNM	-1.056	1.419	1.403	15	14	15	44
ARL14EP	NDNE	0.228	0.693	-0.336	12	7	5	24
WDR59	NDNE	1.793	0.723	-0.237	3	8	8	19
CYP51A1	END	0.599	1.011	-0.145	10	12	9	31
TXNDC15	NDNE	0.859	0.757	-0.072	8	9	10	27
ARL6IP6	NDNE	0.663	1.047	-0.287	6412	7976	4929	19317
BIVM	NDNE	0.924	0.672	-0.689	7	6	3	16
ZNF526	NDNE	1.378	0.471	-0.343	4	3	4	11
LOXL3	NDNE	0.964	0.86	0.400	6	10	13	29
LRRC34	CNM	-0.257	0.998	0.957	14	11	14	39

Comparison of gene-specific scores

The ESPP score was compared to the most recent gene-specific scores, LOEUF and CoNeS, to test their correlation and determine whether combining scores might better predict Mendelian disease genes.

5.4.3 Results of the comparison of ESPP and LOEUF on the essentiality/disease genes spectrums

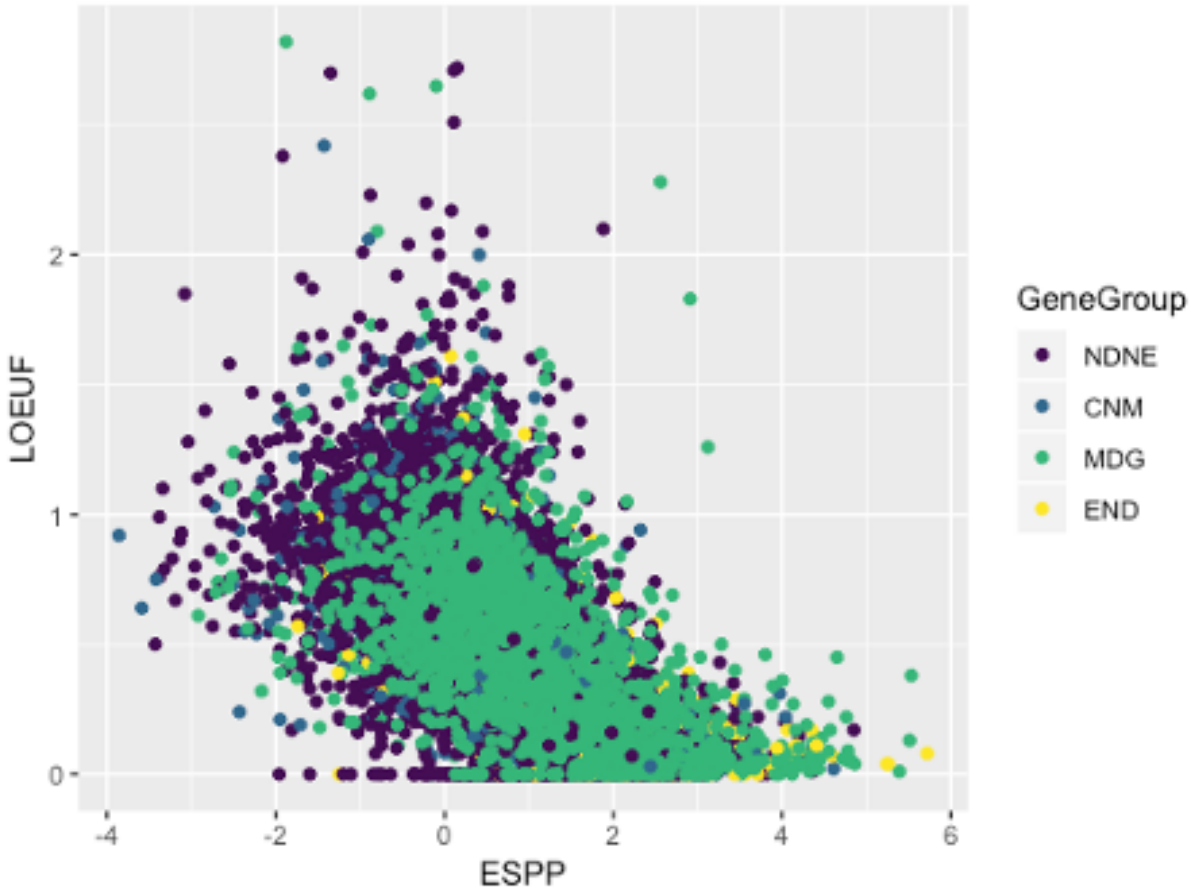


Figure 5-3 The distribution of genes using LOEUF versus ESPP scores based on the updated Spataro et al. gene groups (90).

Figure 5-3 shows that LOEUF and ESPP display the expected correlation, given the differing direction of the scores, although the R-square values shows that the variability in ESPP is explained by 38% variability of LOEUF (P = 0.00) in 11711 observations. Further analysis was done to explore how these scores are distributed in relation to the Spataro groups.

Among 12080 aligned genes with ESPP and LOEUF scores, 11711 genes that have LOEUF and ESPP values after discarding NAs were retrieved. According to Abramovs et al., the hard cut-off in LOEUF for essential genes is 0.35 (genes scoring < 0.35 are classed as essential). The number of genes that scored less than 0.35 in LOEUF was 2222 genes.

Further, 312 out of 2222 genes were successfully classified as END (14 %), 605 as MDG (27%), 306 as CNM (13.8%), and the rest (999 genes) as NDNE (45%) (Table 5-3).

Table 5-3 The distribution of genes which scored < 0.35 as the hard threshold of LOEUF for the most constrained genes

LOEUF hard threshold for most intolerant genes	**Observations of genes which scored < 0.35**	**NDNE**	**CNM**	**MDG**	**END**
< 0.35	2222	999	306	605	312
Percentage	18.5% of total matched data	45%	13.8%	27%	14%

Table 5-4 The distribution of genes scored > 0, > 1 and > 2 by ESPP

ESPP	**Observations of genes scored > 0.00**	**NDNE**	**CNM**	**MDG**	**END**
> 0.00	9189	5072	1030	2370	720
Percentage	77 % of total matched data	55 %	11 %	26 %	7.8 %
> 1.00	5004	2497	618	1327	565
Percentage	41 % of total matched data	50 %	12 %	27 %	11 %
> 2.00	1894	793	232	602	270
Percentage	16 % of total matched data	42 %	12 %	32 %	14 %

If we take zero as the cut-off of END, the number of genes scored > 0 by ESPP in this study's data is 9189. The ESPP score successfully classifies 720 genes as END (7.8%), while 2370 were classified as MDG (26%), 1030 as CNM (11%), and 5072 as NDNE (55%) (Table 5-4).

If we assume that 1 is the ESPP cut-off for candidate essential genes, the total number of genes scored > 1 by ESPP is 5004. Within this gene set, the Spataro classification has 565 genes as END (11%), 1327 as MDG (27%), 618 as CNM (12%), and 2497 as NDNE (50%) (Table 5-4).

If we assume that the cut-off for essential genes according to ESPP is > 2, the total number of genes scored > 2 by ESPP is 1894 genes. More specifically, 270 genes were classified as END (14%), 602 as MDG (32%), 232 as CNM (12%), and 793 as NDNE (42%) (Table 5-4). However, a comparison of ESPP with LOEUF to identify which one is better for predicting END/MDG groups according to the Spataro classification was not conclusive. The reason could be that LOEUF defined the constraint genes as any gene that is haploinsufficient with pLI > 0.9, and those showing a strong depletion of pLoF variation (first LOEUF decile). The number of genes in this category was 2222, and surprisingly, most of the genes falling into the range of LOEUF < 0.35 were classified by Spataro et al. (90) as NDNE accounting for 45% and only 14 % assigned as END. This suggests that the LOEUF score might not have a linear interpretation with groups classified as per Spataro et al., and the NDNE class within the same groups (which this author recognises as likely to contain novel Mendelian genes) is in need of more investigation (Figure 5-4, Figure 5-5).

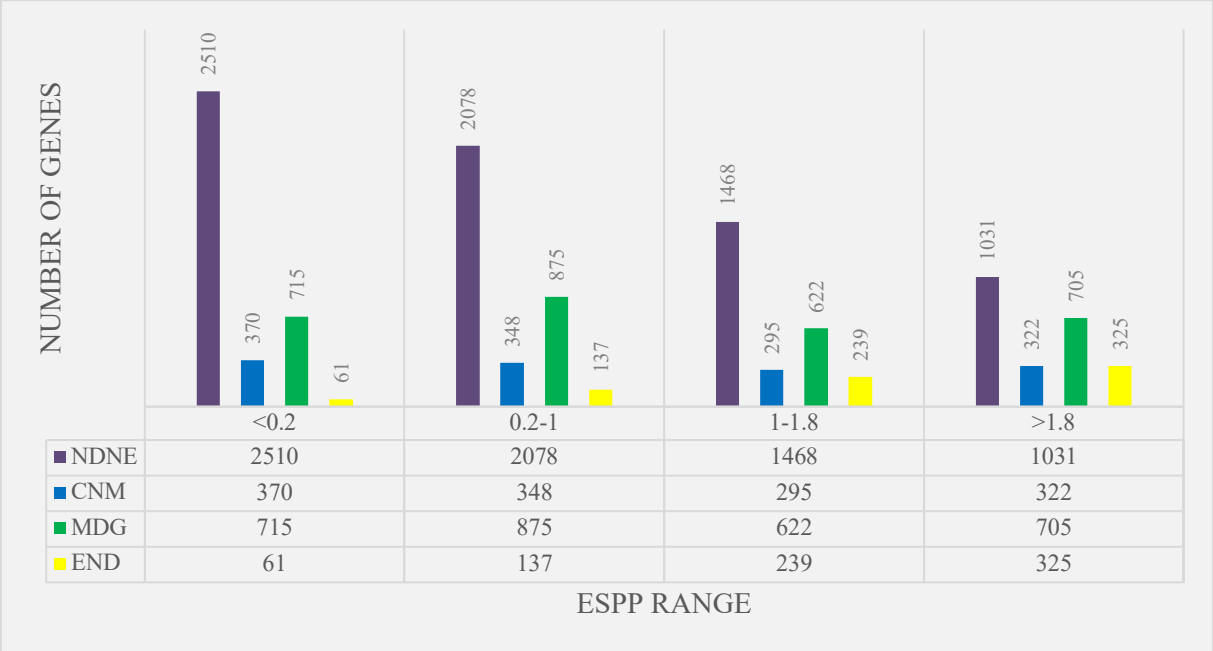


Figure 5-4 Percentage of genes among each gene group of Spataro et al. classification (90) according to the ESPP score

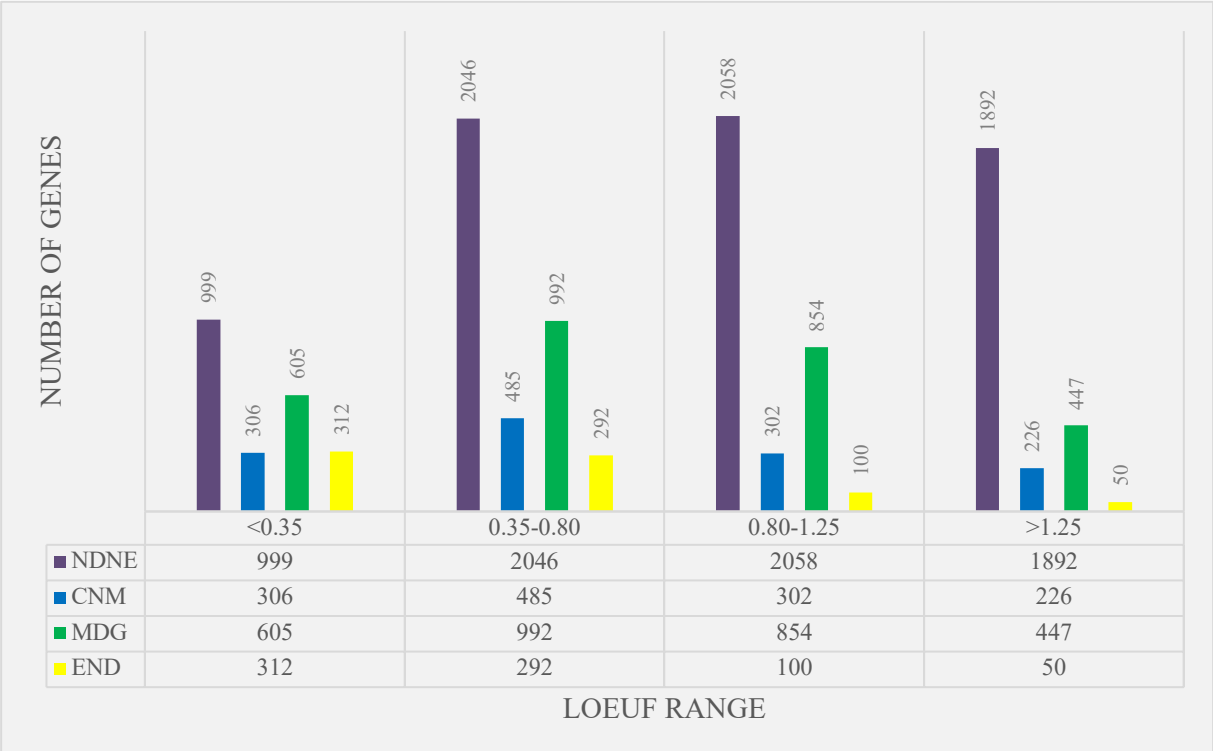


Figure 5-5 Percentage of genes among each gene group of Spataro et al. classification (90) according to the LOEUF score.

5.4.4 Results of the comparison of ESPP and CoNeS on essentiality/disease genes spectrums

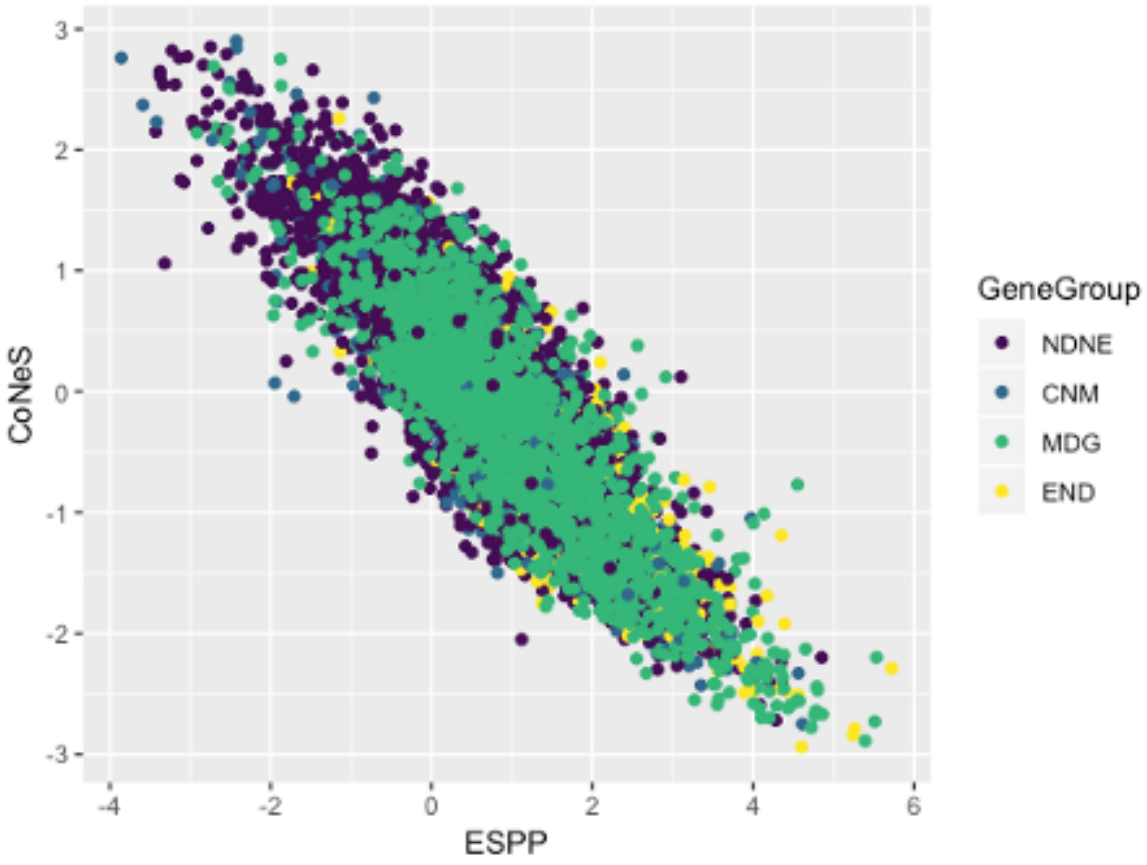


Figure 5-6 The distribution of genes using ESPP Vs CoNeS scores based on the updated Spataro et al. gene groups (90).

The gene-level score (CoNeS) has been designed to predict genes under strong negative selection and therefore, predict constrained genes. Here, this score was compared with that of the ESPP to investigate which might be more useful in identifying essential/disease genes. Figure 5-6 shows that the data of ESPP aligned better with CoNeS than LOEUF as the R-square value shows that the variability in ESPP is explained by 80% variability of CoNeS, and the significance of $P < 0.00$.

Here, there is a clear negative correlation between ESPP and CoNeS, which was expected as constrained genes that scored low as per CoNeS are scored high as per ESPP (Figure 5-6, 5-

4, and 5-7). However, genes within the MDG group are closer to essential genes for both scores. Moreover, the regression analysis between the two scores for 11711 observations shows that there is strong negative correlation between the two scores; R-square = 0.80 and P = 0.00, which means that the variation of ESPP is explained by 80% of the variation in CoNeS (Figure 5-6).

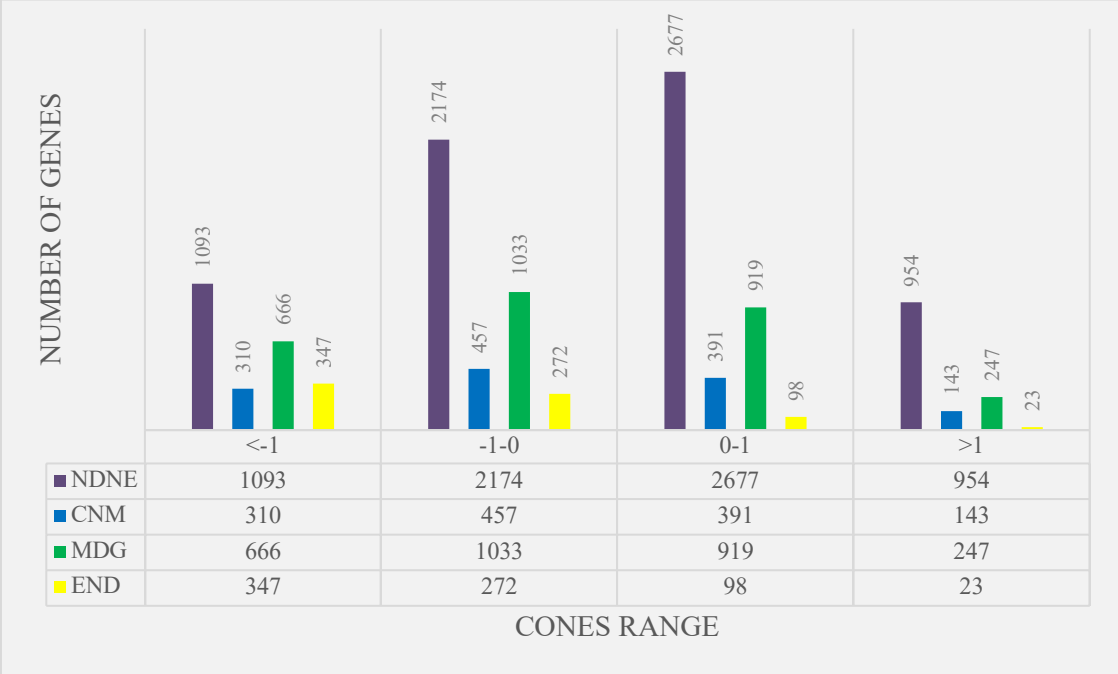


Figure 5-7 Percentages of genes among each gene group of the Spataro et al. classification according to the CoNeS score (90).

Among the 7251 genes classified as constrained genes (scored < 0.2 as per CoNeS), 655 genes were classified as END (9%), 1936 as MDG (27%), 855 as CNM (~12%), and 3808 as NDNE (~53%) (Table 5-5).

Table 5-5 The distribution of genes that scored < 0.2 as the hard threshold of CoNeS for the most constrained genes

CoNeS hard threshold for most intolerant genes	**Observations of genes scored < 0.2**	**NDNE**	**CNM**	**MDG**	**END**
< 0.2	7251	3808	855	1936	655
Percentage	62% of total matched data	53%	12%	27%	9%

There was a slight improvement when the cut-off of essentiality was increased from 0 to 1 in classifying END and MDG genes by ESPP, and the percentage went up when the cut-off was set at 2. Further, the percentage of prioritised genes by ESPP and CoNeS are very close, which explains the strong correlation between them (Table 5-4, 5-5). Figure 5-8 demonstrates that the distribution of ESPP and CoNeS are close, while LOEUF has a distinct distribution as compared to the other two scores.

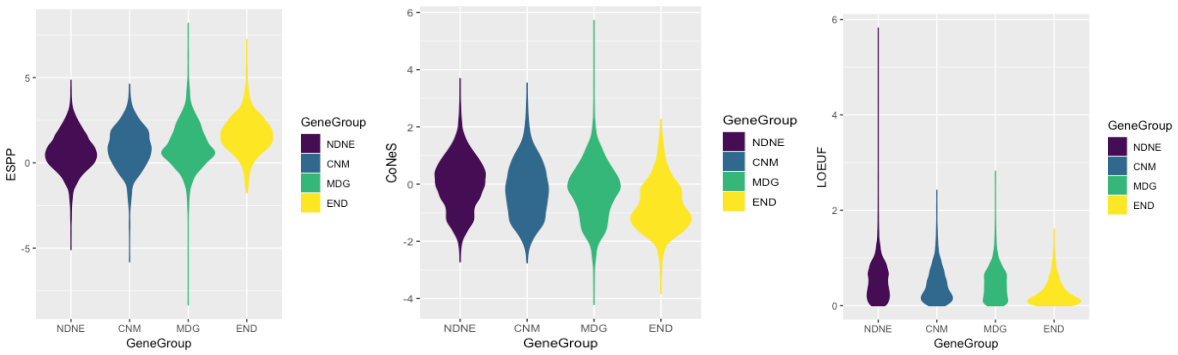


Figure 5-8 Violin charts of ESPP (A), CoNeS (B), and LOEUF (C) showing the distribution of genes among each score range.

5.4.5 Results of the comparison of LOEUF and CoNeS

The results of the regression analysis between LOEUF and CoNeS for 11711 observations showed that 60% of the variations in LOEUF are explained by variations in CoNeS (R-square = 0.60, P = 0.00). Further, there is a clear positive relationship between the two scores as END/MDG genes fall within low score levels in both scores (Figure 5-9).

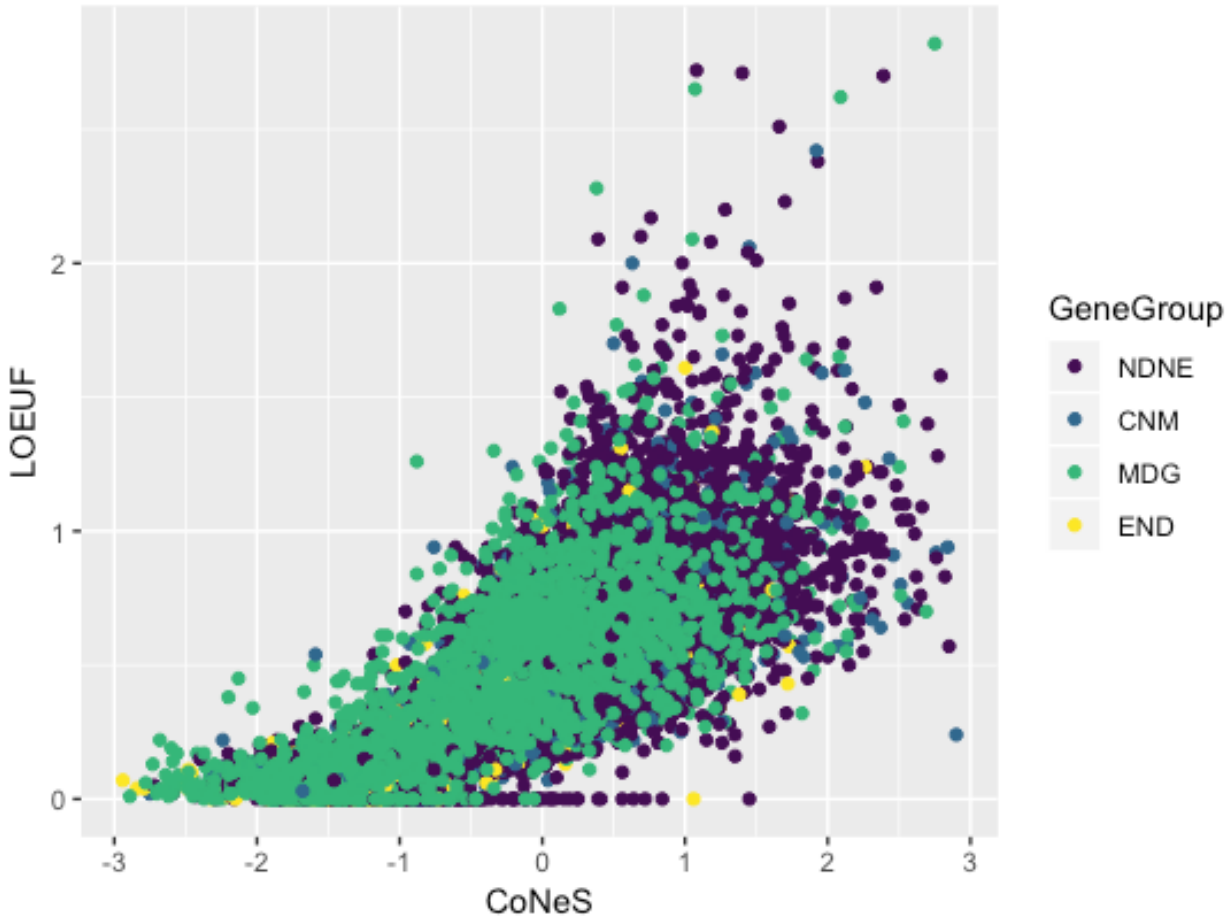


Figure 5-9 Results of the comparison of LOEUF and CoNeS

To predict which score fits better within the Spataro et al. classification (90), genes that scored low in both scores among the 11711 genes were examined. As mentioned earlier, the genes that scored < .2 as per CoNeS are 7251 in total. More specifically, CoNeS classified 655 as END (9%), 1936 as MDG (~27%), 855 as CNM (~12%), and 3808 as NDNE (~53%).

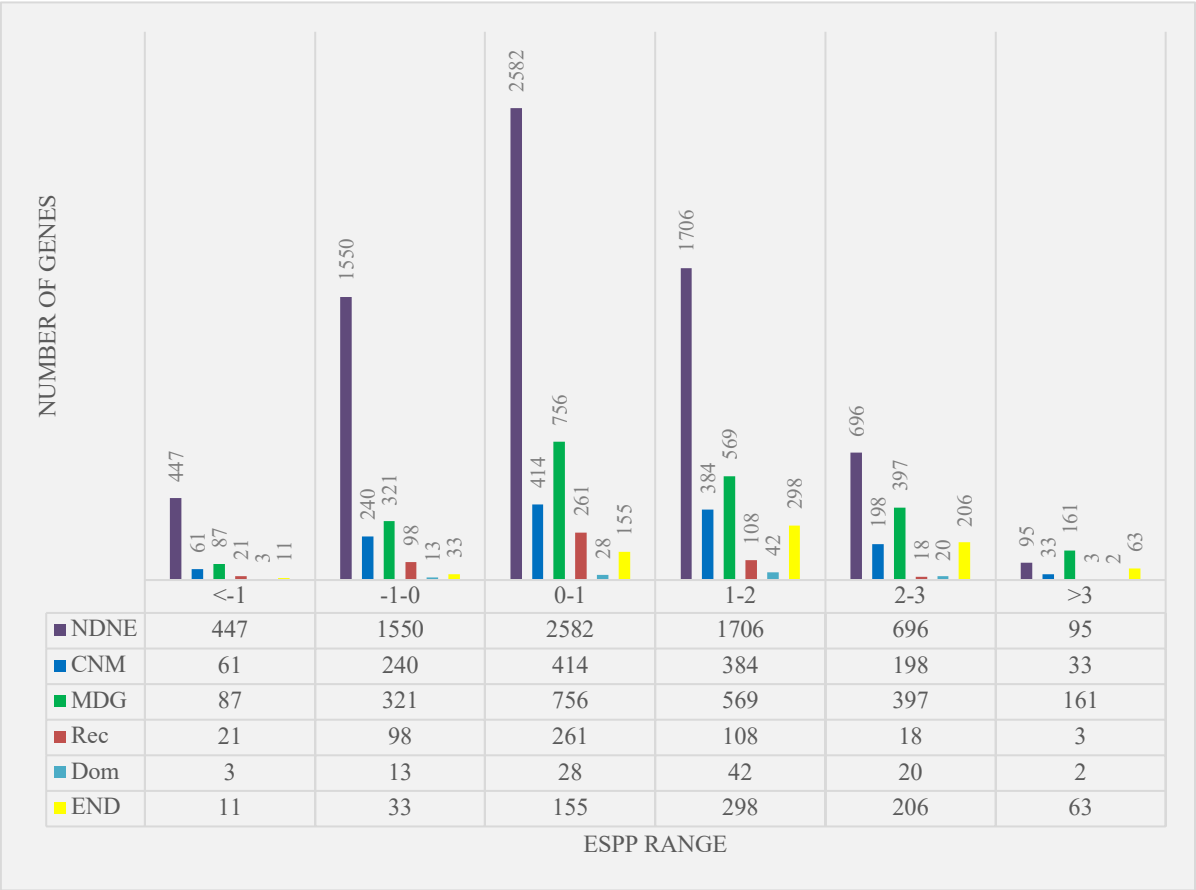
Regarding LOEUF, the number of genes scored less than 0.35 by LOEUF were 2222 genes. A total of 312 genes out of 2222 were successfully classified as END (14 %), 605 as MDG (27%), 306 as CNM (13.8%), and the rest (999 genes) as NDNE (45%) (Table 5-6). Here, LOEUF showed better prediction of END genes than CoNeS and for the MDG group, there were no major differences (Table 5-6).

Table 5-6 The distribution of genes that scored < 0.35 as per LOEUF and genes that scored < 0.2 as per CoNeS for the most constrained genes.

Score threshold	Observations	NDNE	CNM	MDG	END
LOEUF < 0.35	2222	999	306	605	312
Percentage	19 % of total matched data	45%	13.8%	27%	14%
CoNeS < 0.2	7251	3808	855	1936	655
Percentage	62 % of total matched data	53%	12%	27%	9%

5.4.6 Results of prediction for Dominant and Recessive genes using ESPP

The number of matched data of dominant and recessive genes with ESPP without NAs is 9931 genes, in which 108 genes were dominant and 509 were recessive.
Moreover, the ESPP score provides quite a clear separation for dominant genes as shown in Figure 5-10. Almost 60% of dominant genes fall in the range of ESPP > 1, and 85% fall within ESPP > 0. However 75% of recessive genes are towards NDNE genes and scored less than 1. The direction of the recessive genes is not quite clear, suggesting the need for constructing another score to prioritise recessive genes as they are easier to identify by utilizing data enriched with homozygosity.



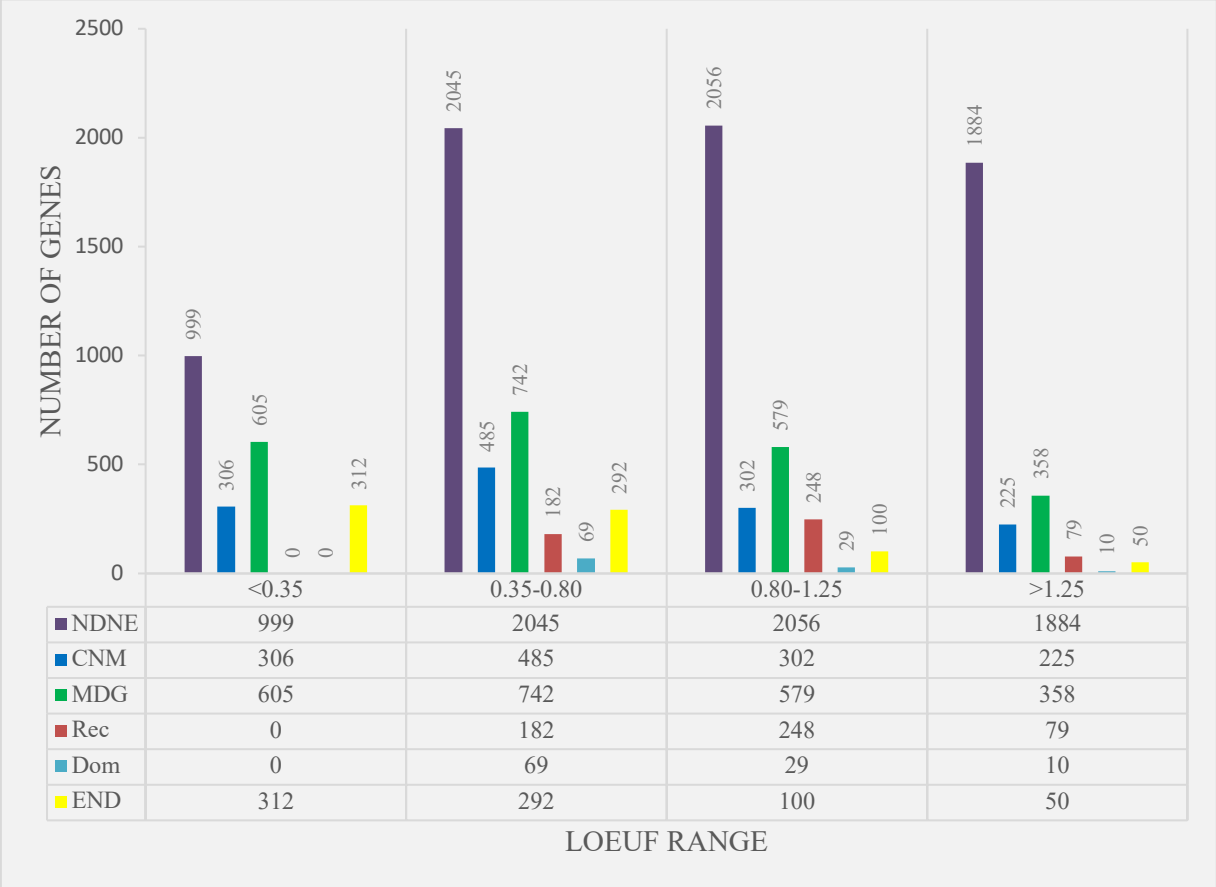
	<-1	-1-0	0-1	1-2	2-3	>3
NDNE	447	1550	2582	1706	696	95
CNM	61	240	414	384	198	33
MDG	87	321	756	569	397	161
Rec	21	98	261	108	18	3
Dom	3	13	28	42	20	2
END	11	33	155	298	206	63

Figure 5-10 Prediction of dominant and recessive genes using ESPP scores

5.4.7 Results of predicting dominant and recessive genes using LOEUF

After aligning the data of LOEUF with dominant and recessive gene lists, almost 11957 genes were retained after exclusions of NAs. In total, there were 509 recessive genes and 108 dominant genes as scored by LOEUF (Figure 5-11).

Moreover, 64% of dominant genes scored < 0.8 as per LOEUF, while 27% of dominant genes scored between 0.8 and 1.25, the least percentage of which (9%) scored > 1.25. Further, 65% of recessive genes scored > 0.8 as per LOEUF, and the rest (36%) are scored < 0.8 (Figure 5-11). Therefore, the direction of recessive and dominant genes using the LOEUF scores support the direction of those genes using the ESPP score—the dominant genes go toward essential genes as per both scores, while the recessive toward the non-essential end.



	<0.35	0.35-0.80	0.80-1.25	>1.25
NDNE	999	2045	2056	1884
CNM	306	485	302	225
MDG	605	742	579	358
Rec	0	182	248	79
Dom	0	69	29	10
END	312	292	100	50

Figure 5-11 Prediction of dominant and recessive genes using LOEUF scores

5.4.8 Results of predicting dominant and recessive genes using CoNeS

The total number of genes that aligned with the dominant and recessive genes list and had a CoNeS score is 10002 genes. Of these, 78% of dominant genes and 58% of recessive scored < 0.2 as per CoNeS, while 32% of recessive genes scored between 0.2 and 1. The distribution of dominant and recessive genes by CoNeS score seems different as compared to ESPP and LOEUF as the majority of dominant and recessive genes went in the direction of essential genes (Figure 5-12). The reason for this could be that most of the CoNeS data falls into the category of < 0.2.

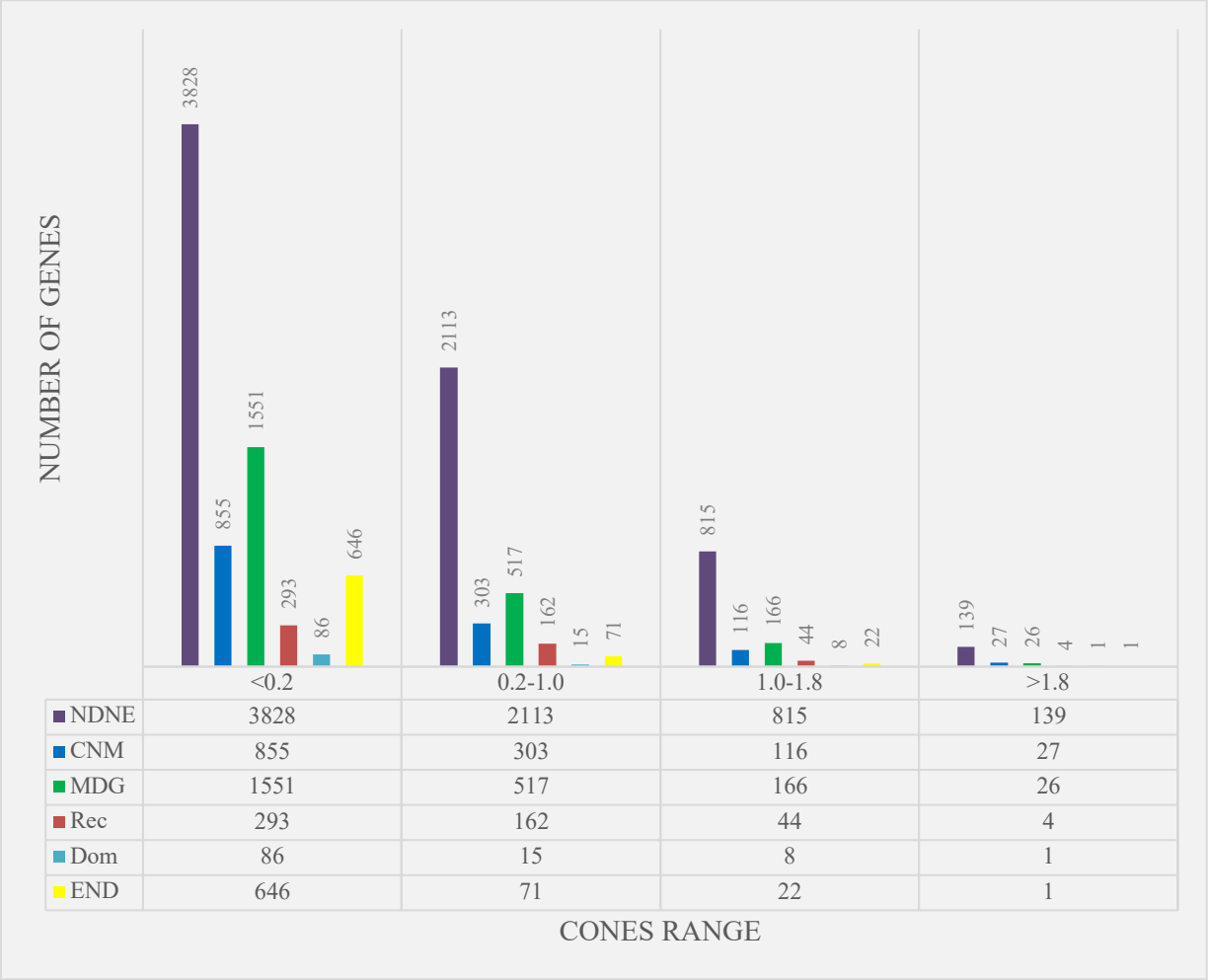


Figure 5-12 Prediction of dominant and recessive genes using CoNeS

Table 5-7 List of 50 candidate disease genes using ranked scores comprising genes scored high by ESPP and low by LOEUF and CoNeS, and that were not classified as MDG/END.

Name	Group	ESPP	LOEUF	CoNeS	ESPP Rank	LOEUF Rank	CoNeS Rank	SumRanks
ANKRD17	CNM	4.612	0.062	-2.810	23	30	14	67
SUPT5H	NDNE	4.084	0.074	-2.616	61	51	30	142
KDM2A	NDNE	3.818	0.045	-2.321	88	10	65	163
USP34	NDNE	4.281	0.091	-2.805	44	107	15	166
XPO1	NDNE	3.752	0.051	-2.181	94	14	106	214
SUPT6H	NDNE	4.003	0.087	-2.394	73	92	52	217
NAV1	CNM	3.204	0.08	-2.297	245	68	72	385
TNPO1	NDNE	3.220	0.056	-2.011	236	21	169	426

CTNND2	CNM	3.220	0.098	-2.098	184	138	135	457
DOT1L	CNM	3.220	0.118	-2.047	81	235	150	466
AP2A2	CNM	3.220	0.108	-2.250	221	186	84	491
CDC42BPB	NDNE	3.220	0.12	-2.126	120	249	123	492
RBM25	NDNE	3.220	0.077	-1.884	165	60	274	499
ZNF521	NDNE	3.220	0.098	-2.239	275	138	87	500
DIP2C	CNM	3.220	0.149	-2.379	26	420	55	501
XPO7	NDNE	3.220	0.12	-2.120	170	249	125	544
PBRM1	CNM	3.220	0.129	-2.062	103	297	146	546
CAND1	CNM	3.220	0.101	-1.936	182	154	222	558
BTAF1	NDNE	3.220	0.128	-2.163	173	289	112	574
CHD5	NDNE	3.220	0.157	-2.258	39	461	82	582
KPNB1	CNM	3.220	0.092	-1.877	199	111	281	591
SUPT16H	NDNE	3.220	0.12	-2.213	259	249	90	598
UBAP2L	NDNE	3.220	0.103	-1.953	230	163	206	599
SETD1A	NDNE	3.220	0.136	-2.194	162	343	102	607
INO80	NDNE	3.220	0.081	-1.891	273	75	268	616
UNC13A	CNM	3.220	0.158	-2.326	111	467	63	641
DDX46	NDNE	3.220	0.144	-2.038	106	386	159	651
ARFGEF1	NDNE	3.220	0.143	-1.959	83	378	204	665
IPO5	NDNE	3.220	0.122	-1.986	256	261	186	703
PLCL2	CNM	3.220	0.088	-1.935	386	94	224	704
SBNO1	CNM	3.220	0.081	-2.183	524	75	105	704
IPO7	NDNE	3.220	0.106	-1.928	304	175	231	710
DHX15	CNM	3.220	0.105	-1.781	155	170	396	721
SMARCC2	NDNE	3.220	0.117	-1.805	159	226	360	745
USP24	NDNE	3.220	0.147	-2.011	180	406	168	754
PRR12	NDNE	3.220	0.051	-1.972	556	14	195	765
U2SURP	NDNE	3.220	0.077	-1.851	401	60	308	769
KDM3B	NDNE	3.220	0.06	-1.692	239	25	547	811
HELZ	NDNE	3.220	0.048	-1.846	491	11	316	818
KSR2	CNM	3.220	0.131	-1.975	321	314	191	826
CKAP5	NDNE	3.220	0.162	-2.002	183	496	171	850
PTPRD	CNM	3.220	0.112	-2.080	514	197	141	852
HIPK1	CNM	3.220	0.121	-1.757	190	256	430	876

TRIM33	CNM	3.220	0.086	-1.932	561	87	229	877
SAP130	NDNE	3.220	0.159	-2.288	327	476	76	879
SF3B3	NDNE	3.220	0.178	-2.053	117	615	148	880
TNPO2	NDNE	3.220	0.129	-1.720	108	297	481	886
CHD9	NDNE	3.220	0.17	-2.043	204	542	153	899
MED12L	NDNE	3.220	0.184	-2.033	95	655	161	911
RTF1	NDNE	3.023	0.122	-1.851	343	261	307	911

The genes listed in Table 5-7 were ranked as candidate genes for Mendelian diseases that score high by ESPP and low by LOEUF and CoNeS but were not classified as END/MDG. Interestingly, seven out of 11 genes that were ranked as candidate genes by ESPP (Table 4-4) are listed in this table. Here, the *CNOT1* gene was excluded as it has recently been found to be causal. Moreover, *RYR3* was ranked 146th in the list, which is considered high ranking among a list of almost 11711 genes. More specifically, the *RYR3* gene has been investigated earlier, and it was shown that it might be a candidate disease gene (refer to Table 4-4). Below, the functions of the genes in this table are considered to identify the strength of the evidence that some may be disease gene candidates (Table 5-8).

Table 5-8 Functions of 50 genes that were prioritised by sumRanks of ESPP, CoNeS, and LOEUF and not classified as END/MDG.

Name	**Group**	**ESPP**	**Full Name**	**Notes on gene function (OMIM)**
ANKRD17	CNM	4.612	ANKYRIN REPEAT DOMAIN-CONTAINING PROTEIN 17	Refer to table 4-4
SUPT5H	NDNE	4.084	SPT5 HOMOLOG, DSIF ELONGATION FACTOR SUBUNIT	Refer to table 4-4
KDM2A	NDNE	3.818	lysine demethylase 2A	Plays a role in histone demethylase activity (H3-K36 specific) and unmethylated CpG binding
USP34	NDNE	4.281	Ubiquitin specific peptidase 34	Refer to table 4-4

XPO1	NDNE	3.752	Exportin 1	CRM1 was described by Stade et al. as an essential nuclear export factor in S.cerevisiae and they suggested renaming it XPO1 (162).
SUPT6H	NDNE	4.003	SPT6 Homolog, Histone Chaperone and Transcription Elongation Factor	Refer to table 4-4
NAV1	CNM	3.204	Neuron navigator 1	The exact function of this protein is not known
TNPO1	NDNE	3.220	Transportin 1	Plays an essential role in atherosclerotic coronary artery disease (CAD) (163)
CTNND2	CNM	3.361	catenin delta 2	'myogenic transcription factors regulate synapse specific transcription of RAPSYN protein'(164)
DOT1L	CNM	3.927	DOT1 like histone lysine methyltransferases	Nucleosomal H3-specific methyltransferase (165)
AP2A2	CNM	3.256	Adaptor related protein complex 2 subunit alpha 2	Not known
CDC42BPB	NDNE	3.603	CDC42 binding protein kinase beta	CDC42BPB binds the kinase domains of MRCK-alpha and MRCK-beta to impede their catalytic function (166)
RBM25	NDNE	3.431	RNA binding motif protein 25	It works as a splicing factor RBM25 to control MYC activity, which plays an important role in acute myeloid leukaemia (167)
ZNF521	NDNE	3.121	Zinc finger protein 521	Unknown function
DIP2C	CNM	4.564	Disco interacting protein 2 homolog C	Refer to table 4-4
XPO7	NDNE	3.415	Exportin 7	Nuclear export factor with broad substrate apecificity
PBRM1	CNM	3.705	Polybromo 1	Chromatin-remodelling complex (168)
CAND1	CNM	3.363	Cullin associated and neddylation dissociated 1	Associated with Tributyl phosphate TBP in nuclear extracts(163,164)

BTAF1	NDNE	3.393	B-TFIID TATA-box binding protein associated factor 1	BTAF1 controls DNA-dependant ATPase activity that separates the TBP from DNA(171)
CHD5	NDNE	4.346	Chromodomain helicase DNA binding protein 5	Refer to table 4-4
KPNB1	CNM	3.314	Karyopherin subunit beta 1	KPNB1 regulates nuclear import in the interphase (172)
SUPT16H	NDNE	3.161	SPT16 homolog, facilitates chromatin remodelling subunit	Splicing factor (173)
UBAP2L	NDNE	3.234	Ubiquitin associated protein 2 like	Independent BMI1/RNF2 (Polycomb complex protein/E3 ubiquitin-protein ligase RING2 is an enzyme that is encoded by the RNF2 gene in humans) complex that inhibits INK4A/ARF, which are cyclin dependent kinase inhibitors (CKIs) (168,169,170)
SETD1A	NDNE	3.438	SET domain containing 1A, histone lysine methyltransferase	This gene was found to be causal for two disorders: 1. Early-onset epilepsy with/without developmental delay (177); 2. Neurodevelopmental disorder with speech impairment and dysmorphic facies (178)
INO80	NDNE	3.123	INO80 complex ATPase subunit	Possible association with inherited immunoglobulin class-switch recombination deficiency pending confirmation (179)
UNC13A	CNM	3.668	Unc-13 homolog A	1. There is a possible association with Polymorphism in amyotrophic lateral sclerosis 2. Possible association with congenital myasthenic syndrome 3. Possible association with a dyskienetic movement disorder associated with delayed development and behavioural abnormalities (180)

DDX46	NDNE	3.684	DEAD-box helicase 46	DDX46 was found to prevent cell growth in human colorectal cancer cell line (181)
ARFGEF1	NDNE	3.915	ADP ribosylation factor guanine nucleotide exchange factor 1	Unknown function
IPO5	NDNE	3.166	Importin 5	Nuclear transport factor (182)
PLCL2	CNM	2.966	Phospholipase C like 2	Phospholipase catalytic activity(183)
SBNO1	CNM	2.797	Strawberry notch homolog 1	Unknown function
IPO7	NDNE	3.076	Importin 7	Unknown function
DHX15	CNM	3.454	DEAH-box helicase 15	Works as a viral RNA sensor to induce interferon-stimulated genes (184)
SMARCC2	NDNE	3.444	SWI/SNF related, matrix associated, actin dependent regulator of chromatin subfamily c member 2	Coffin-Siris syndrome 8 (179,180)
USP24	NDNE	3.368	Ubiquitin specific peptidase 24	It has been associated with Parkinson's disease (187)
PRR12	NDNE	2.773	Proline rich 12	Unknown function
U2SURP	NDNE	2.941	U2 snRNP associated SURP domain containing	Unknown function
KDM3B	NDNE	3.217	Lysine demethylase 3B	Diets-Jongmans syndrome (188)
HELZ	NDNE	2.832	Helicase with zinc finger	Plays a significant role in RNA metabolism in different tissues (189)
KSR2	CNM	3.050	Kinase suppressor of ras 2	Plays a role in regulating MEKK3 and COT activity that is expressed mainly in the kidney and brain (184,185)
CKAP5	NDNE	3.362	Cytoskeleton associated protein 5	Acts as a microtubule polymerase (192)
PTPRD	CNM	2.809	Protein tyrosine phosphate receptor type D	Has a role in PTPase activity against phosphorylated test substrate (193)
HIPK1	CNM	3.341	Homeodomain interacting protein kinase 1	Expression of HIPK1 found to be significantly increased in the breast cancer cell line (194)

TRIM33	CNM	2.770	Tripartite motif containing 33	Works as ubiquitin ligase (195)
SAP130	NDNE	3.047	Sin3A associated protein 130	One of the SIN3A complexes, which repress transcription (196)
SF3B3	NDNE	3.628	Splicing factor 3b subunit 3	Unknown function
TNPO2	NDNE	3.681	Transportin 2	Has a significant role to export cellular mRNA (197)
CHD9	NDNE	3.302	Chromodomain helicase DNA binding protein 9	It is a chromatine remodelling protein (198)
MED12L	NDNE	3.734	Mediator complex subunit 12L	Nizon-Isidor syndrome (199)
RTF1	NDNE	3.023	RTF1 homolog, Paf1/RNA polymerase II complex component	Plays a role in gene expression regulation (200)

Table 5-8 shows the highest ranked genes by ESPP, CoNeS, and LOEUF that were not classified as END/MDG and the potential clinical significance of each gene. Upon investigating the literature and OMIM database, it is clear that a number of genes seem to have nuclear functions. Those genes, therefore, have the possibility to be unrecognised essential genes or causal roles in disease. Here, four genes were identified that have recently been confirmed as causal: *SETD1A* was discovered as causal in 2020, and the rest in 2019 (highlighted in red in Table 5-8). Thus, they have not been classified as MDG as per the Spataro et al. classification (90).

More specifically, first, while *SETD1A* was classified as NDNE, it was scored high as per ESPP at 3.43 which indicates that this gene might be causal. Further, this gene has been linked with two diseases: early-onset epilepsy with/without developmental delay (177) and a neurodevelopmental disorder with speech impairment and dysmorphic facies (178). Second, *SMARCC2* was also classified as NDNE and scored 3.44 as per ESPP, which indicates that this gene might be a candidate gene. Investigation of this gene found that it is causative for Coffin-Siris syndrome 8 (185). Third, *KDM3B* was classified as NDNE with a high ESPP score of 3.21, indicating that it might be a candidate gene for Mendelian diseases. Further, it was found to cause Diets-Jongmans syndrome (188). Last, *MED12L* was classified as NDNE and scored high as per ESPP at 3.73 which indicates that this gene

might be causal. The investigation of *MED12L* showed that this gene is the cause of Nizon-Isidor syndrome (199).

5.5 Discussion

Combining genome and transcriptome sequencing data has enhanced diagnosis by the improved discovery of rare variants with functional effects. However, the analysis of transcriptome data is challenging owing to the fact it is affected by the environment, state of a disease, and technical variations. Therefore, identifying when an effect is genetic and the impact that this effect has beyond the normal population range is quite challenging. In this regard, Mohammadi et al. have developed a method called ANEVA, and its extension, ANEVA-DOT, to be able to quantify genetic variation in gene dosage in the general population, and to recognise genes where a patient seems to have a heterozygous variant with an unexpected strong effect on gene expression. This will hopefully allow single transcriptome comparisons to pre-existing reference data without the technical and reverse causation noise in total gene expression analysis (201). However, the majority of their analyses have been limited to only a small portion of variants that stimulate alterations in the transcriptome, like splice alterations and total loss of expression. In this context, a promising data type is allelic expression, these data quantify the paternal and maternal haplotype expressions of a gene (201). 'However, a quantitative framework for interpreting this data type to identify rare pathogenic variants has been lacking' (201).

Meanwhile, GeVIR and LOEUF showed a potential to rank AR genes, as they demonstrated the intolerance of genes to various types of variants—missense and LoF (144). However, LOEUF was highly biased towards long genes compared to GeVIR. To prioritise genes based on the level of tolerance using both missense and LoF, produced a composite score was produced by combining LOEUF and GeVIR into VIRLoF, which showed a better performance (144).

Moreover, the comparison of ESPP and LOEUF showed that ESPP performs better in prioritizing MDG using the Spataro et al. classification (90). However, regarding GeVIR and VIRLOF, the comparisons were not done due to unavailability of the data.

Further, the results of testing candidate genes prioritised by ESPP using the SHGP data showed that genes that scored more than 2 as per ESPP were 89% MDG and 11% NDNE. The NDNE gene group is worth more investigation as it is likely to contain candidate genes for rare diseases as suggested by the integrated scores. This study had anticipated that the

large NDNE gene group was likely to contained undiscovered disease genes and thus, it was the focus of much of this investigation. For example, upon investigating genes that have were classified as NDNE, the *ZNF219* gene was found to be causal for colobomatous microphthalmia (202). Other genes in SHGP data that have been classified by ESPP as NDNE were found to be candidate genes using autozygosity mapping; this has to be confirmed using GenCC.

In the comparison of ESPP, LOEUF and CoNeS scores, the results of ESPP and LOEUF comparison showed quite a weak correlation. However, the performance of ESPP was quite similar to LOEUF in classifying essential genes, and ESPP appeared to show better performance in recognizing MDG genes within the Spataro groups. Here, an ESPP cut-off > 2.00 was chosen as this threshold showed better alignment with known MDG/END genes (Table 5-3, Table 5-4).

Additionally, CoNeS showed a strong negative correlation with ESPP. It is worth looking into this data in more depth. Further, ESPP showed better performance in classifying MDG/END than CoNeS (Table 5-4, Table 5-5).Regarding LOEUF and CoNeS, there was a positive correlation between both scores as 60% of variance in LOEUF data was explained by CoNeS (Table 5-6). However, CoNeS and LOEUF had similar performances in prioritizing MDG, while the latter was better at predicting END genes

From the results of predicting the directions of dominant and recessive genes using ESPP, LOEUF and CoNeS, ESPP and LOEUF showed the same direction, which supports that dominant diseases go toward the essential end in both scores. However, the direction that these genes took as per the CoNeS score remained quite unclear.

5.6 Conclusions

In this chapter, ESPP performance was tested across multiple databases. The results were promising as two genes that were prioritised by ESPP as candidate genes for Mendelian diseases were found to be causal. The first gene, *CNOT1*, was recently shown to have a disease variation underlying holoprosencephaly disease in two published articles. The second gene, *RYR3*, was shown to cause intellectual disability with a *de novo* splice donor site mutation. Thus, the argument that many undiscovered rare disease variants might be splice variants is supported. Moreover, 15% of the genes that were classified by ESPP as NDNE might be candidate genes worth investigating further. Ultimately, the ESPP score

showed better separation of dominant genes than recessive ones—a fertile area for future research.

Further, ESPP performed better in prioritising MDGs, dominant and recessive genes than LOEUF and CoNeS. Using the sum of ranks of the three scores for ESPP, LOEUF and CoNeS, the results were interesting as four genes of the 50 highest ranked NDNE/CNM genes have recently been found to be causal. The first gene, *SETD1A* has been implicated in two diseases: epilepsy provoked by *de novo* variants, and a novel neurodevelopmental syndrome caused by dominant *de novo* LoF variants (172,171). The second, *SMARCC2,* was proven to be involved in Coffin-Siris syndrome, which can be caused by three heterozygous variants in this gene, the majority of which are novel and proven to be *de novo* variants (185). The third is *KDM3B* that has been proven to cause Diets-Jongmans syndrome, provoked by heterozygous missense variants that have been identified in 17 individuals (188). The fourth is *MED12L* that causes Nizon-Isidor syndrome by deletions and duplication variants (199). In this context, predicting four genes as causal from NDNE/CNM genes using the sumRanks of the three scores suggest that combining the three scores might provide better predictions of Mendelian disease genes.

Chapter 6 Conclusions and Future Work

6.1 Conclusion

The preliminary results of the systematic literature review suggest that there were no single gene-level scores capable of predicting Mendelian disease genes. Therefore, a combination of evidence on genes essentiality will help in building a model predicting genes with potential disease variants. However, individual genes might contain variants that do not cause disease. Thus, distinguishing this type of sequence variant from a deleterious one remains a significant challenge.

Based on the findings, it is worth checking any suspect gene for Mendelian disorders by using the ESPP score, which explains the variation in the data better than other available gene-specific scores. Further, to improve the performance of this study's model, validating ESPP score, providing better classification for the human genes with the understanding of gene essentiality is crucial, which is a fertile area for future studies.

6.1.1 Update on the disease genome

The area of disease genomics is growing rapidly with the ability to sequence the whole genome at a high speed and lowering costs, and the identification of faulty genes among thousands of sequenced genes being a necessary mission. However, despite the development of WES and WGS, the diagnostic rate of rare diseases is only 30% (203). Several obstacles prevent the increase in the rate of rare disease molecular diagnosis. Some of these are bioinformatics, incomplete penetrance, and non-coding variants (203). Further, the quality of a variant depends on the genomic region, type of variant, and depth of sequencing coverage. Moreover, in variant analysis, variant type is an important factor. Typically, the variants considered first are the coding ones; however, splicing variants and indels might affect both coding/non-coding regions (203). The most deleterious variants are LoF variants (stop-gain, frameshift, and essential splice site), but they are rare as compared to other types of variants. Meanwhile, missense variants might be seen in disease genes of healthy individuals, making their interpretation challenging. Splice variants might stimulate regulatory domains within mRNA—especially those regulating splicing—and might also impact translation (203). However, many of the *in silico* tools do not perform well in assessing variants outside the canonical splice sites. Nevertheless, recently, splicing libraries

have been enhancing the understanding of splicing variants, and it, thus, might be useful to train future splicing models. More specifically, splice, one of the most recent predictive tools, showed a high performance at identifying splicing. In this context, Jaganathan et al. predicted that cryptic splice mutations that validate at a high rate through RNA-seq are highly likely to be deleterious, and are a causal variant of monogenic diseases (204).

While synonymous variants are usually not of interest, it has been shown that these variants are not always non-deleterious as they might also cause aberrant splicing either at the splice sites or by interrupting other splice regulatory regions (203). Therefore, undiscovered disease variants might be at splice sites, and this might provide an explanation of limited diagnosis of Mendelian disease.
Moreover, the results of the sumRanks of ESPP, LOEUF and CoNeS showed that most of the highest ranked genes were NDNE, and a few had unknown functions or critical functions that have not been identified to cause diseases. Thus, identifying the functions of such genes and reclassifying most NDNE genes will help in identifying more disease genes in the future

6.2 Plans for future work

Understanding the biology of the human population depends on understanding human variation. Here, certain unique populations like the Saudi population—as their genomic data is enriched for homozygosity due to population demographics—can be utilised to guide the prediction of recessive disease genes by classifying genes enriched for homozygous variants in this population. However, this has not been possible using other population data due to reduced homozygosity genome-wide. Nevertheless, the high percentage of consanguineous marriages in the Saudi population facilitates the study of recessive genes and their role in disease processes. In this context, constructing a classifier to assess the tolerance of all genes to homozygous, and apparently deleterious, variants is worth the effort. Thus, genes that show high tolerance to homozygous deleterious variants without an obvious phenotypic consequence should be considered low priority in the hunt for recessive disease genes. Meanwhile, previous research examined 253 genes that contain confirmed LoF variants that were found to be homozygous in at least a single individual (12). These LoF-tolerant genes tend to have fewer PPIs, reduced conservation, and are enriched for olfactory receptor genes (which were excluded from subsequent analysis), and depleted for genes involved in

embryonic development and crucial cellular processes compared to the rest of the genome. The remaining 213 LoF-tolerant genes were compared to 858 known recessive-disease genes using a linear discriminant model based on human–macaque conservation and proximity to recessive disease genes in PPI networks. Here, MacArthur et al. used a sample of genes homozygously inactivated in the 1000 Genomes project. They considered these genes as LoF-tolerant, so they can be used as a comparison group, which can be utilised to delineate the functional properties that discriminate these genes from severe recessive disease genes. Further, they also found that LoF-tolerant genes have different functional characteristics—they are less conserved and have fewer PPI than the average genome(12), enabling them to generate a recessive disease gene probability score for every gene in the genome. Here, MacArthur et al. noted that while it is useful in prioritising candidate genes for follow up, they lacked the power to definitively delineate LoF-tolerant and recessive disease genes.

However, the unique genetic makeup and population history of the Saudi population would lend significantly increased power to predict recessive genes as compared to the study by MacArthur et al.—which used just 185 individuals from the 1000 genomes project—due to the greater sample size and higher levels of homozygosity within the Saudi population.

In a general sense, identifying likely candidate recessive disease genes and getting a better understanding of the landscape of deleterious and tolerated homozygous variation in the human population will be immensely valuable in addressing the challenges of variant interpretation, while improving variant classification and enhancing diagnostic yields from sequencing projects, both within the Saudi population and beyond. Further, considering patterns of homozygosity alongside detailed patient phenotyping information will provide valuable insights into the function of many poorly characterised recessive disease genes.

More specifically, as shown in this study, integrating information of more gene-specific scores might improve the detection of Mendelian disease genes as most of the composite scores like VIRLOF, ISPP and CoNeS showed better performances than individual scores. Therefore, work is required in order to validate the derived composite ESPP score presented in the thesis. In particular, performing a quantitative analyses that formally assess whether the ESPP is more effective than other measures at predicting gene essentiality and diagnosing monogenic disorders.

Appendix A

Scoping search approach

1.Cochrane Library
(“gene level” classifier) 0 result, (“gene level” score) 6 irrelevant results, (“gene level “approach) 6 irrelevant result, (“disease gen*”)5 irrelevant result, (gene* essentiality) 0 results, (eliminating false-positive variants) 0 result, (gene specific filtering) 21 irrelevant result, (“gene-specific” filtering) 1 irrelevant result, (gene level metrics) 20 irrelevant results, (“next generation” sequencing) 1 irrelevant result.
2.Prospero
(gene level scores) 0 results; (disease genome) 0 results; (gene essentiality)0 results; (gene-specific filtering) 0 results; (next-generation sequencing) 0 results, disease gen* 28 irrelevant results, (eliminating false-positive variants) 0 results.
3.The Trip (Turning Research into Practice) database
Using PICO question and Filtering by evidence type (systematic review): 0 results.
Using the same filter for other keywords: (“gene level” classifier) 0 results, (“gene level” score) 4 results.
4.Evidence Search (NICE)
Using filter results by secondary evidence- systematic review.
(Using Gene specific metrics to facilitate identification of disease genes) 33 irrelevant systematic reviews, (“gene level” classifier) 1 irrelevant systematic review, (“gene level” score) 2 irrelevant systematic reviews, (“gene level “approach) 3 irrelevant systematic reviews,(“gene level”) 4 irrelevant systematic review, (“disease gen*”) 1 irrelevant systematic review, gene* essentiality, (gene* essentiality) 3 irrelevant systematic reviews, (eliminating false-positive variants) 13 irrelevant systematic reviews, (gene specific filtering) 10 irrelevant systematic reviews, (“gene-specific” filtering) 0 result, (“gene level” metrics) 2 irrelevant systematic reviews, (“next generation” sequencing) 30 irrelevant systematic reviews.
5. EBSCOhost Platform (CINAHL, MEDLINE, SportDiscus)
Using (Advance search in all fields: disease AND gen* AND variant AND “gene-level” AND priori* AND filter* AND systematic review, publication date range: not specified, content type: Any, Discipline: Medicine; Sciences; Statistics, Language: English, show only: Scholarly articles, including peer-reviewed, Exclude from results: Newspaper articles, Dissertations/Thesis). This filter was used to get a reasonable number of related articles. The result was 63 articles, which look relevant, but no systematic review was found. 1

Meta-analysis was found about (A Meta-Analysis Strategy for Gene Prioritization Using Gene Expression, SNP Genotype, and eQTL Data), which is relevant but not in the same area.

6. OvidSP Platform (Medline, Embase, EBM Reviews).

Using (Advance search in all fields: disease AND gen* AND variant AND “gene-level” AND priori* AND filter* AND systematic review. content type: Any, Discipline: Medicine; Sciences; Statistics, Language: English, show only: not specified, Exclude from results: Newspaper articles, Dissertations/Thesis. Here, 133 results look related, but no systematic review was identified.

Bibliography

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science (80-). 2001;291(February):1304–51.
2. Ponting CP, Hardison RC. What fraction of the human genome is functional? Genome Res. 2011;21(11):1769–76.
3. Cordell HJ. Epistasis : what it means , what it doesn ' t mean , and statistical methods to detect it in humans. Hum Mol Genet. 2002;11(20):2463–8.
4. Botstein D, Risch N. Discovering genotypes underlying human phenotypes : past successes for mendelian disease , future approaches. Nat Genet. 2003;33(march):228.
5. Eric D. Green, First name. "Physical map" NIH, 10 May 2021, https://www.genome.gov/genetics-glossary/Physical-Map.
6. SM P. Genetic Linkage Analysis. Arch Neurol. 1999;56(6):667–72.
7. Matise TC, Chen F, Chen W, Vega FMD La, Hansen M, He C, et al. A second-generation combined linkage – physical map of the human genome. Genome Res. 2007;12:1783–6.
8. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics. 1990;6(3):575–7.
9. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Fundam concepts Genet ics Link. 2016;9(6):477–85.
10. Sharma, Vinukonda. (2018). Re: What is the difference between QTL and Association mapping?. Retrieved from: https://www.researchgate.net/post/What_is_the_difference_between_QTL_and_Association_mapping/5b75469db93ecd037c485d35/citation/download.
11. Bartha I, Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat Rev Genet. 2017;19(1):51.
12. MacArthur D, Balasubramanian S, Frankish A. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science (80-) [Internet]. 2012;335(6070):1–14. Available from: http://www.sciencemag.org/content/335/6070/823.short
13. Liu M, Watson LT, Zhang L. Classification of Mutations by Functional Impact Type : Gain of Function , Loss of Function , and Switch of Function. 2014. 236–237 p.
14. Gilbert-Diamond D MJ. Analysis of Gene-Gene Interactions. Curr Protoc Hum Genet. 2011;14(1):1–7.

15. Moore J H. The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases. Hum Hered. 2003;56:73–82.
16. Li X, Li W, Zeng M, Zheng R, Li M. Network-based methods for predicting essential genes or proteins : a survey. Brief Bioinform. 2019;00(February):1–18.
17. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP : the Database of Interacting Proteins. Nucleic Acids Res. 2000;28(1):289–91.
18. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-cepas J, et al. STRING v11 : protein – protein association networks with increased coverage , supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(November 2018):607–13.
19. Chatr-aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database : 2017 update. Nucleic Acids Res. 2017;45(December 2016):369–79.
20. Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS : a database for genomes and protein sequences. Nucleic Acids Res. 2002;30(1):31–4.
21. Hermjakob H, Montecchi-palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, et al. IntAct : an open source molecular interaction database. Nucleic Acids Res. 2004;32:452–5.
22. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT : the Molecular INTeraction database. Nucleic Acids Res. 2007;35(November 2006):2006–8.
23. Oti, M. and Brunner H. The modular nature of genetic diseases, Clinical Genetics. 2007. 71: 1-11.
24. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), {Sep,2019}.
25. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Publ Gr. 2010;11(6):446–50.
26. Richter T, Nestler-parr S, Babela R, Khan ZM, Tesoro T, Molsen E, et al. Rare Disease Terminology and De fi nitions — A Systematic Global Review : Report of the ISPOR Rare Disease Special Interest Group. Value Heal. 2015;18(6):906–14.
27. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. Am J Hum Genet. 2017;100(5):695–705.
28. Pogue RE, Cavalcanti DP, Shanker S, Andrade R V., Aguiar LR, de Carvalho JL, et

al. Rare genetic diseases: update on diagnosis, treatment and online resources. Drug Discov Today [Internet]. 2018;23(1):187–95. Available from: http://dx.doi.org/10.1016/j.drudis.2017.11.002

29. Men AE, Wilson P, Siemering K FS. Analysis of Microarray Data Principles of Computational Cell Biology Cancer Diagnostics with DNA Microarrays Handbook of Genome Research. 2008. 280 p.
30. Mu W, Lu HM, Chen J, Li S, Elliott AM. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. J Mol Diagnostics [Internet]. 2016;18(6):923–32. Available from: http://dx.doi.org/10.1016/j.jmoldx.2016.07.006
31. Langaee T, Ronaghi M. Genetic variation analyses by Pyrosequencing. Mutat Res Mol Mech Mutagen. 2005;573:96–102.
32. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008;5(1):16–8.
33. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods. 2007;4(11):903–5.
34. Biesecker LG. Exome sequencing makes medical genomics a reality. Nat Genet. 2010;42(1):13–4.
35. Collins FS, Morgan M, Patrinos A, Watson JD. The Human Genome Project : Lessons from Large-Scale Biology. Science (80-). 2003;300(April):286–91.
36. The 1000 Genomes Project Consortium, Nature 526, 68-74 (01 October 2015) doi:10.1038/nature15393.
37. The 100,000 Genomes Project Protocol v3, Genomics England. doi:10.6084/m9.figshare.4530893.v2. 2017.
38. Lek M, Karczewski KJ, Minikel E V, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature [Internet]. 2016;536(7616):285–91. Available from: http://www.nature.com/doifinder/10.1038/nature19057%0Ahttp://www.ncbi.nlm.nih.gov/pubmed/27535533%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5018207
39. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015;97(2):199–215.
40. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature [Internet].

2009;461(7261):272–6. Available from: http://dx.doi.org/10.1038/nature08250

41. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet [Internet]. 2016;17(6):333–51. Available from: http://dx.doi.org/10.1038/nrg.2016.49
42. Chaitankar V, Karakülah G, Ratnapriya R, Giuste FO, Brooks MJ, Swaroop A. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research [Internet]. Vol. 55, Progress in Retinal and Eye Research. Elsevier Ltd; 2016. 1–31 p. Available from: http://dx.doi.org/10.1016/j.preteyeres.2016.06.001
43. Salomon DR, Ordoukhanian P. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2015;56(2):61–77.
44. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, et al. review The challenges of sequencing by synthesis. Nat Biotechnol. 2009;27(11):1013–23.
45. Pettersson E, Lundeberg J, Ahmadian A. Genomics Generations of sequencing technologies. Genomics. 2009;93(2):105–11.
46. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores , and the Solexa / Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767–71.
47. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J TD. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(6):111.
48. O'Rawe JA, Ferson S, Lyon GJ. Accounting for uncertainty in DNA sequencing data. Trends Genet [Internet]. 2015;31(2):61–6. Available from: http://dx.doi.org/10.1016/j.tig.2014.12.002
49. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol [Internet]. 2018;36(10):983–7. Available from: https://doi.org/10.1038/nbt.4235
50. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics [Internet]. 2015/11/02. 2015 Oct;13(5):278–89. Available from: https://pubmed.ncbi.nlm.nih.gov/26542840
51. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data : Astronomical or Genomical ? PLOS Biol. 2015;13(7): e10:1–11.
52. Maniloff J. Commentary The minimal cell " On being the right size ". Proc Natl Acad Sci. 1996;93(September):10004–6.
53. Iii CAH, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, et al. Global

Transposon Mutagenesis and a Minimal Mycoplasma Genome. Science. 1999;286(December):2165–9.

54. Mccutcheon JP, Mcdonald BR, Moran NA. Origin of an Alternative Genetic Code in the Extremely Small and GC – Rich Genome of a Bacterial Symbiont. PLoS Genet. 2009;5(7):e1000565.
55. Park D, Park J, Gu S, Park T, Shim S. Genomics Analysis of human disease genes in the context of gene essentiality. Genomics. 2008;92(6):414–8.
56. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B AA. Functional profiling of the Saccharomyces cerevisiae genome. Nature. 2002;418(6896):387.
57. Cullen LM AG. Genome-wide screening for gene function using RNAi in mammalian cells. Immunol Cell Biol. 2005;83(3):217–23.
58. Morgens DW, Deans RM, Li A, Bassik MC. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. Nat Biotechnol [Internet]. 2016;34(6):634–6. Available from: http://dx.doi.org/10.1038/nbt.3567
59. Hanna RE, Doench JG. Design and analysis of CRISPR–Cas experiments. Nat Biotechnol [Internet]. 2020;38(7):813–23. Available from: http://dx.doi.org/10.1038/s41587-020-0490-7
60. Gallagher LA, Ramage E, Jacobs MA, Kaul R, Brittnacher M MC. A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. Proc Natl Acad Sci. 2007;104(3):1009–14.
61. Ji Y, Zhang B, Van SF, Warren P, Woodnutt G, Burnham MK RM. Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. Science (80-). 2001;293(5538):2266–9.
62. 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 [Internet]. 2017;(June):1–7. Available from: http://academic.oup.com/bib/article/doi/10.1093/bib/bbx110/4100590/Understanding-the-disease-genome-gene-essentiality
63. Mossotto E, Ashton JJ, Gorman LO, Pengelly RJ, Beattie RM, Macarthur BD, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics. 2019;20(254):1–15.
64. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
65. Gregory M. Cooper DL et al. Single-nucleotide evolutionary constraint scores

highlight disease-causing mutations. Nat methods. 2011;23(1):1–7.
66. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110–21.
67. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013;34(1):57–65.
68. Ng PC, Henikoff S. SIFT : predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
69. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536–43.
70. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations a. Nat Publ Gr. 2010;7(4):248–9.
71. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480–9.
72. Kircher M. A general framework for estimating the relative pathogenicity of human genetic variants. Nat g. 2014;46(3):310–5.
73. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. correspondEnce MutationTaster evaluates disease- causing potential of sequence alterations mrsFAST : a cache-oblivious algorithm for short-read mapping. Nat Publ Gr. 2010;7(8):575–6.
74. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2 : mutation prediction for the deep-sequencing age. Nat Publ Gr. 2014;11(4):361–2.
75. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):54–62.
76. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics [Internet]. 2009 Nov 1;25(21):2744–50. Available from: https://doi.org/10.1093/bioinformatics/btp528
77. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 2011;39(17):37–43.
78. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One. 2012;7(10).

79. Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, et al. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat. 2016;37(1):28–35.
80. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, Mcdonnell SK, Baheti S, et al. REVEL : An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877–85.
81. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125–37.
82. Li MX, Kwan JSH, Bao SY, Yang W, Ho SL, Song YQ, et al. Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genet. 2013;9(1):1–11.
83. Li MX, Gui HS, Kwan JSH, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res. 2012;40(7).
84. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011;88(4):440–9.
85. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761–3.
86. Ionita-Laza I, Mccallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214–20.
87. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. 2013;9(8).
88. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet [Internet]. 2016;48(4):349–55. Available from: http://dx.doi.org/10.1038/ng.3511
89. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc [Internet]. 2006;359–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1839740&tool=pmcentrez&rendertype=abstract
90. 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.

Hum Mol Genet. 2017;26(3):489–500.
91. Vergara-Lope A, Ennis S, Vorechovsky I, Pengelly RJ CA. Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding RNA and intergenic chromosome regions. Eur J Hum Genet. 2019;3: 1.
92. Niroula A VM. How good are pathogenicity predictors in detecting benign variants ? PLoS Comput Biol. 2019;15(2):1–17.
93. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet. 2010;6(10):1–11.
94. Bean LJH, Hegde MR. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Med. 2017;9(1):9–11.
95. Kitchenham B, Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. Engineering. 2007;2:1051.
96. Khan SU, Niazi M, Ahmad R. Barriers in the selection of offshore software development outsourcing vendors: An exploratory study using a systematic literature review. Inf Softw Technol [Internet]. 2011;53(7):693–706. Available from: http://dx.doi.org/10.1016/j.infsof.2010.08.003
97. Jalal, Samireh C, Wohlin. Systematic Literature Studies: Database Searches vs. Backward Snowballing Samireh.
98. Badampudi D. Experiences from using snowballing and database searches in systematic literature studies.
99. Gehanno J, Rollin L, Darmoni S. Is the coverage of google scholar enough to be used alone for systematic reviews. 2013;(December 2009):0–4.
100. Becker S, Bryman A, Ferguson H (Thomas H. Understanding research for social policy and practice : themes, methods and approaches. Policy; 2012. 430 p.
101. Craswell G, Poore M. Writing for academic success. SAGE; 2012. 248 p.
102. Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418–26.
103. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: A gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics. 2017;33(4):471–4.
104. Rackham OJL, Shihab HA, Johnson MR, Petretto E. EvoTol : a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. 2014;43(5).
105. Samocha KE, Robinson EB, Sanders SJ, Sabo A, Mcgrath LM, Kosmicki JA, et al. A framework for the interpretation of de novo mutation in human disease. 2015;46(9):944–50.

106. Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, et al. De novo mutations in epileptic encephalopathies. Nature [Internet]. 2013;501(7466):217–21. Available from: http://dx.doi.org/10.1038/nature12439
107. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. Genome Biol [Internet]. 2016;17(1):1–11. Available from: http://dx.doi.org/10.1186/s13059-016-0869-4
108. Jiang Y, Li Z, Liu Z, Chen D, Wu W, Du Y, et al. MirDNMR: A gene-centered database of background de novo mutation rates in human. Nucleic Acids Res. 2017;45(D1):D796–803.
109. Newsweekly IT. Hsu JS, Kwan JS, Pan Z, Garcia-Barcelo MM, Sham PC, Li M. Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. Bioinformatics. 2016 Jun 26;32(20):3065-71. 2016;32(20):2016–8.
110. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of Genomic Variants Using a Unified Biological Network Approach. PLoS Comput Biol. 2013;9(3).
111. Mistry D, Wise RP, Dickerson JA. DiffSLC : A graph centrality method to detect essential proteins of a protein-protein interaction network. PLoS One. 2017;12(11):1–25.
112. Ge X, Kwok PY, Shieh JTC. Prioritizing genes for X-linked diseases using population exome data. Hum Mol Genet. 2015;24(3):599–608.
113. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. Nucleic Acids Res. 2015;43(15):1–9.
114. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: An integrative approach to predicting haploinsufficient genes. Bioinformatics. 2017;33(12):1751–7.
115. 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. Proc Natl Acad Sci [Internet]. 2015;112(44):13615–20. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1518646112
116. Quinodoz M, Royer-bertrand B, Cisarova K, Alessandro S, Gioia D, Superti-furga A, et al. REPORT DOMINO : Using Machine Learning to Predict Genes Associated with Dominant Disorders. Am J Hum Genet [Internet]. 2017;101(4):623–9. Available from: https://doi.org/10.1016/j.ajhg.2017.09.001
117. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–29.
118. Encode Consortium, Carolina N, Hill C. For Junk DNA. Nature. 2013;489(7414):57–

74.
119. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153–7.
120. Eilertson KE, Booth JG, Bustamante CD. SnIPRE: Selection Inference Using a Poisson Random Effects Model. PLoS Comput Biol. 2012;8(12).
121. 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;8(11):1–9.
122. Alyousfi D, Baralle D, Collins A. Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes : a systematic review. Brief Funct Genomics. 2018;18(1):23–9.
123. Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, et al. Genetic variation in an individual human exome. PLoS Genet. 2008;4(8).
124. Durbin RM, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature [Internet]. 2010;467(7319):1061–73. Available from: https://doi.org/10.1038/nature09534
125. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Project: data management and community access. Nat Methods [Internet]. 2012;9(5):459–62. Available from: https://doi.org/10.1038/nmeth.1974
126. Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfied MJ, et al. Integrating Genomics into Healthcare : A Global Responsibility. Am J Hum Genet. 2019;104(1):13–20.
127. Zhang Z, Ren Q. Why are essential genes essential ? - The essentiality of Saccharomyces genes. Microb Cell. 2015;2(8):280–7.
128. Risch N. Genetic linkage: Interpreting lod scores. Science (80-). 1992;255(5046):803–4.
129. Kauwe JSK, Goate A. Genes for a ???Wellderly??? Life. Trends Mol Med [Internet]. 2016;22(8):637–9. Available from: http://dx.doi.org/10.1016/j.molmed.2016.05.011
130. Berg JS, Adams M, Nassar N, Bizon C, Lee K, Schmitt CP, Wilhelmsen KC EJ. An informatics approach to analyzing the incidentalome. Genet Med. 2013;15(1):36–44.
131. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM PM. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2009;18(12):883–9.
132. Al. GJ et. An introduction to statistical learning: with applications in R. Vol. XIV.;

2013. p. 426.
133. Bioresource TN, Project G. Whole-genome sequencing of rare disease patients in a national healthcare system. 2019;
134. Wei JJ, Lander ES, Sabatini DM. Identification and characterization of essential genes in the human genome. 2015;350(6264):1096–101.
135. Cacheiro P, Muñoz-Fuentes V, Murray SA, Dickinson ME, Bucan M, Nutter LM, Peterson KA, Haselimashhadi H, Flenniken AM, Morgan H, Westerberg H. Human and mouse essentiality screens as a resource for disease gene discovery. bioRxiv. 2019 Jan 1:678250. 2019;1–51.
136. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature [Internet]. 2020;581(7809):434–43. Available from: https://doi.org/10.1038/s41586-020-2308-7
137. Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. Available at http://www.orpha.net Accessed (17 July,2019).
138. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM CN. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet. 2009;84:524–33.
139. Alyousfi D, Baralle D, Collins A. Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data. Brief Bioinform. 2020;00(September 2019):1–8.
140. R Core Team. R: A language and environment for statistical computing. RFoundation for Statistical Computing. R-3.5.1 edn (R Foundation for Statistical Computing, Vienna, Austria, 2018).
141. Cui L, Siouve E, Becavin C, Depardieu F, Bikard D. Genome-wide CRISPR-dCas9 screens in E . coli identify essential genes and phage host factors. PLoS Genet. 2018;14(11):1–28.
142. Collins A. The genomic and functional characteristics of disease genes. Briefings in bioinformatics. 2014 Jan 13;16(1):16-23.
143. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Wang Q, Collins RL, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of- function intolerance across human protein-coding genes. bioRxiv. 2019;
144. Abramovs N, Brass A, Tassabehji M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. Nat Genet [Internet]. 2020;52(1):35–9. Available from: http://dx.doi.org/10.1038/s41588-019-

0560-2

145. G. S. The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease. IEEE Pulse. 2015;6(6):22–6.
146. Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6(12).
147. Rapaport F, Boisson B, Gregor A, Béziat V, Boisson-Dupuis S, Bustamante J, et al. Negative selection on human genes causing severe inborn errors depends on disease outcome and both the mode and mechanism of inheritance. bioRxiv. 2020;
148. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294–6.
149. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–6.
150. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics [Internet]. 2014 Apr 1;30(7):1003–5. Available from: https://doi.org/10.1093/bioinformatics/btt637
151. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul;25(14):1754–60.
152. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–9.
153. Franco E De, Watson RA, Weninger WJ, Wong CC, Flanagan SE, Caswell R, et al. REPORT A Specific CNOT1 Mutation Results in a Novel Syndrome of Pancreatic Agenesis and Holoprosencephaly through Impaired Pancreatic and Neurological Development. Am J Hum Genet. 2019;985–9.
154. Kruszka P, Berger SI, Weiss K, Everson JL, Martinez AF, Hong S, et al. REPORT A CCR4-NOT Transcription Complex , Subunit 1 , CNOT1 , Variant Associated with Holoprosencephaly. Am J Hum Genet. 2019;104(5):990–3.
155. The National Genomics Research and Healthcare Knowledgebase Genomics England. Vol. 5. 2019.
156. Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, et al. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. Cell Rep [Internet]. 2015;10(2):148–61. Available from: http://dx.doi.org/10.1016/j.celrep.2014.12.015

157. The Gene Curation Coalition GenCC. https://thegencc.org, date accessed[5 May,2021].
158. AlMuhaizea M, AlMass R, AlHargan A, AlBader A, Medico Salsench E, Howaidi J, et al. Truncating mutations in YIF1B cause a progressive encephalopathy with various degrees of mixed movement disorder, microcephaly, and epilepsy. Acta Neuropathol [Internet]. 2020;139(4):791–4. Available from: https://doi.org/10.1007/s00401-020-02128-8
159. Ahmed A, Wang M, Bergant G, Maroofian R, Zhao R, Alfadhel M, et al. Biallelic loss-of-function variants in NEMF cause central nervous system impairment and axonal polyneuropathy. Hum Genet [Internet]. 2021;140(4):579–92. Available from: https://doi.org/10.1007/s00439-020-02226-3
160. Richard EM, Polla DL, Assir MZ, Contreras M, Shahzad M, Khan AA, et al. Bi-allelic Variants in METTL5 Cause Autosomal-Recessive Intellectual Disability and Microcephaly. Am J Hum Genet [Internet]. 2019;105(4):869–78. Available from: https://doi.org/10.1016/j.ajhg.2019.09.007
161. Bend R, Cohen L, Carter MT, Lyons MJ, Niyazov D, Mikati MA, et al. Phenotype and mutation expansion of the PTPN23 associated disorder characterized by neurodevelopmental delay and structural brain abnormalities. Eur J Hum Genet [Internet]. 2020;28(1):76–87. Available from: http://dx.doi.org/10.1038/s41431-019-0487-1
162. Stade K, Ford CS, Guthrie C, Weis K. Exportin 1 (Crm1p) is an essential nuclear export factor. Cell. 1997;90(6):1041–50.
163. Zhang X, Sun R, Liu L. Potentially critical roles of TNPO1, RAP1B, ZDHHC17, and PPM1B in the progression of coronary atherosclerosis through microarray data analysis. J Cell Biochem. 2019;120(3):4301–11.
164. Rodova M, Kelly KF, VanSaun M, Daniel JM, Werle MJ. Regulation of the Rapsyn Promoter by Kaiso and δ-Catenin. Mol Cell Biol. 2004;24(16):7188–96.
165. Feng Q, Wang H, Ng HH, Erdjument-bromage H, Tempst P, Struhl K, et al. Methylation of H3-Lysine 79 Is Mediated by a New Family of HMTases without a SET Domain University of North Carolina at Chapel Hill. Current. 2002;12(02):1052–8.
166. Ng Y, Tan I, Lim L, Leung T. Expression of the human myotonic dystrophy kinase-related Cdc42-binding kinase γ is regulated by promoter DNA methylation and Sp1 binding. J Biol Chem [Internet]. 2004;279(33):34156–64. Available from: http://dx.doi.org/10.1074/jbc.M405252200
167. Ge Y, Schuster MB, Pundhir S, Rapin N, Bagger FO, Sidiropoulos N, et al. The

splicing factor RBM25 controls MYC activity in acute myeloid leukemia. Nat Commun [Internet]. 2019;10(1). Available from: http://dx.doi.org/10.1038/s41467-018-08076-y

168. Lemon B, Inouye C, King DS, Tjian R. Selectivity of chromatin-remodelling cofactors for ligand-activated transcription. Nature. 2001;414(6866):924–8.
169. Yogosawa S, Makino Y, Yoshida T, Kishimoto T, Muramatsu M, Tamura T. Molecular Cloning of a Novel 120-kDa TBP-Interacting Protein 1 polymerases I , II , and III , and thus appears to be required for all nuclear transcription (1). The to interact selectively with different classes of transcriptional regulators , and certa. Biomed Biophys Res Commun. 1996;617(229):612–7.
170. Yogosawa S, Kayukawa K, Kawata T, Makino Y, Inoue S, Okuda A, et al. Induced expression, localization, and chromosome mapping of a gene for the TBP-interacting protein 120A. Biochem Biophys Res Commun. 1999;266(1):123–8.
171. Chicca JJ, Auble DT, Pugh BF. Cloning and Biochemical Characterization of TAF-172, a Human Homolog of Yeast Mot1. Mol Cell Biol. 1998;18(3):1701–10.
172. Wiese C, Wilde A, Moore MS, Adam SA, Merdes A, Zheng Y. Role of importin-β in coupling ran to downstream targets in microtubule assembly. Science (80-). 2001;291(5504):653–6.
173. LeRoy G, Orphanides G, Lane WS. Requirement of RSF and FACT for transcription of chromatin templates in vitro. Science (80-). 1998;282(5395):1900–4.
174. Mammalian Gene Collection (MGC) Program Team. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences Mammalian Gene Collection (MGC) Program Team*. Pnas [Internet]. 2002;99(26):16899–903. Available from: http://www.pnas.org/content/99/26/16899.full.pdf%5Cnhttps://www.ncbi.nlm.nih.gov/nuccore/BC000312.2
175. Bordeleau ME, Aucagne R, Chagraoui J, Girard S, Mayotte N, Bonneil É, et al. UBAP2L is a novel BMI1-interacting protein essential for hematopoietic stem cell activity. Blood. 2014;124(15):2362–9.
176. Voncken JW, Roelen BAJ, Roefs M, De Vries S, Verhoeven E, Marino S, et al. Rnf2 (Ring1b) deficiency causes gastrulation arrest and cell cycle inhibition. Proc Natl Acad Sci U S A. 2003;100(5):2468–73.
177. Yu X, Yang L, Li J, Li W, Li D, Wang R, et al. De Novo and Inherited SETD1A Variants in Early-onset Epilepsy. Neurosci Bull [Internet]. 2019;35(6):1045–57. Available from: https://doi.org/10.1007/s12264-019-00400-w
178. Kummeling J, Stremmelaar DE, Raun N, Reijnders MRF, Willemsen MH,

Ruiterkamp-Versteeg M, et al. Characterization of SETD1A haploinsufficiency in humans and Drosophila defines a novel neurodevelopmental syndrome. Mol Psychiatry. 2020;1–12.

179. Kracker S, Di Virgilio M, Schwartzentruber J, Cuenin C, Forveille M, Deau MC, et al. An inherited immunoglobulin class-switch recombination deficiency associated with a defect in the INO80 chromatin remodeling complex. J Allergy Clin Immunol. 2015;135(4):998-1007.e6.
180. Tan HHG, Westeneng HJ, van der Burgh HK, van Es MA, Bakker LA, van Veenhuijzen K, et al. The Distinct Traits of the UNC13A Polymorphism in Amyotrophic Lateral Sclerosis. Ann Neurol. 2020;88(4):796–806.
181. Li M, Ma Y, Huang P, Du A, Yang X, Zhang S, et al. Lentiviral DDX46 knockdown inhibits growth and induces apoptosis in human colorectal cancer cells. Gene [Internet]. 2015;560(2):237–44. Available from: http://dx.doi.org/10.1016/j.gene.2015.02.020
182. Yaseen NR, Blobel G. Cloning and characterization of human karyopherin β3. Proc Natl Acad Sci U S A. 1997;94(9):4451–6.
183. Sampson B a, Misra R, Benson S a. Identification and Characterization of a New Gene. Genet Soc Am. 1989;103(1972):97–103.
184. Wang P, Zhu S, Yang L, Cui S, Pan W, Jackson R, et al. Nlrp6 regulates intestinal antiviral innate immunity. 2015;350(6262):826–31.
185. Machol K, Rousseau J, Ehresmann S, Garcia T, Nguyen TTM, Spillmann RC, et al. Expanding the Spectrum of BAF-Related Disorders: De Novo Variants in SMARCC2 Cause a Syndrome with Intellectual Disability and Developmental Delay. Am J Hum Genet. 2019;104(1):164–78.
186. Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu YF, McSweeney KM, et al. Whole-exome sequencing in undiagnosed genetic diseases: Interpreting 119 trios. Genet Med. 2015;17(10):774–81.
187. Oliveira SA, Li Y, Noureddine MA, Zu S, Qin X, Pericak-vance MA, et al. Identification of Risk and Age-at-Onset Genes on Chromosome 1p in Parkinson Disease. 2005;252–64.
188. Diets IJ, van der Donk R, Baltrunaite K, Waanders E, Reijnders MRF, Dingemans AJM, et al. De Novo and Inherited Pathogenic Variants in KDM3B Cause Intellectual Disability, Short Stature, and Facial Dysmorphism. Am J Hum Genet. 2019;104(4):758–66.
189. Wagner DS, Gan L, Klein WH. Identification of a differentially expressed RNA helicase by gene trapping. Biochem Biophys Res Commun. 1999;262(3):677–84.

190. Channavajhala PL, Rao VR, Spaulding V, Lin LL, Zhang YG. hKSR-2 inhibits MEKK3-activated MAP kinase and NF-κB pathways in inflammation. Biochem Biophys Res Commun. 2005;334(4):1214–8.
191. Channavajhala PL, Wu L, Cuozzo JW, Hall JP, Liu W, Lin LL, et al. Identification of a Novel Human Kinase Supporter of Ras (hKSR-2) That Functions as a Negative Regulator of Cot (Tp12) Signaling. J Biol Chem [Internet]. 2003;278(47):47089–97. Available from: http://dx.doi.org/10.1074/jbc.M306002200
192. Brouhard GJ, Stear JH, Noetzel TL, Al-Bassam J, Kinoshita K, Harrison SC, et al. XMAP215 Is a Processive Microtubule Polymerase. Cell. 2008;132(1):79–88.
193. Krueger NX, Streuli M, Saito H. Structural diversity and evolution of human receptor-like protein tyrosine phosphatases. EMBO J. 1990;9(10):3241–52.
194. Kondo S, Lu Y, Debbas M, Lin AW, Sarosi I, Itie A, et al. Characterization of cells and gene-targeted mice deficient for the p53-binding kinase homeodomain-interacting protein kinase 1 (HIPK1). 2003;1.
195. Dupont S, Zacchigna L, Cordenonsi M, Soligo S, Adorno M, Rugge M, et al. Germ-layer specification and control of cell growth by ectodermin, a Smad4 ubiquitin ligase. Cell. 2005;121(1):87–99.
196. Fleischer TC, Yun UJ, Ayer DE. Identification and Characterization of Three New Components of the mSin3A Corepressor Complex. Mol Cell Biol. 2003;23(10):3456–67.
197. Shamsher MK, Ploski J, Radu A. Karyopherin β2B participates in mRNA export from the nucleus. Proc Natl Acad Sci U S A. 2002;99(22):14195–9.
198. Salomon-Kent R, Marom R, John S, Dundr M, Schiltz LR, Gutierrez J, et al. New Face for Chromatin-Related Mesenchymal Modulator: n-CHD9 Localizes to Nucleoli and Interacts With Ribosomal Genes. J Cell Physiol. 2015;230(9):2270–80.
199. Nizon M, Laugel V, Flanigan KM, Pastore M, Waldrop MA, Rosenfeld JA, et al. Variants in MED12L, encoding a subunit of the mediator kinase module, are responsible for intellectual disability associated with transcriptional defect. Genet Med [Internet]. 2019;21(12):2713–22. Available from: http://dx.doi.org/10.1038/s41436-019-0557-3
200. Warner MH, Roinick KL, Arndt KM. Rtf1 Is a Multifunctional Component of the Paf1 Complex That Regulates Gene Expression by Directing Cotranscriptional Histone Modification. Mol Cell Biol. 2007;27(17):6103–15.
201. Mohammadi P, Castel SE, Cummings BB, Einson J, Sousa C, Hoffman P, et al. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. Science (80-). 2019;366(6463):351–6.

202. Patel N, Khan AO, Alsahli S, Abdel-Salam G, Nowilaty SR, Mansour AM, et al. Genetic investigation of 93 families with microphthalmia or posterior microphthalmos. Clin Genet. 2018;93(6):1210–22.
203. Seaby EG, Ennis S. Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies. Brief Funct Genomics. 2020;00(00):1–16.
204. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell [Internet]. 2019;176(3):535-548.e24. Available from: http://dx.doi.org/10.1016/j.cell.2018.12.015

Publications

OXFORD

Briefings in Bioinformatics, 2017, 1–7

doi: 10.1093/bib/bbx110
Paper

Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation

Reuben J. Pengelly, Alejandra Vergara-Lope, Dareen Alyousfi, M. Reza Jabalameli and Andrew Collins

Corresponding author: Andrew Collins, Genetic Epidemiology and Genomic Informatics, Faculty of Medicine, University of Southampton, Duthie Building (808), Tremona Road, Southampton, SO166YD, UK. Tel.: þ44(0)2381206939; E-mail: arc@soton.ac.uk

Abstract

Despite the identification of many genetic variants contributing to human disease (the ‘disease genome’), establishing reliable molecular diagnoses remain challenging in many cases. The ability to sequence the genomes of patients has been transformative, but difficulty in interpretation of voluminous genetic variation often confounds recognition of underlying causal variants. There are numerous predictors of pathogenicity for individual DNA variants, but their utility is reduced because many plausibly pathogenic variants are probably neutral. The rapidly increasing quantity and quality of information on the properties of genes suggests that gene-specific information might be useful for prediction of causal variation when used alongside variant-specific predictors of pathogenicity. The key to understanding the role of genes in disease relates in part to gene essentiality, which has recently been approximated, for example, by quantifying the degree of intolerance of individual genes to loss-of-function variation. Increasing understanding of the interplay between genetic recombination, selection and mutation and their relationship to gene essentiality suggests that gene-specific information may be useful for the interpretation of sequenced genomes. Considered alongside additional distinctive properties of the disease genome, such as the timing of the evolutionary emergence of genes and the roles of their products in protein networks, the case for using gene-specific measures to guide filtering of sequenced genomes seems strong.

Key words: disease genome; gene essentiality; gene-specific filtering; next-generation sequencing

Introduction

Next-generation sequencing of exomes (the protein-coding regions of the sequence) or whole genomes from clinical patient samples typically yields tens of thousands of coding DNA variants. The volume and complexity of these data present many challenges for identification of underlying disease causal mutations. The diagnostic rate for rare diseases by whole exome sequencing is generally in the range 25–50%, varying by phenotype [1, 2]. Chong et al. [3] indicate that the genes underlying 50% (3152) of all known Mendelian phenotypes are still unknown, and many more Mendelian conditions have yet to be recognized. Establishment of gene–disease relationships is complicated by pleiotropy where genetic loci harbour multiple variants associated with multiple and sometimes distinct traits. Therefore, many gene–disease relationships remain poorly understood, even for single-gene disorders and, particularly, for common, complex diseases where multiple causal gene variants of small effect are difficult to recognize.

Part of the difficulty with the interpretation of genome sequences arises through the large number of plausibly damaging variants, which are tolerated [4]. A ‘healthy’ human genome is estimated to contain 100 loss-of-function (LoF) variants [5]. Efforts to predict disease causal variation amongst sequenced genomes usually focus on the properties of individual DNA variants. To predict variant pathogenicity, a number of metrics have been developed, based on, for example, conservation scores, changes in amino acid sequence or predicted effect on protein function [for example, Sorting Intolerant From Tolerant (SIFT), PolyPhen and GERP (Genomic Evolutionary Rate Profiling)] [6–8]. Both SIFT and PolyPhen use sequence homology of related proteins to predict whether an amino acid change might damage protein function. The conservation of the specific base though evolution is considered through multiple sequence alignment across species. The SIFT algorithm uses only homology for prediction, whereas PolyPhen also considers whether an amino acid change occurs in an important functional or structural site in the protein. GERP considers evolutionary constraint at specific positions in the sequence using a maximum likelihood approach to compute evolutionary rates. There are also evolutionary and functional prediction tools such as CADD (Combined Annotation Dependent Depletion) [9], which integrates predictive scores from multiple annotations into one metric.

However, many variants scored as apparently damaging by these methods are likely to be tolerated and firmly establishing a molecular diagnosis from a sequenced genome can be challenging. Therefore, the use of gene-specific measures alongside these variant-specific annotations has been suggested as a way to improve the ability to identify causal mutations [10].

Understanding the nature of the ‘disease genome’, which we define as the set of genes which contain coding variation and/or have associated non-coding regulatory variation contributing to disease, is important for developing strategies which best exploit gene-specific information. The complex mechanisms underlying the creation and persistence of the disease genome and pathogenic variation therein are not clearly understood but depend substantially on interactions between genetic recombination, selection and mutation. The pattern of linkage disequilibrium (LD) is an outcome of these processes and may have a close relationship to the disease genome [11, 12]. Increased understanding of how these underlying processes define patterns of disease variation will go some way towards resolving the causes of disease in individual genomes [13]. The interplay between these processes and outcomes, and how they underlie disease variation in the genome, is fascinating and the main focus of this review.

Selection and the disease genome

The role of selection in shaping genomes is defined in three ways: (1) ‘Hard’ selective sweeps where new, advantageous, mutations are driven to fixation by positive selection, (2) ‘Soft’ selective sweeps in which there is more gradual fixation of weakly beneficial variation by positive selection and (3) by negative (purifying) selection in which there is elimination of deleterious mutations. The relative impact of negative versus positive selection in shaping the human genome is uncertain [14]; however, most mutations affecting phenotypes must be deleterious [15]. Lohmueller et al. [16] stress the relative importance of negative over positive selection acting on the genome. Although selective sweeps tend to locally reduce genetic variation, they are not considered to be a dominant factor explaining patterns of variability across the genome. Variants contributing to disease most likely arise by random mutation and, at least for highly penetrant monogenic variants, are maintained at low frequencies by purifying selection [17]. However, variants involved in complex traits (common disorders in which a disease allele contributes only a small fraction of disease risk) must be subject to only extremely weak negative selection [18].

The efficiency of selection may be reduced under certain conditions through the mechanism of Hill–Robertson interference (HRI) [19–21]. Considering variants subject to positive selection, there may be a situation where an advantageous mutation arises and starts to spread through the population. However, before this mutation achieves fixation, a second advantageous mutation at a nearby locus emerges in an individual who lacks the first mutation. The two advantageous alleles are effectively in competition. Recombination enables the creation of haplotypes carrying both advantageous alleles, with increased fitness

assuming it is more advantageous to carry both alleles. However, in weakly recombining genomic regions, this haplotype is much less likely to be generated. Therefore, the efficacy of selection acting on linked sites simultaneously can be reduced in the presence of limited recombination.

The impact of HRI in weakly recombining genome regions can also be seen for variants subject to purifying selection. Hussin et al. [13] considered the impact of HRI on the distribution of damaging variants. If there are many sites (for example, damaging non-synonymous variants) in a small genomic region, which has a low recombination rate, HRI may allow potentially deleterious variation to achieve high frequencies [15]. The impact is greater with an increasing number of sites subject to purifying selection. Meiotic recombination acts to break down this interference allowing these sites to segregate independently and form new haplotypes leading to reduction in the accumulation of damaging alleles [22].

Recombination and the disease genome

During meiosis, the creation of DNA double-stranded breaks is followed by repair through homologous recombination. This process enables allele/haplotype shuffling with significant evolutionary advantage through the breakdown of associations between alleles at linked loci (in LD), which arise by genetic drift [22]. The close alignment between the recombination structure and patterns of LD enabled the recognition of the exquisite and remarkable mechanism, which promotes narrow, intense regions of recombination (hotspots). This process involves the binding of histone methyltransferase PR (positive-regulatory) domain containing 9 (*PRDM9*). This mechanism results in histone methylation before creation of a double-stranded break and is associated with biased gene conversion or 'hotspot drive' [14]. Selective bias in favour of the non-recombinogenic allele eventually drives the extinction of the recombination hotspot [23]. However, the highly evolving zinc finger domain of PRDM9 changes the motif it recognizes with subsequent generation of new hotspots [24]. Recombination may also influence the evolution of the genome through GC-biased gene conversion in which there is biased introduction of G and C nucleotides during mismatch repair following recombination [25], and also, there may be biased transmission of the shorter or longer allele of an insertion–deletion polymorphisms (indels) during meiosis [14]. However, Webster and Hurst [14] suggest there is no evidence that these indirect effects of variation in recombination rate across the genome impacts the efficiency of selection.

Meiotic recombination has a significant role in determining the abundance and location of disease associated variation in the genome [13]. Where recombination is absent (and there is

no mutation back to the original allele), a process termed Muller's ratchet [22, 26] has an important impact. In the absence of recombination, deleterious variants arising by mutation cannot be eliminated because the original haplotypes that lack the mutation cannot be re-generated. Suppressed recombination and the build-up of deleterious variation may explain why most Y chromosome genes are inactive [20]. Given the highly variable recombination rates across chromosome regions, the pattern of recombination provides insights into the processes underlying the distribution of disease variation across the genome. Hussin et al. [13] contrasted levels of potentially damaging variation in highly recombining parts of the genome with weakly recombining regions. They provide clear evidence that purifying selection removes damaging variation more efficiently in highly recombining regions.

The possibility that recombination is itself mutagenic has been considered. It is known that recombination underlies sequence structural changes because of non-allelic homologous recombination [14], but there is limited evidence it can introduce point mutations [27, 28]. Schaibley et al. [27] found wide variation in mutation rates related to local GC content, but not to the recombination rate. Overall, the available data suggest that the recombination rate has limited effect on the frequency of mutation.

Mutation and the disease genome

Genes with high mutation rates might appear to be disease candidates simply because multiple patient genomes are likely to contain mutations in these genes. Variability in mutation rate provides a particular challenge to interpretation of sequenced genomes [28]. Mutations arise through copying errors during replication, spontaneous DNA changes and DNA instability [29]. It is known that mutation rates vary widely on different scales from single nucleotides through to whole chromosomes [30]. There are powerful context effects in which the mutation rate is influenced by adjacent nucleotides causing mutation rate variability of >650-fold [31]. For example, CpG dinucleotides constitute <2% of the genome but account for 19% of the *de novo* mutations [29, 32] and are the most mutable sites in the genome [33]. During replication, DNA mispairing is frequent with G-T and A-C mispairing the most common. This creates a 2-fold rate of transitions compared with transversions, when the opposite would be expected if all changes were equally likely [29]. There is also evidence for more cryptic context-independent variation with some sites appearing hyper-mutable [34]. The sequence context of each gene is a powerful predictor of mutation rates. Aggarwala and Voight [35] introduce 'substitution intolerance scores' for genes demonstrating that a heptanucleotide context accounts for >81% of variability in substitution probabilities. They identify mutation-promoting motifs at ApT dinucleotides, CAAT and

TACG sequences. Based on this 7-mer sequence, the substitution intolerance score quantifies the difference between expected and observed functional variants in a gene given the sequence context.

LD and the disease genome

The pattern of LD is broadly conserved among different populations [36] and known to be highly determined by recombination, but 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, although its effects are complex [37]. Remarkable alignment between the structure of the linkage map in centimorgans (which quantifies meiotic recombination over a few generations) and the 'historical' pattern of recombination in LD maps (reflecting accumulated recombination over many generations) has been demonstrated [36]. The X chromosome shows an excess of LD reflecting either reduced recombination or, more significantly, increased selective pressure on the haploid X [38, 39]. Lek et al. [40] note that genes on the X chromosome are significantly more constrained (having fewer rare variants per gene than expected under a selection neutral model) compared with genes on the autosomes.

Gibson et al. [11] and Collins [12] have shown, by constructing LD maps of individual genes from exome data, that there is enrichment of disease variation amongst genes with 'average' levels of LD. This pattern is distinct from genes with strong LD, which are enriched for essential functions (e.g. phosphorylation, cell division, cellular transport and metabolic processes) and genes with weak LD, which are enriched for functions related to sensory perception and some immune functions.

Gene essentiality and the disease genome

Essential genes are critical for cell viability. The degree of gene essentiality is likely to have a direct bearing on the tolerance a gene has for damaging/disease variation. Quantifying gene essentiality is challenging, and the essentiality of individual genes has traditionally been evaluated from mouse knockout experiments for the orthologous genes. Dickerson et al. [41] questioned whether knockouts, which remove the protein-coding region of the gene, are a valid representation, as less severe changes (such as point mutations) are more typical with likely less damaging effects. More recently, a range of techniques, such as large-scale short hairpin RNA screens of diverse cell lines, ChIP-seq (chromatin immunoprecipitation-sequencing) and computational predictions, through integration of gene expression,

molecular alterations and pathways, have been developed [42]. CRISPR-Cas9 genome editing has also emerged as a technique to allow largescale studies into genome-wide essentiality [43, 44]. The latter approach has enabled refined determination of some of the distinct features of essential genes suggesting that protein interaction networks, integrated with gene expression or histone marks, are predictive of gene essentiality.

The substitution intolerance score [35], in which higher scores indicate functionally constrained genes, is a measure correlated with essentiality. As expected, genes that were classed as likely to be essential or ubiquitously expressed scored highly for intolerance of functional variation. Genes related to keratin pathways or with olfactory functions were highly tolerant of functional changes, whilst OMIM disease genes had more intermediate tolerance (Table 1).

Similarly, the loss intolerance probability (pLI) score, described from the ExAC data set of 60706 exomes [40], has been used as an approximation to gene essentiality. pLI defines the probability of a gene being intolerant to variation causing loss of gene function. Lek et al. [40] identified 3230 genes as intolerant (pLI>0.9) and 10374 as tolerant (pLI<0.1). Dominant

Table 1. Some comparative functional and sequence characteristics among gene classes

Characteristic	NDNE	Complex disease genes	Monogenic disease genes	END[a]	References
Gene age	+	++	+++	++++	[45]
Cellular localization of encoded protein	Plasma membrane/ extracellular	Plasma membrane/ extracellular	Plasma membrane/ extracellular	Nuclear localization	[41]
Gene expression, position in protein network	Not ubiquitously expressed	Not ubiquitously expressed, peripheral functions in protein networks	Not ubiquitously expressed, peripheral functions in protein networks	Ubiquitous expression, protein network hub	[45–47]
Degree of connectivity in protein–protein interaction networks	+	++	++	+++	[41]
Intensity of purifying selection	+	+/++?	+++(more for dominant)	++++	[17, 48]
Coding sequence length	++	+++?	+++?	+	[12, 46, 49]
Substitution intolerance score	+	?	++	+++	[35]
Gene intolerance of rare variation	+	++?	++	+++	[40]

Note: +, ++, +++¼relative magnitude of specific gene property. [a]Any damaging mutations likely to be lethal.

disease genes were found to be enriched for LoF intolerant genes, whereas recessive disease genes were found to include a smaller proportion of LoF-intolerant genes. Genes found to be intolerant of LoF variation had almost complete absence of protein-truncating variants suggesting strong purifying selection. The gene-specific pLI metric is positively correlated with degree of interconnectivity in protein–protein networks, and the most constrained pathways include core biological processes (spliceosome; ribosome; proteasome components), whereas olfactory receptors are the least constrained.

A number of studies have evaluated the relationship between gene essentiality and human disease. Tu et al. [46] recognized that essential genes are distinct from other 'non-disease' genes. They compared ubiquitously expressed human genes (housekeeping genes), as a group likely to contain many essential genes, with disease genes and other non-disease genes. Ubiquitously expressed genes are presumed essential for fundamental cellular physiology, but essential genes with more tissue-specific functions will not be included in this set. Essential genes might be regarded as the most severe 'disease' genes in that disruption of function is likely to be developmentally lethal. Housekeeping genes were found to have shorter coding sequence lengths than disease genes consistent with earlier evidence [49] of shorter introns, untranslated regions and coding sequences, suggesting selection for more compact sequences (Table 1). Interestingly, there is some evidence that disease genes are longer on average than other genes [12, 46] (Table 1).

Spataro et al. [17] analysed gene properties based on roles in protein networks, rates of protein evolution and tests of neutrality. They identified three gene groups with distinct degrees of essentiality:

i. Genes that are neither essential nor associated with disease (non-disease, non-essential genes, NDNE), which have the least functional relevance and are under the weakest levels of purifying selection.
ii. Human disease (HD) genes, from a curated version of OMIM (hOMIM) [48], which are functionally relevant but less than essential non-disease (END) genes. These genes are under stronger and longer lasting purifying selection than NDNE genes.
iii. END, based on orthologues of mouse essential genes from knockout experiments. These genes have no association with human disease because functionally relevant mutations are likely to have lethal consequences such as a miscarriage or early death.

We compared two alternative representations of essentiality by evaluating pLI scores [40] in each of the three Spataro et al. [17] gene groups (Table 2). Although there is a significant trend towards higher pLI (greater intolerance of functional variation) from NDNE genes, HD genes through to END genes (correlation $P=0.17$, $P<0.0001$), in line with assumptions about essentiality, there is a wide overlap between the three groups. This suggests only limited consistency between the three-group classification and pLI scores as measures of gene essentiality. Inconsistency in classification could arise in a number of ways. For example, the classification of END genes, based on mouse knockouts, has been criticized [41], and pLI essentiality scores consider only functional variation in coding regions of the genome, whereas disease variation is known to extend to non-coding regulatory regions, particularly for complex diseases. However, integrative analysis of alternative measures of

essentiality may form a basis for the development of models, which enhance recognition of disease variation.

Mendelian or complex trait genes?

It is known that most variants associated with complex traits are regulatory in function, and their target genes are difficult to ascertain requiring challenging functional investigation [50]. Therefore, the understanding of variation underlying complex phenotypes is far less complete than for monogenic disorders. Spataro et al. [17] find that genes with variants for Mendelian disorders, which are also associated with variation underlying complex traits ('Complex-Mendelian' genes), have higher functional relevance in protein networks and higher expression levels than genes associated only with complex traits. In this sense, they might be seen as intermediate between Mendelianonly and complex trait-only genes.

Synthesis

We propose a scheme representing the opposing and interacting processes, which define the disease and non-disease genomes (Figure 1). We assume an underlying increasing measure of gene essentiality in which the most essential genes are those which are required for survival and reproduction such that functional disruption is lethal [17, 46]. The intensity of recombination and selection varies across the spectrum of gene essentiality. The nature of the relationships between these processes is not known, but Figure 1 indicates trends supported by published studies.

Genes with low essentiality tend to have high recombination rates (for example, quantified as centimorgans per kilobase) and are weakly impacted by selection. They have high haplotype diversity and correspondingly weak LD. Genes at this end of the essentiality scale may be more tolerant of mutation and include genes involved in sensory perception, such as genes encoding olfactory receptors [12, 40, 51, 52]. The high recombination rate may enable re-generation of less damaging haplotypes, but residual variation is presumably tolerated and unlikely to contribute to disease.

Genes with high essentiality, however, tend to have low recombination rates, but the impact of selection is intense because, with increasing essentiality, any damaging variation is associated with lethality. As a result, they have limited haplotype diversity and strong LD. Previous studies [40, 51] have found that genes involved in DNA and RNA metabolism, response to DNA damage and the cell cycle may fall into this category. The most essential genes might be regarded as the most severe 'disease' genes.

Genes which contain, or are impacted by, disease variation are suggested to occupy an intermediate place in this scheme. There is evidence that disease genes show intermediate levels of LD [12]. Genes may be exposed to recombination and selection of reduced intensity, which enables retention of some damaging variation associated with disease. The impact of HRI in reducing the efficiency of selection and Muller's ratchet in enabling accumulation of damaging variation may be significant for this class of genes. Arguably genes impacted by variation involved in common diseases might be discriminated from genes involved in more severe monogenic disease through the monogenic forms being closer to the essential gene end of the spectrum.

Table 2. Gene essentiality pLI scores [40] within gene essentiality groups [17]

Gene class [17]	Number of genes [17]	Number of genes with pLI score	1st quartile pLI score	Median pLI score	3rd quartile pLI score	Mean pLI score
NDNE	13135	12062	0.000	0.010	0.475	0.251
HD	3275	3165	0.000	0.041	0.820	0.339
END	1572	1509	0.022	0.704	0.991	0.554

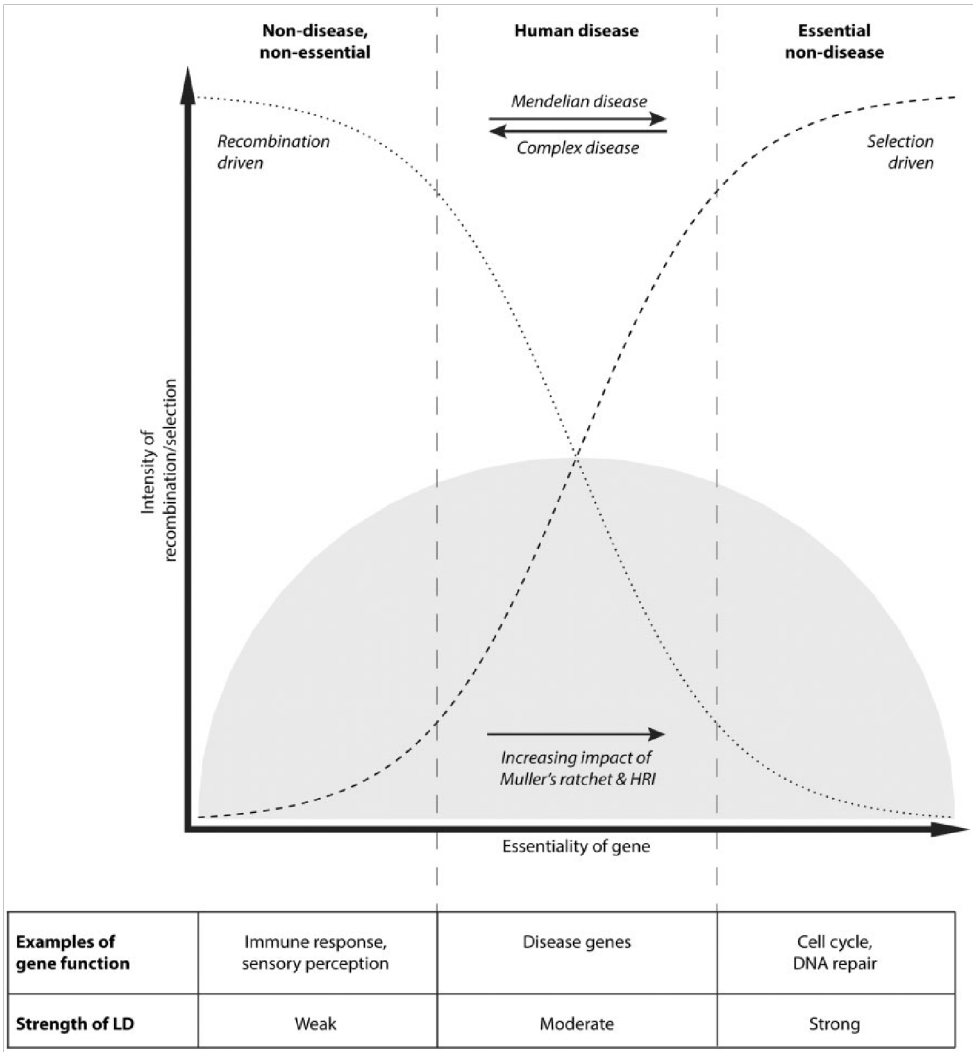


Figure 1. Outline of hypothetical relationships between gene essentiality, recombination (dotted line) and selection (dashed line). Deleterious variation (shaded area) is presumed to be depleted through recombination for 'non-disease, non-essential' gene groups and intense selection for 'essential non-disease' gene groups. Relatively weaker recombination and selection intensities may allow persistence of damaging variation for genes with associated disease variation.

Discussion

The dramatic growth in the number of human genomes sequenced (now likely to be in the hundreds of thousands [40]) is underpinning a developing understanding of genes with diseaserelated variation in their coding or regulatory regions. Increased knowledge of the processes that generate this variation and allow it to persist is likely to improve the efficiency with which patient genomes can be screened to identify the molecular basis of disease. The interplay of selection, recombination and mutation underlies the pattern of disease variation, and understanding these processes may enhance resolution of more cases with monogenic disease. The extent to which these processes can be informative for complex disease is less clear, given the extremely small effect size of the variants involved that have mostly been identified in large genome-wide association studies [50]. However, analyses of gene properties may be enhanced through consideration of additional gene characteristics (Table 1). Gene age was highlighted by Cai et al. [45], where age is defined through models of evolutionary emergence times [53]. Younger genes, for example, are more likely to have primate or human-specific functions contrasting with older genes, which have more ancient phylogenetic origins. They found that Mendelian disease genes tend to be a more ancient group compared with non-disease genes, whilst complex disease genes tended to have intermediate ages.

Coding sequence length is reduced in genes with greater essentiality [49]. These genes are subject to intense selection but have reduced recombination rates (Figure 1). Where the outcome of selection is not lethality, the efficiency of selection may be impacted by HRI, and damaging variation might accumulate by Muller's ratchet. Conceivably, the smaller coding sequence length in these genes reduces the target size and, therefore, the probability of a deleterious mutation occurring in the sequence, offsetting the impact of these processes. Sequence context analysis, for example, through substitution probabilities [35] may provide insights into differential mutation rates across genes and their interaction with other contributing mechanisms.

Further distinctions include degree of connectivity of the protein product [41], position in the protein network [45] and cellular localization [41] (Table 1), although these may be more informative of the essential gene: non-essential gene categorization.

As might be expected, broad gene categories are not independent. Genes that contain lethal null alleles can have nonlethal disease alleles [41] complicating efforts to categorize genes. The use of gene-specific measures to filter sequenced genomes to identify causal variation can only be successful when used alongside variant-specific analyses with conclusions supported by functional tests. However, there is good evidence that integrated

models using emerging approximations for essentiality and gene-specific data on recombination, mutation and selection may contribute to improved molecular diagnostics in the analysis of patient sequence data.

Key Points

- The identification of causal disease variation from patient genome sequences is challenging and confounded by plausibly damaging variation, which is neutral.
- Methods that predict whether a variant is damaging might be misleading, and recent studies have suggested that information about the properties of genes might improve molecular diagnoses.
- There is evidence that genes that have associated disease variation have intermediate essentiality between the extremes of genes of low essentiality (which are tolerant of functional variation) and genes of high essentiality (in which functional variation may be lethal).
- Gene essentiality and its relationship to variable recombination and mutation rates, along with variation in intensity of selection, may provide a basis for developing models, which improve gene-specific predictors of disease variation.

Funding

The authors acknowledge support from the University of Southampton EPSRC Centre for Doctoral Training in Next Generation Computational Modelling (CDT NGCM).

References

1. Smith ED, Radtke K, Rossi M, et al. Classification of genes: standardized clinical validity assessment of gene–disease associations aids diagnostic exome analysis and reclassifications. Hum Mutat 2017;38(5):600–8.
2. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med 2017;9(386):eaal5209.
3. Chong JX, Buckingham KJ, Jhangiani SN, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet 2015; 97:199–215.
4. Itan Y, Casanova JL. Can the impact of human genetic variations be predicted? Proc Natl Acad Sci USA 2015;112(37): 11426–7.
5. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 2012;335(6070):823–8.

6. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.
7. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005;15:901–13.
8. Kumar P, Henikoff S, Ng PC. Predicting the effects of codingnon-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
9. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5.
10. Petrovski S, Wang Q, Heinzen EL, et al. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 2013;9(8):e1003709.
11. Gibson J, Tapper W, Ennis S, et al. Exome-based linkage disequilibrium maps of individual genes: functional clustering and relationship to disease. Hum Genet 2013;132(2):233–43.
12. Collins A. The genomic and functional characteristics of disease genes. Brief Bioinform 2015;16(1):16–23.
13. Hussin JG, Hodgkinson A, Idaghdour Y, et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. Nat Genet 2015;47(4):400–4.
14. Webster MT, Hurst LD. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet 2012;28(3):101–9.
15. Charlesworth B. The effects of deleterious mutations on evolution at linked sites. Genetics 2012;190(1):5–22.
16. Lohmueller KE, Albrechtsen A, Li Y, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genet 2011;7(10):e1002326.
17. SpataroN, RodriguezJA,Navarro A, etal.Propertiesofhumandisease genes and the role of genes linked to Mendelian disorders in complexdiseaseaetiology.HumMol Genet2017;26:489–500.
18. Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 2009; 26(9):2097–108.
19. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res 1966;8(3):269–94.
20. Charlesworth B, Charlesworth D. The degeneration of Y chromosomes. Philos Trans R Soc B Biol Sci 2000;355(1403):1563.
21. Comeron JM, Williford A, Kliman RM. The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity 2008;100(1):19–31.
22. Felsenstein J. The evolutionary advantage of recombination. Genetics 1974;78:737–56.
23. Jeffreys AJ, Neumann R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat Genet. 2002;31(3):267–71.

24. Oliver PL, Goodstadt L, Bayes JJ, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS Genet 2009;5(12):e1000753.
25. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 2009;10:285–311.
26. Muller HJ. The relation of recombination to mutational advance. Mutat Res 1964;1(1):2–9.
27. Schaibley VM, Zawistowski M, Wegmann D, et al. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. Genome Res 2013;23(12):1974–84.
28. Fuentes Fajardo KV, Adams D, Mason CE, et al. Detecting false-positive signals in exome sequencing. Hum Mutat 2012; 33(4):609–13.
29. Se´gurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet 2014;15:47–70.
30. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 2011;12(11):756–66.
31. Carlson J, Scott LJ, Locke AE, et al. Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. bioRxiv 2017;108290.
32. Fryxell KJ, Moon WJ. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol. 2005;22(3):650–8.
33. Cooper DN, Youssoufian H. The CpG dinucleotide and human genetic disease. Hum Genet 1988;78(2):151–5.
34. Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate. PLoS Biol 2009;7(2): e1000027.
35. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet 2017;48:349–55.
36. Lonjou C, Zhang W, Collins A, et al. Linkage disequilibrium in human populations. Proc Natl Acad Sci USA 2003;100(10): 6069–74.
37. Jacobs GS, Sluckin TJ, Kivisild T. Refining the use of linkage disequilibrium as a robust signature of selective sweeps. Genetics 2016;203(4):1807–25.
38. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes. Nat Rev Genet 2006;7(8):645–53.
39. Wang ET, Kodama G, Baldi P, et al. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci USA 2006;103(1):135–40.
40. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of proteincoding genetic variation in 60,706 humans. Nature 2016;536(7616):285–91.
41. Dickerson JE, Zhu A, Robertson DL, et al. Defining the role of essential genes in human disease. PloS One 2011;6(11):e27368.
42. Jiang P, Wang H, Li W, et al. Network analysis of gene essentiality in functional genomics experiments. Genome Biol 2015;16(1):239.

43. Shalem O, Sanjana NE, Hartenian E, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 2014; 343(6166):84–7.
44. Wang T, Wei JJ, Sabatini DM, et al. Genetic screens in human cells using the CRISPR-Cas9 system. Science 2014;343(6166): 80–4.
45. Cai JJ, Borenstein E, Chen R, et al. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol Evol 2009;1:131–44.
46. Tu Z, Wang L, Xu M, et al. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics 2006;7(1):31.
47. Goh KI, Cusick ME, Valle D, et al. The human disease network. Proc Natl Acad Sci USA 2007;104(21):8685–90.
48. Blekhman R, Man O, Herrmann L, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol 2008;18(12):883–9.
49. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet 2003;19(7):362–5.
50. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet 2015;16(4):197–212.
51. Smith AV, Thomas DJ, Munro HM, et al. Sequence features in regions of weak and strong linkage disequilibrium. Genome Res 2005;15(11):1519–34.
52. Pierron D, Corte´s NG, Letellier T, et al. Current relaxation of selection on the human genome: tolerance of deleterious mutations on olfactory receptors. Mol Phylogenet Evol 2013; 66(2):558–64.
53. Domazet-Loso T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol 2008;25:2699–707.

Briefings in Functional Genomics, 00(00), 2018, 1–7

doi: 10.1093/bfgp/ely033

Advance Access Publication Date: 12 October 2018 Review paper

Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review

Dareen Alyousfi, Diana Baralle, and Andrew Collins

Corresponding author: Andrew Collins, Human Development and Health, Faculty of Medicine, University of Southampton, SO16 6YD, UK. Tel: +44 (0)23 81206939; E-mail: arc@soton.ac.uk

Abstract

The evolution of next-generation sequencing technologies has facilitated the detection of causal genetic variants in diseases previously undiagnosed at a molecular level. However, in genome sequencing studies, the identification of disease genes among a candidate gene list is often difficult because of the large number of apparently damaging (but usually neutral) variants. A number of variant prioritization tools have been developed to help detect disease-causal sites. However, the results may be misleading as many variants scored as damaging by these tools are often tolerated, and there are inconsistencies in prediction results among the different variant-level prediction tools. Recently, studies have indicated that understanding gene properties might improve detection of genes liable to have associated disease variation and that this information improves molecular diagnostics. The purpose of this systematic review is to evaluate how understanding gene-specific properties might improve filtering strategies in clinical sequence data to prioritize potential disease variants. Improved understanding of the 'disease genome', which includes coding, noncoding and regulatory variation, might help resolve difficult cases. This review provides a comprehensive assessment of existing gene-level approaches, the relationships between measures of gene-pathogenicity and how use of these prediction tools can be developed for molecular diagnostics.

Key words: gene-specific metrics; disease genome; gene-level scores; gene essentiality; gene-specific filtering

Introduction

The sequencing of whole genomes using next-generation sequencing (NGS) yields vast data sets that present significant analytical challenges for identification of disease-causal variants. It is known that a subset of human genes contains, or is associated with, rare and/or common variation that has a role in disease processes (the 'disease genome'). However, recognition of causal variants among many thousands of mostly neutral variants is a huge challenge and a pressing problem. For example Chong *et al*. [1] state that the genes underlying ~50% of all Mendelian phenotypes remain unknown and many more Mendelian conditions are still to be described. Alongside methods for predicting the potential pathogenicity of individual DNA variants at least 20 gene-specific metrics (scores) have been developed in recent years that may help facilitate recognition of disease-causing variation. An example of one of these methods is residual variation intolerance score (RVIS) that ranks genes by whether they have more or less common functional genetic variation relative to the genome wide expectation [2]. A candidate pathogenic variant found in a gene classed as intolerant of common functional variation might be worthy of follow-up as a potential causal variant. Understanding the properties of the disease genome and integrating existing gene-specific predictors may help in clas- sifying genes based on their specific features to refine molecu- lar diagnosis. Pathogenicity scores for individual DNA variants are often inconsistent in that different methods can provide conflicting evidence on potential pathogenicity. The degree of redundancy in the genome makes the task of picking out causal variation particularly challenging. We recognize that variant prediction tools alone are currently not conclusive and that evidence at the gene-specific level has the potential to enhance the recognition of variant pathogenicity [3].

This systematic review considers the literature related to gene-specific scores and their applicability to improve filtering of genome sequence data. We set out to achieve a satisfactory answer to the following research question: 'Can the use of gene- specific metrics facilitate the identification of disease genes in patient genomes?'

Gene-specific metrics are frequently based on properties of genic coding regions. The extent to which they provide informa- tion on the tendency of a gene to have associated disease-

causal variation outside the coding region is limited. Most of the tools analyzed in this review, with a few exceptions, are concerned with genomic coding variation.

Details of the methodology used in this systematic review are given in the Supplementary Data,Supplementary Figures 1 and 2 and Supplementary Table 1 [3-14].

Findings: key models

Each of the 20 gene-specific approaches identified by the systematic review was classified into 1 of 3 groups according to the main focus of each method. We consider below each of the three groups: (i) Essentiality and conservation, (ii) Haploinsufficiency (HI) and (iii) Selection. Supplementary Tables 2–4 give details of the main methods and scores allocated into each category.

Characteristics of essential and conserved genes

Essential and conserved genes encode proteins that have core biological functions that are essential for an organism's viability. Genes vary in their degree of essentiality and a number of different quantitative scores provide approximations to essentiality. These include predictions of the extent to which a gene is tolerant or intolerant of loss of function (LoF) mutations and estimation of the expected rate of *de novo* mutations (DNMs) [14]. Supplementary Table 2 outlines the key approaches in this category. The RVIS ranks genes by probability of carrying more, or less, functional genetic variation than expected highlighting genes intolerant to common functional variation [2]. Genes with positive scores have more common functional variation, while negative scoring genes are less tolerant having reduced associated common functional variation. Genes containing variation involved in monogenic diseases have lower RVIS scores than other genes.

By examining the evolutionary conservation of protein sequences, Rackham *et al*. [16] developed the Evolutionary inTolerance (EvoTol) score to identify genes that are intolerant to mutation [15]. Because only small areas of a gene may be intolerant, for example protein-coding domains, these subregions might be particularly important domains of essentiality [16].

EvoTol allows identification of intolerant protein subdomains alongside the identification of intolerant genes more generally.

The development of NGS makes possible the identification of newly arising DNMs and their potential roles in rare disease. Recognition of these variants is not without difficulty because of errors in alignment and poorly supported variant calls. Validation by re-sequencing and, in particular, sequencing of additional family members (often the parents of a patient) can help correctly resolve *de novo* variation that might be of disease significance. Such mutations are not considered to play a significant role in the pathogenesis of complex diseases [17]. To accurately estimate the expected rate of DNMs in a given gene, careful

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/ely033/5126507 by University of Southampton user on 30 October 2018

assessment of gene mutability is required.Gene length and local sequence context are essential factors underlying mutation rate differences [17]. Samocha *et al*. calculated per-gene probabilities of mutation that are correlated with observed counts of rare missense variants in the Exome Sequencing Project (ESP) data set. The Samocha *et al*. study extends a model that investigated DNMs in epileptic encephalopathy patients (Epi4K consortium) by considering depth of coverage (i.e. how many sequence reads were present on average per base) and the regional divergence in genes between humans and macaques. Significant numbers of genes with missense variant deficits were observed,compared to expectation from predicted mutation rates, suggesting strong evolutionary constraint removing variants by negative selection [17, 18]. The Samocha *et al*. [17] model utilizes exome sequence data to evaluate the DNM rate (DNMR) by gene set and on a single gene basis; this score is referred to as *de novo* excess (DNE). The metric is predictive of selective constraint in the human genome and identifies 1003 constrained genes known to cause severe human disease [17].It was found that constrained genes contain higher *de novo* LoF mutation rate than expected by chance [17].

The LoFtool measures the ratio of LoF mutations to synonymous mutations for every gene. The performance of the LoFtool, compared to RVIS, DNE Z-score and EvoTol, suggests enhanced prediction of *de novo* haploinsufficient disease-causing genes. The LoFtool represents values as intolerance percentiles; genes that are intolerant to LoF variation have low LoFtool percentiles [15]. The four measures of genic intolerance outlined so far were included by Bartha *et al*. [19] who described them as essentiality scores.

In early 2016, using data from 1000 Genomes Project, Aggarwala *et al*. proposed the Substitution Intolerance Score (SIS) as a gene-level measurement of essentiality. Genes with high SIS scores are functionally constrained, while genes which score low are tolerant of functional changes in the protein that might arise through mutations in the DNA sequence [20].

Another scoring system by Gussow *et al*. [21] evaluates intolerance in genic subregions proposing that more conserved regions within a gene are expected to contain more variants that are pathogenic. Genes are divided into subregions and tiered by intolerance to functional variation.This ‘subRVIS’ score ranks regions using RVIS but with the addition of information on conservation. Regions intolerant to functional variation are scored low by the subRVIS scoring system. The method utilizes the GERP++ [22] score to evaluate evolutionary constraint for bases in each subregion [21].

The Loss Intolerance probability (pLI) score quantifies the likelihood that a gene is intolerant to a mutation that produces LoF in the protein product [23]. The score is derived using the Exome Aggregation Consortium (ExAC) database that is an extensive catalogue of

human genetic diversity. This catalogue identifies one variant every eight bases on average in the exome providing a powerful filter for analysis of candidate deleterious variants in severe Mendelian diseases [23]. Lek *et al*. proposed that genes with high pLI score (pLI >= 0.9) are most intolerant of LoF variation. Genes in this category are the most evolutionarily constrained. The least constrained genes (LoF tolerant) have low pLI scores (pLI < = 0.1) and typically contribute to the least constrained biological pathways, such as sensory perception, where high haplotype diversity is potentially advantageous [23].

It is challenging to assess the relationship between the DNMR and genes involved in disease. In 2017, Jiang *et al*. utilized available DNM data to correct for the background mutation rate seen as one of the main limitations of the Samocha *et al*. [17] model. The problem arises because by sequencing more individuals, more DNMs are inevitably observed in the same gene by chance. Therefore, in a given disease, if a DNM is related to pathogenesis, disease genes might be expected to contain more DNMs than predicted from background rates. This work includes the development of a database that describes the background DNMR acquired from population variation data [24].

Characteristics of haploinsufficient genes

HI occurs whenever there is a missing or damaged copy of a gene leaving a single copy that is insufficient to maintain normal function [3]. HI is mostly caused by LoF mutations and results in dominant diseases. Recognition and prediction of genes that are haploinsufficient can facilitate the filtering of disease genome data wherever the phenotype is likely to have arisen through reduced levels of gene product.

In 2010, Huang *et al*. [3] proposed a deletion-based HI score by identifying differences between HI and haplosufficient genes, aiming to better distinguish pathogenic from benign deletions that help in variant prioritization. The analysis develops a logarithm-of-odds (LOD) score to estimate the probability of a deletion causing a HI phenotype. A high LOD score suggests deletions are likely to be deleterious through HI and therefore potential candidates for causing dominant traits. The score assumes there are no statistical interactions between the genes [3].Previously,and to try to assess the pathogenicity of a deletion, clinicians considered the length of a deletion or the number of genes deleted. The Huang *et al*. [3] score provides a rational basis to classify pathogenic deletions by comparing deletions seen in patients with deletions in controls and calculating the fraction of controls with a deletion at least as deleterious as that seen in the patient.

Distinguishing false-positive disease variants from the genuinely causal variants is crucial for accurate molecular diagnoses. MacArthur *et al*. [25] developed the RECessive (REC)

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/ely033/5126507 by University of Southampton user on 30 October 2018

score for distinguishing genes involved in recessive diseases from genes that are LoF-variation tolerant. A ‘healthy’ genome might contain 100 true LoF variants, the majority in a heterozygous state. Evidence suggests that the average human carries five recessive lethal alleles in single copy in their genome. Consequently, the majority of LoF variants are considered common variants.However,these variants might still have a phenotypic effect [25]. MacArthur *et al*. [25] demonstrated differences in functional and evolutionary features between recessive disease and LoF-tolerant genes, allowing for the development of a predictive model to predict recessive disease variants.

Khurana *et al*. [26] developed the ‘gene position in NETworks’ (NET) indispensability score to investigate relationships between degree of network centrality of a gene and selection within biological networks. They consider a range of biological networks relating to phosphorylation, signaling, protein–protein interaction and regulatory and genetic networks. Genes that are highly connected to many biological networks are the most functionally significant; therefore, mutations in those genes might have serious consequences [26]. However, genes connected to metabolic networks were found to have an excess of duplicated copies through more paralogs with LoF mutations[26].This score was included as a predictor of haploinsufficient genes in the Hsu *et al*. study [5].

Ge *et al*. [27] consider gene-specific pathogenicity using the ratio of non-synonymous to synonymous substitution rates (dN/dS) for X-chromosome genes. Genes with unusually low ratios suggest intolerance to non-synonymous variation, indicating they may be susceptible to disease-related variation. The authors found correlation between genomic regions depleted for missense variation with disease-causal variants [27].

Steinberg *et al*. proposed that study biases existing in many biological networks might affect the ability of previous HI prediction scores to recognize the genuinely haploinsufficient genes. For that reason they constructed a new, unbiased, HI score, the Genome-wide HaploInsufficiency Score (GHIS), which replaces biological networks with co-expression networks [28, 29]. They compared their model with the three preexisting methods (i.e. HI [3], NET [26] and RVIS [2]) and demonstrated that GHIS provides a score for many genes not scored by other methods [28] with enhanced performance at classifying less well-studied genes [28].

Scores have been developed to recognize Mendelian genes with different modes of inheritance. Hsu *et al*. considered Mendelian disease gene characteristics according to their mode of inheritance. HI is an essential characteristic of Mendelian disease genes with an autosomal dominant (AD) mode of inheritance and sensitivity to DNMs was recognized for this group of genes [5]. In contrast, disease genes with autosomal recessive (AR) modes of inheritance tend to have more nonsynonymous variants and regulatory transcript isoforms

[5]. However, the X-linked (XL) pattern of inheritance is associated with fewer non-synonymous and synonymous variants [5]. Based on these findings they create a new approach to prioritize Mendelian disease genes based on their mode of inheritance (AD,AR and XL) termed Inheritance-mode Specific Pathogenicity Prioritization (ISPP) [5]. This score integrates preexisting genespecific prediction methods namely, HI [3], REC [25], RVIS [2], NET [26], DNE [17] and GDI [30] along with numerous genetic properties including global expression from RNA-Seq data, DNA replication time and the noncoding (intronic region) mutation rate [5].

Because the human genome contains an abundance of non-deleterious heterozygous variants, the identification of dominant mutations for monogenic disorders is challenging. Quinodoz *et al*. [31] created DOMINO, a method using machine learning to identify whether a given gene is liable to carry dominant changes.

Inevitably, well-studied genes are over represented in most biological networks used to create scores that predict HI compared to less-studied genes, hence most biological networks are affected by study bias. Therefore, the creation of unbiased HI score becomes particularly important [29]. Recently, Shihab *et al.* produced an integrated machine learning approach called HIPred, merging functional annotations with genomic and evolutionary features to predict HI genes without study bias using data from National Institute of Health (NIH) Roadmap Epigenomics [32] and the Encyclopedia of DNA Elements (ENCODE) [33] project. The performance of this approach is considered to exceed the preexisting HI predictors [29]. Supplementary Table 3 outlines the key approaches in this category.

Characteristics of genes under selection

Genetic variants may be subject to positive selection whereby, if they are advantageous, they may increase in frequency. Negative selection, in contrast, acts to remove deleterious alleles. Scores that quantify the intensity of negative selection acting on genes provide insights into which genes are more likely to have variation that may have damaging consequences. The pattern is complex because some essential genes are not known to have any associated disease variation and are perhaps subject to negative selection at particularly high intensity [34].

Bustamante *et al*. calculate the extent and directionality of Selection operating on a given gene,this score referred to here as ‘Sel’. They first compared fixed sequence differences, both synonymous and non-synonymous,between humans in the sample and chimpanzees over 11.81 Mb region of aligned coding DNA. The ratio of non-synonymous to synonymous differences (divergence) was 23.76%. In contrast, the ratio of non-synonymous to

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/ely033/5126507 by University of Southampton user on 30 October 2018

synonymous polymorphisms in the human subjects was 38.42%. This shows a significant excess of amino acid variation, relative to divergence, consistent with previous work stating that much amino acid variation in the human genome is slightly to moderately damaging [35].

Eilertson *et al*.[36] create a model to identify genes under natural selection with a non-parametric approach (with no assumption of a specific population genetic model) that is robust to demography. This approach, called Selection Inference using Poisson Random Effects, utilizes polymorphism and divergence data from synonymous and non-synonymous sites within genes.

The Gene-level Integrated Metric of negative Selection (GIMS) was created by combining two meta-analyses into a single meta analysis.The first meta-analysis combines comparative genomic metrics (GERP++) [22] and functional genomic metrics (Polyphen2) [37], and the second meta-analysis combines mutation rates (as SNPs/kb) and allele frequencies (as percentage rare) from the 1000 Genomes Project. Meta-analysis was achieved by combining those metrics into GIMS scores for 20,079 genes [38]. Because the majority of genes are under purifying selection, the aim was to quantify the degree of negative selection applied to genes. Conservation and functional scores were initially combined as 'functional genomic metrics' integrated with mutation rates and fraction of rare variants as 'population genetic metrics'. The GIMS score combines these two metrics and provides a unified score per gene. GIMS gives a probability distribution across the entire genome in quantiles. Genes under negative selection are scored low by GIMS [38].

The Gene Damage Index (GDI) is a gene-specific score that predicts the liability of a human protein-coding gene to contain disease-causing mutations considering the influences of selection and genetic drift.In GDI,Combined Annotation Dependent Depletion (CADD) [39] scores are used as the variant-level damage prediction method because this method is efficient at distinguishing between benign and deleterious variants and is strongly dependent on evolutionary conservation [30].Moreover, CADD scores can assess most types of variants while other methods, like Poly-Phen-2 [37] and SIFT [40], can only predict missense variants. To construct the GDI score the cumulative predicted damage in exonic regions of the gene is calculated using the CADD score for each allele compared to the expected score for variants with similar allele frequencies. The homogenized Phred I-score is calculated for each metric to indicate the ranking of the targeted gene relative to all other genes. A low Phred score indicates a human gene with a low GDI and high Phred score indicates a gene susceptible to contain damaging variation. Genes with high GDI tend to be under less intense purifying selective pressure. A low GDI score is associated with highly conserved genes (including genes enriched for ribosome, chemokine signaling proteasome

and spliceosome functions) reflecting essentiality. Such genes tend to be under stronger purifying selection than the median selective pressure acting on human genes [30]. Supplementary Table 4 outlines the key approaches in this category.

Discussion

Considering approaches that score genes according to essentiality and conservation,the DNE score offers some advantages.The main limitation of DNE is its validity only for interpretation of DNMs[5],but it considers more variables related to mutation rate going beyond sequence context compared to other methods like RVIS and Sel. These additional variables include consideration of sequence depth of coverage and regional divergence in genes between humans and macaques independently, which improve the predictive value of this model [17]. The DNE score has been compared to the RVIS and negative selection score Sel. The comparison showed that DNE and RVIS were equally effective emphasizing the benefits predicted from combining the two scores [17].

The strength of Samocha *et al*. model is enhanced by incorporation of the depth of coverage (i.e. how many sequence reads were present on average per base) and the regional divergence in genes between humans and macaques independently. These strengths play a significant role in the improvement of their predictive model. The number of rare synonymous variants in the ESP that comprises a relatively small sample of 6700 exomes [41] is shown to be highly correlated with the probability of a synonymous mutation determined by their model. As rare variant allele frequencies are impacted by sample size evaluation in larger databases such as ExAC would be of interest [41].

EvoTol was compared to the RVIS and the DNE scores and shown to have increased performance at classifying intolerant genes compared to RVIS.EvoTol was shown to be highly sensitive and more powerful to characterize genes with high pathogenicity [16]. Although there was no significant correlation between RVIS and EvoTol, the application of the two scores simultaneously will likely be advantageous [16].

Considering approaches for scoring genes for potential roles in HI phenotypes the HIPred approach has been evaluated against five predictors (HI Score, NET, RVIS, EvoTol and GHIS, Supplementary Tables 2 and 3). HIPred was found to outperform all in predicting HI genes [29]. Using different perspectives across the 26 disease-associated gene lists,Hsu *et al*.[5] estimate the power of several methods that predict gene pathogenicity showing a substantial positive correlation between HI and REC (correlation $r = 0.77$) while the six scores have a moderate relationship with each other ($r = 0.46$). Among these gene scores (DNE, GDI, HI, NET, RVIS and REC) the best predictor of disease predisposing genes was

the REC score [5]. The performance of the ISPP score was significantly superior for prioritizing AR and XL disease-associated genes [5].The REC score is effective at predicting disease-associated genes generally but less successful in discriminating recessive and dominant disease genes [5].

DNE measures the rate of per-gene DNM while RVIS ranks human genes based on the strength and consistency of the purifying selection acting against functional variation. Analysis has shown that GDI and RVIS capture unique sets of reciprocal information from population genetic data [30]. In essence, RVIS reflects selective pressure while DNE is based on DNMR estimates; both methods do not quantitatively estimate the mutational load for a gene in a healthy human population. For this reason, these methods are not optimal for filtering genes with high mutation rates and many residual false positives might be expected. GDI has proved to be the most efficient approach for filtering out false-positive variants in genes known to contain damaging variation [30].

The Ge *et al*. XL scoring system is not limited by previous gene annotation and the dN/dS ratio can be calculated for any protein-coding gene. This score applies to all X-chromosome protein-coding genes and therefore can assess genes for multiple disease phenotypes [27]. Because the intra human dN/dS ratio is not specific to the X-chromosome, the analysis of more genomic data using dN/dS ratio is recommended for future studies to identify genes that may have disease variation [27].

The effort to improve the predictive ability of variant level scores now includes combination of evidence from multiple pathogenicity scores and other data. An example is the 'Mendelian Clinically Applicable Pathogenicity' score [42] that uses machine learning classification based on existing pathogenicity scores and measures of evolutionary conservation. Such a combinatorial approach might usefully integrate evidence in both variant-level and gene-level metrics to improve predictive abilities overall [42].

This work aims to bring together the growing evidence that gene properties, alongside variant scoring systems, can play an important role in filtering disease sequence data. As healthy individuals can have genetic variants that lead to disruption of protein-coding genes (with no clinical phenotype) [25, 28, 29, 43], challenges remain to distinguish which LoF variants are associated with disease phenotypes from those that do not cause any functional disturbance [28]. Data from the 1000 Genomes Project show that on average a healthy person might carry 250–300 LoF SNVs [5].

The ACMG guidelines consider *in silico* predictions of whether a variant is involved in disease, but without specifying which or how many variant interpretation algorithms to use. These data can be used only as 'supporting' evidence for variant interpretation. There are difficulties with respect to validation of these methods, and there is a relatively high error

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/ely033/5126507 by University of Southampton user on 30 October 2018

rate with many pathogenic variants assessed as benign by some methods and many benign variants assessed as pathogenic [44]. The guidelines do not currently consider gene-specific metrics that are the subject of this review but presumably could similarly constitute supporting evidence given alongside stronger independent evidence suggesting role or lack of role in disease. Ultimately, functional validation is optimal although is frequently not timely, practical or reimbursable [44, 45].

The understanding of human genomes is advanced through the accumulation of sequence data in publicly available databases. The ExAC resource provides a potent filter to aid recognition of pathogenic variants in severe Mendelian diseases. Using ExAC for filtering to remove false-positive, but plausibly pathogenic,variantsdecreases the number of candidate proteinaltering variants by 7-fold compared to the smaller ESP database that has fewer exome sequences [23].

Coupled with the previous evidence, another study suggests that the missense Z score that represents genes rather than variants adds more information than variant-specific Polyphen2 and CADD classifications signifying that gene-level scores of constraints provide additional information for evaluating pathogenicity [23]. Furthermore, Huang *et al*. contend that variant-level scores (e.g. SIFT [40] and poly-phen 2 [37]) are limited by lacking the capability to determine, from crossspecies alignments, whether negative selection at a given site is acting in a recessive, additive or dominant mode [3].

The work proposed by Gussow *et al*. [21] was based on dividing the genes into subregions to identify exactly where the pathogenic mutations are likely to present. This study identified an important question: is the whole gene the correct unit by which to judge patterns of intolerance? Future analyses may consider refinements to gene-specific scores that consider within-gene regional patterns of intolerance in more detail.

Another controversial issue is the difficulty in interpretation of benign LoF variants for which the nomenclature is still not unified. It is important to realize that there are overlaps in the interpretation of LoF variants in healthy people.In the literature, all the following categories represent LoF variants in healthy individuals: true variants that do not seriously disrupt gene function, benign LoF variation in redundant genes and nondeleterious or less-deleterious variants that have an impact on risk of phenotype or disease [25].

Because each genic scoring approach considers only a specific property of genetic architecture, each individual score has limitations. For example (i) the REC score does not consider dominant disease-predisposing genes; (ii) Non-Copy Number Variation genetic variants were not included in HI prediction score; (iii) the NET score lacks the systematic comparison of different known disease-associated genes; (iv) the RVIS score does not consider variations in allele frequencies across different populations; (v) the DNE score has

limited applicability for testing DNMs; and (vi) the GDI score only considers mutation profiles [5]. Furthermore, a major limitation of the GHIS score is that the genetic background in individuals is not considered, which is an important issue since genetic variants do not act in isolation and disturbance of individual genes within a single biological pathway might affect the risk of a disease [28]. Accordingly, this analysis that provides a comprehensive review of each prediction scheme may help establish new routes for prioritizing disease-causal variants.

Many advances have been developed to assess whether a gene is tolerant or intolerant to common functional variation. Initially, scores were developed per gene then studies were published showing that dividing the gene into sub-regions might help in allocating the mutation accurately. At that time all scores that measure genic intolerance required disease knowledge,this limitation was addressed by developing a tool with no prior disease knowledge required, an essential step to better predict genic intolerance.

Reviewed here is a range of well-studied gene-specific predictors with various independent genetic properties. It is hoped that recognizing some of the limitations of each score and perhaps combining evidence in both variant-specific scores and gene-wise evidence might enable better prediction since there is currently no single method that is reliably predictive of gene pathogenicity. Therefore, this hopefully might help to overcome one of the main challenges of 100,000 Genomes Project that is variant annotation to prioritize important variants from harmless neutral variants. This review is intended to highlight existing work to identify and explain different gene-specific pathogenicity predictors, while pointing to the gaps in disease gene prioritization and annotation issues to facilitate new scores and better prioritization of disease-causal genes.

Key points

- A wide range of well-established models exists that prioritize genes based on their associated disease variation potential.
- Integration of these strategies to represent individual genes could have a significant impact on our understanding of genic properties and the recognition of disease-related functional variation.
- Evaluation and comparison of these individual scores and the development of integrated models to enhance NGS filtering strategies in disease genomes is a fertile area for future studies.

Supplementary Data

Supplementary data are available online at http://bib.oxford journals.org/.

Funding

D.A. is funded by the Saudi Arabia Cultural Bureau, London, UK.
D.B. is funded through a NIHR Research Professorship.

References

1. Chong JX, Buckingham KJ, Jhangiani SN, *et al.* The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 2015;**97**(2):199–215.
2. Petrovski S, Wang Q, Heinzen EL, *et al.* Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;**9**(8):e1003709.
3. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising andpredicting haploinsufficiency in the human genome. *PLoS Genet* 2010;**6**(10):1–11.
4. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as aknowledge representation for clinical questions. *AMIA Annu Symp Proc* 2006;359–63.
5. Hsu JS, Kwan JSH, Pan Z, *et al*. Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. *Bioinformatics* 2016;**32**(20):3065–71. https://doi.org/10.1093/bioinformatics/btw381.
6. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. *Engineering* 2007;**2**:1051.
7. Khan, SU, Niazi, M, Ahmad, R. Barriers in the selectionof offshore software development outsourcing vendors: an exploratory study using a systematic literature review. *Inf Softw Technol* 2011;**53**(7):693–706.
8. Jalali S, Wohlin C. Systematic literature studies: databasesearches vs.backward snowballing.In*Proceedings of the ACMIEEE international symposium on Empirical software engineering and measurement*, ACM; 29-38.
9. Badampudi D, Wohlin C, Petersen K. Experiences from usingsnowballing and database searches in systematic literature studies.In *Proceedings of the Nineteenth International Conference on Evaluation and Assessment in Software Engineering* ACM; 17.
10. Gehanno JF, Rollin L, Darmoni S. Is the coverage of GoogleScholar enough to be used alone for systematic reviews.*BMC Med Inform Decis Mak* 2013;**13**(1):7.
11. Becker S, Bryman A, Thomas H, *et al. Understanding research for social policy and practice: themes, methods and approaches*. Policy, 2012. Bristol,UK.
12. Craswell G, Poore M. *Writing for academic success*. SAGE, 2012. London,UK.
13. Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* 2014;**30**(9):418–26.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/ely033/5126507 by University of Southampton user on 30 October 2018

14. Pengelly RJ, Vergara-Lope A, Alyousfi D, *et al.* Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. *Brief Bioinform* June 2017;1–7.
15. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a geneintolerance score based on loss-of-function variants in 60 706 individuals. *Bioinformatics* 2017;**33**(4):471–4.
16. Rackham OJL, Shihab HA, Johnson MR, *et al.* EvoTol: a protein-sequence based evolutionary intolerance framework for disease-gene prioritization. *Nucleic Acids Research* 2014;**5**:43.
17. Samocha KE, Robinson EB, Sanders EJ, *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* 2015;**46**(9):944–50.
18. Allen AS, Berkovic SF, Cossette P, *et al.* De novo mutations in epileptic encephalopathies. *Nature* 2013;**501**(7466):217–21.
19. Bartha I, di Iulio J, Venter J, Telenti A. Human gene essentiality. *Nat Rev Genet* 2017;**19**:12.
20. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat Genet* 2016;**48**(4): 349–55.
21. Gussow AB, Petrovski S, Wang Q, *et al.* The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol* 2016;**17**(1):1–11.
22. Davydov EV,Goode DL,Sirota M,*et al.*Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**(12):e1001025.
23. Lek M, Karczewski KJ, Minikel EV, *et al.*, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2017;**536**(7616):285–91.
24. Jiang Y, Li Z, Liu Z, *et al.* MirDNMR: a gene-centered database of background de novo mutation rates in human. *Nucleic Acids Res* 2017;**45**(D1):D796–803.
25. MacArthur D, Balasubramanian S, Frankish A, *et al*. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012;**335**(6070):1–14.
26. Khurana E, Fu Y, Chen J, Gerstein M. Interpretationof genomic variants using a unified biological network approach. *PLoS Comput Biol* 2013;**9**(3):e1002886.
27. Ge X, Kwok PY, Shieh JTC. Prioritizing genes for X-linkeddiseases using population exome data. *Hum Mol Genet* 2015;**24**(3):599–608.
28. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. *Nucleic Acids Res* 2015;**43**(15):1–9.
29. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: anintegrative approach to predicting haploinsufficient genes. *Bioinformatics* 2017;**33**(12):1751–7.
30. Itan Y, Shang L, Boisson B, *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci USA* 2015;**112**(44):13615–20.
31. Quinodoz M, Royer-Bertrand B, Cisarova K, *et al.* REPORT DOMINO: using machine learning to predict genes associated with dominant disorders. *Am J Hum Genet* 2017;**101**(4):623–9.

32. Kundaje A, Meuleman W, Ernst J, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**
(7539):317–29.
33. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*
2012;**489**(7414)57–74. doi:10.1038/nature11247.
34. Spataro N, Rodríguez JA, Navarro A, Bosch E. Propertiesof human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum Mol Genet* 2017;**26**(3):489–500.
35. Bustamante CD, Fledel-Alon A, Williamson S, *et al.* Natural selection on protein-coding genes in the human genome. *Nature* 2005;**437**(7062):1153–7.
36. Eilertson KE, Booth JG, Bustamante CD. SnIPRE: selectioninference using a poisson random effects model. *PLoS*
Comput Biol 2012;**8**(12):e1002806.
37. Adzhubei IA,Schmidt S,Peshkin L,*et al*.A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4)248–9.
38. Sampson MG, Gillies CE, Ju W, *et al.* Gene-level integrated metric of negative selection (GIMS) prioritizes candidate genes for nephrotic syndrome. *PLoS One* 2013;**8**
(11):1–9.
39. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**(3):310–5.
40. Kumar P, Henikoff S, Ng PC. Predicting the effects of codingnon-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**(7):1073–82.
41. Auer PL, Reiner AP, Wang G, *et al.* Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI exome sequencing project. *Am J Hum Genet* 2016;**99**(4):791–801.
42. Jagadeesh KA, Wenger AM, Berger MJ, *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;
48(12):1581–6.
43. Ng PC, Levy S, Huang J, *et al.* Genetic variation in an individual human exome. *PLoS Genet* 2008;**4**(8):e1000160.
44. Bean LJH, Hegde MR. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Med* 2017;**9**(1):9–11.
45. Samson M, Porter N, Orekoya O, *et al.* Progestin and breast cancer risk: a systematic review. *Breast cancer research and treatment* 2016;**155**(1):3–12. doi:10.1007/s10549-015-3663-1.

Downloaded from https://academic.oup.com/bfg/advance-article-abstract/doi/10.1093/bfgp/ely033/5126507 by University of Southampton user on 30 October 2018

Briefings in Bioinformatics, 00(00), 2020, 1–8

doi: 10.1093/bib/bbaa029
Methods Review

Essentiality-specific pathogenicity prioritization gene score to improve filtering of disease sequence data

Dareen Alyousfi, Diana Baralle and Andrew Collins

Corresponding author: Dareen Alyousfi, Faculty of Medicine, Genetic Epidemiology and Bioinformatics Research Group, University of Southampton, Southampton SO17 1BJ, UK. Tel.: +44-7454557786; E-mail: dma1n16@soton.ac.uk

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa029/5809560 by guest on 19 March 2020

Abstract

The causal genetic variants underlying more than 50% of single gene (monogenic) disorders are yet to be discovered. Many patients with conditions likely to have a monogenic basis do not receive a confirmed molecular diagnosis which has potential impacts on clinical management. We have developed a gene-specific score, essentiality-specific pathogenicity prioritization (ESPP), to guide the recognition of genes likely to underlie monogenic disease variation to assist in filtering of genome sequence data. When a patient genome is sequenced, there are frequently several plausibly pathogenic variants identified in different genes. Recognition of the single gene most likely to include pathogenic variation can guide the identification of a causal variant. The ESPP score integrates gene-level scores which are broadly related to gene essentiality. Previous work towards the recognition of monogenic disease genes proposed a model with increasing gene essentiality from 'non-essential' to 'essential' genes (for which pathogenic variation may be incompatible with survival) with genes liable to contain disease variation positioned between these two extremes. We demonstrate that the ESPP score is useful for recognizing genes with high potential for pathogenic disease-related variation. Genes classed as essential have particularly high scores, as do genes recently recognized as strong candidates for developmental disorders. Through the integration of individual gene-specific scores, which have different properties and assumptions, we demonstrate the utility of an essentiality-based gene score to improve sequence genome filtering.

Key words: whole genome sequence; monogenic disease; gene-level metrics; disease genome; gene-specific score; gene essentiality

Dareen Alyousfi is Bachelor of Medicine and Bachelor of Surgery (MBBS) and MSc in genomic medicine, University of Southampton, and is undertaking a PhD in human genetics focusing on development and application of methods for resolving molecular diagnoses from patient sequence data.

Diana Baralle is a Professor of genomic medicine and an Honorary Consultant in clinical genetics, Faculty of Medicine, University of Southampton and a Programme Lead for the MSc in genomic medicine.
Andrew Collins is the Head of the Genetic Epidemiology and Bioinformatics Research Group at the University of Southampton and is involved in nextgeneration sequencing-based studies in population genetics and research into the genetic basis of a number of diseases. Submitted: 7 September 2019; Received (in revised form): 17 February 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

Monogenic diseases include those with a Mendelian pattern of inheritance in families and conditions arising in individuals through de novo pathogenic variants. To resolve the basis of these conditions at a molecular level, it is necessary to under- stand the disease phenotype in terms of the patient's genotype. Where this is achieved, diagnoses can be refined and potential routes for improved clinical management may become available. The Online Mendelian Inheritance in Man (OMIM) database [1] recognizes 3790 genes underlying 5470 Mendelian phenotypes. However, although ∼69% of all known Mendelian phenotypes have a resolved genetic cause, many more Mendelian conditions have yet to be characterized . A recent review, using data from 57 National Health Service (NHS) hospitals in the UK and 26 hospitals in other countries, found only a small fraction of patients with hereditary rare diseases receive a genetic diag- nosis [2]. Even for conditions in which the genetic aetiology is understood, the possibility of making a firm diagnosis may be reduced through the incomplete characterization of the patient phenotype or because incomplete genetic testing is restricted to a set of candidate genes which may not include the causal gene. In some cases, the molecular basis of the condition is determined for a patient only after as many as 16 clinic visits following an average of three misdiagnoses in a process which can last more than 2 years [2].

Recently, the 100 000 Genomes Project [3] was extended to focus on clinical care through a plan to sequence, over 5 years, the genomes of 5 million patients who have phenotypically described rare diseases and cancers.The plan involves the development of core NHS infrastructure, data sharing and clinical training [4]. For the success of this project and the advance of similar initiatives, it is critically important to further develop strategies which improve the interpretation of disease genome data so that causal variation can be distinguished from plausibly damaging, but in fact neutral, variation.

Spataro *et al.* [5] identified five discrete groups of genes which, when ordered by degree of gene essentiality, form the basis for the model proposed by Pengelly *et al.* [6]. The five gene groups described by Spataro *et al.* are: non-disease and non-essential (NDNE),complex non-Mendelian (CNM),complexMendelian (CM), Mendelian non-complex (MNC) and essential non-disease (END). Essential genes are defined as genes responsible for core biological functions in the organism and so are required for cell survival [7]. In the

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa029/5809560 by guest on 19 March 2020

hypothetical model, genes which may contain disease-causal variation occupy a position of intermediate essentiality between the NDNE gene group (genes considered tolerant to functional variation) and the END gene group (genes highly intolerant to functional variation). The latter comprises a gene set defined through mouse knock-out experiments excluding human disease genes that appear in the OMIM list and genes involved in common diseases identified by genome-wide association studies [5]. Essential gene candidates have also been recognized through experiments using technologies such as CRISPR-Cas9 [8]. Essential genes are critical for organism survival such that damaging variation is not tolerated and likely to be maintained only by a selection/mutation balance. These genes encode certain regulators of core cellular functions, and the disruption of these pathways may cause fatal disease [8]. Within the wider set of essential genes,Cacheiro *et al.* [9] distinguished between cellular lethal (CL) genes,which show nearly complete concordance with mouse lethal genes and are essential for both a cell and an organism to survive,and developmental lethal (DL) genes are not essential at a cellular level, but the loss of function (LoF) variation may be lethal at an organism level.

Essentiality-related scores for individual genes include measures such as the residual variation intolerance score (RVIS, [10]) and the probability of LoF Intolerance (pLI, [11]), which both quantify the tolerance of LoF variation. Other scores focus on the degree of conservation, such as the recessive (REC) score [12]; the position of genes in regulatory and other networks (for example the NET score [13]) or consider local sequence context such as the substitution intolerance score (SIS, [14]). Linkage disequilibrium (LD) is another factor related to gene essentiality as highly essential genes tend to have reduced haplotype diversity and therefore strong LD [6]. Groups of genes already known to have associated disease variation [5], include genes which may contribute only to complex disease variation (CNM genes), genes which contribute to both complex and monogenic variation (CM genes) and genes which contribute only to monogenic variation (MNC genes). Through the integration of different gene-level metrics, each of which is broadly related to gene essentiality, we develop a score that facilitates the recognition of genes most likely to include monogenic disease variation.

Materials and methods

The gene-specific scores used

We considered the 10 gene-specific scores, for which genespecific data were made available by the authors, as reviewed by Alyousfi *et al.* [15] as the baseline data set for constructing

and testing the essentiality-specific pathogenicity prioritization (ESPP) score. We also included gene-specific LD scores from LD maps in LD units (LDUs, [16]). These maps were constructed from data from more than 400 whole genome sequence samples from the Wellderly study [17]. Because of the close correlation between LDU and physical gene lengths, the LDU lengths of genes were corrected for physical length by regression to form the 'LDU_res-fit' scores used in the analysis. LDU scores represent the extent of LD per gene such that genes with high values have relatively reduced LD compared with genes with low values. For each of the 11 scores used, the numbers of genes obtained from the sources listed in Table 1 are given in Table 2. Further details on each score are provided in Table 1 and described in depth in Alyousfi *et al.*, [15] and the cited references.

Constructing a combined score

The R Studio statistics software [23], version 1.0.153—2009-2017 RStudio Inc. was used for the analysis. We aligned the list of 18 269 genes with essentiality-related scores from the studies described in Table 1. The aim was to construct an integrated score from individual gene-specific scores to guide the recognition of monogenic disease-causal genes.The majority of selected scores follow the earlier systematic review as scores broadly related to gene essentiality [HI, RVIS, pLI, SIS, NET, REC, DNE, gene damage index (GDI), gene-level integrated metric of negative selection (GIMS) and genome-wide haploinsufficiency score (GHIS)] [15]. Accepting a broad definition of essentiality, the review categorizes some scores (such as RVIS, DNE, SIS and pLI) as 'gene essentiality and conservation' related; others (including HI, NET and GHIS) as 'haploinsufficiency gene scores' and GIMS as a score 'measuring selection'. Based on the hypothetical model proposed by Pengelly *et al.*[6] and work by Bartha *et al.*[24] and Alyousfi *et al.* [15], we assume that haploinsufficient genes and genes strongly impacted by selection tend to be closer to the essential end of the spectrum.

Where the original papers provided Ensemble Gene IDs (GIMS [21], SIS [14] and GHIS [20]: Table 1), the corresponding official gene symbol was substituted (Supplementary_data_set) enabling matching across scores. To construct the ESPP score, we undertook principal component analysis (PCA) to reduce the dimensionality of the data. Data were standardized to mean zero and SD=1 when performing the PCA. The total variance explained by the first principal component for 11 scores was 0.36 (Supplementary Table 1). As the individual contribution of three scores (GDI [19], LDU [16] and REC [12]) was small (<0.2), we undertook a second PCA utilizing the eight remaining scores

Table 1. Description of 11 essentiality-related scores

Essentiality measures	Score magnitude[a]	Score characteristics	Score data	References
DNE: gene constraint score-*de novo* excess	I	Measures constraint for each gene using a mutation model quantifying the difference between observed and expected number of missense variants	Data on 18 860 genes in Supplementary Table 4 of Hsu *et al.* [18]	[13, 18]
GDI	D	Mutational damage by gene in a healthy population (genes susceptible to damage are less likely to underlie monogenic disease)	Data on 18 860 genes in Supplementary Table 4 of Hsu *et al.* [18]	[19, 18]
GHIS	I	Haploinsufficiency prediction using gene co-expression and genetic variation in large sequence data sets (using a support vector machine)	Data on 19 701 genes in Supplementary Table 3 of Steinberg *et al.* [20]	[20]
GIMS	D	Variants in genes under strong negative selection likely to be damaging (lower GIMS: stronger negative selection). Integrates genomic and population genetic metrics	Data for 20 080 genes in Table S1, of Sampson *et al.*, [21]	[21]
HI: deletion-based haploinsufficiency score	I	Haploinsufficient genes contrasted with haplosufficient genes from non-pathogenic copy-number variants. Combines biological properties (genomic, evolutionary, functional and network)	Data for 18 860 genes in Supplementary Table 4 of Hsu *et al.*, [18]	[22, 18]
LDU (LDU_Res-fit)	D	Gene-specific measure of LD (LD units) corrected, by linear regression, for gene size. Low LDU score: strong LD, genes which may be under increased selection	Data on 18 269 genes: Supplementary Data from Vergara-Lope *et al.* [16]	[16, 17]
NET: gene position in networks indispensability score	I	Quantifies gene centrality and indispensability in protein–protein interaction and regulatory networks to assess gene importance	Data for 18 860 genes in Supplementary Table 4 of Hsu *et al.* [18]	[13, 18]
pLI	I	Probability that a gene is intolerant to a LoF mutation: contrasts the observed number of rare variants per gene to the expected number under a selection neutral, sequence-context-based mutational model	Data for 18 226 genes in Supplementary Table 13 of Lek *et al.*, [11]	[11]
REC: recessive score	D	Linear discriminant model based on human–macaque conservation and adjacency to recessive disease genes in a protein–protein interaction network. Classifies genes into LoF tolerant and recessive classes	Data for 18 860 genes in Supplementary Table 4 of Hsu *et al.* [18]	[12, 18]
RVIS	D	Evaluates which genes have more, or less, common functional variation than expected, given their level of apparently neutral variation. Contrasts the number of common missense and truncating variants against all protein-coding variants in a gene	Data for 18 860 genes in Supplementary Table 4 of Hsu *et al.* [18]	[10, 18]
SIS	I	Quantifies the difference between expected and observed number of functional variants in a gene. Considers that the probability a nucleotide substitution occurs at a genomic site depends on the nucleotides flanking the site	Data available for 16 387 genes in Supplementary 3, Table 15 of Aggarwala *et al.*, [14]	[14]

[a]Decreasing (D) and increasing (I) is the direction of the score value such that scores with D have smaller magnitude with increasing essentiality and scores with I have higher magnitude with increasing essentiality.

(Supplementary Table 1) for which the total variance explained was 0.48. These eight scores were used in the computation of the ESPP score calculated as the weighted sum of each of the component scores, with the weights given in the second column of Supplementary Table 1 using the following equation: HI × 0.291+DNE × 0.357+RVIS ×−0.349+NET × 0.273+pLI × 0.353+GIMS × −0.397+GHIS × 0.361+SIS × 0.423. We computed Spearman correlations

(Supplementary Table 2) to measure the strength and direction of association between the eight component scores and the combined ESPP score.

Gene classification

We considered the distribution of ESPP scores within different gene groups. Spataro *et al.* [5] listed 17 982 genes in their Supplementary Table 2 within NDNE, CNM, CM, MNC and

Table 2. Numbers of genes with essentiality score assigned to each group (means of untransformed scores in brackets)

Essentiality measure	NDNE	CNM	MDG[a]	END	Genes with score but no gene group	Totals of genes assigned to groups
Gene totals (Spataro *et al.* [5] and OMIM [1] classification)	10 627	1732	4440	969	0	17 768
DNE	10 482 (0.621)	1730 (0.880)	3769 (1.025)	968 (1.651)	463 (0.564)	16 949
GDI	10 482 (192.264)	1730 (85.421)	3769 (124.199)	968 (189.183)	463 (2487.181)	16 949
GHIS	8971 (0.522)	1557 (0.527)	3448 (0.532)	938 (0.566)	0	14 914
GIMS	10 177 (0.525)	1722 (0.463)	3722 (0.433)	958 (0.322)	371 (0.507)	16 579
HI	10 482 (0.183)	1730 (0.262)	3769 (0.304)	968 (0.411)	463 (0.118)	16 949
LDU	10 627 (−0.008)	1732 (0.237)	3836 (−0.034)	969 (−0.238)	1104 (0.041)	17 164 16 949
NET	10 482 (0.447)	1730 (0.557)	3769 (0.639)	968 (0.733)	463 (0.334)	
pLI	9956 (0.253)	1687 (0.360)	3674 (0.318)	941 (0.579)	356 (0.262)	16 258
REC	10 482 (0.098)	1730 (0.147)	3769 (0.232)	968 (0.200)	463 (0.059)	16 949
RVIS	10 482 (0.091)	1730 (−0.051)	3769 (−0.188)	968 (−0.389)	463 (0.160)	16 949
SIS	9007 (−0.093)	1516 (0.077)	3174 (0.189)	805 (0.619)	0	14 502
ESPP (from eight scores—excluding GDI, LDU, REC)	7076 (0.620)	1330 (0.884)	2914 (1.003)	760 (1.641)	0	12 080

[a]MDG combining CM and MNC genes from [5] and the updated OMIM list. Groups comprise NDNE, CNM, MDG, END and genes with a score but no gene group (genes that have at least one score but were not categorized by Spataro *et al.* [5]).

END categories (Table 2). Because the CM group is relatively small and the focus of this analysis is the recognition of genes implicated in monogenic disease, we combined the genes in the CM and MNC gene groups into a single group: Mendelian disease genes (MDGs). This gene group was further revised to include additional genes implicated in monogenic conditions using an updated list from the OMIM database: https://www.omim.org/static/omim/data/genemap2.txt [1]. The combined set of genes known to be involved in monogenic disorders comprises 4440 genes. Around 300 new rare disease phenotypes are added to OMIM every year and, with the increasing use of genome sequencing, numerous new disease genes are recognized annually [1]. For the ESPP score, we considered the distribution of scores in relation to these four gene groups (Supplementary Table 3 and Figures 1 and 2).

We also considered the set of 82 genes identified by Cacheiro *et al.* [9] as strong candidates for developmental disorders. This is a sub-set of 163 genes classed as potentially

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa029/5809560 by guest on 19 March 2020

developmentally lethal in their Supplementary Table 7. These comprise genes that are highly intolerant to loss-of-function variation (pLI>0.9) [11] or having gnomAD's observed/expected LoF scores with upper bound <0.35 (https://gnomad.broadinstitute.org/) or with a high haploinsufficiency score (HI) [22] and not currently associated with human disease by OMIM [1],Orphanet [25] or the Developmental Disorder Genotype-Phenotype Database (DDG2P) [26]. The gene sub-set is genes known to have *de novo* variants in the 100 000 genomes undiagnosed cases with intellectual disability (47 genes), Deciphering Developmental Disorders [27] cases with variants of uncertain significance in undiagnosed children (44 genes) and 14 genes from the Centre for Mendelian Genomics. The latter Mendelian candidate genes include Tier 1 genes, which have mutations identified in multiple kindreds, fall within a linkage peak, or are associated with phenotypes recapitulated in a model organism, or Tier 2 genes which are strong candidates but mutations are only known in one kindred [9]. Accounting for overlaps, this is a set of 82 genes (Cacheiro *et al.*, [9]: their Figure 4).

Results

A total of 17,768 genes are classified into the four gene groups (Table 2). Considering each of the 11 gene-specific essentiality measures listed in Table 1, individual untransformed scores were available for between 14 502 and 17 164 of these genes depending on the score considered. The mean of the scores for each essentiality measure and gene group is given in Table 2, and considering differences in score direction (Table 1) shows trends mostly consistent with the model of increasing essentiality with the NDNE group appearing least essential and the END group the most essential.

The results of the PCA of 11 scores (Supplementary Table 1) show relatively minor weightings for GDI (0.013),LDU (0.055) and REC (0.187). PCA provides an orthogonal transformation of variables, which may be initially correlated and generates linearly uncorrelated variables; therefore,close correlations between the sub-sets of scores (Supplementary Table 2) reduce the independent contribution of some scores justifying the dimensionality reduction. We therefore excluded the GDI, LDU and REC scores and undertook PCA using the remaining eight variables (Supplementary Table 1), a model which explains a higher proportion of the variance. The ESPP score derived from a linear combination of first principal component weightings (Supplementary Table 1, the combined variance explained is 0.48) with the highest weighting applied to SIS (0.42, [14]) and the lowest to the NET score (0.27, [13], Supplementary Table 2) gives the results of the Spearman correlation analysis for the eight scores and combined ESPP. Correlations are

relatively high throughout and the correlation structure appears to be captured well by the combined ESPP score, which shows a higher correlation than any other scores with DNE, HI and pLI and high correlations with other variables.

Figures 1 and 2 show the breakdown of ESPP scores within a score range with respect to each gene group. There is wide overlap between groups indicating that the properties of

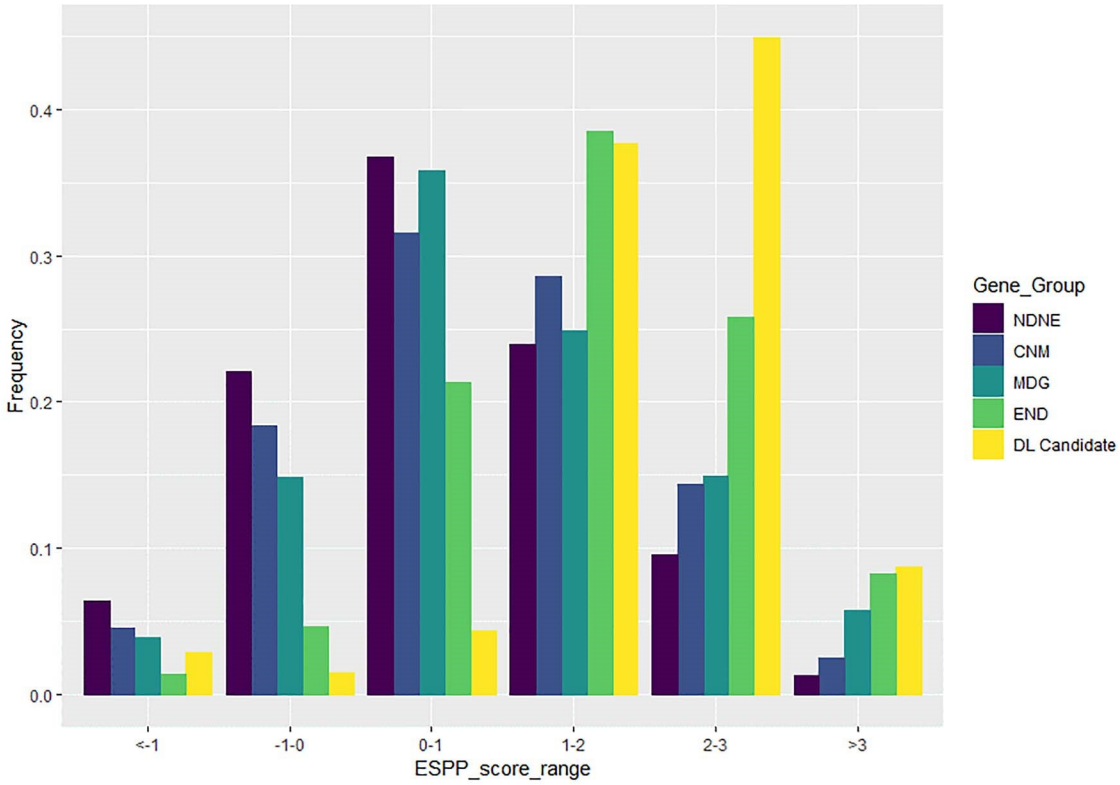


Figure 1. The percentage of genes in each group with ESPP scores from the least essential (NDNE, CNM, MDG to END) genes. Also, included are DL genes from Cacheiro *et al.* [9]. Most genes fall within the ESPP score 1–3 range and most genes with ESPP scores greater than 3 are essential genes.

genes explained by the scores cannot definitively allow genes to be placed into the categories. However, the gene groups are categorized according to current understanding; hence, unrecognized monogenic genes are mis-classified and there is incomplete understanding of human essential genes. Figure 1 shows the proportion of genes in each category within an ESPP score range. There is clear separation between the peaks for NDNE and END genes. Genes with large ESPP scores (exceeding 2) include an excess of END genes (34% of essential genes have ESPP > 2) and ESPP scores of 3 or greater are enriched for monogenic disease genes including 6% of MDG genes and 8% of END genes compared with NDNE and CNM gene groups (1.3 and 2.5%, respectively). In general (Supplementary Table 3), 63% of genes with ESPP > 3 are MDG/END and the proportion increases to 82% for ESPP > 4. High ESPP scores are strongly indicative of genes at the monogenic disease/essential end of the spectrum.

A total of 141 of the 163 DL genes identified by Cacheiro *et al.* [9] have ESPP scores and 70 of the 82 genes (a sub-set of the 141 DL genes) which are recognized as particularly strong candidates for developmental disorders have ESPP scores (Supplementary Table 3, identified as 'DL candidate' in Figure 1): 12 of these genes are in the CNM group, 22 in the END group, 1 in the MDG group and 35 are NDNE. The distribution of ESPP scores for these genes (Figure 1) is highly skewed towards ESPP > 2 in line with expectation that most are strong candidate genes for monogenic disease.

Discussion

The ability to sequence whole genomes is driving a transformation in medicine. However, firmly establishing a molecular diagnosis from genome sequences remains difficult in many cases. As many aspects of gene function are poorly understood and genes may have overlapping functions and a high degree of redundancy, the challenges remain even for highly penetrant monogenic diseases. While scores which are intended to predict the pathogenicity of individual DNA variants are widely used to help interpret genome variation, gene-specific measures are less frequently considered. Here, we integrate several quantitative scores which relate to human genes and broadly reflect the degree of gene essentiality. The scores span diverse properties and gene characteristics including degree of intolerance of genes to functional variation, the position of genes in gene interaction networks and the local sequence context of a gene. The ESPP integrated gene score is related to the gene essentiality framework proposed by Pengelly *et al.* [6] in which monogenic disease genes occupy a position of intermediate essentiality between non-essential and essential genes. Despite individual gene metrics covering a diversity of gene properties, the correlations between scores are relatively high (Supplementary Table 2). By integrating the available gene-level predictors, we establish a simple model to prioritize the recognition of monogenic disease genes. The combined ESPP score is intended to integrate all scores into a single model, which explains a higher proportion of the variance (0.45,Supplementary Table 1) than any individual score.

The PCA of the eight scores that contribute most to the predicted variance (Supplementary Table 1) shows that the NET score has the smallest weighting. This score was developed from genome-wide population genomic data and information on biological networks which preceded the completion of the 1000 Genomes Project [13]. In contrast, the SIS score, which has a high overall contribution, is more recent and used data from 1000 Genomes

Project. Improved understanding as larger numbers of genomes are sequenced is likely to greatly improve the recognition of gene properties relevant to monogenic disease in the coming years.

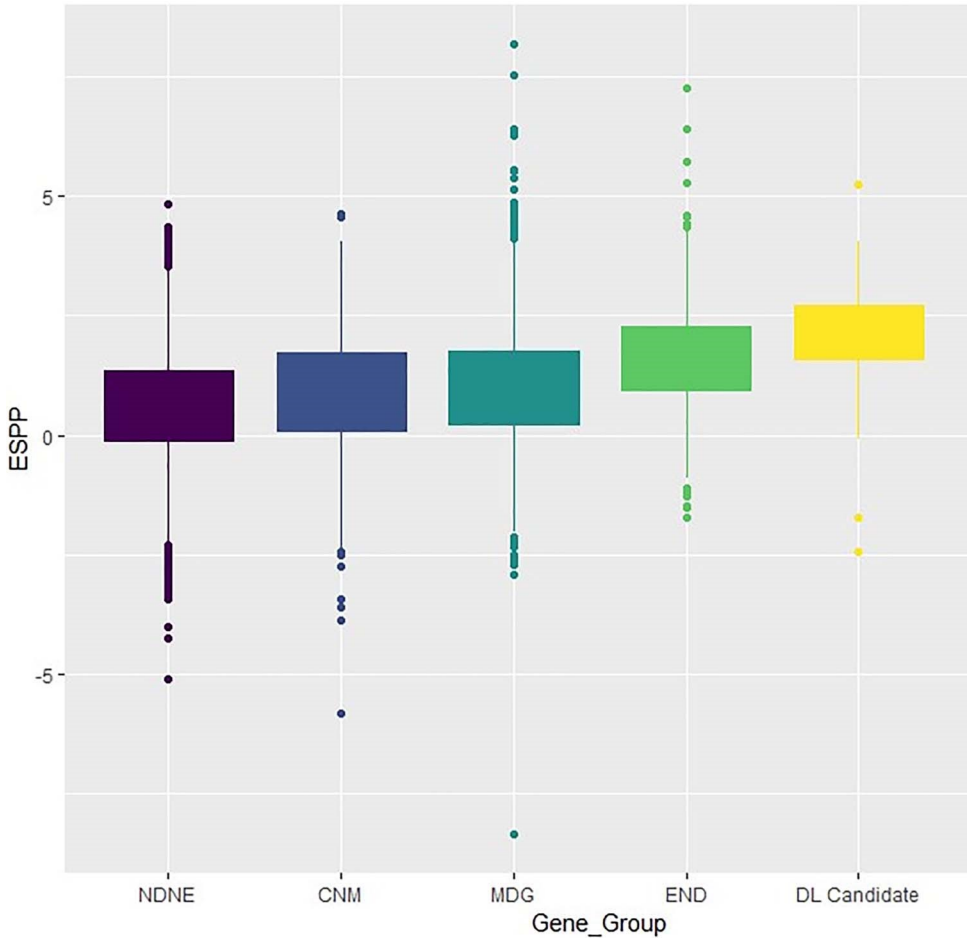


Figure 2. Boxplots representing the median of each group of genes starting from the least essential (NDNE, CNM, MDG to the most essential which is END) and DL from Cacheiro *et al.* [9] based on the ESPP score.

Along with the impact of variability in quality and completeness of individual gene-specific scores, a further difficulty in the interpretation of ESPP scores arises from the incomplete understanding of the gene groups. Recognition of new genes, which have not yet been assigned to the group of genes already known to be involved in monogenic disorders, is the rationale behind this study; so inevitably, genes existing in the gene group classification (Table 2) are expected to include mis-assignments. The expectation is that around 50% of genes involved in monogenic diseases have not yet been discovered and so

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa029/5809560 by guest on 19 March 2020

are currently assigned to gene groups other than MDG; however, with the current rate of advance,the discovery rate for monogenic disease genes seems likely to plateau soon. Another challenge is the difficulty in recognizing essential genes, given that the inactivation of an essential gene is fatal; therefore, the recognition of these genes in humans can only be achieved indirectly through homology or, more recently [28], through techniques such as CRISPR-cas9.

Cacheiro *et al.* [9] consider a sub-division of essential genes through their cross-species gene classification termed 'FUll Spectrum of Intolerance to Loss-of-function variation' (FUSIL). The classification recognizes two classes of essential genes after integrating human, mouse and CRISPR-Cas9 screening data: CL genes essential for a cell and an organism to survive and DL genes not essential at cellular level but where LoF is lethal at the organism level. FUSIL also includes distinct sets of sub-viable and viable genes determined from LoF mice experiments [9]. Their analysis supports the Pengelly *et al.* [6] model which is based on disease gene having intermediate essentiality. Their comprehensively characterized set of developmental candidate genes includes 91% which have ESPP > 1 (Figure 1).

Genes currently classed as NDNE, but with particularly high ESPP scores, are plausible monogenic disease candidates. Supplementary Table 3 shows that scores of ESPP > 3 are indicative of potential monogenic or essential genes.A total of 63%ofgenes with a score of at least 3 are currently classed as MDG or END. Furthermore, 82% of genes with ESPP > 4 are MDG/END. Table 3 shows 11 genes currently assigned to these two categories which have ESPP > 4. They include candidate essential genes currently classified as NDNE (for example *SUPT6H, FRY*) and genes which are known to contain CNM variation but have properties which suggest that they are also candidate monogenic disease genes (for example *RYR3, DIP2C*).

The complexity of disease-gene relationships and the diversity of gene properties limit the ability of individual and integrated scores to fully discriminate certain gene classes. 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

Table 3. Genes with ESPP score>4 not assigned to MDG or END groups

Gene	Group	ESPP score	Full name	Notes on gene function (OMIM)
ANKRD17	CNM	4.612	Ankyrin Repeat Domain 17	May mediate immune responses to bacteria and viruses
DIP2C	CNM	4.564	Disco Interacting Protein 2 Homologue C	May be involved in transcription factor binding
RYR3	CNM	4.039	Ryanodine Receptor 3	Involved in Ca(2+) signalling in neurons in the central nervous system

PLXNA1	NDNE	4.848	Plexin A1	Involved in cortico-motoneuronal connections underlying manual dexterity
CNOT1	NDNE	4.359	CCR4-NOT Transcription Complex Subunit 1	May be involved in transcriptional regulation
CHD5	NDNE	4.346	Cadherin 5	CDH5/beta-catenin signalling appears to control endothelial survival
USP34	NDNE	4.281	Ubiquitin Specific Peptidase 34	May rescue ubiquitinated proteins from proteasomal degradation
FRY	NDNE	4.200	FRY Microtubule Binding Protein	Involved in structural integrity of mitotic centrosomes and maintenance of spindle bipolarity
SUPT5H	NDNE	4.084	SPT5 Homologue, DSIF Elongation Factor Subunit	May control key aspects of neuronal development
PCDH17	NDNE	4.035	Protocadherin 17	May be involved in synaptic function in the central nervous system
SUPT6H	NDNE	4.003	SPT6 Homologue, Histone Chaperone And Transcription Elongation Factor	May regulate transcription through establishment or maintenance of chromatin structure

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 ESPP scores provides a basis for ranking candidates objectively.For example,genes with ESPP scores of two or greater appear particularly interesting in this context. To improve the performance of the model, an effort to integrate additional genomic and functional gene properties [11, 29] alongside improving gene classification given developing knowledge would be a worthwhile basis for future studies.

Key Points

- Integration of gene-specific scores that are related to gene essentiality helps the understanding of genic properties and the recognition of disease-related functional variation.
- The ESPP score is produced to enhance filtering strategies in sequenced disease genomes.
- The identification and characterization of essential genes is a fertile area for future studies to improve disease-gene discovery.
- An apparently pathogenic variant in a gene with a high ESPP score is a candidate disease variant worthy of follow-up. Genes with ESPP scores > 4 are likely to be essential or candidates for monogenic disease variation.

Supplementary Data

Supplementary data are available online at https://academi c.oup.com/bib.

Funding

This work was supported by the Saudi Arabia cultural bureau uk.

Declaration of conflict of interest

No potential conflict of interest was reported by the authors.

References

1. Online Mendelian Inheritance in Man, OMIM®. *McKusickNathans Institute of Genetic Medicine*. Baltimore, MD: Johns Hopkins University, 2019.
2. Ouwehand WH. Whole-genome sequencing of rare diseasepatients in a national healthcare system. *bioRxiv* 2019;1: 507244.
3. Genomics England, Queen Mary University of London, Dawson Hall. Charterhouse Sq., London EC1M6BQ.
4. Stark Z, Dolman L, Manolio TA, *et al.* Integrating genomics into healthcare: a global responsibility. *Am J Hum Genet* 2019;104(1):13–20.
5. SpataroN,RodríguezJA,NavarroA,*etal.*Propertiesofhuman disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum Mol Genet* 2017;26(3):489–500.
6. Pengelly RJ, Vergara-Lope A, Alyousfi D, *et al.* Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. *Brief Bioinform* 2017;20(1):267–73.
7. Zhang Z, Ren Q. Why are essential genes essential?-theessentiality of *Saccharomyces* genes. *Microb Cell* 2015;2(8):280.
8. Wang T, Birsoy K, Hughes NW, *et al.* Identification and characterizationof essential genes in the human genome.*Science* 2015;350(6264):1096–101.
9. Cacheiro P, Muñoz-Fuentes V, Murray SA, *et al.* Human and mouse essentiality screens as a resource for disease gene discovery. *Nature Communications* 2020;31;1–6.
10. Petrovski S, Wang Q, Heinzen EL, *et al.* Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 2013;9(8):e1003709.
11. Lek M, Karczewski KJ, Minikel EV, *et al.* Analysis of protein-coding genetic variation in 60706 humans. *Nature* 2016;536(7616):285.
12. MacArthur DG, Balasubramanian S, Frankish A, *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 2012;335(6070):823–8.
13. Khurana E, Fu Y, Chen J, *et al.* Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 2013;9(3):e1002886.

Downloaded from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbaa029/5809560 by guest on 19 March 2020

14. Aggarwala V, Voight BF. An expanded sequence contextmodel broadly explains variability in polymorphism levels across the human genome. *Nat Genet* 2016;48(4):349.
15. Alyousfi D, Baralle D, Collins A. Gene-specific metrics tofacilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review. *Brief Funct Genomics* 2018;18(1):23–9.
16. Vergara-Lope A, Ennis S, Vorechovsky I, *et al.* Heterogeneity in the extent of linkage disequilibrium among exonic, intronic, non-coding RNA and intergenic chromosome regions. *Eur J Hum Genet* 2019;3:1.
17. Erikson GA, Bodian DL, Rueda M, *et al.* Whole-genome sequencing of a healthy aging cohort. *Cell* 2016;165(4): 1002–11.
18. Hsu JS, Kwan JS, Pan Z, *et al.* Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. *Bioinformatics* 2016;32(20):3065–71.
19. Itan Y, Shang L, Boisson B, *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci* 2015;112(44): 13615–20.
20. Steinberg J, Honti F, Meader S, *et al.* Haploinsufficiency predictions without study bias. *Nucleic Acids Res* 2015; 43(15):e101.
21. Sampson MG, Gillies CE, Ju W, *et al.* Gene-level integrated metric of negative selection (GIMS) prioritizes candidate genes for nephrotic syndrome. *PLoS One* 2013;8(11): e81062.
22. Huang N, Lee I, Marcotte EM, *et al.* Characterising and predicting haploinsufficiency in the human genome.*PLoS Genet* 2010;6(10):e1001154.
23. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013.
24. Bartha I, di Iulio J, Venter JC, *et al.* Human gene essentiality. *Nat Rev Genet* 2018;19(1):51.
25. Orphanet: An Online Database of Rare Diseases and OrphanDrugs. Copyright, INSERM 1997. http://www.orpha.net (17 July 2019, date last accessed).
26. Firth HV, Richards SM, Bevan AP, *et al.* DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am J Hum Genet* 2009;84(4): 524–33.
27. Deciphering Developmental Disorders Study. Prevalenceand architecture of de novo mutations in developmental disorders. *Nature* 2017;542(7642):433.
28. Rousset F, Cui L, Siouve E, *et al.* Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLoS Genet* 2018;14(11): e1007749.
29. Collins A. The genomic and functional characteristics ofdisease genes. *Brief Bioinform* 2014;16(1):16–23.