Developing metrics for prioritisation of candidate disease genes using genetic variation databases

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

2021

Nikita Abramovs

Department of Computer Science

Contents

Word Count: 38,123

List of Figures

Appendix Figures

List of Tables

Appendix Tables

Abstract

Developing metrics for prioritisation of candidate disease genes using genetic variation databases

Nikita Abramovs

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering, 2021

Each human exome contains thousands of protein-altering variants located in more than 19,000 genes. Humans typically have two copies of a gene, and variants that affect one or both gene copies are called heterozygous and homozygous, respectively. If one gene copy is affected by deleterious heterozygous variation and cannot produce normal protein, this could result in a dominant disease. However, some genes can tolerate disruption of one copy, but deleterious homozygous or two heterozygous variants in different copies could still result in a recessive disease. Finally, humans can tolerate the inactivation or deletion of both copies of some genes without developing diseases. Because studied diseases' inheritance patterns are frequently known (e.g. if one of the parents and a child both have a disease, the inheritance pattern is likely to be dominant), clinical researchers want to know a candidate disease-causing variant inheritance pattern to prioritise candidate disease genes for laboratory validation. Although inheritance pattern is a property of disease causing variants, it can be predicted using gene-level properties. The aim of this study was to develop gene-level computational metrics that can be used for this task, and recently created large variant population databases such as Genome Aggregation Database (gnomAD, >137,000 individual exomes/genomes) provided novel data for such studies.

This thesis is written in the journal format and consists of three paper-style result chapters. In the first paper, we analysed deviations from Hardy-Weinberg Equilibrium of rare variants in gnomAD to detect potential disease-causing and heterozygous advantageous variants based on homozygous deficiency in the healthy populations. The second paper developed a gene variation intolerance ranking (GeVIR) system by measuring how unevenly variants in gnomAD were distributed in a gene relative to other genes. Finally, in the third paper, we developed multiple supervised machine learning models based on various gene properties (including GeVIR) and combined

them into a single continuous gene ranking metric that can be used to measure gene predisposition to disease inheritance patterns (DIP).

In conclusion, this thesis contributed to the understanding of variant population data and the application of supervised ML methods to classify candidate disease genes in the context of disease inheritance patterns. The primary outcome of this research was the development of two continuous gene metrics, GeVIR and DIP (available for 19,361 and 15,794 protein-coding genes, respectively), both of which can be used to distinguish dominant, recessive and non-disease genes. We anticipate that these metrics will aid clinical researchers in the prioritisation of candidate disease genes.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to express my gratitude to my supervisors, Dr. Mayada Tassabehji and Prof. Andrew Brass, for encouraging me to pursue a PhD, their guidance and patience, even at my lowest points. I am grateful to the Engineering and Physical Sciences Research Council (EPSRC) for funding this research [EP/N509565/1]. I would like to thank the Genome Aggregation Database (gnomAD) and the groups that provided exome and genome variant data to this resource. Without their data and practice of sharing the unpublished results, this research would not be possible. I wish to express my warmest gratitude to my parents, Marina and Andrey, for their love and unconditional support over these challenging years. Finally, I would like to say special a thank you to my friend Maksim, without whom this journey would be much more lonely.

Abbreviations

1000G	1000 Genomes Project
AB	Allele Balance
ACMG	American College of Medical Genetics and Genomics
AD	Autosomal Dominant
AF	Allele Frequency
AFR	African/African American
AMR	Latino
AR	Autosomal Recessive
ASD	Autism Spectrum Disorder
ASJ	Ashkenazi Jewish
AUC	Area Under the Curve
CADD	Combined Annotation Dependent Depletion
CCR	Constraint Coding Region
CE	Cell Essential
CI	Confidence Interval
CNE	Cell Non-Essential
CV	Cross-validation
DIP	Disease Inheritance Patterns
dN/dS	Ratio of non-synonymous and synonymous substitution rates
EAS	East Asian
EE	Epileptic Encephalopathy
ESP	Exome Sequencing Project
EUR	Europeans
ExAC	Exome Aggregation Consortium
FE	Fold Enrichment
FIN	Finnish
FIS	Functional indispensability score

FN	False Negatives
FP	False Positives
GBM	Gradient Boosting Machine
GDI	Gene Damage Index
GDIT	Gene Discovery Informatics Toolkit
GDP	Gene Dominance Prediction
GERP	Genomic Evolutionary Rate Profiling
GeVIR	Gene Variation Intolerance Rank
GHIS	Genome-wide Haploinsufficiency Score
gnomAD	Genome Aggregation Database
GO	Gene Ontology
GPP	Gene Pathogenicity Prediction
GRP	Gene Recessive Prediction
HD	Human Disease
HetDef	Heterozygote Deficiency
HetExc	Heterozygote Excess
HetExc-	Not Heterozygote Excess
HGNC	HUGO Gene Nomenclature Committee
HI	Haploinsufficient
HS	Haplosufficient
HWE	Hardy-Weinberg Equilibrium
ID	Intellectual Disability
IMPC	International Mouse Phenotype Consortium
ISPP	Inheritance-mode Specific Pathogenicity Prioritization
KNN	K Nearest Neighbour
LDA	Linear Discriminant Analysis
LOEUF	Loss-of-function Observed/Expected Upper bound Fraction
LoF	Loss-of-function
LOFTEE	Loss-of-function Transcript Effect Estimator
LR	Logistic Regression
MAF	Minor Allele Frequency
MD	Mendelian Disease
MD	Mendelian disease
MGI	Mouse Genome Informatics
MHL	Mouse Heterozygous Lethal
ML	Machine Learning

MOEUF	Missense Observed/Expected Upper bound Fraction
NDE	Non-Disease Essential
NDNE	Non-Disease Non-Essential
NFE	Non-Finnish European
NGS	Next Generation Sequencing
NHLBI	US National Institutes of Health Heart, Lung and Blood Institute
OMIM	Online Mendelian Inheritance in Man
OTH	Other
p(HI)	Probability of being haploinsufficient
pLI	Probability of LoF Intolerance
pNull	Probability of being Null
PPI	Protein-Protein Interactions
pRec	Probability of being Recessive
RF	Random Forest
RFECV	Recursive Feature Elimination and Cross-Validation
RVIS	Residual Variation Intolerance Score
SAS	South Asian
Shet	Selective effects of heterozygous protein-truncating variants
SIFT	Sorting Intolerant From Tolerant
subRVIS	Sub-region Residual Variation Intolerance Score
SVM	Support Vector Machine
TP	True Positives
UNEECON	Unified inference of variant effects and gene constraints
VCNAB	Variant Carriers with “Normal” Allele Balance
VIR	Variant Intolerant Region
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
XL	X-Linked

Chapter 1

Introduction and background

1.1 Introduction

The Discovery of Mendelian disease (MD) causing variants and consequently mapping MD phenotypes with genes enables researchers to understand the functions of the latter[1]. This knowledge is essential for developing testing, preventative and treatment methods for rare diseases[1]. Previously, this research was done using a positional cloning technique that required prior knowledge of gene location and function[1]. However, the development of the next-generation sequencing technologies (NGS) in 2010, which did not have these limitations, revolutionised the field and significantly sped up the disease gene discovery[1]. The average number of novel MD gene reports was ~168 and 261 based on five-year statistics before and after NGS development, respectively[1]. However, NGS brought new challenges[2] and based on various estimations the majority of the MD genes (6,000-13,000) are not discovered yet[1,3].

Each human exome contains about 149-182 loss-of-function (LoF) and 10,000-12,000 amino acid altering (missense) nucleotide differences (variants)[4] located in more than 19,000 protein coding genes[5]. To identify a new disease causing gene, researchers have to analyse variant data in a number of affected and control individuals, select genes that contain potentially deleterious variants, and then validate the effect of the variants on protein function by laboratory experiments, which are expensive and time consuming[6]. The larger the list of candidate variants, the more laboratory work is required, but measures have to be taken to avoid excluding the real disease causing variants at this stage. Therefore, computational methods which can prioritise genes and variants are crucial for success.

There are many tools that can predict the effect of variants on human proteins (e.g. Grimm *et al.*[7] compared ten such tools), some of which (e.g. Sorting Intolerant From Tolerant (SIFT)[8]) were developed more than a decade ago. Although these tools show good results on benchmark datasets, their performance on research data is hard to estimate. For example, a recent study modelled all possible missense variants in the gene *TP53* in yeast, and found that 42% of the variants predicted to be deleterious by PolyPhen-2[9] were false positives[10]. Large apparently healthy population databases, such as the genome aggregation database (gnomAD[11]), are used to exclude variants that are too frequent to be disease causing[12]. However, more than half of the missense and LoF variants in the gnomAD database were observed only once in 141,456 individuals[12]. Therefore, even after applying these filters, the studied disease cohort could still contain hundreds of candidate variants that need investigating. To reduce the list further, researchers have to investigate variants at a gene level.

Humans typically have two copies of a gene[13]. Variants that affect one or both gene copies are called heterozygous and homozygous, respectively. If one copy is deleted or inactivated by a pathogenic heterozygous variant and another copy cannot produce enough protein for normal organism function, then a disease phenotype is developed[14]. These genes are called haploinsufficient (HI), whereas all other genes that can tolerate a inactivation or deletion of one copy are called haplosufficient (HS)[15]. However, a pathogenic heterozygous variant that does not result in gene inactivation or deletion might still cause a dominant disease by other molecular mechanisms[16]. For example, pathogenic variant can result in dominant negative effects, increased gene dosage or the production of abnormal proteins with new functions[16]. Recessive diseases develop when both gene copies are affected by one homozygous or two different (compound) pathogenic heterozygous variants[13]. Note that since different pathogenic variants in the same gene can have different effects, a single gene can be associated with both dominant and recessive diseases[17]. Finally, some genes are known to contain homozygous loss-of-function (LoF) variants in healthy individuals and are generally considered unlikely disease candidates (e.g. many olfactory genes fall into this category)[18].

Although diseases are caused by variants and inheritance is a property of a variant, not a gene, disease inheritance correlates with various gene-level properties (e.g. variation intolerance, protein-protein interactions), and, consequently, it could be analysed and predicted on a gene level[17]. From this perspective, we suggest that genes can be classified into three categories based on the number of copies required to be unaffected by pathogenic variants for normal organism function: both (dominant genes), one (recessive genes) and none (non-single-gene disease genes). However, since different pathogenic variants in the same gene can have different effects, a single gene can be associated with both dominant and recessive diseases[17]. Consequently, the first two categories (dominant and recessive) are not mutually exclusive. Therefore, by classifying genes as dominant, recessive and non-disease, in the context of this thesis, we only estimate the predisposition of variants in these genes to be disease-causing and their inheritance pattern based on gene-level properties. However, genes associated with diseases can have benign variants, so gene-level metrics have to be used in combination with variant effect predictors (that are out of the scope of this study) and other evidence (e.g. literature review, variant frequency in population databases) to prioritise candidate disease-causing variants[6]. Nevertheless, because studied disease inheritance patterns are frequently known (e.g. if one of the parents and a child both have a disease, the inheritance pattern is likely to be dominant), knowledge of gene predisposition to dominant or recessive inheritance patterns would be beneficial for candidate disease gene prioritisation.

Gene essentiality represents the severity of consequences of not maintaining the required number of functional gene copies on organism function[19]. Genes essential and non-essential for cell survival were identified in CRISPR/Cas studies by their inactivation in laboratory settings[20], but it is known that consequences in real organisms could be different[19]. Moreover, these experiments were typically performed by inactivation of both gene copies[21], consequently, subsets of heterozygous and homozygous essential genes from these experiments are unknown[11]. Pengelly *et al.*[22] suggested dividing the essentiality scale into three categories: non-disease essential (NDE, i.e. lethal), human disease (HD) and non-disease non-essential (NDNE). However, the severity of human diseases can vary dramatically, is difficult to measure systematically, and can include significant lifespan reduction[23]. For example, Dawes *et al.*[24] stringently curated known human disease genes in the Online Mendelian Inheritance in Man (OMIM) database and categorised ~11% of them as prenatal/infantile lethal. Finally, although some genes are homozygously inactivated in a generally healthy population, it is hard to confidently state that loss of this gene does not affect human health (i.e. categorise them as NDNE)[3]. Consequently, although the consequences of losing gene copies can be categorised (for example, as lethal, pathogenic, and viable), we argue that in the context of all genes, essentiality is probably more linear than categorical property. Nevertheless, estimation of gene essentiality can be used by clinical researchers to prioritise candidate disease genes since they know the severity of their patients' phenotypes.

A number of computational metrics were developed based on analyses of gene-level properties to estimate the severity of having pathogenic variants in a gene (i.e. gene essentiality) and their inheritance pattern[25]. The two common approaches were estimation of gene, or its regions, variation intolerance (also called constraint) using statistical methods based on data from large population variant databases[12,26–33] and developing machine learning models based on various gene properties[3,15,17,18,34–38] (often including variation intolerance metrics[3,17,35–37]). Machine learning allows researchers to analyse and interpret large, complex datasets and is widely used in genetics and genomics[39]. Previous studies predominantly used supervised machine learning methods to estimate gene predisposition to various disease inheritance mechanisms and essentiality[3,15,17,18,34–38], although semi-supervised and unsupervised methods were also tried[12,40]. In supervised learning, a model is created by training a machine learning algorithm on a subset of labelled examples (e.g. genes) to understand correlations between their features (e.g. variation intolerance, protein-protein interactions) and labels (e.g. association with dominant/recessive diseases)[17,39]. The model is then evaluated on another subset of labelled examples to estimate its performance[17,39]. Often, this is done using the cross-validation technique when the labelled dataset is divided into n equal chunks (usually ten), and the model is trained using all chunks except one (90%)[3,15,17,18,34–38]. The remaining chunk (10%) is used for testing, and the procedure is repeated until the model is tested on each chunk[3,15,17,18,34–38]. The variability caused by random partitioning can be reduced by using average results of numerous cross-validation procedure repetitions[15]. Finally, the model is used to predict labels of new examples (e.g. classify all unlabelled genes as dominant/recessive)[17,39]. Considering the existence of curated disease gene databases (e.g. OMIM[41]) that can be used for training, supervised machine learning seems to be an appropriate choice for such tasks.

Gene variation intolerance studies often produce a metric that ranks all studied genes (or their regions) based on the defined statistical formula. However, some studies also applied machine learning on top of it[12,28]. In contrast, metrics produced by machine learning methods reported a gene's probability of belonging to the studied groups (e.g. HI[15,35,37,38]). The efficiency of a gene metric's ability to prioritise novel disease genes is measured based on their performance on known disease genes. Gene variation intolerance metrics were often calculated without investigating the properties of known genes and, therefore, evaluated on all available genes from the studied groups, whereas supervised machine learning studies used subsets of known genes to train the models. Consequently, gene variation intolerance metrics can be used, to some degree, for prioritisation of various gene groups, whereas machine learning based metrics are specifically developed to prioritise some and often only one group of genes.

Although disease inheritance patterns and essentiality are two different concepts, in the context of computational metrics for disease gene prioritisation, these gene characteristics are often interdependent[19]. For example, gene probability of loss-of-function intolerance (pLI) was developed to predict novel haploinsufficient genes based on variant data in 60,706 individuals[12]. However, the set of genes predicted to be highly likely to be haploinsufficient (pLI> 0.9, n = 3,230) contained 90% and 50% of known

HI genes with severe and mild phenotypes, respectively[12]. Cassa *et al.*[28] developed an alternative method (S_{het}) to measure gene intolerance to loss-of-function (LoF) variants using the same data. Evaluation of their metrics showed that both dominant and recessive genes associated with deafness phenotypes were equally intolerant to LoF variants[28]. Therefore, Fuller *et al.*[42] argued that metrics such as pLI and S_{het}, actually represent strength of selection against heterozygous LoF variants that correlates with phenotype severity, but not haploinsufficiency. Most of the supervised machine learning models were developed to categorise genes by inheritance patterns and were often not evaluated in context of essentiality. However, predictions of a model developed by He *et al.*[3], that was trained to prioritise single-disease genes (regardless of inheritance patterns), also correlated with phenotype severity. Considering that some of the gene features used by these models are known to be correlating with essentiality (e.g. variation intolerance metrics), we suppose that predictions of other models might also, to some extent, be biased by disease phenotype severity. Schematically, we visualised this hypothesised correlation between computational gene-level metrics, genes' associated disease inheritance patterns, and degree of essentiality in Figure 1.1 (note that the sizes of elements that represented groups of genes are arbitrary and might not correlate with the actual number of genes in the groups).

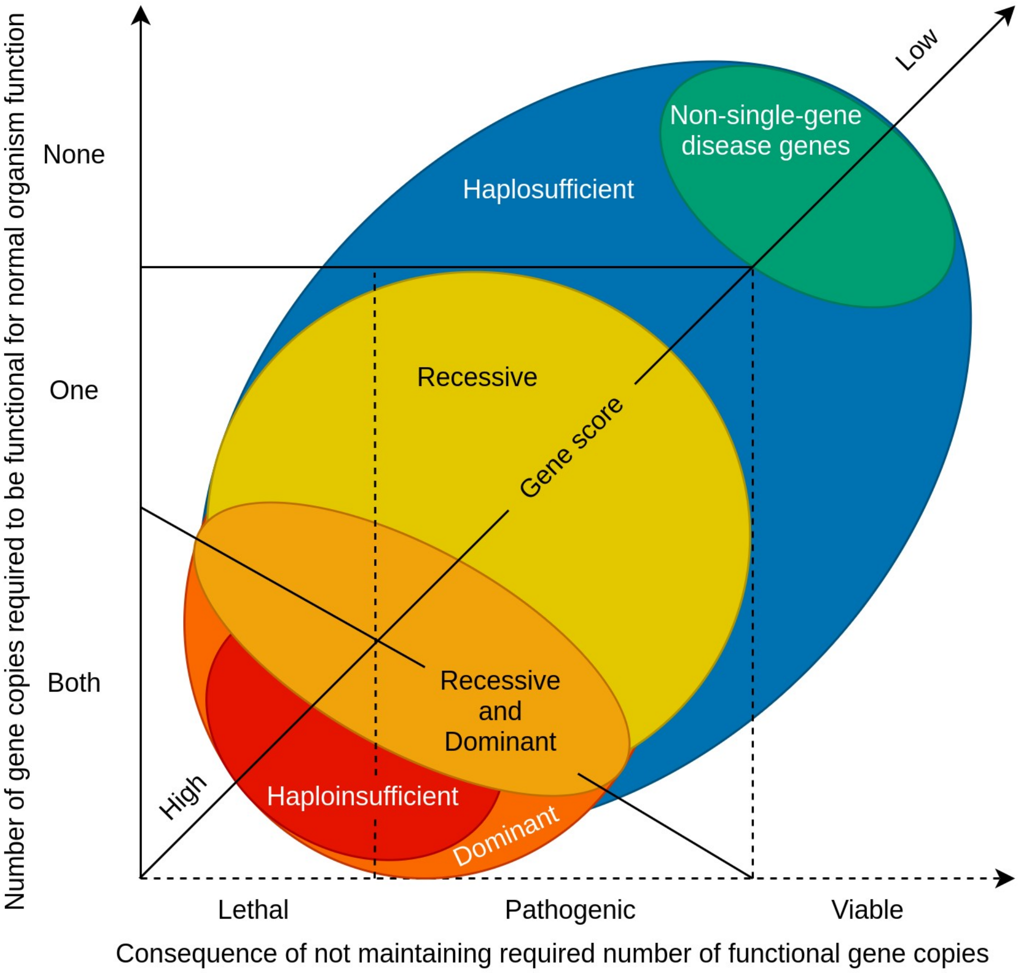

Number of gene copies required to be functional for normal organism function
None
One
Both
Lethal
Pathogenic
Viable
Consequence of not maintaining required number of functional gene copies
Haplosufficient
Non-single-gene disease genes
Recessive
Recessive and Dominant
Haploinsufficient
Dominant
Gene score
High
Low

Figure 1.1: Gene categorisation based on disease inheritance and severity (essentiality).

Illustration of a hypothetical correlation between computational gene scores (e.g. variation intolerance metrics), disease inheritance patterns, and disease severity. The gene group element sizes used are arbitrary.

The scope of this study is computational metrics that can be used for the prioritisation of candidate disease genes, specifically, methods of measurement of gene variation intolerance and classification by inheritance patterns using supervised machine learning models. Since the former are used as features in the latter, enhancement of the variation intolerance metrics can also improve supervised machine learning model performance. Both types of metrics are widely used by researchers working on human genome sequencing projects in a clinical setting and, therefore, have a high impact in the field of genome analysis and disease gene discovery (e.g. Huang *et al*. (2010)[15] 470+ citations, Petrovski *et al*. (2013)[26] 690+ citations according to Dimensions data in August 2021[43]). Next, we provide a brief introduction to Next Generation Sequencing (NGS) technology that led to the creation of variant population databases, and review previous variation intolerance and supervised gene classification methods to highlight under-researched areas studied in this thesis.

1.2 Next generation sequencing technology

Sequencing the first human genome required ~4 years, was completed in 2003, and by different estimates cost from a half to 1 billion dollars[44]. Nowadays, Next Generation Sequencing (NGS) technology allows coding regions (exomes) of the human genome to be sequenced within a day for less than 1000 dollars[45,46]. This is achieved by dividing DNA into millions of small pieces (sequencing reads), that are sequenced and mapped to the reference genome in parallel[45]. Comparisons with the reference genome allows the discovery of nucleotide differences (variants) in an analysed genome, whereas the balance between reference and alternative alleles allows the identification of variant zygosity[47]. However, due to the small size of sequenced reads, and the parallel nature of the process, they sometimes might be mapped to wrong parts of the reference genome that can result in incorrect variant calls, especially in repetitive and GC-rich regions[45]. To reduce these errors, each DNA region is sequenced multiple times and the number of stacked sequence reads at each DNA position is called "sequencing depth" or "coverage" of this position[45]. Higher read depth results in more accurate variant calls, but also increases the cost of the process[45]. Since whole genome sequencing (WGS) is expensive and investigation of some or all protein coding parts of the genome (exomes) is often sufficient for a required analysis, whole exome sequencing (WES; ~1% of the genome) is widely used[5,48]. However, it has now been shown that WGS can be more accurate than WES, even with lower read depth (mean coverage of 39 and 73 was investigated, respectively), especially in the detection of structural variants that affect large parts of the genome[49].

Human genome sequencing is an essential part in the study of genetic diseases as it allows the discovery of novel pathogenic variants[45]. The analysis normally requires comparison of WGS or WES of a group of affected and unaffected (i.e. control) individuals[50]. Candidate pathogenic variants are expected to be absent or significantly less frequently seen in controls[50]. These variants are further filtered by literature analysis and laboratory investigations[50].Variant datasets of individual genomes are available in databases such as dbGaP, but access to these data at an individual level is often restricted for the general public[51]. However, comparisons between affected and control genomes can be performed at a population level using aggregated, and consequently anonymous, WGS and WES datasets with calculated variant frequencies and heterozygous/homozygous numbers[11,12]. The larger the population, the more unique variants it contains, thus allowing the more precise comparison of affected (disease) and unaffected individual genomes to be performed[11,12]. Since the same control datasets can be reused in multiple disease studies, a large and publicly available dataset of healthy individuals was required to advance the field in the NGS era.

1.3 Variant population databases

One of the first such population level datasets was the US National Institutes of Health Heart, Lung and Blood Institute (NHLBI) Exome Sequencing Project (ESP) that consisted of 6,515 WES of European and African American control individuals from heart, lung and blood disorder studies[52(p515)] (Table 1.1 sumarises the main variant population databases). However, the ESP database could not be used for the analysis of non-coding variants and for filtration of benign variants specific to other populations. This limitation was addressed by the 1000 Genomes project that aimed to collect WGS data of healthy individuals from various ethnicities[4,53,54]. The final release of this dataset contained 2,504 individuals from 26 populations of European, African, American, South Asian, and East Asian ancestries[4]. Although the 1000 Genomes population was smaller than the ESP one, its size was sufficient to detect 99% of the known variants with allele frequency greater than 1%[4]. However, a larger dataset was still required to allow filtration of possibly benign rare variants. Thus, the Exome Aggregation Consortium (ExAC) project was launched in 2014 that aimed to aggregate control datasets from multiple studies[12]. The first release of the ExAC dataset consisted of 60,706 WES of individuals from various ethnicities and included both the 1000 Genomes and the majority of the cohorts used to create the ESP datasets[12]. Currently, the largest publicly available database is the genome aggregation database (gnomAD) v2.1.1, that is a continuation of the ExAC project, and consists of 125,748 WES and 15,708 WGS[11]. However, the largest WGS dataset is gnomAD v3.1.1 (76,156 individuals), that also contains a larger number of African American individuals than gnomAD v2.1.1 (20,744 and 8,128 individuals, respectively)[11,55]. Overall, the size of publicly available large population databases is rapidly increasing and, since sequencing costs are decreasing, WGS is becoming the preferred method for sequencing new human genomes in both a research and clinical setting.

Table 1.1: Summary of variant population databases.

Used abbreviations: Whole Exome Sequencing (WES), Whole Genome Sequencing (WGS), The US National Institutes of Health Heart, Lung and Blood Institute (NHLBI), African/African American (AFR), Latino (AMR), East Asian (EAS), Europeans (EUR), Finnish (FIN), Non-Finnish European (NFE), South Asian (SAS), Ashkenazi Jewish (ASJ), Other (OTH).

Name	Individuals	Major Populations	Reference
NHLBI Exome Sequencing Project (ESP5400)	5,363 (WES)	1,864 (AFR) and 3,499 (EUR).	Exome Variant Server (2011)[56]
1000 Genomes Project (1000G Phase 1)	1,092 (combination of WES with low coverage WGS)	185 (AFR), 242 (AMR), 286 (EAS), and 379 (EUR).	The 1000 Genomes Project Consortium (2012)[54]
NHLBI Exome Sequencing Project (ESP6500)	6,515 (WES)	2,217 (AFR), and 4,298 (EUR).	Fu *et al.* (2013)[52(p515)]
1000 Genomes Project (1000G Phase 3)	2,504 (combination of WES with low coverage WGS)	661 (AFR), 347 (AMR), 504 (EAS), 503 (EUR), 489 (SAS).	The 1000 Genomes Project Consortium (2015)[4]
Exome Aggregation Consortium (ExAC)	60,706 (WES)	5,203 (AFR), 5,789 (AMR), 4,327 (EAS), 3,307 (FIN), 33,370 (NFE), 8,256 (SAS), and 454 (OTH).	Lek *et al.* (2016)[12]

Genome Aggregation Database (gnomAD v2.1.1)	125,748 (WES), 15,708 (WGS)	WES: 8,128 (AFR), 17,296 (AMR), 5,040 (ASJ), 9,197 (EAS), 10,824 (FIN), 56,885 (NFE), 15,308 (SAS), 3,070 (OTH); WGS: 4,359 (AFR), 424 (AMR), 145 (ASJ), 780 (EAS), 1,738 (FIN), 7,718 (NFE), and 544 (OTH).	Francioli *et al.* (2018)[57] Karczewski *et al.* (2020)[11]
Genome Aggregation Database (gnomAD v3.1.1)	76,156 (WGS)	20,744 (AFR), 7,647 (AMR), 1,736 (ASJ), 2,604 (EAS), 5,316 (FIN), 34,029 (NFE), 2,419 (SAS), and 1,661 (OTH).	Francioli *et al.* (2019)[58]

Although large population variant databases are widely used as control datasets for NGS analysis, it is important to note that they are not free of disease causing variants, as highlighted by a few studies[59–61]. In fact, ExAC and gnomAD databases were created by combining both affected and control cohorts (only ~42.5% of individuals in gnomAD v2.1.1 were controls), but individuals with known severe paediatric disorders, as well as their parents, siblings and children, were not included in the dataset[11,12]. However, even apparently healthy control individuals might harbour disease causing variants for the following reasons. Firstly, they might carry recessive disease causing variants in a heterozygous state[59]. Secondly, young healthy individuals might carry variants that cause late onset diseases (e.g. Breast-ovarian cancer, familial 1(MIM:604370) and 2 (MIM:612555))[59]. Thirdly, some variants result in disease only in a proportion of carriers[62]. This phenomenon is called incomplete penetrance and is widely observed among cancer variants[62]. Finally, even highly penetrant disease causing variants might be tolerated due to some unknown compensatory mechanisms (e.g. other protective variants), although very rarely[63]. An analysis of 874 known disease genes in 589,306 genomes found only 13 individuals who somehow tolerated pathogenic variants for 8 diseases[63]. Therefore, unless a variant is expected to cause a severe early onset disease, it might be present in large population databases, but its allele frequency is expected to be very low[59]. Whiffin *et al.*[64] attempted to estimate maximum allele frequency from which variants could be filtered as non-disease-causing in the ExAC datasets. However, their statistical framework required statistics of disease penetrance that can be hard to precisely estimate for all genes, and is not available for many diseases[64]. According to the American College of Medical Genetics (ACMG) standards and guidelines, all variants with allele frequency greater than 5% in large population databases should be classified as benign, whereas a frequency that is greater than expected is one out of two required factors for such classification[65]. Nevertheless, data from variant population databases is widely used for calculation of gene variation intolerance metrics.

1.4 Gene variation intolerance metrics

Traditionally, variation intolerant genes were identified using phylogenetic statistical methods that examined ratios of non-synonymous and synonymous substitution rates (d*N*/d*S*) in homologues genes of closely related species[66], for example humans and chimpanzees[67]. However, the creation of large population databases (Table 1.1) resulted in the development of methods that were based mostly on using human data, which outperformed phylogenetic approaches (Table 1.2)[27].

Table 1.2: Summary of methods used to measure variation intolerance of genes or their sub-regions.

Name	Data	Method summary	Reference
Residual variation intolerance score (RVIS)	ESP6500, ExAC, gnomAD v2.0	Studentized residuals calculated on ratios between numbers of common missense and loss-of-function variants (allele frequency (AF) > 0.001) and all variants (including synonymous) in genes.	Petrovski *et al.* (2013)[26]
Missense z-score	ESP6500, ExAC, gnomAD v2.1	Z-scores calculated on ratios between numbers of observed and expected (estimated based on codons mutability) rare (AF < 0.001) missense variants in genes.	Samocha *et al.* (2014)[27]
Gene damage index (GDI)	1000G (phase 1)	Sum of products of gene variants allele counts (AF < 0.5) multiplied by their combined annotation dependent depletion (CADD)[68] scores with each variant score normalised by median CADD score of variants with similar AF in all genes.	Itan *et al.* (2015)[31]
The probability of being loss-of-function (LoF) intolerant (pLI)	ExAC, gnomAD v2.1	Expectation-maximisation algorithm applied on observed (filtered with loss-of-function transcript effect estimator (LOFTEE)[11]) and expected (calculated with Samocha *et al.*[27] framework) rare (AF < 0.001) LoF variants to cluster genes into three groups (haploinsufficient (pLI), recessive (pRec), and LoF tolerant (pNull)).	Lek *et al.* (2016)[12]

Sub-region residual variation intolerance score (subRVIS)	ESP6500	Petrovski *et al.*[26] RVIS method was applied on functionally important gene regions (protein domains).	Gussow *et al.* (2016)[30]
Selective effects of heterozygous protein-truncating variants (S_{het})	ExAC	Bayesian approach was used for estimation loss-of-function intolerance based on cumulative allele frequency of rare (AF < 0.001) loss-of-function variants in genes.	Cassa *et al.* (2017)[28]
Regional missense constraint	ExAC	Samocha *et al.*[27] framework was used to divide genes into regions with different missense intolerance relatively to other regions in the same genes.	Samocha *et al.* (2017)[29]
Constraint coding regions (CCRs)	gnomAD v2.0	Genes were divided into regions completely free of missense and loss-of-function variants sorted based on their length measured in nucleotides.	Havrilla *et al.* (2018)[33]
Loss-of-function observed/expected upper bound fraction (LOEUF)	gnomAD v2.1	Upper bound fraction of 90% confidence interval calculated on observed/expected loss-of-function variant ratio calculated using LOFTEE and Samocha *et al.*[27] framework similarly to pLI study[12].	Karczewski *et al.* (2020)[11]

1.4.1 Functional variation intolerance metrics

The first two methods were developed by Petrovski *et al.*[26] and Samocha *et al.*[27] and were based on “orthogonal” ideas, as the studies hypothesised that genes with fewer *common* and *rare* functional variants should be more intolerant to variation, respectively. Both studies analysed variant data from the ESP database and used 0.001 minor allele frequency (MAF) as a threshold to separate *rare* and *common* variants[26,27]. However, the studies used different methods for statistical calculations, variant categorisation, and normalisation for gene length[26,27].

Petrovski *et al.*[26] grouped common missense and LoF variants together in each gene, regressed them on the number of all protein coding variants (including synonymous) to normalise for gene length and used studentized residuals as a metric named Residual

Variation Intolerance Score (RVIS). Samocha *et al*[27] analysed rare synonymous, missense and LoF variants separately in each gene, compared the number of observed variants with estimated expected number of variants of each type to normalise for gene length, and used z-score statistics as a metric. The number of expected variants in each gene was estimated as the sum of all sequence codons probabilities to mutate to other codons grouped by variation types (e.g. missense)[27]. Codon mutation rates were estimated by analysis of orthologous intergenic regions between humans and chimpanzees[27]. The expected variant numbers were also adjusted based on local coverage and divergence between human and macaques in each gene[27].

Although Samocha *et al.*[27] calculated z-scores for three large variant type groups (LoF, missense, and synonymous), most of the genes did not show significant difference between expected and observed number of synonymous and LoF variants. In the case of synonymous variants, this result was expected as they generally do not have an impact on proteins and, therefore, should not be under strong selection [27]. However, the lack of LoF intolerant genes was a result of insufficient sample size (ESP database contained 6,503 individual exomes), since LoF variants occur less frequently than synonymous or missense ones[27]. Consequently, the missense z-score metric was the main outcome of the study[27]. Comparison of missense z-scores and RVIS scores showed a similar performance in the context of prioritisation of known haploinsufficient genes and genes with *de novo* LoF variants in individuals with diagnosed autism spectrum disorders[27].

Different LoF variants in a protein are generally expected to have the same deleterious effect on it due to the nonsense-mediated decay mechanism, which targets mRNAs containing premature termination codons for degradation to avoid the production of abnormal proteins that could be damaging[18]. However, the consequences of missense variants may vary from being completely benign to “as damaging as LoF”[29]. Various tools were developed to predict pathogenic missense variants, most of which rely to varying degrees on evolutionary conservation that varies within proteins, in particular some regions that encode functional domains are known to be more conserved[30]. Therefore, several studies attempted to incorporate predictions provided by variant prioritisation methods into gene scores[31], or detect gene regions that were more intolerant to variation and, consequently, more likely to harbour pathogenic variants to complement these methods[29,30,33].

Itan *et al.*[31] hypothesised that genes with fewer damaging variants in a healthy population were more likely to be disease causing. They analysed missense and LoF variants together and developed gene damaging index (GDI) scores based on variant frequencies in the 1000 Genomes database and their deleteriousness estimated with the combined annotation dependent depletion (CADD) scores[68]. GDI outperformed RVIS and missense z-scores at prioritising known disease genes, which showed that incorporation of variant damage prediction scores into gene variation intolerance metrics calculations could improve the performance of the latter[31].

1.4.2 Loss-of-function variation intolerance metrics

Estimation of LoF intolerant genes became possible with the creation of the ExAC database, that contained 60,706 individual exomes[12]. The database was released with missense and synonymous z-scores calculated using the Samocha *et al.*[27] framework, and a novel metric developed to measure gene probability of LoF intolerance (pLI). Briefly, pLI scores were calculated as follows. First, Loss-Of-Function Transcript Effect Estimator (LOFTEE) was used to exclude low-confidence LoF variants that might be tolerated for various reasons (e.g. located close to the end of a transcript). Then, ratios of observed to expected numbers of rare (at an allele frequency AF < 0.001) LoF variants in each gene were calculated using the Samocha *et al* framework[27]. Finally, expectation-maximisation algorithm was used to cluster genes into three groups based on observed/expected ratios: haploinsufficient (pLI), recessive (pRec), and tolerant (pNull). The clustering was performed with the assumption that observed/expected ratios in haploinsufficient, recessive, and tolerant genes should be ≤0.1, ≤0.5 and ~1, respectively. Note that in the case of the haploinsufficient and recessive groups, the rate of expected deficiency of LoF variants was based on observed mean ratios in known disease gene lists (Clinical Genome dosage sensitvity map[69] and Blekhman *et al.*[70] studies, respectively). Although the analysis produced three metrics (pLI, pRec, and pNull), the authors stated that only pLI was "valuable" and suggested the use of a threshold of 0.9 to select extremely LoF intolerant and, therefore, likely haploinsufficient genes (3,230/18,225 genes analysed). This subset contained nearly all known severe haploinsufficient genes, but less than half of the dominant disease genes. Nevertheless, the majority of these genes (72%) were not linked with known disease phenotypes and could be of clinical interest.

Note that pLI was calculated by assigning equal weights to all variants with AF < 0.001[12]. Cassa *et al.*[28] suggested that since multiple rare variants were expected to have the same effect on the population as one relatively frequent variant, they should be weighted respectively. They developed an alternative score to measure heterozygous intolerance to LoF variation (S_{het}), which was based on cumulative frequency of LoF variants. However, since they also used the Samocha *et al.*[27] framework to estimate expected number of LoF that was designed for rare variants, they had to exclude genes with a high cumulative frequency of LoF variants. Consequently, they analysed a lower number of genes than in the pLI study (15,998 and 18,225, respectively), but reported a similar number of constraint (S_{het}>0.1) genes (2,984 and 3,230, respectively). They demonstrated in various assays that S_{het} gene scores could be effectively used to distinguish autosomal dominant and recessive genes, with up to 96% positive predictive value when a binary threshold of S_{het}>0.04 was used on a dataset of 504 clinical exomes. Although the authors developing the S_{het} were advised by the pLI authors, a direct comparison of these methods was not performed.

The ExAC database evolved into a larger database named gnomAD (60,706 and 141,456 individuals, respectively) that was released with novel variation intolerance scores, in addition to previously used missense z-scores and pLI[11]. These novel scores were calculated as 90% confidence intervals on expected/observed values calculated similarly to the previous scores[12]. The upper confidence interval was suggested to be used as an intolerance metric named LoF observed/expected upper bound fraction (LOEUF)[11]. The evaluation of LOEUF showed that: (i) genes with low scores were enriched with known haploinsufficient and cell essential genes (~5.5 and ~2.5 fold-enrichment in the first decile, respectively); (ii) genes with high scores were enriched with olfactory and cell non-essential genes (~4.8 and ~2.5 fold-enrichment in the last decile, respectively); (iii) genes with middle scores were to some extent enriched with known autosomal recessive genes (~1.7 fold-enrichment in the fifth and sixth deciles)[11].

Although the same method was used to calculate missense observed/expected upper bound fraction scores (i.e. MOEUF), missense metrics were not evaluated in the study[11]. Karczewski *et al.*[11] acknowledged some important limitations of their metric, especially that usage of the LOEUF metric alone for estimation of LoF intolerance of short genes (~30% of all analysed protein coding genes) could be misleading, since the gnomAD database was not large enough to confidently state that lack of LoF variants in these genes was not a random event (i.e. short genes even with zero observed LoF variants in gnomAD were classified as LoF *tolerant* by LOEUF). Nevertheless, the Karczewski *et al.*[11] study demonstrated that a single-value metric can be developed and used to prioritise not only haploinsufficient genes, but to some extent also to distinguish recessive and non-disease gene groups.

1.4.3 Regional variation intolerance metrics

Both RVIS and missense z-score frameworks were used to develop regional variation intolerance scores[29,30]. Gussow *et al.*[30] attempted to divide genes by exon and protein domain boundaries, and used the RVIS method to measure variation intolerance of the regions, but discovered that only division of the gene by protein domains was effective. Samocha *et al.*[29] attempted to use the missense z-score method to find new boundaries that would result in a significant deficiency of missense variants in some regions relative to others within the same gene. Scores developed by both studies only partially covered the human exome[29,30]. Domain information used to calculate regional RVIS was available for 41.5% of the coding sequence of 16,611 genes[30], whereas a statistically significant difference of missense z-scores in various parts of a gene was detected only in 2,700 genes[29]. Almost in all cases, the Samocha *et al.*[29] method resulted in genes being divided into two or three regions. Nevertheless, both studies reported that known pathogenic variants were significantly more frequently observed in detected variation intolerant regions, and showed that the developed methods would complement existing variant prediction scores[29,30].

Havrilla *et al.*[33] took the idea that functionally important gene regions should contain less variation than others to the extreme and hypothesised that the most constrained gene regions should not contain any variation in the general population. Briefly, they used missense and LoF variants in the gnomAD database as boundaries to divide genes into functional variant free regions and developed a map of constrained coding regions (CCR). The resulting regions were sorted by length, and they suggested that 95th or 99th percentile thresholds should be used to distinguish constrained regions. Although they analysed all regions in 17,639 genes, only 6,909 (39.2%) and 1,415 (8.0%) of them contained at least one region ranked above the 95th and 99th percentile, respectively. Similarly to the previous studies of regional intolerance[29,30], they showed that the known pathogenic variants were enriched in predicted constraint regions (CCR ≥ 95th percentile). Moreover, CCR surpassed existing variant effect prediction methods in an assay of prioritisation of *de novo* variants observed in individuals with neurodevelopmental disorders. They also suggested that the CCR map could be converted into gene scores for a limited number of genes (6,909) by ranking them based on numbers of constrained regions. Comparison of these gene-level scores with pLI, missense z-scores and RVIS showed that CCR prioritised different genes, but their performance in the context of prioritisation of known disease causing genes was not performed. Moreover, we note that the difference in performance between the metrics could be partially caused by bias towards longer genes, as they by definition were more likely to contain more constrained regions, and normalisation of scores according to gene length was not performed. Havrilla *et al.*[33] also acknowledged that the CCR map might be ineffective if genuine constrained regions were affected by sparse benign or pathogenic variants in the general population, in particular with recessive disease causing variants, that could be present in the general population in a heterozygous state. Nevertheless, their study showed that variation intolerance could be measured by using much smaller regions than in previous studies[29,30].

1.4.4 Summary and under-researched areas

We reviewed nine studies that used variant population data to develop genes or their sub-regions variation intolerance metrics[11,12,26–31,33]. The majority of the metrics were calculated based on rare (AF < 0.001) variant data[11,12,27–29], and none considered variant zygosity. Consequently, existing metrics primarily measured gene heterozygous intolerance (i.e. can be used to prioritise candidate HI/dominant genes), although recessive genes were also reported to be more intolerant than unknown or non-essential genes in multiple studies[11,26,31]. Variant population databases such as gnomAD consisted of predominately healthy individuals[11], so we hypothesise that the balance between heterozygous and homozygous individuals in the case of recessive disease-causing variants might be shifted towards excess heterozygous due to the presence of unaffected carriers (heterozygous) and absence of affected individuals (homozygous). Therefore, incorporating variant *heterozygous excess* status into variation intolerance metric calculations might improve its ability to prioritise recessive disease genes. Variants with *heterozygous excess* can be detected with Hardy-Weinberg equilibrium that previously was used to detect sequencing errors in variant population databases[71,72]. However, no attempt was made to use Hardy-Weinberg equilibrium to analyse genuine variants with *heterozygous excess* caused by natural selection and exclusion of individuals with severe childhood diseases in large population databases.

The methods for variation intolerance metrics calculations varied depending on the used variant types. Three metrics (pLI[12], S_{het}[28], and LOEUF[11]) measured gene intolerance to LoF variants. However, LoF variants are rare, and existing variant population databases are not large enough to confidently measure LoF intolerance of genes with short coding sequences[11]. Three metrics (RVIS[26], missense z-score[27], and GDI[31]) were calculated using missense or missense and LoF variants. Since missense variants are observed ~55-80 times more frequently than LoF in a typical individual[4], these metrics mainly depended on missense variant data. However, unlike LoF variants, most of which are expected to trigger nonsense-mediated decay mechanisms[18], the consequences of missense variants can vary dramatically depending on their location in a gene[29]. Therefore, three remaining metrics (subRVIS[30], regional missense constraint[29], and CCRs[33]) were developed to detect gene regions that were more intolerant to variation than others (e.g. located in functionally important domains[30]). However, little effort was made to aggregate information of regional intolerance on a gene level (e.g. CCRs were counted on a gene level but not normalised by gene length[33]). Therefore, incorporating variant distribution information (i.e. how unevenly variants are located in genes) into variation intolerance metric calculations might improve their performance if missense variant data is used.

1.5 Supervised machine learning models used for gene categorisation

The fundamental assumption of supervised machine learning (ML) is that a subset of data used to train a model represents the final data on which the model will be applied[39]. Previous studies used various gene subsets to train the models depending on the gene categorisation task (Table 1.3). However, all existing models were binary and eventually applied on all genes with enough known properties (features), used by models to make predictions, to detect a maximum number of candidate genes for the studied categories (i.e. all studies analysed most of the known protein-coding genes)[3,15,17,18,34–38]. Therefore, the final gene dataset could contain genes from categories that were not adequately represented or completely missing in the subsets used to train the models. Although the performance of supervised ML models could be affected due to misclassification of these genes, most studies ignored some gene categories. For this review, existing models were divided into three groups based on their approach to gene categorisation: haploinsufficient (HI)/haplosufficient (HS), pathogenic/non-pathogenic, and autosomal dominant (AD)/recessive (AR). Note that Hsu *et al.*[36] and He *et al.*[3] developed multiple independent models (Table 1.3).

Table 1.3: Summary of supervised machine learning methods developed for gene categorisation.

Used abbreviations: Protein-Protein Interactions (PPI), Haploinsufficient (HI), Haplosufficient (HS), Loss-of-function tolerant (LoF-tolerant), Autosomal Dominant (AD), Autosomal Recessive (AR), X-linked (XL), Mendelian Disease (MD), Cross-validation (CV), Area Under the Curve (AUC), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM), Gradient boosting machine (GBM), Random Forest (RF), Genomic Evolutionary Rate Profiling (GERP).

Name	Training, validation and testing genes	Classifier	Main features	AUC (CV*)	Reference
Probability of being haploinsufficient (p(HI))	287 HI, 679 HS	LDA	dN/dS between human and macaque; promoter conservation (GERP); embryonic expression; network proximity to HI genes.	0.83*	Huang *et al.* (2010)[15]

Probability of being recessive (p(rec))	213 LoF-tolerant, 858 AR	LDA	dN/dS between human and macaque; PPI network proximity to recessive disease genes.	0.81*	MacArthur *et al.* (2012)[18]
Functional indispensability score (FIS)	140 LoF-tolerant, 115 essential	LR	dN/dS between human and chimp; a number of networks in which gene is present; degree of centrality in various gene networks; average heterozygous AF of missense and synonymous variants (separately) in the 1000 Genomes pilot data populations.	0.91*	Khurana *et al.* (2013)[34]
Genome-wide haploinsufficiency score (GHIS)	297 HI, 297 HS	SVM	dN/dS between human and macaque; distance to HI genes (20 highest link weights were used) in co-expression networks; a ratio of the number of common (AF > 0.1%) and rare non-synonymous variants from ESP data; a ratio of gene expression in foetal to adult tissue.	0.67*	Steinberg *et al.* (2015)[35]
Inheritance-mode specific pathogenicity prioritization (ISPP) AD/AR/XL score	876 AD, 1500 AR, 182 XL and 100,000 random selected control sets of genes	RF	14 genomic, variation and functional gene properties; predictions from existing supervised machine learning models: p(HI), p(rec), FIS; existing variation intolerance gene scores: RVIS, GDI, missense z-scores.	0.75* (AD), 0.73* (AR), 0.85* (XL)	Hsu *et al.* (2016)[36]
DOMINO	291 AD, 694 AR (training, CV); 26 AD, 73 AR (validation)	LDA	Numbers of direct PPI with AD genes; variation intolerance scores (ExAC): p(Rec), missense z-scores, a ratio of splice LoF and synonymous variants; promoter conservation (PhyloP); mRNA half-life>10h.	0.91*, 0.92 (validation)	Quinodoz *et al.* (2017)[17]
HIPred	298 HI, 386 LoF-tolerant	GBM	dN/dS ratios between human and various species; cell-type-specific interactomes; variation intolerance scores (ExAC): missense z-scores, synonymous z-scores.	0.89*	Shihab *et al.* (2017)[37]

Episcore	287 HI, 574 HS	RF	Epigenomic data from various tissues.	0.88*	Han *et al.* (2018)[38]
Gene pathogenicity /dominance/ recessive prediction (GPP, GDP, GRP)	**GPP**: 630 (850) MD, 630 (850) LoF-tolerant (testing); **GDP**: 1,243 AD+ “AD,AR”, 1,584 AR; **GRP**: 1,985 AR+ “AD,AR”, 842 AD	RF	Custom gene variation intolerance scores; variant damage prediction scores (calculated based on aggregated numbers of variants predicted to be pathogenic in population databases (mostly ExAC) by various tools); protein-protein interactions in STRING. 201, 183 and 183 features were used to build GPP, GDP and GRP models, respectively.	0.87 (GPP), 0.81* (GDP), 0.81* (GRP)	He *et al.* (2019)[3]

1.5.1 Models that classify genes as haploinsufficient and haplosufficient

The HI/HS approach was used in the probability of being haploinsufficient (p(HI))[15], genome-wide haploinsufficiency score (GHIS)[35], HIPred[37], and Episcore[38] studies. The HI and HS training genes were selected based on the annotation in disease databases and deleterious variants in control populations, respectively [15,35,37,38]. Three out of four studies (p(HI), GHIS, Episcore) considered genes affected by disruptive heterozygous/homozygous structural variants in the control population to be HS[15,35,38]. However, Hsu *et al.*[36] argued that such a set of HS genes could not be considered representative, as it did not include profiles of genes unaffected by structural variants. Shihab *et al.*[37] used homozygous LoF variants instead of structural variants to define HS genes, but we argue that such a set also cannot be considered representative, since it does not include recessive disease genes that are also HS.

Moreover, Quinodoz *et al.*[17] insisted that models trained only on HI genes might be less effective in predicting genes associated with other types of dominant phenotypes (e.g. gain-of-function). None of the studies that used the HI/HS categorisation approach claimed that their models could distinguish between HI and other dominant phenotypes[15,35,37,38]. Moreover, in two studies (p(HI) and GHIS), AD genes were used as positive samples to test the models[15,35]. Therefore, we consider that the HI/HS gene categorisation approach is unsuitable for supervised ML since the HS set of genes by definition is extremely non-homogeneous and includes dominant, recessive and non-disease genes.

1.5.2 Models that classify genes as pathogenic and non-pathogenic

The pathogenic/non-pathogenic approach was used in the probability of being recessive (p(rec))[18], functional indispensability score (FIS)[34], and gene pathogenicity prediction (GPP)[3] studies. Non-pathogenic genes were selected based on homozygous LoF variants in control individuals, whereas pathogenic gene sets varied in different studies[3,18,34]. MacArthur *et al.*[18] used recessive disease genes (p(rec)), Khurana *et al.*[34] used essential genes in which homozygous LoF variants result in lethal or infertility phenotypes (FIS), He *et al.*[3] used Mendelian disease genes with various inheritance patterns (GPP). However, several subsequent studies highlighted that p(rec) and FIS scores were similar or more effective than p(HI) scores in their ability to prioritise HI and AD genes[35–37]. Therefore, although these studies aimed to prioritise different groups of genes (essential, recessive and all Mendelian disease genes, respectively), technically, all of these models were trained to distinguish non-pathogenic (LoF-tolerant) from some subset of pathogenic genes. From this perspective, GPP was trained on the most representative set of pathogenic genes. Moreover, although the GPP model performance was not compared with p(rec) and FIS models, considering the 6-7 year interval between these studies, it is likely to be the most effective due to the more up-to-date features and training data.

1.5.3 Models that classify genes as autosomal dominant and recessive

The AD/AR classification approach was used in Inheritance-mode specific pathogenicity prioritisation (ISPP)[36], DOMINO[17] and GPP[3] studies. Similarly to HI/HS metrics, DOMINO used a single ML model to distinguish dominant and recessive genes[17], whereas ISPP and GPP studies developed independent models to produce dominant and recessive scores[3,36]. These studies provide a potential solution for the prioritisation of at least one major group of genes (AD) and, therefore, will be reviewed separately.

Hsu *et al.*[36] proposed a novel approach to classify genes by inheritance modes instead of haploinsufficiency. They developed three models to predict AD, AR and X-linked (XL) disease genes. The procedure used to create ML models was not clearly described in their paper, but it included extensive feature tuning and usage of 100,000 random selected control sets of genes. However, the number of genes predicted (probability ≥ 0.5) to be AD (867), AR (1,479) and XL (179) in the whole gene dataset (18,859) was suspiciously similar to the number of genes used to train the models (876, 1,500, 182, respectively). Moreover, there were sudden drops in probability values right after 0.5 thresholds in all gene lists (e.g. AD genes ranked 867 and 868 had AD probabilities 0.51 and 0.34, respectively). Hsu *et al.*[36] did not provide specific lists of AD, AR and XL genes used to train the models, but stated that they used gene annotations from Clinical Genomic Database (CGD)[73]. Since then, this database was updated and contained 1,259 AD, 2,177 AR, and 230 XL genes on March 2021. Consequently, a thorough check of whether the genes used to train the models and those predicted were the same was impossible. In our study, we performed an analysis using a more up-to-date version of the CGD database, and most of the genes predicted to be AD, AR and XL by ISPP were known disease genes (94.0%, 94.1% and 96.6%, respectively). Therefore, ISPP models are probably extremely over-fitted and basically predicts only genes from the training dataset to be disease causing with probability ≥ 0.5.

Hsu *et al.*[36] computed AD, AR and XL gene ranks by sorting genes by predicted probabilities and suggested estimating gene inheritance based on the highest rank. To further validate the ISPP models performance, they demonstrated that ISPP AD and AR scores were statistically significantly (P ≤ 2.206E-5) different in several recessive gene sets (n = 128, n = 107, n = 206) not used in cross-validation. They also demonstrated the applicability of this approach on a small set of preselected novel disease genes (5 AD, 5 AR, and 13 XL). Although such evaluation did show that ISPP scores had some predictive power, we argue that it was not sufficient to understand their usefulness, especially considering that the performance of other metrics on these datasets was not shown.

Quinodoz *et al.*[17] developed a model (DOMINO) that can be used to prioritise AD dominant genes by training it on a manually curated subset of AD and AR genes. The training set consisted of genes associated exclusively with AD or AR non-cancer clinical phenotypes reported in two or more pedigrees. Genes associated with both AD and AR diseases (n = 78) were excluded from the training dataset. However, the evaluation of DOMINO scores on the final gene dataset (n = 17,998) showed that these genes were almost equally likely to be predicted as AD (55.1%) and AR (44.9%), which was considered to be evidence of the absence of an artifactual bias in the model. The evaluation also showed that almost all of the genes from a set of "well-known false positives for rare conditions in genome-wide screens"[74] (i.e. non-pathogenic) had less than 0.2 probability of being AD based on DOMINO model predictions[17]. However, we note that more than half of the genes from the final dataset (n = 10,198, 56.7%) were assigned AD probability less than 0.2[17]. We state that it would be reasonable to assume that, generally, non-pathogenic gene properties should have more extreme values than AR genes (e.g. more tolerant to LoF variation, have fewer network interactions). Therefore, AR genes should be predicted more likely to be AD than non-pathogenic genes. However, this might not be true for some of the features used in the DOMINO model.

For example, the second most informative feature (weight = 19.2%) used in the DOMINO model was one of the three interdependent ExAC LoF intolerance metrics (i.e. all three metric values sum to 1 for each gene), that represented gene probability of being recessive[12,17]. Quinodoz *et al.*[17] considered using the other two ExAC metrics that represented gene probability of being haploinsufficient and null (i.e. non-pathogenic), but they were filtered out by the feature selection procedure. The training dataset consisted of AD and AR genes with average ExAC probabilities of being recessive of 0.297 and 0.689, respectively. Therefore, the model was trained that genes with low metric values were more likely to be AD.

However, the final dataset also contained non-pathogenic genes that, similarly to AD genes, could have low ExAC probabilities of being recessive and high probabilities of being null[12], that were not considered by the model[17]. Moreover, three features of the

DOMINO model with a cumulative weight of 47.5% were based on the number of gene protein-protein interactions (PPI) with AD genes in various PPI sub-networks in the STRING database[75]. Although it is known that non-pathogenic genes have significantly less PPI than AR genes[18,34], the degree to which non-pathogenic genes had less PPI with AD genes than AR genes, was not assessed in this or previous studies. Therefore, although the DOMINO model could be used to prioritise candidate AD genes, its ability to distinguish AR and non-pathogenic genes was not evaluated and could be low since model features were selected for a different task.

He *et al.*[3] developed the gene dominance prediction (GDP) and gene recessive prediction (GRP) models that were supposed to be used in combination with the gene pathogenicity prediction (GPP) model to identify gene inheritance. All models were developed using the same ML algorithm (random forest), but with different feature sets. However, while the GPP model was trained and tested on different datasets, GDP and GRP models were developed using four-fold cross-validation called "cross-testing". The final models were trained on the best performing subset of the training genes (i.e. 75% of the training data). Disease genes used to train GDP and GRP models were not provided and, since the models were independent, these could be two different subsets. Considering that the training dataset consisted of most genes with well-known disease inheritance patterns (n = 2,827), this approach made GDP and GRP scores practically incomparable for future studies due to the lack of known disease genes that were not used to train the models and can be used for testing. Moreover, in the training of both models, genes associated with both inheritance types (AD, AR) were used as negative samples. We find this approach controversial for two reasons.

First, Quinodoz *et al.*[17] showed that genes associated with both inheritance types were almost equally likely to have properties of each class. Therefore, incorporating these genes into the negative training set might be similar to labelling some of the positive samples as negatives. Second, predictions of the independent models are harder to interpret since they do not sum to one in each gene. He *et al.*[3] suggested that GPP scores should be given higher priority than GDP and GRP. However, ~43% (2,116/4,942) of the genes predicted to be dominant were also predicted to be recessive (GPP $\geq$ 0.5 and

GDP ≥ 0.5 and GRP ≥ 0.5). Although the more likely mode of inheritance can be chosen by comparing GDP and GRP probabilities in each gene, the effectiveness of this approach was not evaluated in the original study[3]. For example, this might be suitable for genes with a high probability of being dominant or recessive, since only ~11% (117/1,091) of the genes predicted to be dominant with probability ≥ 0.8 were also predicted to be recessive with the same probability threshold. However, ~35% (553/1,559) of the genes that were predicted to be pathogenic (GPP ≥ 0.5) and had GDP ≥ 0.5 and < 0.6, also had GRP probabilities in the same range. Moreover, GDP and GRP model performance was not compared with previous solutions, although He *et al*.[3] reused DOMINO model features to solve the same gene classification problem (AD/AR). Therefore, the effectiveness of GDP and GRP models is unclear and hard to verify due to the lack of sufficient number of disease genes that were not used to train the models.

1.5.4 Machine learning algorithms used to build the models

The choice of machine learning algorithm depends on various factors of a task, including but not limited to the size of available training data, number of features, sensitivity to over-fitting, ability to handle sparse/missing/various types of data, and the importance of explainability of results[76]. The gene classification models discussed were built using various ML algorithms (Table 1.3)[3,15,17,18,34–38], but the most commonly used ones were linear discriminant analysis (LDA)[15,17,18] and random forest (RF)[3,36,38], each of which was used in three studies. In five studies (p(rec), FIS, GHIS, ISPP, Episcore), methodological reasons for selection of the used ML algorithms (LDA, logistic regression (LR), support vector machine (SVM), RF, RF, respectively) were not explained[18,34–36,38]. Three studies (p(HI), HIPred, DOMINO) highlighted that they selected ML algorithms (LDA, gradient boosting machines (GBM), LDA, respectively) due to their ability to estimate relative feature weights[15,17,37]. Shihab *et al.*[37] (GHIS) and He *et al.*[3] (GPP) stated that they selected GBM and RF, respectively, due to their ability to handle heterogeneous features and samples with missing feature values.

In five studies (p(rec), FIS, ISPP, DOMINO, GPP), the performance of alternative algorithms on the same feature data was not evaluated[3,17,18,34,36]. Huang *et al.*[15] (p(HI)) compared LDA and SVM algorithms performance. They observed similar performance, and selected LDA as a more straightforward solution. Steinberg *et al.*[35] (GHIS) compared linear and radical SVM kernels performance. Radical kernel performed significantly better on the training data (well-studied genes) and significantly worse on the testing data (candidate disease genes) than a more simplistic linear kernel. Therefore, the latter was used in the final GHIS model. Shihab *et al.*[37] investigated three approaches: 1) train GBM directly on all features; 2) use the various base kernel for different feature groups and then train SVM on a composite kernel (i.e. multiple kernel learning); 3) select the best algorithm for each feature group (complete list of considered ML classifiers was not specified, but it included naive Bayes, RF, and SVM) and then LR on their predictions (i.e. stacking). Parameters of individual classifiers in all cases were optimised via 10-fold cross-validation (CV). The first and the simplest approach (also used by all other studies[3,15,17,18,34–36,38]) performed the best. Han *et al.*[38] (Episcore) compared RF and two SVM algorithms, and selected the former due to its superior performance. Considering that the simplest solution was often the best, or the only one investigated/reported, we concluded that the development of supervised ML methods for gene classification was driven by improvements of feature and training data. Therefore, any ML algorithm that fits the feature data used and provides feature weights can be used in future studies.

1.5.5 Features used to build the models

In the early studies (p(HI), p(rec), FIS)[15,18,34], network-based features were the most influencial (p(HI), FIS) or, in the case of p(rec), one of two features used for gene classification. Various approaches were used to convert network interaction data into gene-level features. Huang *et al.*[15] and MacArthur *et al.*[18] measured proximity to known positive samples (HI and recessive genes, respectively). Khurana *et al.*[34] analysed gene interactions in various networks and used a degree of centrality in regulatory genetic and metabolic networks, and a total number of networks in which the gene was present, as features. Steinberg *et al.*[35] measured distance to positive samples (HI), but counted only 20 links with the highest weights. Quinodoz *et al.*[17] counted direct interactions with positive samples (AD) in three protein-protein interaction sub-networks (based on combined, experimental, and text-mining evidence of interaction) and also used upper limits (8, 3, 3, respectively).

Interestingly, in all studies where gene labels from the training data were used to construct features[15,17,18,35], only interactions with positive samples were analysed. Hsu *et al.*[36] model evaluation showed that p(rec) was more effective at prioritising HI genes than p(HI), although the former analysed interactions with recessive genes[18]. Therefore, analysing the difference between gene interactions with positive and negative samples (e.g. AD/HI and AR) is an unexplored and potentially fruitful area for further work.

Despite the informativeness of network-based features for gene classification, their usage was also criticised[35]. Steinberg *et al.*[35] showed that known disease genes were more studied based on the number of mentions in PubMed papers. Consequently, these genes were also represented in more gene networks and had many interactions in manually constructed networks. Their evaluation of previous models, that were heavily influenced by network-based features (p(HI) and FIS), highlighted that they were less effective for the prediction of less-studied genes (e.g. candidate disease genes with *de novo* variants in Autism cohorts). Steinberg *et al.*[35] suggested limiting usage of gene networks to those based on co-expression data that they found to be unbiased. Steinberg *et al.*[35] did not provide feature weight distribution in their model (GHIS), but in a Shihab *et al.*[37] model (HIPred), features based on variant data in a large population database (ExAC) cumulatively were ~2 times more influential than those based on cell-type specific interactomes.

Nevertheless, non-co-expression network-based features were still used in gene classification models (DOMINO[17], GPP[3]) developed after the Steinberg *et al.*[35] study. For example, in the DOMINO model, the cumulative weight of features based on protein-protein interaction data was 47.5%[17]. To demonstrate that their model can effectively predict less-studied genes, Quinodoz *et al.*[17] constructed a validation dataset consisting of disease genes discovered after the PPI network data used was released[75]. The model showed excellent performance on this dataset of 0.92 AUC. DOMINO also could effectively predict candidate disease genes with two or more protein-altering variants in different individuals in cohorts with intellectual disability and epilepsy. These genes were statistically significantly enriched in genes with a high probability (≥0.95) of being AD, with 18.9 and 43.1 enrichment scores respectively[17]. Therefore, it could be possible to develop unbiased models using various gene network data by limiting the influence of the number of interactions on feature values (e.g. usage of upper limits as in the Quinodoz *et al.*[17] study).

Features based on variant data in large population databases were not used or had minor significance in the early studies (p(HI), p(rec), FIS)[15,18,34] because population databases were not large enough at that time. However, in most models developed after the release of large population databases (e.g. ESP[56], ExAC[12]), and the development of self-sufficient gene variation intolerance metrics (e.g. RVIS[26], missense z-scores[27], ExAC p(LI)/p(Rec)[12]), this group of features became one of the most influential[3,17,35–37]. For example, in the DOMINO model, these features had a cumulative weight of 37.8% and were outperformed only by network-based features with a cumulative weight of 47.5%[17]. However, in the HIPred model, which used only co-expression network data, this group was the most influential with a cumulative weight of at least 45% based on the top ten features, and missense z-score was the most informative feature with a weight of 34% (the second-best feature weight was only 9%)[37]. Authors of supervised ML models also tried to develop custom features based on this data[3,17,35–37], but they were usually less effective than self-contained gene variation intolerance metrics published as independent studies. For example, the DOMINO model used the ratio of splice donor to synonymous variants (weight = 5.7%)[17] in addition to the ExAC p(Rec) metric (weight = 19.2%), that was based on splice acceptor/donor and stop gained variant data[12]. Hsu *et al.*[36] (ISPP) and Shihab *et al.*[37] (HIPred) also reported that a number of variants of various types (e.g. missense) were informative features, but we argue that these values must not be used as features since they, by design, are biased by gene coding sequence length. A notable exception could be the non-synonymous variation depletion score ("NoVaDs") developed by Steinberg *et al.*[35] as a feature for the GHIS model that outperformed the RVIS metric. NoVaDs was based on a similar principle as RVIS, calculated as a ratio between common non-synonymous and all variants (including synonymous) in each gene[26]. However, NoVaDs used the same numerator with a number of rare non-synonymous variants as a denominator and, therefore, was not affected by codon usage differences between genes[35]. Overall, the development of novel gene variation intolerance metrics can be a fruitful area for future research since they are free from study bias and can be used both by themselves and as features in ML models to prioritise candidate disease genes.

Features based on evolutionary data, such as ratio of non-synonymous and synonymous substitution rates (d*N*/d*S*) in homologues genes between humans and other species[67] (primarily primates) and promoter conservation scores (measured by average GERP[77]/PhyloP[78] scores), played a significant role in early studies (p(HI), p(rec))[15,18]. In the p(HI) model (based on four features, Table 1.3), the cumulative contribution of d*N*/d*S* and promoter conservation (GERP) features was similar to a network-based feature, whereas[15], in the p(rec) model (based on two features), the d*N*/d*S* feature contribution was higher than a network-based one[18]. The d*N*/d*S* features were used by most models (p(HI), p(rec), FIS, GHIS, ISPP (indirectly by using other model scores as features), HIPred, GPP)[3,15,18,34–37]. However, their weights in more recent models decreased[37], probably due to the usage of variant based features (e.g. missense z-scores) that were superior in prioritisation of disease genes[27]. In the HIPred model, two d*N*/d*S* features were among the top ten features, but their cumulative weight was more than three times lower than the missense z-score[37]. In the DOMINO model, which also used missense z-score as a feature, d*N*/d*S* features did not survive the feature selection procedure[17]. However, DOMINO, similarly to the p(HI) model, used promoter conservation (PhyloP) as a feature (weight = 11.4%)[17], possibly because features based on variant data did not represent constraint of non-coding gene regions. Although protein-coding evolutionary data was less informative than variant data for prioritising disease genes[27], these two sources could still complement each other in combined features[3]. For example, in the GPP model, 4 out of 10 top features were calculated based on damage prediction of variants in the ExAC database by various tools that used evolutionary conservation data[3].

Features based on network, variant, and evolutionary data had the largest contribution in most models (e.g. 96.7% in DOMINO[17]), but other data sources were also used to construct features. For example, the Episcore model was developed using only epigenomic data to construct features[38]. However, the Episcore model performance was not compared with any other reviewed model, all of which used features from various data sources. Moreover, the studies used different feature selection procedures (e.g. Huang *et al.*[15] aimed not to use features from the same data sources if possible) and, consequently, the number of features used could differ dramatically in various models. For example, GPP and GDP models used 201 and 183 features, respectively, to prioritise pathogenic and AD genes[3], whereas p(rec) and DOMINO models used only 2 and 8 features, respectively, for the same tasks[17,18]. Considering that some features could be biased, interpretability of model predictions is essential, and predictions of models developed with fewer features are easier to explain. Therefore, future research should primarily focus on improving the most influential features based on network, variant and evolutionary data.

1.5.6 Challenges in evaluation of model effectiveness

The evaluation and comparison of various ML models developed for gene classification is a difficult task for the following four reasons.

First, studies used different gene categorisation approaches (Table 1.3), and it is unclear to what extent it is fair to evaluate model performance on gene categories that they were not trained to predict. For example, Quinodoz *et al.*[17] proposed the novel AD/AR gene categorisation approach, and DOMINO model performance was not compared with previous models trained to predict HI genes (p(HI) and GHIS).

Second, models were trained on various gene sets (Table 1.3), and in some studies, genes used to train the final models were not reported (p(HI), GDP, GRP, ISPP)[3,15,36]. These genes have to be excluded from any fair comparison, since models can perform exceptionally well on their training data. For example, the Episcore model predicted 286 of 287 HI genes from the training dataset to be HI with probability > 0.6[38]. Moreover, we previously showed that ISPP models were probably extremely over-fitted to their training data. Therefore, fair and simultaneous comparison of multiple models is often impossible, or can be performed only on small sets of known disease genes unseen by all models.

Third, almost all models were evaluated using the cross-validation (CV) method (Table 1.3, note that GPP was trained and tested on different sets[3]), but these results must be interpreted with caution because training sets were often different, and these genes were also used to select features or tune model parameters. All studies used 10-fold CV[15,17,18,34–38], except He *et al.*[3] who used 4-fold CV to evaluate the GDP and GRP models. The analysis was repeated analysis multiple times (e.g. 30 times[15,18,37]) to reduce variability in results, and average Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) scores were reported[3,15,17,18,34–38]. Note that Quinodoz *et al.*[17] and He *et al.*[3] stated that they used CV to test their models (DOMINO, GDP, GRP). However, we argue that they misused ML terms as testing set by definition can only be used to assess the performance of the final model[79], whereas they also used this data for feature selection[3,17]. In studies where the comparison with other models was performed at this stage, the novel model always outperformed all competitors[36,37]. Nevertheless, since it was the most common evaluation technique used in all studies, we reported these results in Table 1.3.

Finally, the models were often tested on gene categories that were not used to train the models (e.g. p(HI) was developed to predict HI genes, but tested on AD and AR gene sets[15]) or on candidate disease genes that could be false positives[37]. The latter assessment was critical since many models relied on gene network data in which well-known disease genes were over-represented and could, consequently, be less effective at predicting novel disease genes[35]. Orthological mouse genes in which heterozygous knockout result in severe/lethal phenotypes, was often considered as candidate HI/AD disease genes (p(HI), GHIS, HIPred)[15,35,37], but it is important to note that human and mouse phenotypes might not be identical[80]. Moreover, Shihab *et al.*[37] showed that although mouse orthologue datasets could consist of less-studied genes based on the median number of publications in PubMed, at least in their study, the difference was not statistically significant. Another group of less-studied genes used to test the models were genes affected by *de novo* variants in disease cohorts (e.g. developmental disorders or epilepsy (GHIS, DOMINO, HIPred, Episcore))[17,35,37,38]. However, these gene sets were often small and expected to contain false-positive disease genes (e.g. approximately half of *de novo* LoF variants in Autism cohorts is expected to be benign[81]) and, therefore, might be unreliable for model effectiveness estimation. For example, in the Steinberg *et al.*[35] study, GHIS model outperformed RVIS by ~0.07 and ~0.10 AUC on 50 and 49 candidate disease gene sets, respectively. However, in the Shihab *et al.*[37] study, RVIS outperformed the GHIS model by ~0.17 and ~0.06 AUC on 55 and 61 candidate disease gene sets, respectively. Both studies selected candidate disease genes based on LoF *de novo* variants in autism cohorts from various studies and tested models on similar-sized datasets, but the results were dramatically different[35,37]. Therefore, tests on small gene sets must be interpreted with caution, but especially when less studied candidate disease gene sets were used.

1.5.7 Summary and under-researched areas

We reviewed nine studies that developed supervised ML models for gene categorisation[3,15,17,18,34–38]. Four studies used the HI/HS categorisation approach (p(HI), GHIS, HIPred, Episcore)[15,35,37,38] that was criticised for the usage of non-representative training data (e.g. in three studies the HS set did not include profiles of genes unaffected by structural variants)[36]. Moreover, considering that HI was a subset of dominant genes, three more recent studies (ISPP, DOMINO, GPP) argued that genes should be categorised based on inheritance patterns (AD/AR)[3,17,36]. However, we showed that the ISPP study models were probably extremely over-fitted, whereas the GPP study models (GDP and GRP) results were not tested (only validated using cross-validation) and were hard to interpret. Moreover, in both the ISPP and GPP studies, the exact gene list used to train the inheritance specific models were not reported[3,36], making them incomparable for future studies. The DOMINO model was trained on a manually curated set of AD and AR genes that can be reused in future studies and validated/tested on various other known or candidate dominant genes[17]. However, the DOMINO model was built with the assumption that all genes had to be categorised as AD or AR (i.e. ignored the existence of non-pathogenic genes)[17] and, consequently, we argue that it might not be effective at distinguishing AR and non-pathogenic genes. Previously, MacArthur *et al.*[18] tried to develop such a model (p(rec)), but a subsequent study showed that it could not distinguish between AD and AR genes[36] and, therefore, was classifying genes as pathogenic and non-pathogenic. This categorisation approach was also used in the remaining two studies (FIS and GPP)[3,34], where the GPP model was trained on a more representative and larger set of pathogenic genes. Therefore, with the existing models, categorisation of genes into three major groups (AD, AR, non-pathogenic) can be performed by first using the GPP model to distinguish pathogenic and non-pathogenic genes, and then using the DOMINO model to classify pathogenic genes as AD or AR. However, it is unknown how these two types of gene categorisation models work together, especially their ability to distinguish AR genes from other groups.

Therefore, further studies need to either develop a single model or combine predictions from multiple models that can classify genes into three major groups by assessing the likelihood of a gene harbouring AD, AR and non-pathogenic variants. Developing and combining two binary models (AD/AR and pathogenic/non-pathogenic) can be more feasible as it will allow direct comparison with existing binary models (DOMINO and GPP, respectively), where training datasets can be re-used. Future models have to be evaluated on various groups of genes from the final gene list, including those that models were not trained to distinguish between (e.g. AD/AR models also have to be evaluated on non-pathogenic genes). The development of advanced gene features can improve future model performance. Specifically, future AD/AR models might benefit from novel network-based features that will be based on the difference between numbers of interactions with positive (AD) and negative (AR) samples (previous studies analysed interactions only with positive samples[15,17,18]). Considering that most of the top features in the GPP model were based on variant data from large populations (some in combination with evolutionary data)[3], future pathogenic/non-pathogenic models might benefit from novel gene variation intolerance metrics. These metrics usually can be used on their own to prioritise disease genes and as features in future AD/AR models. Therefore, in this study we will consider developing novel features as steps towards the creation of more efficient gene classification models that have to be able to categorise genes as AD, AR, and non-pathogenic.

1.6 Aim and objectives

One of the reasons why the identification of novel human Mendelian disease genes is a challenging task is that many genes can tolerate deleterious variants in one or even both copies without developing pathogenic phenotypes. The aim of this study was to develop computational metrics that can be used to prioritise candidate Mendelian disease genes (i.e. classify genes as disease and non-disease) and distinguish them by probable inheritance patterns (i.e. classify disease genes as dominant and recessive). The review of existing methods highlighted several under-researched areas. Specifically, metrics based on variation data from large population databases were calculated without considering variant zygosity and their location within genes. Moreover, all existing supervised machine learning models were developed to prioritise candidate disease genes from only one out of three groups (dominant, recessive, or non-disease). For example, several models that used protein-protein interaction data considered interactions only with one group of genes (e.g. only dominant) to classify genes, and very little attention has been paid to how models trained to predict genes from various groups work together. Based on these gaps in previous studies, we defined the following objectives whose accomplishment could result in the development of improved metrics for candidate disease genes prioritisation:

1. Examine whether rare variants with deficiency of homozygous cases in a large population are more frequently observed in known recessive disease genes and, consequently, can be used as a feature for detection of novel recessive disease genes (addressed in Chapter 2).

2. Develop metric(s) that measure the randomness of variant distribution within genes and investigate its correlation with disease gene inheritance patterns (addressed in Chapter 3).

3. Develop metric(s) that can predict novel dominant and recessive disease genes by simultaneous analysis of protein-protein interactions with known genes from both groups (addressed in Chapter 4).

4. Combine features/metrics developed and proved to be effective in previous objectives with other gene biological properties to develop the novel supervised model(s) for gene classification that can be used to categorise genes into three classes: dominant, recessive and non-disease (addressed in Chapter 4).

1.7 Thesis structure

This thesis is written in the journal format since the fulfilment of the objectives resulted in three self-contained studies with little content overlap, two of which were published at the time of submission. The rest of this thesis is structured as follows:

Chapter 2 measured deviations from Hardy-Weinberg Equilibrium for relatively rare variants (AF<0.05) in gnomAD to understand if variants with heterozygote excess are enriched in known recessive disease genes and, consequently, can be used to predict novel recessive disease genes. This study is published, and the thesis version has only minor stylistic differences.

Chapter 3 developed a gene variation intolerance ranking (GeVIR) system by measuring how unevenly variants were distributed in a gene relative to other genes. This study is published in a short letter format with supplementary methods and notes, and the thesis version was re-written in a regular journal article format. The results in both versions are the same, but they are described in more detail in the thesis version.

Chapter 4 first developed multiple supervised machine learning models based on gene protein-protein interactions with known dominant, recessive disease genes, and candidate disease genes based on variation intolerant scores. Then predictions from these models and various other gene properties (including GeVIR scores) were used as features to develop two supervised machine learning models that classified genes as pathogenic/non-pathogenic and dominant/recessive. Finally, predictions from these two models were combined into a single continuous gene ranking metric that can be used to measure gene predisposition to disease inheritance patterns (DIP). This study is unpublished but is formatted as a journal article and is ready for publication.

Chapter 5 summarises research contributions, provides directions for future work and makes concluding remarks.

1.8 Publications and presentations

The work presented in this thesis has resulted into two publication:

- Abramovs, N., Brass, A., and Tassabehji, M. (2020). Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Front. Genet.* 11. doi:10.3389/fgene.2020.00210.

- Abramovs, N., Brass, A., and Tassabehji, M. (2020). GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Nat. Genet.* 52, 35–39. doi:10.1038/s41588-019-0560-2.

Both papers were co-authored by my supervisors May Tassabehji and Andrew Brass. NA, MT, and AB conceived and designed the research. NA executed the analysis. NA and MT performed the primary writing. MT and AB supervised all aspects of the research, reviewed, and edited the manuscript.

In addition, the work described in Chapter 3 was also presented at:

- Abramovs N. Gene Variation Intolerance Rank – an *in silico* approach to prioritise disease candidate genes. *Research Symposium*. 2019. At The University of Manchester, School of Computer Science.

- Abramovs N. and Tassabehji M. GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Video Call*. Feb 20, 2020. At The National Human Genome Research Institute Genome Sequencing Program Methods Working Group (chaired by Monkol Lek).

Chapter 2

Hardy-Weinberg equilibrium in the large scale genomic sequencing era

2.1 Abstract

Hardy-Weinberg Equilibrium (HWE) is used to estimate the number of homozygous and heterozygous variant carriers based on its allele frequency in populations that are not evolving. Deviations from HWE in large population databases have been used to detect genotyping errors, which can result in extreme heterozygote excess (HetExc). However, HetExc might also be a sign of natural selection since recessive disease causing variants should occur less frequently in a homozygous state in the population, but may reach high allele frequency in a heterozygous state, especially if they are advantageous. We developed a filtering strategy to detect these variants and applied it on genome data from 137,842 individuals. The main limitations of this approach were quality of genotype calls and insufficient population sizes, whereas population structure and inbreeding can reduce sensitivity, but not precision, in certain populations. Nevertheless, we identified 161 HetExc variants in 149 genes, most of which were specific to African/African American populations (∼79.5%). Although the majority of them were not associated with known diseases, or were classified as clinically “benign”,

they were enriched in genes associated with autosomal recessive diseases. The resulting dataset also contained two known recessive disease causing variants with evidence of heterozygote advantage in the sickle-cell anemia (*HBB*) and cystic fibrosis (*CFTR*). Finally, we provide supporting *in silico* evidence of a novel heterozygote advantageous variant in the chromodomain helicase DNA binding protein 6 gene (*CHD6*; involved in influenza virus replication). We anticipate that our approach will aid the detection of rare recessive disease causing variants in the future.

2.2 Introduction

The Hardy-Weinberg Equilibrium (HWE) is an important fundamental principle of population genetics, which states that “genotype frequencies in a population remain constant between generations in the absence of disturbance by outside factors”[82]. According to HWE, for a locus with two alleles *A* and *a* with corresponding frequencies p and q, three genotypes are possible *AA*, *Aa*, and *aa* with expected frequencies p^2 , 2pq, q^2 , respectively[71]. However, various factors, including mutation, natural selection, non-random mating, genetic drift, and gene flow can cause deviations from HWE[71]. Positive and negative assertive mating might result in deviations from HWE due to heterozygote deficiency or excess respectively, although the latter is more rarely observed in humans[83]. Non-random mating due to geographical location might be a common cause of deviations from HWE due to heterozygous deficiency in large populations of different ethnicities[84]. If a population consists of several sub-populations and individuals randomly mate within, but not between sub-populations, then homozygous alleles in the overall population will be observed more frequently than expected by HWE (“Wahlund effect”)[85]. A technical cause of deviations from HWE, sometimes observed in population studies, is sequencing errors[71,72]. Previous studies found that variants deviated from HWE mainly due to heterozygote excess (60– 69% of the cases)[71,72] and deviations were 11 times more frequently observed in unstable genomic regions such as segmental duplications and simple tandem repeats[71], that are prone to sequencing errors. These issues were addressed in the Genome Aggregation Database (gnomAD v2.1.1)[11], currently the largest publicly available population variant database (137,842 predominantly healthy individuals from seven ethnic populations). Variants with extreme heterozygote excess in the database were excluded by gnomAD, whereas those located in repeat regions were marked as dubious.

A factor causing deviations from HWE that has not been investigated on a large scale, is natural selection. Although individuals with known severe pediatric diseases were excluded from gnomAD[11], some disease causing variants persisted[60]. For example, the African specific (∼91% of the carriers) ENST00000335295.4:c.20A>T (rs334, note that variants of the same type that are located at the same position share the same rs number in dbSNP[86]) variant in *HBB* gene, is a known recessive pathogenic variant which causes sickle-cell disease (MIM:603903)[87], but it is present in four African individuals (who could have sickle-cell disease) in a homozygous state in gnomAD. Moreover, this variant is present in a heterozygous state in ∼9% (1,113/12,482) of African individuals (i.e., unaffected carriers), which is significantly more (∼2.5 times) than the expected number (∼439 individuals) according to HWE (P = 1.38E-07), for that number of homozygous individuals. The presence of a recessive disease causing variant at a high frequency in populations may also be due to over-dominant selection, i.e., a heterozygous variant provides some advantage to carriers[88], as is the case for ENST00000335295.4:c.20A>T variant in *HBB* gene, which provides carriers protection from malaria (MIM:611162)[89]. This example illustrates that variants deviating from HWE due to heterozygote excess may also be recessive disease causing and possibly heterozygote advantageous. Here we developed a variant filtering strategy to detect novel potential disease causing variants that might deviate from HWE due to natural selection, and applied it to population data from gnomAD.

2.3 Methods

2.3.1 Collecting Gene and Variant Datasets

The gene dataset, with disease phenotype and inheritance data from Online Mendelian Inheritance in Man (OMIM) database[41], was obtained from Gene Discovery Informatics Toolkit (GDIT)[24] and consisted of 19,196 protein coding genes. Population variant data with clinical annotation (ClinVar)[90] was obtained from gnomAD[11] via API (https://gnomad.broadinstitute.org/api, accessed January 2020). The database consisted of 137,842 individuals from seven populations: Non-Finnish European (NFE, n = 64,603), Latino/Admixed American (AMR, n = 17,720), South Asian (SAS, n =

15,308), Finnish (FIN, n = 12,562), African/African American (AFR, n = 12,487), East Asian (EAS, n = 9,977), and Ashkenazi Jewish (ASJ, n = 5,185)[11]. The initial variant dataset consisted of more than 17 million unique variants in 18,214 genes whose symbols in the GDIT dataset were found in gnomAD. At the time of the analysis performed in this chapter (January 2020), gnomAD did not store Ensembl transcript versions, which is an issue since they are required to report variants in HGVS-compliant format[91]. The transcript versions were later added to the gnomAD, and we updated variant transcript information via gnomAD API (accessed April 2022). There should be no inconsistency since the gnomAD database, and Ensembl Variant Effect Predictor (VEP)[92] versions were the same (2.1.1 and 85, respectively). Still, we note that everything except transcript versions is reported as it was at the time of the original analysis (January 2020).

2.3.2 Filtering Initial Variant Dataset

Variants which satisfied the following criteria were selected for initial analysis of deviations from HWE: (i) Variant is located in the canonical transcript [as defined in gnomAD who used GENCODE[93] v19 annotation]; (ii) Variant is located on an autosomal chromosome; (iii) Variant is protein coding, i.e., has one of the following Variant Effect Predictor (VEP)[92] version 85 consequences: *"transcript_ablation"*, *"splice_acceptor_variant"*, *"splice_donor_variant"*, *"stop_gained"*, *"frameshift_variant"*, *"stop_lost"*, *"start_lost"*, *"transcript_amplification"*, *"inframe_insertion"*, *"inframe_deletion"*, *"missense_variant"*, *"protein_altering_variant"*, *"splice_region_variant"*, *"incomplete_terminal_codon_variant"*, *"start_retained_variant"*, *"stop_retained_variant"*, *"synonymous_variant"*; (iv) Variant Allele Frequency (AF) is >0.001 in at least one population; (v) Variant site is covered in ≥80% of the individuals in each seven populations; (vi) Variant is "PASS" quality in exome and genome datasets (if present in both); (vii) Variant site does not contain frequent alternative variants that could compromise statistical results of the biallelic HWE test (sum of AFs of all alternative variants seen at the same chromosomal position in the same population must be <0.001).

2.3.3 Statistics and Measuring Deviations From HWE

The original code to measure statistical significance of variant deviation from HWE, developed by Wigginton *et al.*[94], calculated *P* (two-sided) as the probability of observed sample plus the sum of all probabilities of more extreme cases. However, Graffelman and Moreno[95] later showed that mid *P*, calculated by adding only half of the probability of observed sample to the sum of all probabilities of more extreme cases, was less conservative (i.e., mid *P* is always smaller than two-sided *P*) and showed better potential for testing deviations from HWE of rare variants. Therefore to create a python implementation of Graffelman and Moreno[95] method, we modified Wigginton *et al.*[94] code to return mid *P*. Variants deviating from HWE with mid $P \leq 0.05$ were considered to be statically significant. For all other cases two-sided Fisher's exact test was used (SciPy python package[96]), and the results were reported as *P* and fold-enrichment (FE), defined as the ratio of the two proportions.

2.3.4 Selecting Candidate Disease/Heterozygote Advantageous Variants

Variants that satisfied the following criteria were selected into a final dataset of candidate disease/heterozygote advantageous variants: (i) Variant AF is ≤0.05 in each of the ethnic populations [more common variants are classified as "benign" according to American College of Medical Genetics and Genomics (ACMG) guidelines[65]; and are more likely to deviate from HWE due to genotyping errors[97]]; (ii) Variant has statistically significant ($P \leq 0.05$) excess of heterozygotes in at least one ethnic population; (iii) Variant has excess of heterozygotes in each population (not required to be statically significant). This filter was added as variants with heterozygote excess in one ethnic population but not in the others, might be a result of gene flow; (iv) Variant is not located in a segmental duplication[98] or tandem repeat region[99] (loci obtained from UCSC Genome Browser[71,100]; (v) 50% of heterozygote variant carriers in the overall population have allele balance (AB), defined as the proportion of reads that support the minor allele[a], between 0.4 and 0.55 (AB thresholds are justified in the Results section). After applying these filters the resulting dataset consisted of 299 variants located in 267 genes.

[a] https://gatk.broadinstitute.org/hc/en-us/articles/360036479772-FilterVcf-Picard-

A recent investigation of the deviation from HWE of the CCR5-Δ32 (rs333, ENST00000343801.4:c.554_585del) allele in gnomAD showed that excess of heterozygotes can be caused by misclassification of homozygous individuals as heterozygous with high AB[97]. To minimize the number of false positive candidate disease/heterozygote advantageous variants, HWE statistics for them were recalculated considering heterozygous individuals with AB > 0.8 as homozygous, which is a more conservative than the 0.9 AB threshold used in the original study[97]. AB data was not available for each ethnic population, so an assumption was made that novel homozygous individuals were distributed among populations in the same proportions as heterozygotes. After excluding variants that were no longer deviating from HWE due to heterozygote excess, the final dataset consisted of 161 variants located in 149 genes (Appendix Table A.1). HWE statistics of these variants was recalculated using gnomAD v3[58] data (71,702 whole genome samples mapped to build GRCh38), which contained a larger AFR population (∼1.7 times larger; 21,042 individuals, all other populations were smaller than in gnomAD v2.1.1). Chromosome coordinates were mapped with LiftOver[100].

2.4 Results

After applying initial filters on variant data from seven ethnic populations (Figure 2.1a), the resulting dataset consisted of 382,506 unique variants (803,584 if counted in each population separately, Figure 2.1b) located in 16,871 genes. Exclusion of rare variants (AF < 0.001) from the analysis reduced the possible impact of population size (Figure 2.1a) on the number of variants analyzed (Figure 2.1b). For example, the Finnish (FIN) populations was ∼5 times smaller than the Non-Finnish European (NFE) population (12,562 and 64,603 individuals, respectively), but had a similar number of unique variants (85,553 and 92,458 variants, respectively). However, population size had a significant effect on the ability of the HWE test to detect Heterozygote Excess (HetExc) deviation of rare variants: the larger the population, the smaller the AF threshold after which statically significant HetExc can be reported. The minimal HetExc AF thresholds (i.e., assuming complete absence of homozygotes) are shown on Figure 2.1c, note the negative correlation with population sizes shown in Figure 2.1a.

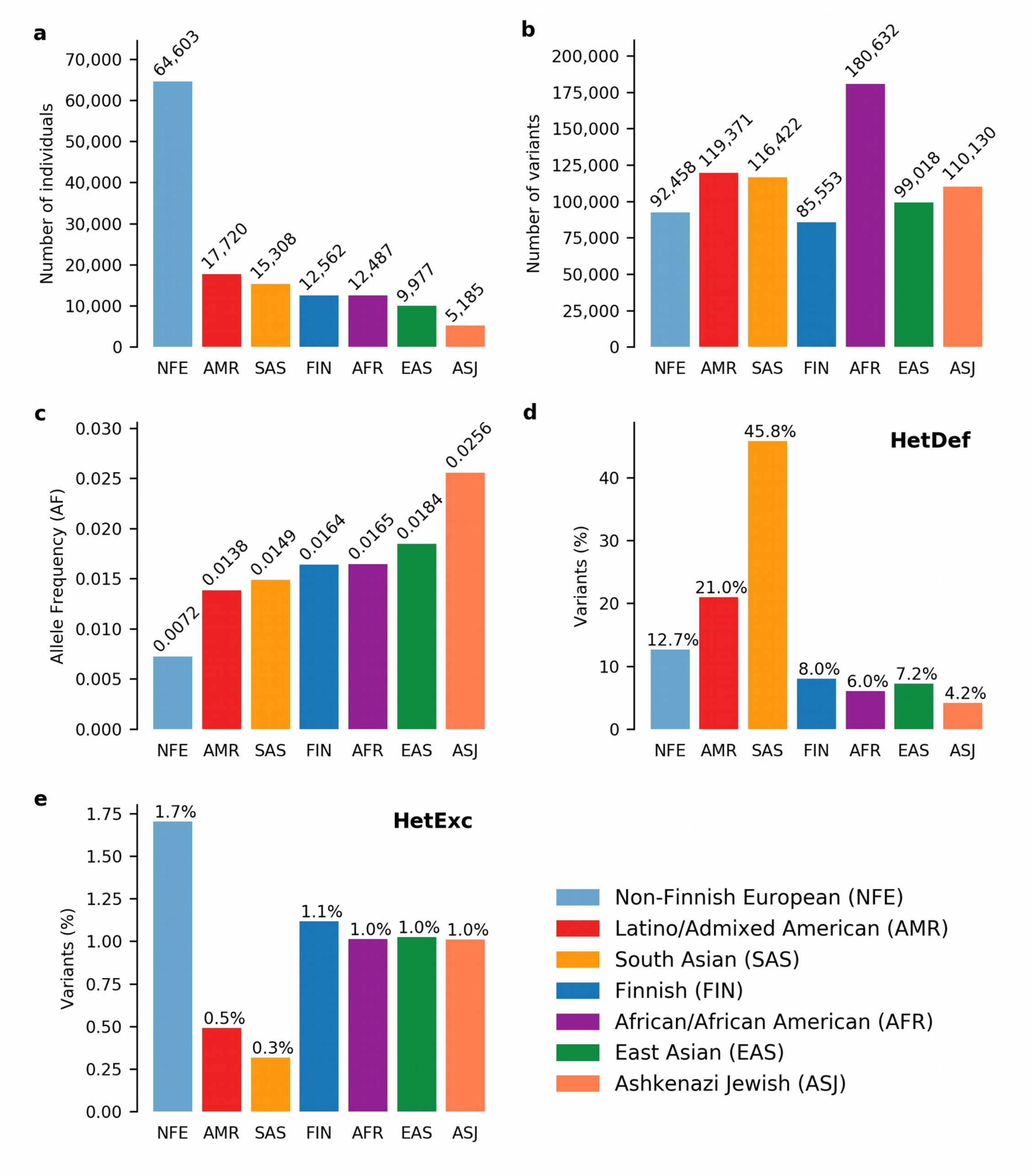


Figure 2.1: Deviations from Hardy-Weinberg Equilibrium (HWE) in 7 ethnic gnomAD populations.

Another factor that could affect detection of HetExc variants, was the degree to which HWE assumptions were satisfied in each population. For example the “random mating” assumption would be violated in populations with a high degree of consanguineous marriages or consisting of individuals from several countries, and would result in a higher proportion of variants deviating from HWE due to heterozygote deficiency (HetDef) (i.e., “Wahlund effect”). To some degree, all populations deviated more frequently from HWE due to HetDef than HetExc (Figure 2.1d,e). The largest proportion of HetDef variants were observed in South Asian (SAS) and Latino/Admixed

American (AMR) populations, 45.8% and 21.0%, respectively. Consequently, these populations also had the lowest proportion of HetExc variants, 0.3% and 0.5%, respectively. The lowest proportion of HetDef variants was observed in the Ashkenazi Jewish (ASJ) population. However, even in this population, variants deviated from HWE due to HetDef ~4 times more frequently than due to HetExc, 4.2% and 1.0%, respectively. Interestingly, the African/African American (AFR) population had the second lowest percentage of HetDef variants (6.0%), which outscored the FIN population (8.0%), considered as a homogeneous isolate. The largest proportion of HetExc variants was in the NFE population (1.7%, 1,574 variants), which had the smallest AF threshold for HetExc detection (AF = 0.0072, Figure 2.1c). Despite this, the AFR population still had the largest absolute number of HetExc variants (1,829). Therefore, overall population variant shift from HWE toward HetDef (i.e., the majority of the variants have higher than expected homozygous AF) decreased the number of statistically significant HetExc variants (especially in SAS and AMR), which can also be seen in Figure 2.2 for relatively rare variants (AF < 0.1).Number of **(a)** individuals and **(b)** variants in each populaton. **c)** Minimum variant allele frequency (AF) required for statistically significant heterozygote excess according to HWE, in the absence of homozygous individuals in each population. Percentage of variants (raw numbers are shown in **b**) deviating from HWE due to **(d)** heterozygote deficiency or **(e)** heterozygote excess in each population.

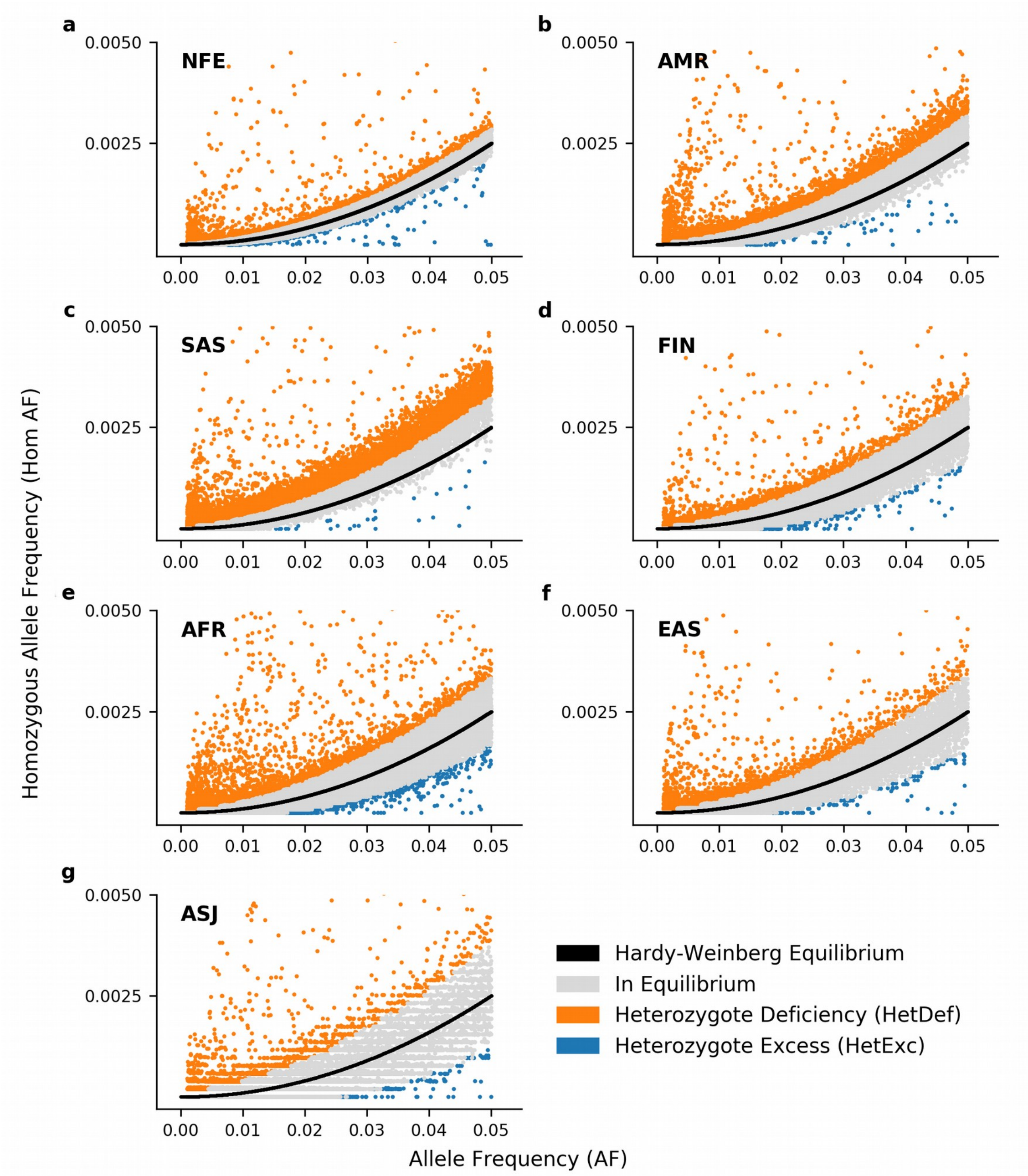


Figure 2.2: Comparison of observed ratio between variant Allele Frequency (AF) and homozygous AF with expected ratio according to Hardy-Weinberg Equilibrium (HWE) in 7 ethnic gnomAD populations.

Populations: Non-Finnish European (**a**, NFE), Latino/Admixed American (**b**, AMR), South Asian (**c**, SAS), Finnish (**d**, FIN), African/African American (**e**, AFR), East Asian (**f**, EAS), and Ashkenazi Jewish (**g**, ASJ). Black line represents expected ratio between AF and expected homozygous AF according to HWE. Variants where deviation from HWE are not significant (*P* > 0.05) are shown in grey, whereas those that deviate from HWE due to heterozygote deficiency or excess are shown in orange and blue respectively. Only variants with $0.001 \leq$ AF ≤ 0.05 and homozygous AF ≤ 0.005 are shown.

This initial analysis has been performed on variants that remained following the gnomAD sequence quality filtering process and might therefore be assumed to be real. However, variant databases are known to contain errors that could give a significant HetExc signal. To explore this we developed a set of more stringent filters. In particular, variant properties that could produce a false positive HetExc signal were investigated. For this analysis, variants present in multiple populations were counted once, and variants with AF > 0.05 in at least one population were excluded. At this stage, only variants that had an excess of heterozygotes in all populations, and were statistically significant in at least one population were classified as HetExc.

Firstly, to investigate the correlation between HetExc and chromosomal regions prone to sequencing errors, variants were divided into three groups: (i) “segmental duplication” (2,676), (ii) “tandem repeat” (1,182), and (iii) all others named “Ref ” (40,801). HetExc variants were significantly more frequent in the “segmental duplication” (FE = ∼2.5, *P* = 2.0E-08) group than in the “Ref ” group, whereas the proportion of HetExc variants in the “tandem repeat” and “Ref ” groups were almost the same (Figure 2.3a and Appendix Table A.2). Therefore, HetExc of variants located in segmental duplications might be a result of genotyping errors.

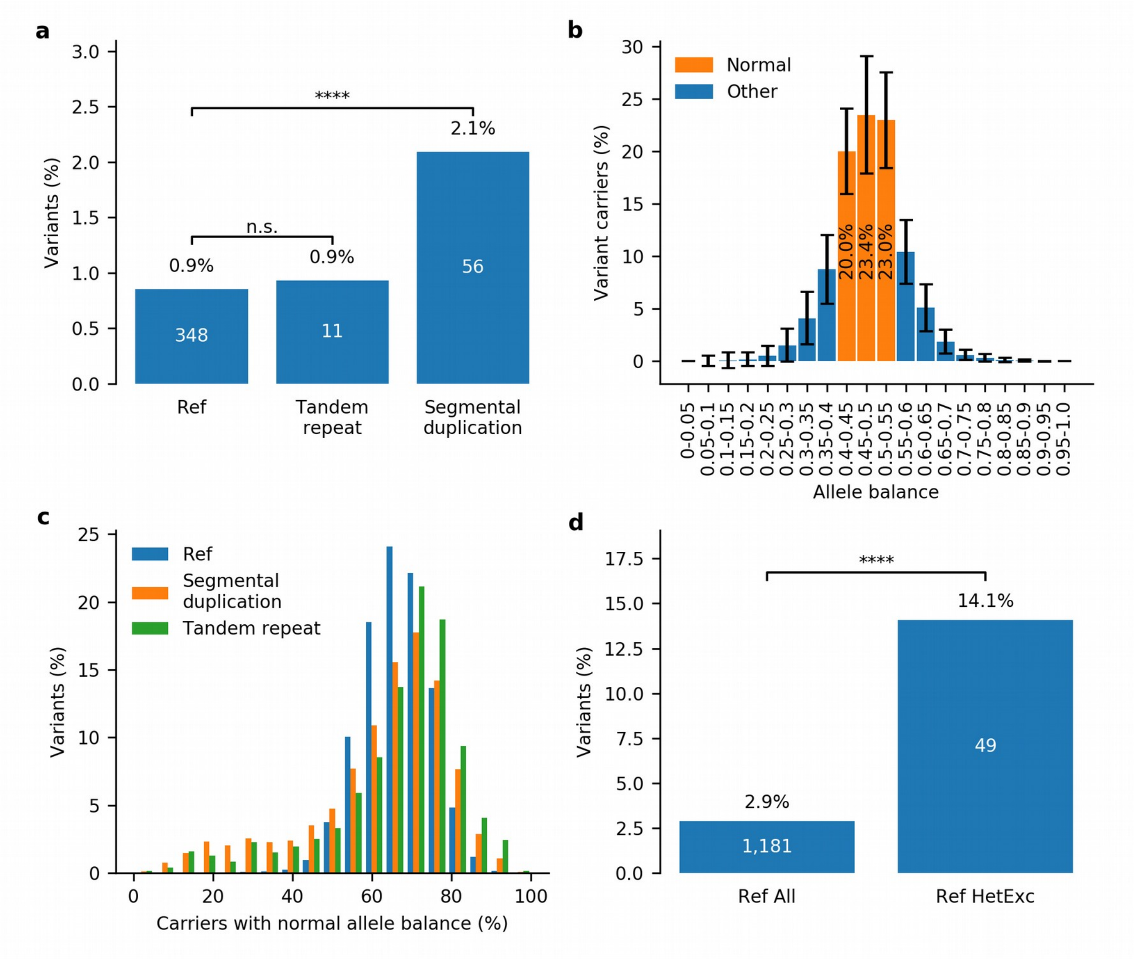


Figure 2.3: Impact of tandem repeats, segmental duplications and allele balance on the probability of variant deviation from Hardy-Weinberg Equilibrium (HWE) due to heterozygote excess (HetExc).

a) Percentage of variants deviating from HWE due to HetExc that are located in tandem repeat, segmental duplication regions or the reference ("Ref," all other regions) group. **b)** Distribution of allele balance (AB) between variant carriers in variants from "Ref" group (error bars indicate standard deviation). For each variant these statistics are aggregated into a single metric that represents cumulative percentage of Variant Carriers with Normal (0.4–0.55) Allele Balance (VCNAB, e.g., 20.0% + 23.4% + 23.0% = 66.4%). **c)** Distribution of variants with various VCNAB percentages in "Segmental duplication", "Tandem repeat" and "Ref" groups. **d)** Percentage of variants with VCNAB < 50% in the whole "Ref" group and a subset of variants with statistically significant excess of heterozygotes in "Ref" group. ∗∗∗∗Indicates statistical significance of $p \leq 0.0001$.

Secondly, to investigate the correlation between HetExc and allele balance (AB), which is a known indicator of systematic genotyping errors[101], the AB profile of an average gnomAD variant was required. In gnomAD, variant AB data is stored as a number of variant carriers (converted to percentages here) in 20 AB groupings (from 0 to 1, 0.05 group size). Figure 2.3b shows the distribution of AB between variant carriers in variants from the "Ref " group. For an average variant, the majority of variant carriers (66.4%) had an AB between 0.4 and 0.55 and were named "Normal" here, because it was close to the expected normal 0.5 ratio for heterozygous variants. To aggregate variant AB data from 20 groups into a single numeric metric, it was measured as percentage of variant carriers with "normal" allele balance (VCNAB), calculated as the number of heterozygote variant carriers with AB 0.4-0.55 divided by the total number of heterozygote variant carriers. Both "segmental duplication" and "tandem repeat" groups had more variants with high and low VCNAB [80% Confidence Interval (CI) = 33.9-79.6% and 41.9-81.4%, respectively] than the "Ref " group (80% CI = 55.6-77.0%), which indicates that variants in these regions are more prone to genotyping errors and were excluded from further analysis (Figure 2.3c). The minimal VCNAB threshold for "PASS" quality variants was defined by a lower bound fraction of 95% CI calculated for variants from the "Ref " group (CI = 49.3-82.2%), rounded to 50% (i.e., half of the variant carriers must have AB in the range 0.4-0.55). Only 2.9% of variants in the "Ref " group would not pass this filter, but the fail rate among HetExc variants would be ~4.9 times higher (*P* = 1.9E-17, Figure 2.3d and Appendix Table A.3). Therefore, variants with low AB (VCNAB < 50%) might be enriched with genotyping errors and were also excluded from further analysis.

Finally, HWE statistics for HetExc variants that were not located in segmental duplication or tandem repeat regions and had VCNAB ≥ 50% (299 variants in 267 genes) were recalculated considering heterozygous individuals with AB > 0.8 as homozygous. 161 variants in 149 genes that were still HetExc according to the updated HWE statistics were selected as candidate recessive disease causing genes (Appendix Table A.1). These HetExc variants were then compared with a group of variants that survived the same filtering process, but did not have an excess of heterozygotes (HetExc-). The HetExc- and HetExc groups consisted of 39,430 and 161 variants (50,365 and 161 if counted in seven ethnic populations separately, Figure 2.4a) in

11,842 and 149 genes, respectively. Most of the HetExc variants were present in African/African American populations (128/161, ~79.5%), which was significantly more than expected (FE = ~1.7, P = 3.0E-05) based on the proportion in the HetExc- group (18,957/39,430), whereas all other populations had significantly less than expected HetExc variants ($P \leq 0.001$) except EAS and ASJ (Appendix Table A.4). Both HetExc- and HetExc groups contained a similar proportion of missense and synonymous variants (Figure 2.4b and Appendix Table A.5).

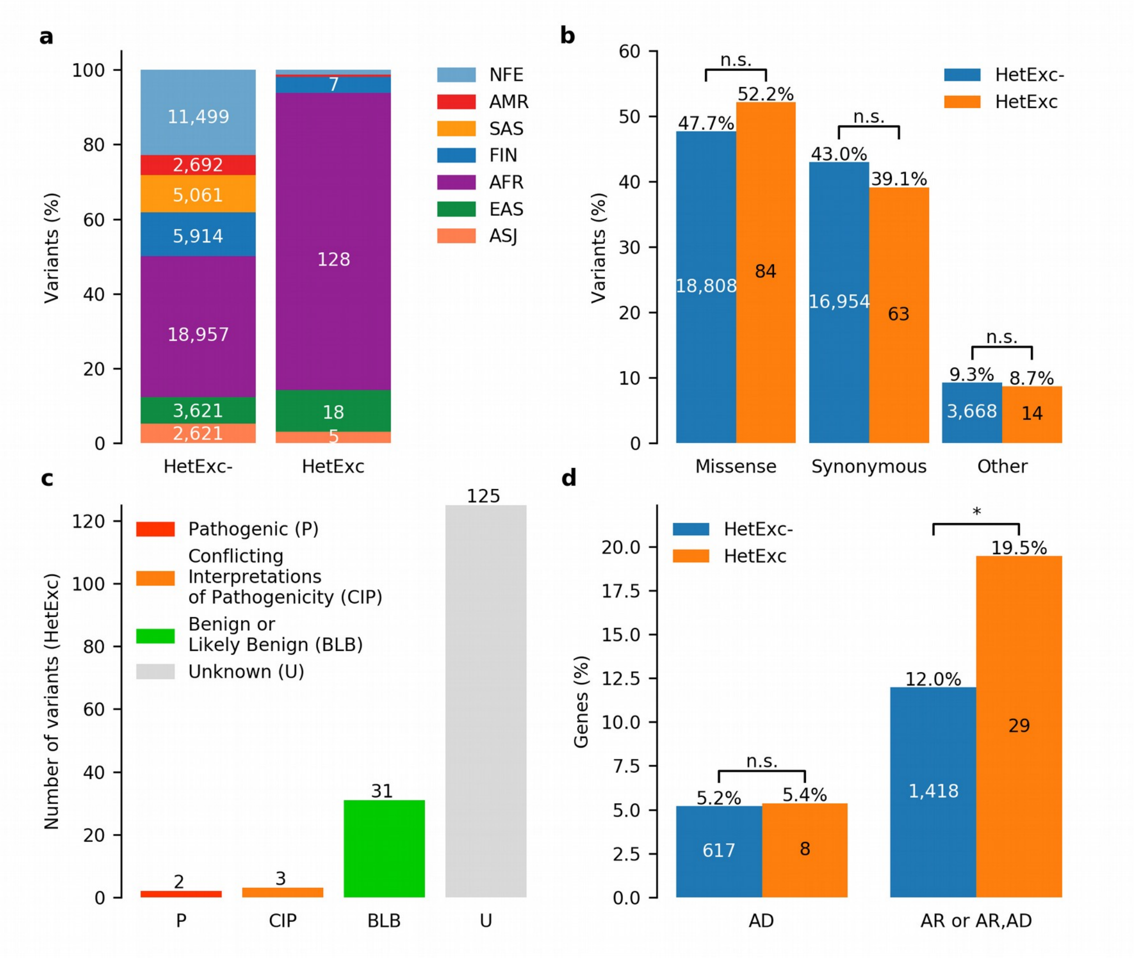


Figure 2.4: Potential recessive disease causing variants identified by deviation from Hardy-Weinberg Equilibrium (HWE) due to excess of heterozygotes (HetExc).

a) Distribution of variants deviating and not deviating from HWE due to excess of heterozygotes (HetExc and HetExc-, respectively) in 7 ethnic gnomAD populations. **b)** Proportions of missense, synonymous and other protein coding variants in HetExc and HetExc- datasets. **c)** ClinVar status (e.g., pathogenic/benign) of HetExc variants. **d)** Known disease associated genes with at least one variant in HetExc and HetExc-datasets grouped by inheritance pattern: autosomal dominant (AD), autosomal recessive (AR) or both. ∗Indicates statistical significance of $p \leq 0.05$.

To determine which of the HetExc candidate recessive disease causing variants were already known, their clinical significance in the disease variant database (ClinVar[90]) was analyzed. The majority of HetExc variants (125/161, ~77.6%) were not present in ClinVar, whereas the majority of those that were present in ClinVar (31/36, ~86.1%) had a “Benign” or “Likely benign” status (Figure 2.4c). The only two variants with “Pathogenic” status were ENST00000335295.4:c.20A>T (rs334) in *HBB* (causes recessive sickle cell disease MIM:603903; carriers are protection from malaria, MIM:611162) and ENST00000003084.6:c.1521_1523delCTT (rs1801178) in *CFTR* [causes recessive cystic fibrosis disease, MIM:219700; hypothesized to be protective from cholera[102] or tuberculosis[103]]. However, genes with at least one HetExc variant were significantly more frequently associated with known autosomal recessive (AR) diseases than genes containing only HetExc- variants (FE = ∼1.6, *P* = ∼0.02, Figure 2.4d and Appendix Table A.6). HetExc variant enrichment in known AR genes adds evidence that some of the selected variants might deviate from HWE due to natural selection and could have some disease association. However, only seven (∼5.5%) of these variants were also HetExc in gnomAD v3: ENST00000335295.4:c.20A>T (rs334) in *HBB*, ENST00000373233.3:c.7210G>C (rs61292917) in *CHD6*, ENST00000267622.4:c.3118A>G (rs34805848) in *TRIP11*, ENST00000374695.3:c.9540G>A (rs62642506) in *HSPG2*, ENST00000246186.6:c.1691G>A (rs751887) in *MMP24*, ENST00000298317.4:c.626C>T (rs35157957) in *RPUSD4* and ENST00000359486.3:c.441C>T (rs73887968) in *TCF20*. Note that 3/7 (∼42.9%; *HBB*, *TRIP11*, and *HSPG2*) genes are associated with known recessive diseases (Appendix Table A.1). Since gnomAD v3 mostly consisted of individuals that were not present in gnomAD v2.1.1, this adds evidence that these seven variants are not deviating from HWE by chance.

2.5 Discussion

Analysis of deviations from the HWE on a large genomic dataset has shown that all populations, but especially South Asian (SAS) and Latino/Ad-mixed American (AMR), were more frequently deviating due to heterozygote deficiency (HetDef) than heterozygote excess (HetExc). A higher rate of HetDef variants in SAS and AMR populations is in line with previous reports[72,104], possibly due to the large number of consanguineous marriages in these regions [e.g., 38% of SAS population in the Exome Aggregation Consortium (ExAC)[12]]. However, our findings that HetDef is a major cause of deviations from HWE in all populations is contrary to previous studies (Chen *et al.*[72] and Graffelman *et al.*[71], which used more strict *P* thresholds (0.001 and 0.0001, respectively) and reported that deviations from HWE were more frequently observed due to HetExc. However, previous studies focused on error detection in older and smaller datasets, some of which were corrected in gnomAD. Graffelman *et al.*[71] analyzed 104 Japanese individuals in the 1000 Genomes database[4], where the minimal statistically significant HetExc AF threshold (0 homozygous and $P < 0.001$) was ∼0.23[71]. Only 11/382,506 variants analyzed in our study were that frequent and had no homozygous individuals reported, nine of which were located in segmental duplication or tandem repeat regions. We observed a higher rate of HetExc variants in these regions, as well as those that had low allele balance, which correlates with previous work[71,101]. Chen *et al.*[72] analyzed "open reading frame" genes and selected only one variant per gene where AF was closer to 0.50 (584 variants in total) in ExAC (60,706 individuals). However, this approach resulted in the exclusion of rare variants that were analyzed in this study and might be more affected by the Wahlund effect (i.e., more likely to be HetDef). Moreover, some of the HetExc variants detected in previous studies were marked as non-pass quality or were no longer HetExc in gnomAD, possibly due to differences in variant filtering and genotype calling procedures. For example, ENST00000369356.4:c.1801C>T (rs1778112) variant in *PDE4DIP* was present in the heterozygous state in ∼91% of individuals in the 1000 genomes database, but was never observed as homozygous and was assigned "non-pass" quality in gnomAD. Another example, the *BRSK2* variant ENST00000382179.1:c.551+6delG (rs61002819) was HetExc in ExAC (P = 1.9E-15), but not in gnomAD (P = 0.13). Therefore, a higher rate of HetDef variants in our study could be explained by a larger population size and a different variant dataset, as well as improvements in variant filtering and genotype calling procedures.

Analysis of HetExc variants (Appendix Table A.1), selected as recessive disease causing candidates, led to somewhat contradictory results, which should be interpreted with caution. Enrichment of HetExc variants in the African/African American (AFR) population was unexpected, and might indicate more extensive natural selection or be a sign of systematic genotype errors in this population. Enrichment of HetExc variants (32/161) in genes associated with known autosomal recessive diseases supports the hypothesis that some of these variants could be causing recessive diseases, whereas the presence of a large proportion of synonymous variants (11/32) and the assigned "Benign" or "Likely benign" status of the majority of the known variants (21/32 in CinVar, 17/21 were "Benign" or "Likely benign") in this group provides evidence against it. Moreover, despite applying our extensive filtering strategies, many of the HetExc variants might still be deviating from HWE due to genotype errors or by chance due to insufficient population size. The latter might be an explanation for some AFR variants that were HetExc in gnomAD v2.1.1, but not in the new v3 release, which had a larger AFR population. However, the ENST00000003084.6:c.1521_1523delCTT (rs1801178) variant in *CFTR* also was not HetExc in gnomAD v3 and was observed as homozygous in 4 out of 32,299 NFE individuals (and 2 heterozygotes with AB > 0.8), whereas in v2.1.1 only 1 out of 64,603 NFE individuals was homozygous. Therefore, the difference between the number of homozygote in gnomAD v2.1.1 and v3 might also be explained by other factors, such as differences between genotype calling procedures for exome and genome data.

Nevertheless, the presence of known pathogenic and heterozygote advantageous variants such as ENST00000335295.4:c.20A>T (rs334) in *HBB* and ENST00000003084.6:c.1521_1523delCTT (rs1801178) in *CFTR* suggests that some of the other 161 HetExc variants might also be functionally significant. Especially, the ENST00000373233.3:c.7210G>C (rs61292917) variant in *CHD6* gene variant, which was HetExc in both versions of the gnomAD database and was predicted to be deleterious by *in silico* tools (SIFT[8] = 0; PolyPhen-2[9] = 0.961). Moreover, it was more frequently (FE = ~5.21, *P* = 1.19E-04) seen in African than African American populations in the 1000 genomes database (Appendix Table A.7), similar to the ENST00000335295.4:c.20A>T (rs334) variant in *HBB* (FE = ~3.42, *P* = 1.49E-05), which suggests that these variants might be under purifying selection in populations that moved out of Africa (i.e., they might be disease causing, but advantageous only in Africa, which is known in the case of the ENST00000335295.4:c.20A>T (rs334) variant in *HBB*). *CHD6* is associated with the rare Hallermann-Streiff syndrome (HSS, MIM:234100)[105], and is known to act as transcriptional repressor of different viruses including influenza and papiloma virus[106,107]. Interestingly, ENST00000373233.3:c.7210G>C (rs61292917) variant has a much lower AF in the African population (AF = 0.066), than ENST00000335295.4:c.20A>T (rs334) variant (AF = 0.120) in the 1000 genomes database. Considering *CHD6* is extremely intolerant to variation (missense z-score[27] = 4; LOEUF[11] = 0.07), ENST00000373233.3:c.7210G>C (rs61292917) variant is more enriched in the African population compared with ENST00000335295.4:c.20A>T (rs334) (i.e., possibly due to stronger purifying selection), which suggests that ENST00000373233.3:c.7210G>C (rs61292917) might be disease causing even in the heterozygous state.

Our study highlighted that the ability of HWE to detect candidate recessive disease causing variants is mainly limited by both the quality of genotype calls and the size of available exome/genome variant data, whereas absence of information about sub-populations (e.g., Africans and African Americans) and a high level of inbreeding (e.g., SAS) could reduce sensitivity, but not precision, of the approach in certain populations. We anticipate that improvements in sequencing technologies and variant filtering software should reduce the number of false positive HetExc variants in the future. In fact, false positive HetExc variants that survived our strict quality filters, might aid the development of more efficient sequencing filtering strategies by helping to understand new patterns of genotype errors. The size of the largest population analyzed in this study (NFE = 64,603 individuals) allowed us to detect statistically significant HetExc only amongst variants with AF ≥ ∼0.0072 (∼33% of 61,077 variants with AF = 0.001–0.05). Consequently, some common recessive disease causing variants were missed even if homozygous individuals were completely absent in the population. For example, HetExc of the ENST00000374855.4:c.448G>C (rs1800546) variant in *ALDOB* [causes recessive hereditary fructose intolerance, MIM:229600[108]] was not statistical significant (*P* = ∼0.3), despite being observed in the heterozygous state in 627 NFE individuals (AF = ∼0.005). As the number of sequenced exomes and genomes is rapidly growing, this problem may soon be addressed. Indeed, the United Kingdom National Health Service is planning to sequence 1 million genomes by October 2023 with a wider ambition to increase this number to 5 million[109]. If the NFE population was 1 million, then the AF threshold would drop to ∼0.0018 (∼73% of 61,077 variants with AF = 0.001–0.01), whereas with 5 million individuals it would be possible to detect statistically significant HetExc in all variants with AF ≥ ∼0.0008. Therefore, it might be possible to use HWE strategies to detect rare recessive disease causing variants in the near future.

In this study, we explored the use of HWE to identify potential recessive disease causing variants in a large mainly healthy population database by developing a bespoke filtering strategy to detect variants where an excess of heterozygotes in a population could be a result of natural selection. Overall, this approach showed potential, especially for the AFR population, successfully identifying some variants in recessive diseases that are known to be heterozygote advantageous, and providing novel candidates for further investigation. A natural progression of this work would be validation of genotype calls of HetExc variants to understand possible causes of genotype errors and analysis of the biological effect of true positive HetExc variants to determine their potential health implications. We also anticipate that this approach will become more robust in the future as the size and quality of available genomic data increases.

Chapter 3

GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes

3.1 Abstract

With large scale population sequencing projects gathering pace there is a need for strategies that progress disease gene prioritisation. Metrics that provide information about a gene and its ability to tolerate protein altering variation can aid clinical interpretation of human genomes and advance disease gene discovery. Previous methods analysed total variant load in a gene, but not their distribution pattern within a gene. Utilising data from 138,632 exome/genome sequences, we developed Gene Variation Intolerance Rank (GeVIR), to produce a continuous gene level metric for 19,361 genes that is able to prioritise both dominant and recessive Mendelian disease genes, outperforming missense constraint metrics and comparable, but complementary, to loss-of-function constraint metrics. GeVIR is also able to prioritise short genes, for

which loss-of-function constraint cannot be confidently estimated. The majority of the most intolerant genes identified here have no defined phenotype and are candidates for severe dominant disorders.

3.2 Introduction

Large databases of human genetic variation from large-scale genome sequencing projects such as gnomAD (Genome Aggregation Database containing 138,632 individuals)[11] are an essential resource in genomic medicine to help prioritise disease causing variants, which are expected to be rare or not present at all in healthy individuals[60]. Consequently, variant load varies greatly between genes and a deficiency of variants could be a sign of intolerance, probably due to the process of selective constraint, therefore this feature could be used to identify new disease genes. Residual Variation Intolerance Score (RVIS) was one of the first attempts to measure variant load by assuming that genes with a smaller proportion of common missense and loss-of-function (LoF) variants, could be more intolerant to variation[26]. A more recent approach used gene codon mutability to estimate the expected number of LoF or missense variants within a gene, and compare it with the observed number within large populations (Exome Aggregation Consortium (ExAC) containing 60,706 individuals)[27]. These ratios were used to calculate gene constraint metrics, missense z-score and probability of LoF Intolerance (pLI), which have been used to prioritise genes intolerant to deleterious heterozygous LoF or missense variation[12]. Both methods were also used to detect regions within a gene which showed greater intolerance to variation than the whole gene, by using functional domains and exon boundaries[30], or measuring statistical significance of variant deficiency to define region borders[29]. GnomAD v2.1 (released in October 2018[57]) contains new gene LoF and missense constraint metrics based on confidence intervals (CI) of observed over expected variants in genes, together with recalculated pLI and missense z-scores[11]. The authors suggest the use of loss-of-function observed/expected upper bound fraction (LOEUF) instead of pLI as a metric for LoF intolerant genes and provided, but did not evaluate, the same metric based on missense variant data, referred to here as MOEUF, by analogy with LOEUF[11].

The increasing size of publicly available variant databases from large predominantly healthy populations of different ethnicities, provides us with an intriguing opportunity to analyse relatively small gene coding regions devoid of variants, as an alternative to existing methods that measure variant deficiency at a gene level or within large coding regions. Since some disease causing variants are dominant, highly penetrant, and tend to cluster in functionally important coding regions[110], it is reasonable to hypothesise that genes important for human development might have regions that are so sensitive to variation they are never observed in healthy individuals. A recent study investigating regions between non-synonymous protein coding variants, referred to as Constrained Coding Regions (CCRs), in a large healthy population gnomAD (v2.0.1) found that CCRs are enriched with known pathogenic variants and could complement existing gene intolerance metrics[33]. However, only the longest CCRs that were present in a limited set of genes (6,909) were considered to be descriptive enough for gene prioritisation, and gene level metrics were not developed.

Here we propose an alternative approach to measure gene intolerance to variation and produce ranking scores, Gene Variation Intolerance Rank (GeVIR), which is based on analysing the length, evolutionary conservation and number of Variant Intolerant Regions (VIRs). Similar to CCR we define VIR as a region between non-synonymous protein coding variants in gnomAD (v2.0.1)[111], but process them differently to develop a more useful continuous gene level metric, which considers all VIRs within a gene and therefore provides scores for a much larger set of genes (19,361). We evaluate our GeVIR method by comparing it with gnomAD gene constraint metrics[57] on genes associated with Mendelian diseases[41], null genes identified in apparently healthy individuals[112], mouse[113,114] and cell essential genes[20].

3.3 Methods

3.3.1 Protein coding gene list

The gnomAD database (version 2.0.1) that was used (https://github.com/macarthur-lab/gnomad_browser) contained gene annotation from GENCODE (v19). Gene nucleotide and peptide sequences were obtained from Ensembl Biomart (human assembly GRCh37 or hg19, v93) (http://grch37.ensembl.org/biomart/martview/, accessed September 2018) and gene constraint metrics from gnomAD (https://gnomad.broadinstitute.org/, accessed October 2018). Only genes with canonical transcripts that start with methonine, end with a stop codon, with protein coding nucleotide sequence divisible by three, and without reported issues in gnomAD gene constrain metrics were used. The gene list consisted of 19,361 genes (18,326 were linked with HUGO Gene Nomenclature Committee (HGNC)[115] approved symbols (accessed November 2019) by Ensembl IDs), 1,009 of which were marked as outliers (e.g. “too many missense variants”) in the gnomAD gene constraint metrics study[11].

3.3.2 Variant Intolerant Regions (VIRs)

Variants in canonical transcripts of 19,361 genes were analysed by looking at regions between two variants (referred to here as start and stop of a VIR) predicted by Variant Effect Predictor (VEP, v85) to affect the amino acid sequence of a protein, namely: “stop_gained”, “frameshift_variant”, “stop_lost”, “start_lost”, “inframe_insertion”, “inframe_deletion”, and “missense_variant”. Only variants with filter status “PASS”, “LCR” (low-complexity) or “SEGDUP” (segmental duplication) were analysed. If start and stop variants were located in different exons, protein-coding nucleotides from multiple exons were concatenated into a single region. To take into account regions between a start codon and the first variant or the last variant and a stop codon in a gene, extra “fake” variants were added at the start and stop codons of all transcripts that were analysed. We excluded 24,276 regions in which start or stop variant consequence predicted by VEP were inconsistent with transcript exon chromosome location (e.g. “missense” variant outside exon boundaries). The remaining dataset contained 4,323,481 regions which we called Variant Intolerant Regions (VIRs).

3.3.3 VIR properties

We measured three properties of VIR: length, mean coverage (gnomAD exomes) and conservation (GERP++[77]) (Figure 3.1). VIR length and conservation were measured at the amino acid level. Regions between variants in two adjacent amino acids were included in the analysis, but had zero length and conservation score. VIR coverage was measured as the mean exome coverage of nucleotides including those affected by region start and stop variants. Exome rather than genome coverage was measured because the majority of samples in gnomAD were exome sequenced (123,136/138,632), and exome coverage tends to be less stable and hence could better highlight potential bias in the variant load. As variant absence may be a consequence of low coverage, strict filters were used to separate high- and low- covered VIRs. Autosomes, on average, had higher coverage than allosomes and different coverage thresholds were used to mark approximately 80% of the VIRs in both groups as high coverage: ≥50 for autosomes (3,427,010/4,221,872, ~81.17%) and ≥45 for allosomes (82,033/101,609, ~80.73%). Note that 1,570,941 VIRs were formed by variants in adjacent amino acids and therefore had zero length.

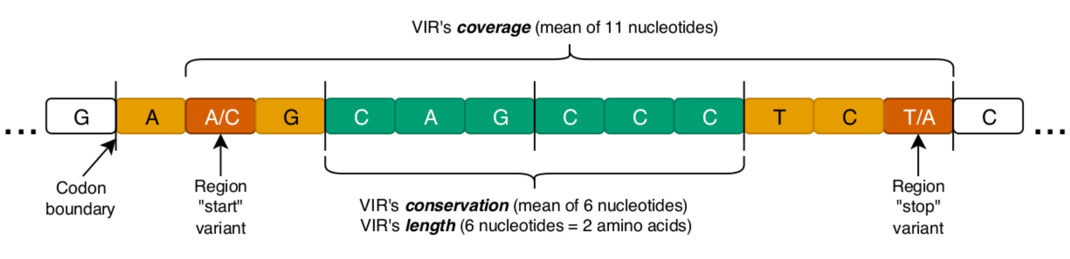

VIR's *coverage* (mean of 11 nucleotides)
...
G
A
A/C
G
C
A
G
C
C
C
T
C
T/A
C
...
Codon boundary
Region "start" variant
VIR's *conservation* (mean of 6 nucleotides)
VIR's *length* (6 nucleotides = 2 amino acids)
Region "stop" variant

Figure 3.1: Boundaries of nucleotide regions used to calculate VIR properties.

3.3.4 Pathogenic variants and evolutionary conservation of VIRs

Variants with known disease association status (e.g. “Pathogenic”, “Likely pathogenic” or “Pathogenic/Likely pathogenic”) were downloaded from ClinVar[90] (accessed August 2018) and re-annotated with Ensembl VEP (v90). Only variants predicted by VEP to be “missense_variant” (29,433), “stop_gained” (15,991) or “frameshift_variant” (21,471), and that were located in 19,361 genes with mapped VIR loci were analysed. First, 26,291 pathogenic variants were excluded as they were located in VIRs with low coverage or in VIRs with zero length (that is, regions between two variants in adjacent codons), as the latter had an evolutionary conservation score of zero. Then all VIRs with high coverage in 2,935 genes with at least one pathogenic variant were grouped by their length in five bins (1-5, 6-10, 11-15, 16-20, 21 or more amino acids), and the number and proportion of pathogenic variants that were located in VIRs in each bin were counted. Missense and loss-of-function (LoF) variants were analysed separately because the LoF variant effect on a protein was expected to be less dependent on its location in the protein sequence. To compare the pathogenic variant load in VIRs of different lengths, a “pathogenic variants per amino acid” metric was produced by dividing the number of pathogenic variants observed in VIRs in each bin by the summed length of all VIRs in that bin. To calculate 95% confidence intervals for "pathogenic variants per amino acid" metric, the analysis was repeated 10,000 times on a randomly generated set of 50% of the 2,935 genes containing pathogenic variants (that is, bootstrapping) using the resample method from Python scikit-learn module[116]. To measure correlation between VIR length and evolutionary conservation, we grouped VIRs with high coverage in 5 bins (containing 1-5, 6-10, 11-15, 16-20, 21 or more amino acids; 1,938,102 VIRs in 18,491 genes in total) and displayed the distribution of VIR conservation scores using boxplots.

3.3.5 Gene Variation Intolerance Rank (GeVIR) calculation

To develop a gene level metric based on VIRs, relative weights of VIRs based on their length were first calculated. It was assumed that longer VIRs should have larger weights, as they might be less likely to be caused by random variant distribution. The number of VIRs (with high coverage) of each length were calculated and sorted based on length, in ascending order. We then calculated the weight (*W*) of each length (*l*) based on the ratio between all VIRs and the number of VIRs with length *l* or longer:

$$W_l = \frac{\sum_{i=0}^{l_{max}} VIR_i}{\sum_{i=l}^{l_{max}} VIR_i}$$

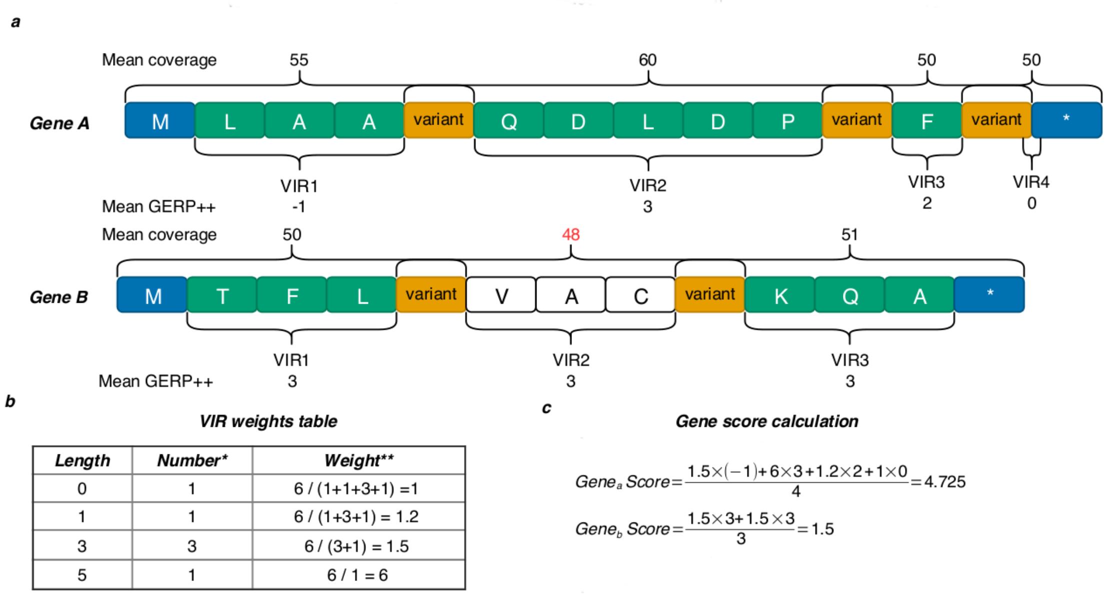
For the number and weights of VIRs of each length, see Appendix Tables B.1 and B.2 for autosomes and allosomes, respectively.

A score for each gene ($Gene_iScore$) was calculated as the sum of its high covered VIR weights (W_j) adjusted by their conservation ($GERP_j$) and normalised by the total number of regions in a gene including low covered regions ($Nregions_i$). GERP++ scores between -1 and 1, but not 0 (that is, adjacent variants or no evolutionary conservation data), were rounded to 1 or -1 to avoid extreme penalties in VIR weight *(W)* adjustment due to multiplication:

$$GERP=\begin{cases}1, & if \quad 0<GERP<1\\ -1, & if \ -1<GERP<0\\ GERP, & otherwise\end{cases}$$

$$Gene_i\,Score=\sum_j (W_j * GERP_j)/Nregions_i$$

VIRs with low coverage were counted in normalisation (*Nregions$_i$*), to avoid genes with a large proportion of VIRs with low coverage receiving high GeneScores owing to the presence of a few highly covered VIRs with large weights. GeVIR workflow is explained by example in Figure 3.2.



Length	Number*	Weight**
0	1	6 / (1+1+3+1) =1
1	1	6 / (1+3+1) = 1.2
3	3	6 / (3+1) = 1.5
5	1	6 / 1 = 6

$$Gene_a\,Score=\frac{1.5\times(-1)+6\times3+1.2\times2+1\times0}{4}=4.725$$

$$Gene_b\,Score=\frac{1.5\times3+1.5\times3}{3}=1.5$$

Figure 3.2: GeVIR workflow.

a) Two genes (A and B) containing a total of five variants (orange boxes), which divide genes A and B into four and three VIRs, respectively, six of which are valid (five shown in green and one (VIR4) with zero length). In gene B, VIR2 has low mean sequence coverage (48, shown in white) and is classified as non-valid. **b)** Calculation of valid region length weights based on their frequency in all genes (longer regions are rare and have a greater weight) *VIRs with mean coverage <50 (45 for X and Y chromosomes, analyzed separately) are not counted. **Weight is calculated as the total number of VIRs divided by the number of VIRs with this length or less. **c)** Gene score calculations for A and B; each valid region weight in a gene is multiplied by its GERP++ score, and their sum is divided by the total number of regions in the gene (valid and non-valid).

GeVIR ranking is based on these gene scores sorted in descending order. According to GeVIR, gene A is more intolerant to variants than is gene B.

All 19,361 genes were sorted based on their *GeneScores* in descending order and percentiles calculated to generate a Gene Variation Intolerance Rank (GeVIR), where a lower rank indicates higher intolerance to variation (Appendix Table B.3).

3.3.6 Disease and null and essential genes

Disease associated genes were downloaded from the Online Mendelian Inheritance in Man (OMIM) database[41] (accessed November 2018). Only associations with confident phenotype mapping number (3) were included. Only genes which were associated exclusively with dominant (phenotype keywords: "autosomal dominant") or recessive (phenotype keywords: "autosomal recessive") diseases were labeled as "AD" (n = 790) and "AR" (n = 1,585), respectively. Phenotypes that were marked as unconfirmed – "?", non-diseases – "[]" or susceptibility – "{}" were excluded, as were genes that were associated with both AD and AR diseases. A set of genes with at least two different high-confidence LoF variants found in a homozygous state in at least one individual in Exome Aggregation Consortium (ExAC) data set ("null", n = 328)[12], as well as cell essential (n = 663) and non-essential (n = 865) gene lists, according to CRISPR/Cas studies[20], were obtained from the MacArthur laboratory GitHub repository (https://github.com/macarthur-lab/gene_lists, accessed February 2019). Mouse essential genes were obtained from the MGI database via MouseMine[113,114] website (http://mousemine.org/, accessed February 2019). All genes from heterozygous mouse models with a phenotype term containing the word "lethal" and that were linked to human orthologs (http://www.informatics.jax.org/downloads/reports/HMD_HumanPhenotype.rpt, accessed February 2019) were exported (n = 388). Note that genes which were not present in our list of 19,361 protein coding genes were excluded from all gene data sets.

3.3.7 GeVIR comparison with gnomAD gene constraint metrics

GeVIR performance was compared with three gnomAD gene constraint metrics: loss-of-function observed/expected upper bound fraction (LOEUF)[11], missense observed/expected upper bound fraction (referred to here as MOEUF, by analogy with LOEUF) and missense z-score[27]. LOEUF, MOEUF and missense z-scores for the 19,361 genes were obtained from gnomAD (https://gnomad.broadinstitute.org/downloads, accessed October 2018). To convert constraint metrics into ranked lists, gene lists were sorted (LOEUF, MOEUF in ascending, missense z-score in descending order) and gene "weak" percentiles were computed, so that the most variation intolerant genes have low percentiles. To investigate possible complementarity of GeVIR and LOEUF metrics a "combined" ranked list (VIRLoF) was created, in which GeVIR and LOEUF ranks were summed for each gene, the list was re-sorted and percentiles were recalculated. We evaluated ranked lists on six datasets: AD, mouse heterozygous lethal, cell essential, AR, cell non-essential, and null. The Area Under the Curve (AUC) method from the Python scikit-learn[116] module was used to measure performance on AD, mouse heterozygous lethal, cell essential, cell non-essential and null gene datasets. Genes from the AR data set were expected to be enriched in the middle of the ranked lists. To investigate this, we divided gene ranked lists into deciles and calculated ratios between AR genes in each decile and all AR genes. As a "good" ranked gene list should prioritise AD over AR genes we evaluated balance between AD and AR genes at each cumulative percentile X. We considered AD genes with percentiles ≤X as true positives (TP), AR genes with percentiles ≤X as false positives (FP) and AD genes with percentiles >X as false negatives (FN). Based on these values we calculated precision, recall and F1 score for the AD gene class:

$$precision=\frac{TP}{TP+FP} \qquad recall=\frac{TP}{TP+FN} \qquad F1=2*\frac{precision*recall}{precision+recall}$$

The F1 score could be used as a representative metric only for approximately the first third of the genes because, after some unknown threshold, AR genes had to be prioritised over null genes, which would result in higher numbers of false positives and thus lower precision and F1 scores. Therefore, the peak F1 score was used for gene ranking method comparisons. Performance was measured using F1 score of AD genes instead of accuracy or AUC, because there were ~2 times fewer AD than AR genes in the data set. Similarity between GeVIR and other ranked gene lists was investigated by calculating the proportion of genes prioritised by both GeVIR and each other method at each cumulative percentile. The relationship between gene ranks and protein length (based on the canonical transcript) was investigated by measuring Spearman's rank correlation coefficient (SciPy[96] package) and the median protein length comparison at each decile for all ranked gene lists.

Note that the evaluation was performed on a data set that included 1,009 genes containing extremely large or small numbers of variants (that is, outliers) according to gnomAD gene constraint metrics[11], and genes without constraint metrics received the lowest ranks (1 gene in missense z-score and 482 genes in LOEUF lists). However, we included outlier genes in GeVIR score calculations to provide metrics for a greater number of genes, as some of them are well known disease genes (for example, the CFTR gene that is associated with cystic fibrosis) and their presence did not decrease the performance of GeVIR. Evaluation of gene scores on a data set excluding outliers (18,352 genes) was also carried out and showed the same overall trends.

3.3.8 GeVIR and LOEUF comparison of most variation intolerant genes

We selected 2,989/19,361 genes (~15%) that had LOEUF <0.35 (a hard threshold suggested by the authors for most LoF intolerant genes at https://macarthurlab.org/2018/10/17/gnomad-v2-1/) and the same number of genes ranked at the top of the lists obtained by GeVIR and VIRLoF. Overlapping genes were investigated by using Venn diagrams to compare all genes, AD and AR genes in the sets. To compare "prioritised gene similarity" with "method performance" we also calculated the precision, recall, and F1 score for the AD gene class. We then selected 1,317 genes present only in the GeVIR set, 1,317 genes present only in the LOEUF set, and 1,672 genes present in both sets, which were analysed by DAVID 6.8[117,118] (accessed May 2019) for functional enrichment in gene ontology (GO) terms (such as biological process, cellular component, molecular function) and biological pathways (KEGG).

3.3.9 GeVIR comparison with Constrained Coding Regions (CCRs)

Although no gene level metric was provided by the CCR study[33], the authors compared it with gene constraint metrics by sorting genes based on the number of CCRs at the 95th or greater percentile content within them. CCR data was downloaded from https://s3.us-east-2.amazonaws.com/ccrs/ccr.html (accessed December 2018), and genes present in our study with at least one CCR at the 95th percentile were selected (7,000 genes) and sorted, based on the number of CCRs, to create a ranked gene list. The same number of genes ranked at the top by GeVIR were selected; the cumulative number of AD and AR genes as well as cumulative F1 score for the AD gene class and protein length of prioritised genes, were compared.

3.4 Results

A scoring method to measure and rank gene intolerance, GeVIR, was developed by systematically analysing the coding sequence of 19,361 genes to determine the length, number, and evolutionary conservation of VIRs. In total ~3.5 millions VIRs with high sequence coverage were identified and used to generate gene ranking scores. To illustrate the hypothesis behind GeVIR, that disease genes might harbour VIRs containing no variation in healthy individuals, we analysed the distribution of pathogenic variants from ClinVar[90] and control variants from gnomAD in all the genes. Examples of variant distribution patterns in four disease genes with different modes of inheritance are demonstrated in Figure 3.3. Note that in *TCF4* and *LITAF* (both associated with autosomal dominant (AD) diseases), the majority of pathogenic missense variants, as expected, reside inside or close to relatively long VIRs, whereas in *ARSA* (associated with an autosomal recessive (AR) disease) VIRs are shorter and their location does not correlate with pathogenic variants. However, not all AD disease genes follow this pattern. For example, there are clearly visible long VIRs in *TARDBP*, but all known pathogenic variants are clustered outside them, which highlights that VIRs might be used to prioritise the AD gene, even when they cannot be used to prioritise all pathogenic variants in the gene.

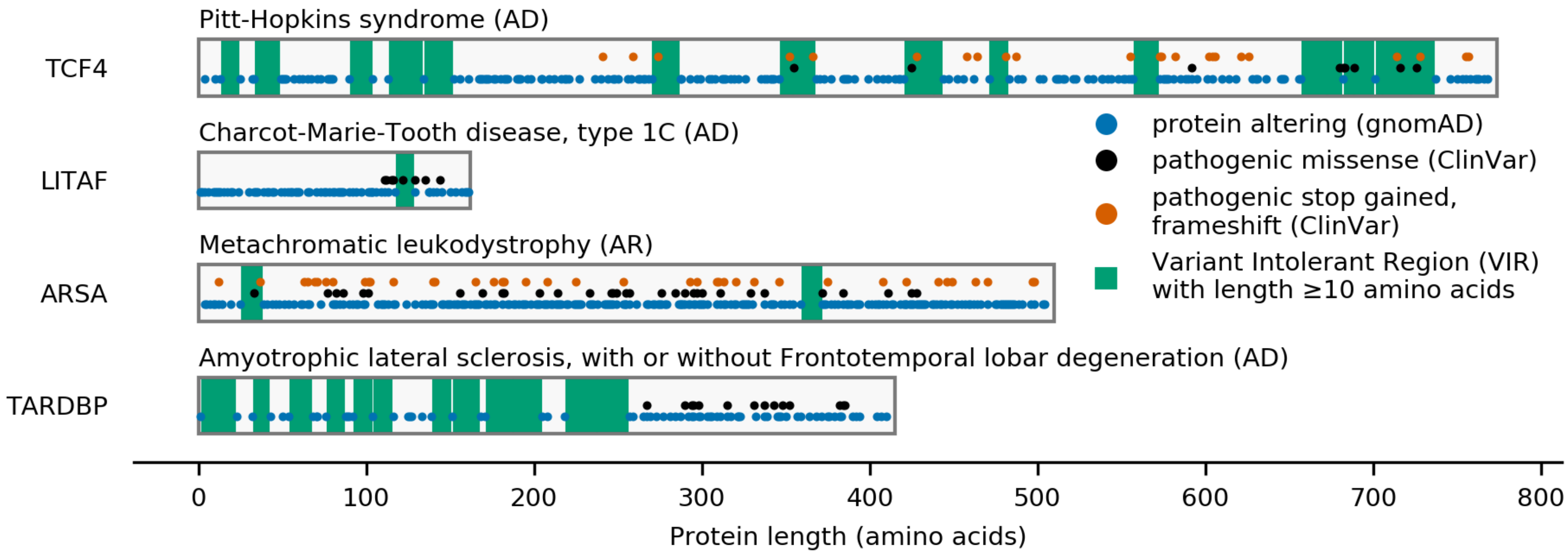


Figure 3.3: Examples of variant distribution patterns in four disease genes.

Comparison of distribution of pathogenic missense, stop gained and frameshift variants from ClinVar with non-synonymous protein coding variants from gnomAD in 4 genes TCF4, LITAF, ARSA and TARDBP associated with Mendelian diseases. Variant Intolerant Regions (VIR) can be seen as “gaps” in the genes (VIRs with length ≥10 amino acids are marked with green rectangles).

To test our hypothesis on a genome-wide scale, we first analysed the location of 66,895 known pathogenic variants (missense, stop gain and frameshift) from ClinVar in 19,361 genes with valid canonical transcripts and no reported variant calling issues in gnomAD[11]. Only 2,935 of the genes contained at least one pathogenic variant inside VIRs (40,604 ClinVar variants were located in regions which contained no non-synonymous variants in gnomAD), and were further analysed. Although ~91% of the VIRs were small (between 1 to 5 amino acids), only ~53% of pathogenic missense variants were located within them (Figure 3.4a,b). Moreover, pathogenic missense variants were observed ~3.7 times (two-sided Fisher's exact test $P < 2.23\text{E-}308$, Appendix Table B.4) more often inside the largest VIRs (length > 20 amino acids) (Figure 3.4c). In contrast, pathogenic stop gained and frameshift variants, expected to cause loss-of-function regardless of their location within a gene, were observed ~1.56 times (two-sided Fisher's exact test $P < 2.93\text{E-}29$, Appendix Table B.4) more often in the smallest VIRs than in the largest VIRs (Figure 3.4c). There was also a positive linear correlation between VIR length and evolutionary conservation, measured by mean GERP++ score[77] (Figure 3.4d). Overall our results show that long conserved regions depleted of variation are potentially a marker for some disease causing genes.

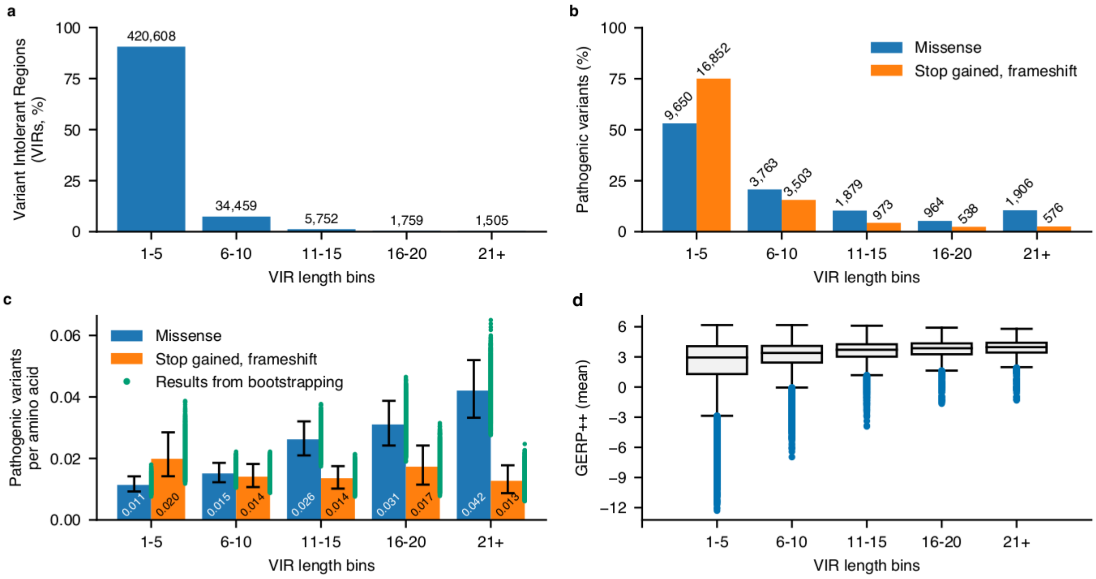


Figure 3.4: Correlation between length of variant intolerant regions (VIRs), location of pathogenic variants, and evolutionary conservation.

a) Distances in amino acids between two non-synonymous variants (i.e. VIRs) from gnomAD in n = 2,935 genes containing at least one missense or LoF (stop gained or frameshift) pathogenic variant. **b,c)** Distribution of pathogenic missense and LoF variants **(b)** inside VIRs and **(c)** normalised by cumulative region length. In **(c)** the central values represent pathogenic variants per amino acid ratio computed on n = 2,935 genes, whereas error bars represent 95% confidence interval from bootstapping (n = 10,000 iterations of resampling with replacement on 50% of 2,935 genes (i.e. n = 1,467), each iteration result is shown on overlaid dot plots). **d)** Correlation between VIR length and evolutionary conservation (GERP++ scores) of 1,938,102 VIRs with high coverage in 18,491 genes.

We developed a method to calculate gene scores based on length, conservation and number of VIRs. Briefly, we counted the number of VIRs of each length (only with high coverage) in the canonical transcripts of all genes analysed (19,361), and calculated the weight of each VIR length based on its frequency. A score for each gene was calculated as a sum of all its VIR weights, adjusted by average GERP++ score, and divided by the total number of VIRs in a gene (including those with low coverage). VIR weights were adjusted by GERP++ scores to increase the impact of regions in which absence of variation was supported by evolutionary conservation (Appendix Figure B.1 shows the distribution of VIR GERP++ scores) as well as to penalise those where lack of variation could be a result of sequencing or variant filtering errors (i.e. VIR weight is subtracted from the overall gene score). GERP scores between -1 and 1, but not 0 (i.e. adjacent variants or no evolutionary conservation data), were rounded to 1 or -1 to avoid extreme penalties from multiplication. Finally, genes were sorted using this score to create a Gene Variation Intolerance Rank (GeVIR).

To evaluate our ranking method, we considered its performance on three groups of genes with potentially high, medium and low intolerance to variation (Table 3.1, Appendix Figure B.2a-e). Genes associated exclusively with Autosomal Dominant (AD) (n = 790) and Autosomal Recessive (AR) (n = 1,585) diseases in OMIM[41] were used for high and medium variant intolerant groups, respectively. The least intolerant group (null) consisted of genes with at least two different high-confidence homozygous LoF variants which were observed in healthy population studies (n = 328)[12]. It is important to note that some apparently healthy individuals in gnomAD might be affected by some less severe disorders (e.g. cataracts), or older age onset ones, or display no/mild phenotypes in disorders with variable penetrance. GeVIR scores were also evaluated on three further groups: human orthologs of mouse heterozygous lethal knockouts (n = 388)[113,114], human cell essential (n = 663) and non-essential (n = 865) genes according to CRISPR/Cas studies[20]. GeVIR scores were compared with recently released gnomAD gene constraint scores (missense z-score, MOEUF and LOEUF)[11].

Table 3.1: Comparison of GeVIR gene ranking with gnomAD constraint metrics on 19,361 genes.

Autosomal Dominant (AD) and Autosomal Recessive (AR) gene groups consisted of genes associated with diseases with only one mode of inheritance in OMIM[41]. Human cell essential and non-essential gene groups were based on CRISPR/Cas screens[20]. Metrics performance was measured with Area Under the Curve (AUC). Assays in which lower AUC indicates better performance are marked with *. In the AD/AR classification assay, AD class F1 score is calculated at each percentile (cumulative) considering AD genes as true positives and AR genes as false positives and performance is reported as peak F1 score with gene percentile in brackets. Results of metrics which outperformed others were highlighted in bold. If VIRLoF showed the best results, then the second best performing metric was also highlighted in bold.

Assay	Gene number	GeVIR	GeVIR (without GERP++)	Missense z-scores	MOEUF	LOEUF	VIRLoF
AD	790	**73.54**	72.80	68.72	68.21	72.40	**74.38**
Mouse het lethal knockout	388	74.92	74.10	69.71	69.45	**76.03**	**77.18**
Cell essential	663	**72.62**	69.97	66.81	68.22	68.70	72.13
Nulls*	328	27.10	32.88	**24.69**	27.89	31.18	27.38
Cell non-essential*	865	29.61	37.07	**27.26**	27.61	27.92	**27.08**
AD/AR classification	790 AD, 1,585 AR	62.34 (29.94)	61.48 (25.44)	59.13 (35.40)	60.27 (32.82)	**64.15** **(27.17)**	**65.74** **(25.84)**

First, GeVIR score performance on the most and least intolerance groups was measured using Area Under the Curve (AUC). High AUC means that genes from a group are located closer to the top in the ranked list, whereas low AUC represents the opposite trend. GeVIR outperformed both missense metrics (missense Z score and MOEUF) at prioritising AD (AUC = 73.54%), mouse heterozygous lethal (AUC = 74.92%) and cell essential genes (AUC = 72.62%), by ~4-6% AUC, whereas LOEUF showed a similar performance to GeVIR on AD (AUC = 72.40%) and mouse heterozygous lethal genes (AUC = 76.03%), but was worse at prioritising cell essential genes (AUC = 68.70%). GeVIR also showed good performance at deprioritising null genes (AUC = 27.10%) and was outperformed only by missense z-scores (AUC = 24.69%), due to better ranking in the second half of the gene list. Interestingly, LOEUF was the worst in this assay despite being based on LoF variant data (AUC = 31.18%). However, GeVIR was also slightly outperformed by the other metrics at deprioritising cell non-essential genes (AUC = 29.61%), with missense z-scores showing the best results (AUC = 27.26%). These results should be interpreted with caution since some of the genes in null and cell non-essential groups can still be associated with disease. In fact, despite opposite trends in ranking cell essential and non-essential genes, there was no statistically significant enrichment or depletion of AD or AR genes (two-sided Fisher's exact test $P > 0.05$, Appendix Table B.5) in both groups when compared to overall proportions (AD = ~4.1%, AR = ~8.2%). In comparison, there was ~6 fold enrichment of AD genes in mouse heterozygous lethal genes (P = 1.95E-37), and no enrichment of AR genes (P = 0.22). Surprisingly, the Null gene group was depleted for AR (P = ~0.0005), but not AD genes (P = 0.78). This shows that gene essential/non-essential status deduced from *in vitro* cell culture assays[20] might indicate the severity of a possible phenotype, but not its inheritance pattern, probably because CRISPR/Cas experiments are typically performed to generate complete (homozygous in diploids, null in haploids) gene knockouts[21]. Moreover, CRISPR/Cas experiments do not investigate the broad effect of gene loss on the health of a whole organism, or possible gain of function mechanisms, which might be another cause of AD diseases in genes tolerant to homozygous LoF variants.

Second, GeVIR performance at distinguishing AD from AR genes was evaluated. To do this, AD genes were considered as true positives, AR genes as false positives and an F1 score for AD class was calculated cumulatively at each percentile (i.e. consider all genes with ≤ percentile rank), which would represent the balance between the number of correctly prioritised AD and incorrectly prioritised AR genes (Table 3.1, Appendix Figure B.2g). In the first quartile, GeVIR and LOEUF performed similarly and surpassed both missense scores. In the second quartile, LOEUF showed slightly better results with a peak F1 score of 64.15% at the 27th percentile, whereas for GeVIR the peak F1 score was 62.34% at the 30th percentile. After the 40th percentile, MOEUF had higher F1 scores, because it did not prioritise AR genes after AD genes. Therefore, LOEUF performed slightly better than GeVIR at distinguishing AD from AR genes, however, GeVIR score is based on all protein altering variants, the majority of which are missense, and GeVIR performed better than both missense scores.

Third, since AR genes are expected to be less intolerant than AD genes, they should be enriched in the middle of the ranked gene list. To assess this, AR gene distribution was analysed in deciles (Figure 3.5a, Appendix Table B.6), since the ideal threshold which would separate them from most and least intolerant genes is not known. The null hypothesis is that each decile of all genes analysed (~1,936) should contain approximately 10% of all AR genes (~158). Our results showed that up to the 3rd decile all scoring methods performed similarly (percent content of AR genes was increasing), but after this threshold the trends were different. GeVIR and LOEUF, prioritised AR genes with on average ~1.77 (two-sided Fisher's exact test P = 1.61E-34) and ~1.72 (P = 2.99E-31) fold enrichment at the 40th - 60th percentiles, respectively, whereas missense z-score and MOEUF fold enrichment was only ~1.14 (P = 0.015) and 1.28 (P = 8.86E-7). In the last 30 percentiles, GeVIR and LOEUF had on average ~2.2 (P = 9.74E-32) and ~1.9 (P = 1.79E-22) times fewer AR genes than expected, respectively, whereas missense z-score and MOEUF had around the expected number of AR genes ($P > 0.05$), but MOEUF had ~1.7 (P = 5.34E-7) less than the expected in the last decile. Regarding GeVIR performance on null genes, this analysis showed that GeVIR performed best at distinguishing them from AR genes (i.e. AR genes were ranked closer to the middle), whereas existing missense metrics cannot be used to effectively rank AR genes.

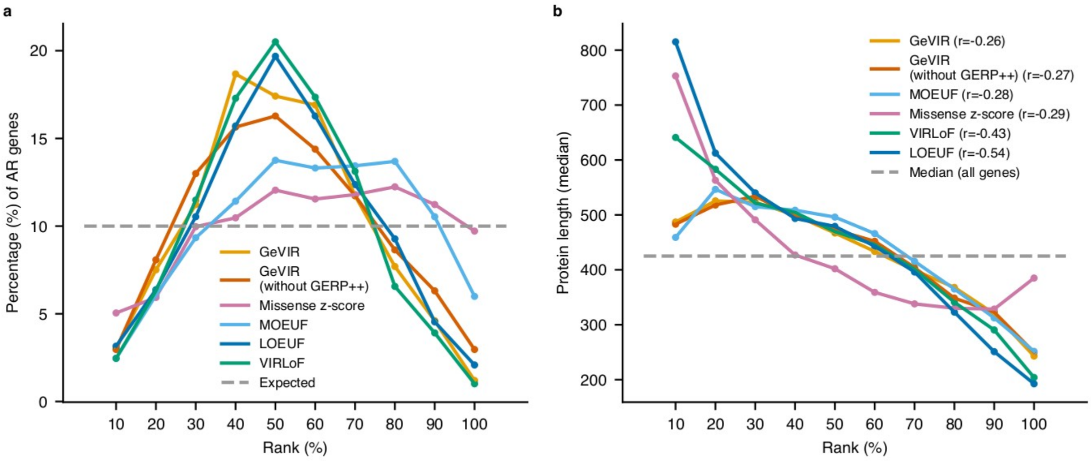


Figure 3.5: Comparison of GeVIR gene ranking with gnomAD constraint metrics on 19,361 genes.

a) Percentage of genes associated exclusively with Autosomal Recessive (AR) diseases out of all AR genes in OMIM (n = 1,585) in each rank decile. **b)** Median protein length (amino acids) in each rank decile. Correlation between protein length and gene rank was measured with Spearman's rank correlation coefficient.

Fourth, all gene scoring methods, to some extent, tend to prioritise longer genes and an analysis of protein length of the ranked genes showed that LOEUF was the most biased towards longer genes (Spearman r = -0.54), especially in the first decile where median protein length was ~1.91 times longer than overall median (425 amino acids) (Figure 3.5a, Appendix Figure B.3 represents this data with notched boxplots). Overall GeVIR was the least biased method (r = -0.26), with genes in the first decile being ~1.15 times longer than expected. Note that the evaluation was performed on a dataset which included 1,009 genes containing extremely large or small numbers of variants (i.e. outliers) according to gnomAD gene constraint metrics[11], and genes without constraint metrics received the lowest ranks (1 gene in missense z-score and 482 genes in LOEUF lists). However, we included outlier genes in GeVIR score calculations to provide metrics for a larger number of genes, since some of them are well known disease genes (e.g. *CFTR* associated with cystic fibrosis) and their presence did not decrease the performance of GeVIR. Evaluation of gene scores on a dataset excluding outliers (18,352 genes) was also carried out and showed the same overall trends (Appendix Figure B.4).

Fifth, to investigate the impact of evolutionary conservation on GeVIR ranks, a metric without GERP++ adjustments was computed and evaluated in the same assays (Table 3.1, Appendix Figure B.2). Exclusion of GERP++ weights had a minor impact on prioritising AD and mouse heterozygous lethal genes (<1% AUC difference) and, despite decrease in prioritising cell essential genes (-2.65% AUC), it still slightly outperformed all existing methods in this assay (~1-3% AUC). Although, GeVIR without GERP++ performed worse at deprioritisation of Null (+5.78 AUC) and cell non-essential genes (+7.46 AUC), it still had an enrichment of AR genes at the 40th - 60th percentiles (~1.5 times, P = 2.27E-19, Figure 3.5a). Finally, it performed slightly better at AD/AR gene classification than missense metrics (~+1-2% F1 score) and was less biased by gene length than all other metrics except GeVIR (with GERP++) (Figure 3.5b). Null and cell non-essential genes probably had shorter VIRs in general due to larger number of variants and consequently their weights relied more on GERP++ adjustments. Another reason for decreased performance were genes in which lack of variation were probably caused by sequencing or variant filtering errors. Overall, evolutionary conservation adjustments are an important part of our GeVIR method, but nevertheless, it still works well without it and is superior to existing methods which estimate expected number of missense variants (MOEUF, missense z-score).

Finally, since GeVIR was predominately based on missense variants data and, unlike other missense scores, ranked AR genes similarly to LOEUF with an enrichment at the 40th - 60th percentiles, we investigated whether combination of GeVIR and LOEUF result in a more optimal gene ranking. The combined metric (VIRLoF) was created by adding GeVIR and LOEUF ranks for each gene and resorting the resulting ranked list. The combined ranking outperformed both GeVIR and LOEUF at priorotising AD, mouse heterozygous lethal, and Null genes (Table 3.1, Appendix Figure B.2a,b,d). Moreover, VIRLoF prioritised AR genes closer to the middle of the list than either GeVIR or LOEUF (Figure 3.5a) and, consequently, was the most efficient at distinguishing between AD and AR genes (peak F1 score of 65.74% at the 26th percentile Table 3.1, Appendix Figure B.2g). This shows that GeVIR could be used with LOEUF to improve disease gene prioritisation.

The combined rank of GeVIR and LOEUF (VIRLoF) outperformed all other variant based gene constraint metrics in most assays, however GeVIR and LOEUF had the lowest proportion of genes in common amongst the top ranked genes (Appendix Figure B.2h). To investigate this difference we selected 2,989 genes (~15%) with the lowest LOEUF (< 0.35; a recommended hard threshold for most intolerant genes[57]), and the same number of genes ranked as the most intolerant by GeVIR and VIRLoF. At this threshold, only 1,637/4,306 (~38.8%) of the genes prioritised by GeVIR or LOEUF were in common (Figure 3.6a). The two methods agree more on AD genes (250/452, 55.3%) (Figure 3.6b), than on AR genes (35/168, 20.8%) (Figure 3.6c). This explains why the combined method (VIRLoF) showed a better performance with ~82% precision and ~48% recall, despite GeVIR and LOEUF having nearly identical performance with ~77-78% precision and ~44-45% recall (Figure 3.6d). Therefore, genes considered highly intolerant to variation by GeVIR were different from those prioritised by LOEUF.

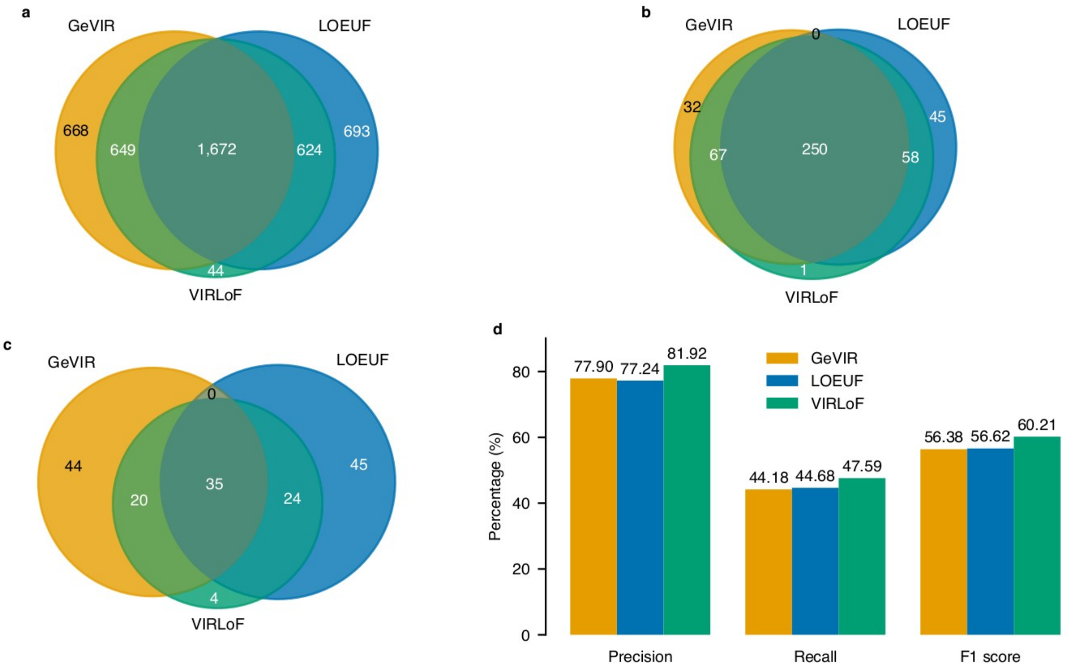


Figure 3.6: Comparison of GeVIR, LOEUF and VIRLoF performance on the most variant intolerant genes.

2,989/19,361 genes (~15%) with upper observed/expected Loss-of-Function (LOEUF) scores < 0.35 (blue), the same number of genes with the highest GeVIR (orange) and VIRLoF (green) scores. Venn diagrams show overlaps between **(a)** all 2,989 genes, **(b)** genes exclusively associated with Autosomal Dominant (AD), or **(c)** Recessive (AR)

OMIM genes. **d)** Method performance is compared, considering AD genes as True Positives and AR genes as False Positives, using precision, recall and F1 metrics of AD class.

To understand which categories of genes were prioritised by GeVIR compared with LOEUF, "functional annotation term" enrichment analysis was performed using DAVID v6.8[117,118] on 3 gene lists: 1,317 genes prioritised only by GeVIR, 1,317 genes prioritised only by LOEUF and 1,672 genes prioritised by both (Figure 3.6a). Table 3.2 presents the most significantly enriched Gene Ontology (GO) biological process terms and KEGG pathways (see Appendix Table B.7 for all results). Genes prioritised only by GeVIR were involved in mRNA processing, positive regulation of macromolecular metabolic processes, and MAPK signaling pathways, whereas genes prioritised only by LOEUF were involved in multicellular organism development, especially neuron development. Genes prioritised by both GeVIR and LOEUF, similar to GeVIR alone, were involved in regulation of macromolecular metabolic processes, various RNA related processes, and MAPK signaling pathways. These latter genes were also involved in histone modification and dopaminergic synapse, a trend not observed with genes prioritised by either GeVIR or LOEUF alone. Genes prioritised only by GeVIR were generally shorter than those prioritised by both GeVIR and LOEUF (Table 3.2), even if they were involved in the same biological pathways (e.g. spliceosome – only GeVIR: 42 genes, median length = 221 (amino acids); GeVIR and LOEUF: 51 genes, median length = 559), whereas genes prioritised only by LOEUF were the longest (e.g. multicellular organism development – 506 genes, median length = 966). Therefore, GeVIR might prioritise potentially important short genes for which LOEUF might not be confidently estimated due to an insufficient gnomAD population size (an acknowledged limitation of LOEUF[11]). Mouse heterozygous lethal (MHL) and cell essential (CE) genes were significantly more enriched in genes prioritised by both GeVIR and LOEUF (MHL: Genes = 121/1,672, Fold Enrichment (FE) = 3.6, two-sided Fisher's exact test *P* = 7.42E-27. CE: Genes = 162/1,672, FE = 2.8, *P* = 3.07E-25), than in genes prioritised only by GeVIR (MHL: Genes = 52/1,317, FE = 2.0, *P* = 3.13E-5. CE: Genes = 100/1,317, FE = 2.2, *P* = 5.07E-11) or LOEUF (MHL: Genes = 59/1,317, FE = 2.2, *P* = 2.82E-7. CE: Genes = 62/1,317, FE = 1.4, *P* = 2.52E-2), when compared to expected proportions in the 19,361 studied genes (388 MHL and 663 CE genes). This result shows that genes intolerant to both LoF and missense variants are more likely to be essential, and confirms that GeVIR and LOEUF metrics complement each other.

Table 3.2: The most significantly enriched Gene Ontology (GO) terms (Biological Process) and KEGG pathways in genes prioritised by GeVIR, LOEUF or both. The functional enrichment analysis was performed using DAVID 6.8, statistical significance was reported by False Discovery Rate (FDR). Genes (%) is a number and proportion of genes from the analysed list found in DAVID and associated with corresponding GO term or KEGG pathway. Mean, median and standard deviation (SD) of protein length were calculated separately (see Appendix Table B.7 for full report).

Gene group	Source	Term	Genes (%)	Fold enrichment	FDR	Protein length (amino acids)		
						Mean	Median	SD
Only GeVIR	GO	mRNA metabolic process	130 (10.0)	2.72	6.36E-23	300	223	240
Only GeVIR	GO	Positive regulation of macromolecule metabolic process	333 (25.6)	1.64	2.77E-19	400	355	280
Only GeVIR	GO	Protein targeting to membrane	57 (4.4)	4.21	1.63E-17	241	151	224
Only GeVIR	KEGG	Spliceosome	42 (3.2)	3.64	1.66E-10	295	221	253
Only GeVIR	KEGG	Ribosome	38 (2.9)	3.22	1.64E-07	144	137	62
Only GeVIR	KEGG	MAPK signaling pathway	48 (3.7)	2.19	3.75E-04	358	339	180
Only LOEUF	GO	Multicellular organism development	506 (38.8)	1.48	1.09E-22	1,228	966	1,749
Only LOEUF	GO	Neuron projection morphogenesis	99 (7.6)	2.59	2.87E-15	1,234	986	772
Only LOEUF	KEGG	Axon guidance	27 (2.1)	3.09	4.43E-04	1,106	1,030	360
Only LOEUF	KEGG	Focal adhesion	34 (2.6)	2.40	4.50E-03	1,450	1,360	708
GeVIR & LOEUF	GO	Regulation of macromolecule metabolic process	928 (55.9)	1.72	2.34E-95	829	630	631
GeVIR & LOEUF	GO	Cellular macromolecule metabolic process	1,166 (70.2)	1.50	8.82E-95	832	640	644

GeVIR & LOEUF	GO	Histone modification	152 (9.2)	3.72	5.35E-46	1,130	805	917
GeVIR & LOEUF	GO	RNA splicing	126 (7.6)	3.39	8.21E-33	687	568	406
GeVIR & LOEUF	KEGG	Dopaminergic synapse	53 (3.2)	3.52	4.31E-14	805	525	619
GeVIR & LOEUF	KEGG	Spliceosome	51 (3.1)	3.26	9.09E-12	713	559	460
GeVIR & LOEUF	KEGG	mRNA surveillance pathway	40 (2.4)	3.73	6.13E-11	738	583	571
GeVIR & LOEUF	KEGG	MAPK signaling pathway	73 (4.4)	2.45	2.09E-10	896	697	617

To compare GeVIR with recently published CCRs[33] at a gene level, we first excluded genes not analysed in the CCR study (18,012 genes remained). Although, no gene level metrics were produced, the authors compared their method with other gene constraint metrics (pLI and missense z-score) by ranking their genes based on the number of CCRs at ≥95th percentile[33]. We used the same approach and compared the top 7,000 genes with CCRs ≥ 95% with the same number of genes with the highest GeVIR scores to understand which metric is more efficient at prioritising AD genes. Although GeVIR and CCR methods both initially prioritised a similar number of AD genes, GeVIR started outperforming CCR after the first ~3,500 genes (Figure 3.7a). Moreover, GeVIR prioritised fewer AR genes, considered as false positives in this analysis (Figure 3.7b). Consequently, GeVIR also had a higher peak F1 score of 62.38% at the 32th percentile compared with a CCR peak F1 score of 57.25% at the 38th percentile (Figure 3.7c). Note that GeVIR also prioritised AR genes over Null genes after the 40th percentile of the full list of 19,361 ranked genes (Figure 3.5a), which explains the decrease in F1 score observed after the 28th percentile. The remaining 11,012 genes in common (i.e. ~61%) contain only CCRs below the 95th percentile, and a method to rank them was not provided by the authors[33], which makes the CCR ranking approach less universal than GeVIR. Moreover, CCR ranking was significantly biased towards longer genes since the suggested approach of counting the number of CCRs does not involve normalisation by gene length (Figure 3.7d). In this assay, the difference in performance between GeVIR with and without GERP++ adjustments were minor (e.g. 0.7% F1 score decrease Figure 3.7c), which highlights that GeVIR is a novel approach of using variant distribution data to assess gene constraint. Therefore, GeVIR outperforms CCRs at prioritising the most intolerant genes, and provides ranks for a much larger list of genes with less size bias.

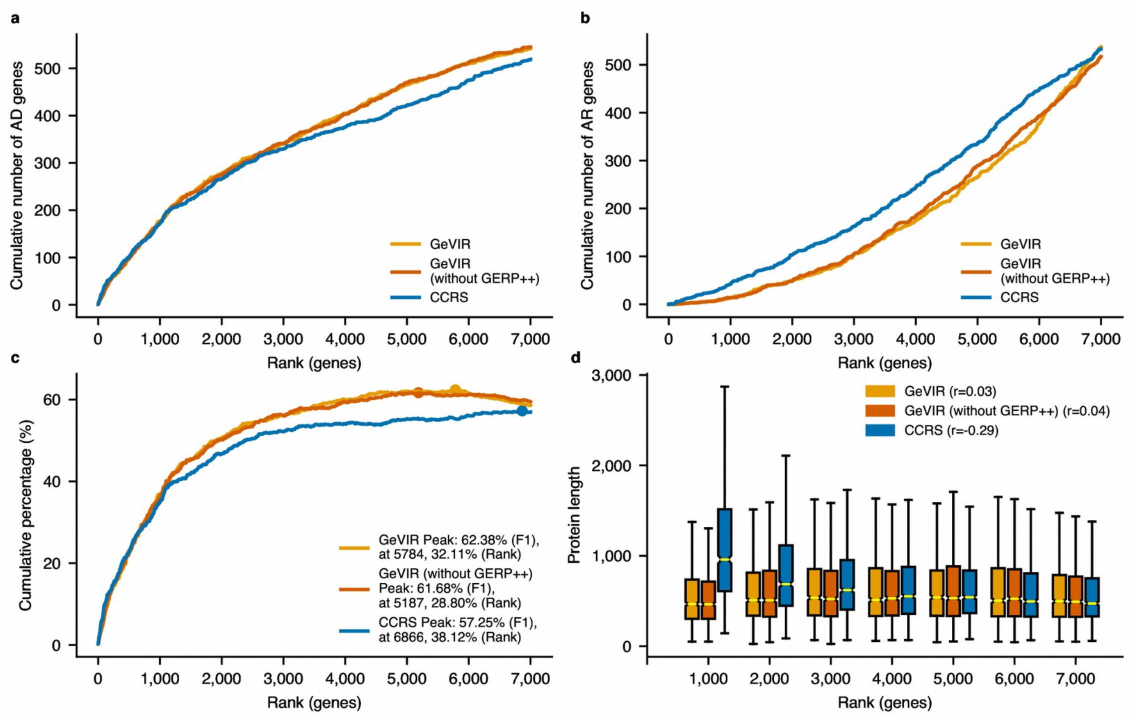


Figure 3.7: Comparison of top genes ranked by GeVIR with a list of genes sorted by number of CCRs at 95th or greater percentile (7,000 genes).

a) Cumulative number of genes associated exclusively with AD diseases in OMIM (n = 770). **b)** Cumulative number of genes associated exclusively with AR diseases in OMIM (n = 1,553). **c)** AD class F1 score calculated at each subset of top genes (cumulative) considering AD genes as true positives and AR genes as false positives. **d)** Gene canonical transcript protein length in each thousand ranked genes (that is 1–1,000, 1,001–2,000 ... 6,001–7,000). Standard notations are used for elements of the box plot (that is, upper or lower hinges: 75th or 25th percentiles; inner segment: median, notches are calculated using a Gaussian-based asymptotic approximation; and upper or lower whiskers: extension of the hinges to the largest or smallest value at most 1.5 times of interquartile range). Outliers are not shown due to the presence of genes with extreme protein length (for example TTN, ~36,000 amino acids) in the data set, which would distort the figure. Correlation between protein length and gene rank was measured with Spearman’s rank correlation coefficient.

3.5 Discussion

Large-scale genome analysis has shown that genes in the human genome differ in their ability to tolerate disruptive genetic variation, and identifying those involved in disease is not a trivial task. Different methodologies have been used to classify genes based on their mutation intolerance, many are frequently used to aid interpretation of patient genome sequencing data, for example ExAC pLI and missense z scores. In this study we present an alternative method, GeVIR, to measure gene variation intolerance based on analysing variant distribution in large populations obtained from the gnomAD database, and show that it outperforms existing missense constraint metrics. Incorporation of evolutionary conservation data (GERP++) into GeVIR significantly improved the model, especially its ability to rank less variation intolerant genes (e.g. AR and null groups). In fact, GeVIR performance was comparable to existing LoF constraint metrics, but since these methods prioritised substantially different genes, we found that they were complementary and could be combined to produce a more efficient score (VIRLoF) for ranking genes according to their intolerance to variation. Comparisons with other commonly used methods also revealed that GeVIR and LOEUF were best at ranking AR genes closer to the middle of the gene list, distinguishing them from AD and null genes. Thus, these metrics allow genes to be prioritised in general, not only to separate the most intolerant genes from all others, thereby producing a spectrum of intolerance.

Although GeVIR is based on the same principles as the CCR method[33], published during preparation of data for this study, GeVIR can be used to rank a larger list of genes since it takes into account variant distribution patterns in the whole gene and, consequently, is able to some extent to prioritise AR over null genes. Moreover, GeVIR prioritises the most intolerant genes more precisely, so fewer AR genes reside amongst top ranked genes, which suggests that GeVIR detects fewer false positive “important” regions (i.e. those due to low coverage or sequencing issues) than the CCR method. An explanation for this might be the different methods used to analyse the regions, which, in GeVIR, include higher coverage thresholds, exclusion of non-canonical transcripts, evolutionary conservation adjustments and normalisation by total number of regions in a gene. Our analyses show that it is important to differentiate between gene (e.g.

GeVIR, LOEUF) and variant (e.g. CCR) prioritisation methods. Approximately half of the ClinVar pathogenic missense variants studied were located in short regions (1-5 amino acids), however, gene variation intolerance metrics can still be used to prioritise these disease genes. Moreover, as seen with the gene *TARDBP* (GeVIR = ~1%; Figure 3.3), even in genes with multiple long VIRs all known disease causing variants could be located outside them. We could speculate that deleterious changes inside such regions might result in more severe or even lethal phenotypes, a hypothesis that warrants further investigation in conditions such as AD amyotrophic lateral sclerosis. This highlights that VIRs might be used to prioritise AD genes, even when they cannot be used to prioritise all pathogenic variants in the gene, and the overall pattern of distribution should also be taken into account. Therefore, candidate missense variants located inside long VIRs/CCRs might be interpreted as deleterious, but the opposite assumption should not be made.

Existing gene scores based on exome LoF variant data (pLI) are biased towards longer genes[38], which was also confirmed in this study for the more recent constraint metric LOEUF[11]. However, GeVIR showed a potential to prioritise short genes, which might have important biological functions (e.g. mRNA metabolic processes). Similar to other constraint metrics based on variant data, GeVIR is not biased towards more studied genes, a criticism known to be associated with gene scores that rely on gene networks[35]. Previously, ExAC pLI scores were created using unsupervised clustering algorithms based on three major gene categories: haploinsufficient, recessive, and null[12]. Although GeVIR and LOEUF did not use any prior knowledge about gene classes, and therefore might result in a more “natural” gene ranking, both methods supported this assumption. However, there are currently no supervised gene classification models that can classify genes into these three categories. In the future, GeVIR and LOEUF might be used as features for such models, as we have demonstrated their applicability for this task and ability to complement each other by combining them into one metric, VIRLoF.

The key assumption of GeVIR is that the longer the region devoid of any protein altering variants within a gene, the more intolerant it is, however, there are limitations to this assumption. Although we considered VIR coverage and evolutionary conservation in GeVIR calculations, some of the well covered and conserved VIRs might still be a result of sequencing errors. Even the presence of a single sufficiently long VIR in a gene might be enough for GeVIR to predict it as highly intolerant to variation (i.e. false positive). Improvements in sequencing technologies and variant filtering software should reduce these issues in the future. GeVIR might underestimate intolerance of genes associated with phenotypes which could be present in gnomAD population (e.g. old-age onset, cancer)[59], the same as other gene constraint metrics derived from the same data[11]. Since, long unaffected regions (e.g. >20 amino acids; 3,389 VIRs detected in this study) are significantly less common than medium length regions (e.g. 8-12 amino acids; 56,117 such VIRs detected here), they have a much larger impact on GeVIR scores. Therefore, GeVIR might underestimate intolerance of functionally important genes with rare and sparse variants in healthy individuals (i.e. false negative). This limitation might be overcome in future by more rigorous selection of "healthy" individuals in exome and genome sequencing population databases.

The number of sequenced exomes and genomes is rapidly growing, but the effect of sample size on our method is hard to predict. We expect that VIRs in a larger database will be smaller due to the larger number of variants, some of which might occur in important regions due to reduced penetrance or presence of some affected individuals in the population. However, region length weights used to calculate GeVIR are relative and we also expect that in a larger database all regions between variants will be smaller and thus relative weights might not change significantly. Nevertheless, in the future our method might require some adjustments to region weights based on bordering variant allele frequencies (AF) or exclusion of singletons, if a population is sufficiently large. However, this limitation is also true, to some extent, for other gene or regional variation intolerance metrics, which utilise missense variant data and, at the moment, do not account for variant AF[27,33]. Regardless, our study shows that the analysis of variant distribution patterns allows better estimation of gene intolerance than variant load in the case of missense variants.

In conclusion, we show that GeVIR can be used to prioritise candidate genes intolerant to missense variants, especially if they are short. For example *LITAF*, a short protein coding gene (162 amino acids) associated with the AD Charcot-Marie-Tooth disease type 1C caused by missense variants (Figure 3.3), is ranked ~35% by GeVIR, which is ~2 times closer to the top most intolerant genes than any of the current gnomAD gene constraint metric rankings (missense z-score = ~67%, MOEUF = ~72%, LOEUF = ~72%). The LoF metric LOEUF is recommended to prioritise candidate genes intolerant to LoF variants, especially if they are large, although, GeVIR might be more useful when investigating LoF in short genes. VIRLoF can be used when a single metric is required, as it shows the best performance of all the variant-based gene constraint metrics assessed. In fact, the top 10% most intolerant genes according to VIRLoF (1,936 genes) contain ~37% (291/790) of known AD genes, and only ~2.5% (39/1585) of known AR genes (two-sided Fisher's exact test FE = ~15, *P* = 3.52E-84). This suggests that heterozygous *de novo* deleterious variants in these genes are likely to result in pathogenic phenotypes. Although ~70% of these genes (1,357/1,936) are not yet linked to any phenotype in OMIM, they are significantly enriched (FE = ~4.1, *P* = 4.74E-12) with mouse heterozygous lethal genes (43/1,357), compared with unknown genes in the remaining 90% of the genes (108/13,981). We hypothesise that genes which are intolerant to both missense and LoF variants, are crucial for human development and that deleterious variants in them may result in severe dominant disorders or embryonic lethality. The availability of a continuous gene variation intolerance ranking system based on variant distribution should aid interpretation of genome sequencing data in a clinical setting, and progress human disease gene discovery.

Chapter 4

DIP: using machine learning to classify genes across the spectrum of disease inheritance patterns

4.1 Abstract

Metrics that predict whether loss or disruption of one or both gene copies can cause disease are useful to aid the discovery of novel autosomal dominant (AD) and recessive (AR) diseases and associated genes. Recent gene constraint studies showed that gene intolerance to variation continuously correlated with associated disease inheritance patterns, but the most tolerant genes were deficient in genes associated with dominant or recessive diseases and, therefore, could be segregated into a third "non-disease" causing gene group. However, existing supervised machine learning solutions to this problem were built on the assumption that genes had to be classified into two groups (e.g. AD/AR or haploinsufficient/haplosufficient), and predictions were difficult to interpret in ambiguous cases with no clear mode of inheritance. Here, we present a novel gene level metric that continuously ranks 15,794 autosomal genes by Disease Inheritance Patterns (DIP), which was developed by combining multiple supervised machine

learning models with gene variation intolerance metrics. DIP performance was comparable with existing metrics in distinguishing AD from AR genes. However, it more effectively prioritised disease genes in general, resulting in a more optimal ranking of genes across the spectrum of disease inheritance patterns. The first and last five percentiles were significantly enriched with AD and cell non-essential genes (approximately 4.6 and 4.2 times, respectively), whereas AR genes were 3 times more frequently seen in the middle ranks (41-51%). Although, perfect categorical classification of genes by mode of inheritance might not be possible, continuous metrics can provide a better estimation of a gene's predisposition to certain modes of inheritance, especially in ambiguous cases.

4.2 Introduction

The large amount of data produced by next generation sequencing technologies gave impetus to the development of computational methods that could facilitate its interpretation and aid discovery of novel disease genes. Initially, the research was focused on the development of methods that could prioritise candidate pathogenic variants amongst the thousands of rare variants observed in every individual[119,120]. However most of these methods were predicting variant deleteriousness regardless of disease mode of inheritance (i.e. dominant or recessive) and, therefore, additional metrics were required to further prioritise candidate pathogenic variants to facilitate discovery of novel disease genes[17]. Although inheritance pattern is a property of disease-causing variants, it can be predicted using gene-level properties[17]. Consequently, genes can be classified as dominant/recessive, which indicates the predisposition of candidate pathogenic variants in these genes to these inheritance patterns[17].

Recent statistical methods, that measured gene intolerance to variation using data from a large population database, has shown that genes can be ranked continuously and highlighted the existence of at least three major groups of genes: dominant (including haploinsufficient genes), recessive, and tolerant genes (i.e. unlikely to be associated with diseases, e.g. olfactory genes)[11,121]. An alternative three-group gene classification strategy was suggested by Pengelly *et al.*[22] who hypothesised that disease genes might have medium essentiality, whereas the most and the least essential genes might not be associated with diseases since deleterious variants in them might be lethal and tolerated, respectively. However, previous studies that used supervised machine learning methods for gene classification aimed to classify all genes into only two groups (e.g. haploinsufficient and haplosufficient or autosomal dominant and recessive)[3,15,17,35–38], and were reviewed elsewhere[25]. Some studies developed separate models to prioritise genes with different disease inheritance patterns, but these scores were not integrated into a single continuous metric[3,36]. Moreover, until recently, potentially non-disease genes were not analysed as a separate class[3]. Although methods to prioritise haploinsufficient or dominant genes might to some extent assign lower scores to non-disease genes rather than recessive genes, this behaviour should not be assumed since the models were not optimised for this task. Therefore, metrics provided by existing supervised machine learning gene classification methods could be used to distinguish one group of genes from all others but, unlike statistical variation intolerance metrics, they might not be continuous.

Here, we show that continuous gene ranking by Disease Inheritance Patterns (DIP) can be achieved by combining two supervised models trained to distinguish autosomal dominant from recessive genes (ADR), and disease from non-disease genes (DND). DIP scores were compared with existing supervised machine learning methods developed to prioritise autosomal dominant and Mendelian disease genes, DOMINO[17] and Gene Pathogenicity Prediction (GPP)[3] scores, respectively. Finally, we discuss the issues highlighted by comparison of supervised machine learning models with statistical gene variation intolerance metrics (VIRLoF[11,121]), that led to our decision to use the latter to resort genes predicted to be AD by ADR by DIP gene ranking.

4.3 Methods

4.3.1 Gene datasets

The list of autosomal protein-coding genes was obtained from the DOMINO study (n = 17,898, obtained from https://wwwfbm.unil.ch/domino/download.php, version February 2019, accessed March 2020)[17] and mapped to the official gene symbols according to HUGO Gene Nomenclature Committee (HGNC, accessed March 2020)[122]. Gene symbols that were not found in HGNC (official, alias and previous symbols were checked) or those that could be mapped to multiple official symbols were excluded (n = 17,857 genes remained, termed the “main” gene dataset). Genes in the main dataset were annotated with labels from the following sources.

The model developed to distinguish disease from non-disease genes (DND) was trained on the disease (n = 588) and non-disease (i.e. loss-of-function tolerant, n = 587) genes from the Gene Pathogenicity Prediction (GPP) training dataset[3]. The model developed to distinguish autosomal dominant from recessive genes (ADR) was trained and validated on the sets of autosomal dominant (AD, n = 291 and 26, respectively) and autosomal recessive (AR, n = 693 and 73, respectively) from the DOMINO training and validation datasets[17]. Disease Inheritance Patterns (DIP) scores were calculated for genes in the main set, that were not used to train or validate DND and ADR models (n = 15,794, termed the “evaluation” gene dataset). However, when DIP performance was compared with DOMINO, GPP and VIRLoF, only genes from the evaluation dataset that were present in all metric lists were used (n = 15,278, termed the “common evaluation” gene dataset).

The performance of all metrics on the evaluation gene dataset was analysed based on the distribution of the genes from the following groups (the number of genes that were present in the common evaluation dataset is reported in square brackets): (i) OMIM disease genes (AD, n = 465 [447], AR, n = 696 [686], “AD and AR”, n = 224 [222], and all Mendelian Disease (MD) genes with clinically relevant phenotypes, n = 1,587 [1,556]) as annotated in Gene Discovery Informatics Toolkit (GDIT)[24,41]; (ii) Cell essential (n = 598 [581]) and cell non-essential (n = 691 [667]) genes according to CRISPR–Cas studies[20] (obtained from the MacArthur Lab GitHub repository: https://github.com/macarthur-lab/gene_lists, accessed February 2020); (iii) Severe haploinsufficient (HI) genes from the regional missense constrained study (n = 28 [26])[29]; (iv) Genes with statistically significant enrichment (false discovery rate ≤ 0.05) of *de novo* mutations in individuals with diagnosed autism spectrum disorder, epileptic encephalopathy or intellectual disability from the gene4denovo database (referenced as “ASD, EE or ID *de novo”,* n = 160 [157])[123]; (v) Genes linked to lethal phenotypes with heterozygous (n = 109 [104]) and homozygous (n = 2,912 [2,847]) knock-out mice from the Mouse Genome Informatics (MGI)[124] or International Mouse Phenotype Consortium (IMPC)[125,126] databases as annotated in GDIT; (vi) Olfactory genes (n = 289 [279], obtained from the MacArthur Lab GitHub repository)[127].

4.3.2 Genomic and evolutionary features

DIP scores were created using a diverse set of genomic and evolutionary features, the majority of which were obtained from the previous studies[11,17,121,128,129]. Gene intolerance to loss-of-function (LoF) was represented by four features: (i) loss-of-function observed/expected upper bound fraction (LOEUF)[11]; (ii) Ratio between splice donor and synonymous variants (Donor/Syn)[17]; (iii) Ratio between cumulative length of exons affected by splice acceptor variants and total length of gene canonical transcript (Acceptor/Length, single exon transcripts were ignored) in gnomAD (v2.0.1)[11]; (iv) Cumulative allele frequency of structural variants expected to result in LoF in the gnomAD control population (SV-LoF, obtained from https://gnomad.broadinstitute.org/downloads, accessed June 2020)[128]. Gene intolerance to missense variation was represented by gene variation intolerance rank (GeVIR)[121] and unified inference of variant effects and gene constraints (UNEECON)[129] metrics. Both GeVIR and UNEECON also used evolutionary conservation data in combination with variant data from gnomAD, but GeVIR was mostly influenced by variation distribution patterns in gnomAD. A separate evolutionary based feature was used to measure conservation near transcription start site (+/- 500 bp)[17]. Finally, a binary feature was used to distinguish genes with shorter mRNA half-life (≤10 hours) in mouse embryonic stem cells, that were shown to be more often associated with AD diseases[17]. Apart from Acceptor/Length and SV-LoF, all features were obtained from the referenced studies and median feature values were used for the missing genes.

4.3.3 Protein-protein interaction network-based features

Four features were calculated on protein-protein interaction (PPI) data from the STRING v10 database[75] (accessed December 2019), three of which were used by the autosomal dominant from recessive model (ADR) and one by the disease from non-disease model (DND). ENSEMBL protein IDs used in the STRING database were mapped to HGNC gene IDs using mapping file obtained from ENSEMBL BioMart[130] and interactions with genes outside the main gene dataset (*G*) were excluded. STRING PPIs were grouped into 27 subsets based on the interaction source ($s \in$ *{"combined_score", "experimental", "textmining"}*) and confidence score (examined range: *{≥100c | c ∈ [1,2,...,9]}).* The performance of each feature was assessed on all PPI subsets, and final feature values for each $g \in G$ were calculated on the PPI subset with the best results.

Both PPI-N1 and PPI-N2 (used by ADR model) features were probabilities of genes being AD predicted by two models (named the same as features). Each of the models used K nearest neighbour (KNN) classifier[116], which was trained on two features that represented gene interactions with AD and AR genes from the ADR training dataset. The PPI-N1 model was based on the idea that if a gene interacts more with known AD than AR genes, then the gene itself is more likely to be AD than AR. For each $g \in G,$ PPI-N1 features (N^1) were calculated as the proportion of known AD or AR genes (two separate features) from the training dataset (*T*) in the set of genes with which *g* interacts (G_g):

$$N^1(g)=\frac{|G_g \cap T|}{|G_g|}$$

The PPI-N2 model was based on the idea that if a gene interacts more with genes that, regardless of their own status (e.g. AD, AR, unknown etc.), interacts more with known AD than AR genes, then the gene itself is more likely to be AD than AR. For each $g \in G,$ PPI-N2 features (N^2) were calculated as the sum of the PPI-N1 features (N^1) of all genes with which *g* interacts (G_g):

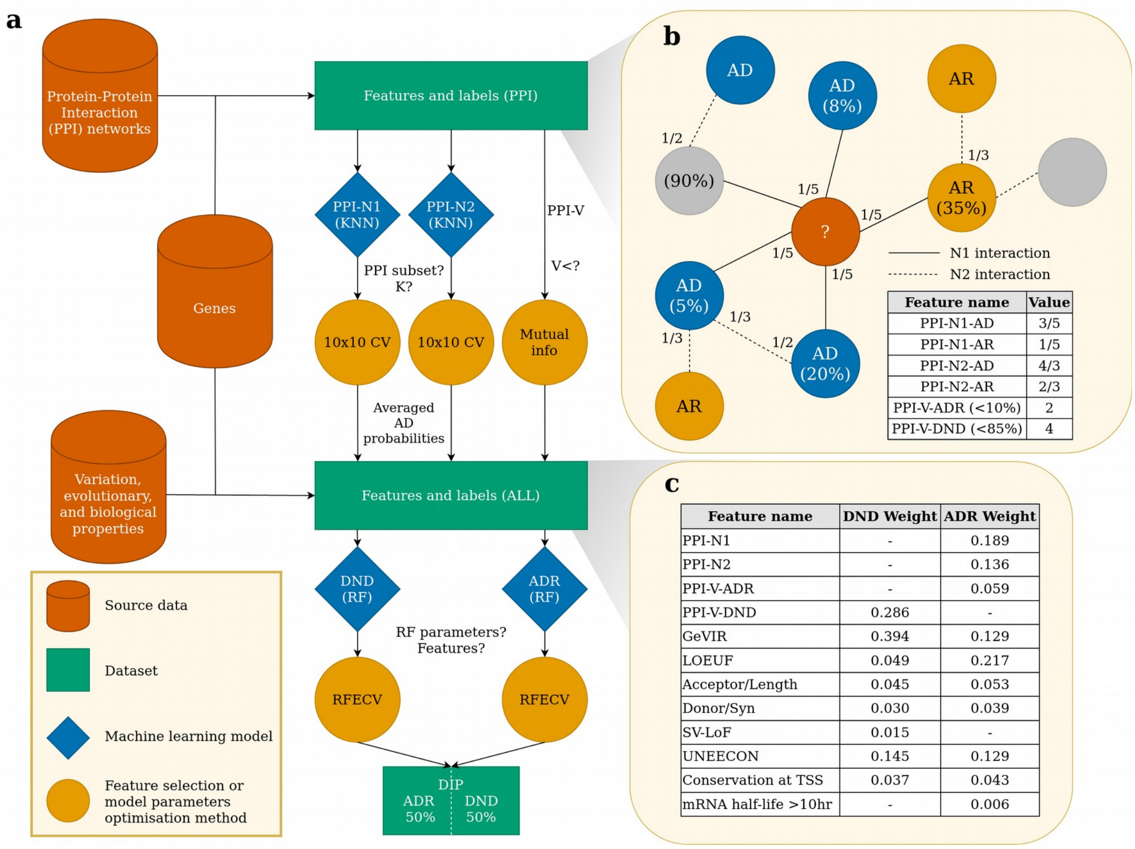
$$N^2(g)=\sum_{g^1}^{G_g} N^1(g^1)$$

The PPI-N1 and PPI-N2 model were evaluated with various numbers of nearest neighbours (KNN classifier parameters, the examined range: *{k ∈ [1,2,...,150]})* and average performance in a 10x10 cross-validation scenario (F1 score of the minor (AD) class) was measured to select the optimal *k* and PPI subsets. Both PPI-N1 and PPI-N2 models showed the best performance on the same PPI subsets (s = "textmining", c = 5) with k = 22 and 50, respectively. Final PPI-N1 and PPI-N2 features were created by training the models on the whole training dataset, so AD predicted probabilities for training genes were replaced with the averaged probabilities from the 10x10 cross-validation evaluation.

PPI-V-ADR and PPI-V-DND features (used by ADR and DND models, respectively) were based on the assumption that if a gene had many direct interactions with genes that could be AD or disease causing according to VIRLoF (i.e. intolerant to both missense and LoF variation), then the gene itself is more likely to be AD or disease causing, respectively. For each $g \in G$, PPI-V (V) features were calculated as the number of genes with VIRLoF less than a certain threshold (G_v) in the set of genes with which g interacts ($\mathrm{G_g}$):

$$V(g) = |G_g \cap G_v|$$

The optimal VIRLoF thresholds (examined range: $\{5v \mid v \in [1,2,\ldots,20]\}$) and PPI subsets for candidate PPI-V-ADR (s = “textmining”, c = 6, v = 2) and PPI-V-DND (s = “textmining”, c = 5, v = 17) features were selected by measuring mutual information on ADR and DND training datasets, respectively. The procedure for calculating DIP ranking is summarised in Figure 4.1.



Feature name	Value
PPI-N1-AD	3/5
PPI-N1-AR	1/5
PPI-N2-AD	4/3
PPI-N2-AR	2/3
PPI-V-ADR (<10%)	2
PPI-V-DND (<85%)	4

Feature name	DND Weight	ADR Weight
PPI-N1	-	0.189
PPI-N2	-	0.136
PPI-V-ADR	-	0.059
PPI-V-DND	0.286	-
GeVIR	0.394	0.129
LOEUF	0.049	0.217
Acceptor/Length	0.045	0.053
Donor/Syn	0.030	0.039
SV-LoF	0.015	-
UNEECON	0.145	0.129
Conservation at TSS	0.037	0.043
mRNA half-life >10hr	-	0.006

Figure 4.1: Procedure for calculating Disease Inheritance Pattern (DIP) ranking. **a)** General workflow. PPI-N1 and PPI-N2 models were built with K nearest neighbour classifiers (KNN). DND and ADR models were built with random forest (RF) classifiers, their features and parameters were selected simultaneously with recursive feature elimination and cross-validation (RFECV) method. **b)** Example of gene network based features calculation. PPI-N features are used in PPI-N1 and PPI-N2 models which predictions are then used in ADR model, whereas PPI-V features are directly used in ADR and DND models. Percentages in brackets are VIRLoF scores. **c)** Feature weights in DND and ADR models.

4.3.4 Calculation of DIP scores

Both DND and ADR models were based on a random forest (RF) classifier[116]. The feature selection and RF parameter tuning was performed simultaneously with the scikit-learn recursive feature elimination and cross-validation (RFECV) method[116]. All non-network based features and model specific network based features were examined with all combinations of the following RF parameters: (i) The number of trees ($\{t \in$

[100, 200, 300, 400, 500]}); (ii) The maximum number of features (*{f ∈ ['auto', None]}*); (iii) The maximum depth of the tree (*{d ∈ [2, 4, 6, None]}*); (iv) The pair of dependent parameters representing the minimum numbers of samples required to be the internal and leaf nodes (*{n ∈ [(2, 1), (4, 2), (8, 4), (16, 8)]}*); (v) Bootstrap method of tree creation (i.e. partial usage of the dataset for tree creation) on/off (*{b ∈ [True, False]}*). The optimal RF parameters for DND (*t = 200, f = None, d = 6, n = (4, 2), b = True*) and ADR (*t = 100, f = 'auto', d = None, n = (4, 2), b = True*) models were selected based on the F1 score of the less represented class (non-disease and AD genes, respectively). The best performing models were built using all features except *"mRNA half-life ≤10 hours"* and *"SV-LoF"* in DND and ADR, respectively. The final DND and ADR models were trained on the whole training datasets and used to calculate non-disease and AD probabilities for all genes studied (n = 15,794). Probabilities for training genes were replaced with the averaged probabilities obtained from the 10x10 cross-validation evaluation.

The DIP ranking of 15,794 genes, that were not used to train or validate DND and ADR models, was calculated by first sorting all genes based on the DND model predicted probabilities of genes being non-disease (ascending), and then resorting the first half based on the ADR model predicted probabilities of genes being AD (descending). DND model predictions were prioritised over ADR, since DND was trained on the more representative dataset (i.e. both disease genes with various inheritance patterns and non-disease genes), whereas ADR was built on the assumption that all genes had to be associated with AD or AR disease. The top ranked genes predicted to be AD by the ADR model (probability ≥ 0.5) were resorted based on their VIRLoF scores, which were not biased by training data and therefore might better represent ranking of the most important genes. To facilitate interpretation of DIP ranks, fold enrichment of known AD, AD and AR, and AR genes was calculated for each gene using the same parameters employed in our GeVIR study (i.e. in a range up to ±5% ranks, the first and last ranked genes were calculated only on 0-5% and 95-100% ranges, respectively)[121].

4.4 Results

Disease Inheritance Pattern (DIP) gene ranking was created by combining predictions from two models (both based on random forest classifiers), that were trained to distinguish disease from non-disease (DND) genes and autosomal dominant from recessive (ADR) genes, but the highest ranked genes (predicted to be AD by ADR) were sorted based on VIRLoF scores. ADR and DND were developed as alternatives to DOMINO and GPP models, respectively. ADR features included predictions from two other models (both based on K nearest neighbour classifiers), that were trained to distinguish AD and AR genes using protein-protein interaction (PPI) network data by analysing the first and the second level neighbours, named PPI-N1 and PPI-N2, respectively. These models were compared with the approach used in DOMINO for calculating network-based features. ADR performance was compared with DOMINO on the training and validation datasets, common for both models, and tested on the highest ranked genes (first decile and quartile, i.e. expected to be AD) from the common evaluation dataset that included all genes for which all compared metrics (ADR, DND, DOMINO, GPP and VIRLoF) had scores. In the latter case, ADR and DND predictions and were combined and used for ranking the first and the last half of the genes, respectively. DND had priority over ADR in conflict cases and was compared with GPP on the least ranked genes (last quartile, i.e. expected to be non-disease genes). Finally, DIP (i.e. combination of ADR, DND and VIRLoF scores) was compared with other metrics in the context of their ability to continuously rank genes by inheritance patterns.

PPI-N1 and PPI-N2 models, that predicted AD probabilities, were used as features in the ADR model, and were compared with the approach for the "network-based features" calculation used in DOMINO (i.e. counting direct interactions with AD genes from the training dataset with some upper limits)[17]. The most important DOMINO feature was calculated by counting PPIs with STRING combined confidence ≥500[17]. We found that the maximum achievable performance of this method on the training data was 64.95% F1 score when all genes with ≥4 AD PPIs were predicted as AD. This result could be achieved when all AD gene interactions were counted and all training genes were analysed to find the optimal threshold. However, the PPI-N1 and PPI-N2 model average performance in a 10x10 cross-validation scenario (i.e. 90% of the known AD and AR genes were used to train the models) on the same data were 66.54% and 65.52% F1 scores, respectively. Moreover, both PPI-N1 and PPI-N2 models showed the best performance of 67.50% and 66.44% F1 scores, when PPIs with STRING textmining confidence ≥500 was used. Based on the average predicted AD probabilities, PPI-N1 and PPI-N2 models agreed more on correctly predicted AD genes (161/220, 73.18%), than on misclassified AR genes (43/127, 33.86%), which explains why both features were beneficial for the ADR model (i.e. survived the feature selection procedure). It should be noted that the DOMINO model used three network-based features (based on combined, textmining, and experimental STRING data) where parameters (confidence thresholds and upper limits) were optimised to work effectively with other features[17], and consequently might not perform the best on their own. Nevertheless, this evaluation showed that our method for network-based features calculation could be more efficient for AD gene prediction than the one used by DOMINO.

ADR model performance was compared with DOMINO on the training (average results from 10x10 cross validation) and validation datasets using metrics for the AD class that was less represented in both datasets (29.57% and 26.26%, respectively). On the training dataset, ADR and DOMINO models achieved similar F1 scores (76.26% and 75.84%, respectively), but the ADR model was more precise (82.55% and 74.10%, respectively), whereas DOMINO had higher recall (71.38% and 77.66%, respectively). On the validation dataset, ADR and DOMINO models had the same recall (88.46%), but the ADR model was again more precise than DOMINO (76.67% and 65.71%, respectively) and, therefore, also had higher F1 scores (82.14% and 75.41%, respectively). However, it should be noted that the validation dataset consisted of only 99 genes and contained AD genes (n = 26) that were significantly (P=0.011, two-sided Mann-Whitney U test with continuity correction) more intolerant to variation than AD genes (n = 291) in the larger training dataset (median VIRLoF 7.73% and 13.38%, respectively). Therefore, the performance of both models on the validation dataset could be affected by selection of more severe, and consequently easier to predict, AD genes (i.e. it should not be assumed that both models can predict 88.5% of all AD genes). Nevertheless, evaluation on the validation dataset confirmed the observation made on the training dataset, that the ADR model could be more precise than DOMINO. Also on both training and validation datasets the ADR model obtained higher ROC AUC

(93.05% and 93.07%, respectively) than DOMINO (91.15% and 92.04%, respectively), which suggests that ADR model probabilities predicted might result in a slightly better gene ranking in general.

Genes from the evaluation dataset (n = 15,794) were ranked by both ADR (first half) and DND (second half) model probabilities, that predicted 23.05% (n = 3,641) and 42.41% (n = 6,699) of these genes to be AD and non-disease (probability ≥0.5), respectively. These groups were expected to represent genes from the opposite sides of the spectrum, but they overlapped and 9.26% (n = 337) of the genes predicted to be AD by ADR were also predicted to be non-disease by DND. This group of genes, as expected, was significantly deficient in AR genes (fold enrichment (FE) = 0.07, two-sided Fisher's exact test[96] P = 2.07E-05), but was not enriched with AD genes (FE = 0.81, P = 7.42E-01), and was significantly enriched with cell non-essential genes (FE = 1.76, P = 1.15E-02). Note that while the DND model was trained to distinguish disease genes with any mode of inheritance from potentially non-disease genes, the ADR model was trained with a naive assumption that all genes had to be AD or AR. Consequently, when the ADR model was used to classify all genes, it might to some extent also prioritise non-important genes over strong AR candidates. Therefore, although some of these genes could be AD, others could be non-disease and a decision was made to rely on the DND model in conflict cases to more precisely prioritise disease genes in general.

The performance of the ADR model performance was compared with other metrics on the first quartile of genes (n = 3,820/15,278) from the common evaluation dataset (Table 4.1), which was close to the number of genes predicted to be AD by DOMINO (n = 3,807, 24.92%). In the context of AD and AR gene classification, ADR performed slightly better (precision = 70.39%, recall = 60.63%, F1 = 65.14%) than DOMINO (precision = 69.11%, recall = 59.06%, F1 = 63.69%) and VIRLoF (precision = 69.54%, recall = 54.14%, F1 = 60.88%), but much better than GPP (precision = 45.06%, recall = 53.02%, F1 = 48.72%). The latter showed that models that are designed to distinguish disease from non-disease genes, such as GPP, may not be effective in prioritising AD over AR genes. Note that the performance of both ADR and DOMINO on this dataset

(F1 = 65.14% and 63.69%, respectively) was lower than on the training (F1 = 76.26% and 75.84%, respectively) and validation (F1 = 82.14% and 75.41%, respectively) datasets. This could be because AD genes in the evaluation dataset were significantly more tolerant to variation (P=0.002, two-sided Mann-Whitney U test with continuity correction[96]; median VIRLoF = 19.69%), compared with those used to train the model (median VIRLoF = 13.38%). However, both ADR and DOMINO successfully prioritised 92.31% and 100% of the genes from the severe haploinsufficient (HI) group (n = 26) respectively, which were much more intolerant to variation (median VIRLoF = 2.65%). Although, ADR and DOMINO performance at this threshold was similar, only 76.47% (n = 2921) of the genes prioritised by ADR and DOMINO were in common. The genes prioritised exclusively by ADR (n = 899, 23.53%) were more intolerant to variation (median VIRLoF = 21.44%) than those prioritised exclusively by DOMINO (median VIRLoF = 35.67%), but had fewer protein-protein interactions (STRING combined score >= 500 median number of interactions were 38 and 50, respectively). ADR also prioritised genes essential in cell culture assays (n = 289, 49.74%), those linked to lethal phenotypes in heterozygous (n = 66, 63.46%) and homozygous (n = 1480, 51.98%) knock-out mice, and especially those enriched with *de novo* mutations in individuals with diagnosed autism spectrum disorder (ASD), epileptic encephalopathy (EE), or intellectual disability (ID) disorder (n = 116, 73.89%). In contrast, there was a deficiency of cell non essential genes (n = 66, 9.90%), and no olfactory receptor genes resided in the first quartile. However, it should be noted that the difference between ADR and DOMINO or VIRLoF was not significant in all gene groups examined (Table 4.1), except for cell non-essential genes (two-sided Fisher's exact test P = 1.13E-02), a gene group not prioritised by VIRLoF (n = 39, 5.85%).

Table 4.1: Number and percentage of genes from various groups in the first quartile (n = 3,820) of the evaluation dataset (n = 15,278) sorted by DND, DOMINO, GPP and VIRLoF metrics.

The proportion of genes in the ADR group was compared with other metrics using a two-sided Fisher's exact test (P indicates statistical significance; n = number). Used abbreviations: autosomal dominant (AD), autosomal recessive (AR), Mendelian disease (MD), autism spectrum disorder (ASD), epileptic encephalopathy (EE), intellectual disability (ID), haploinsufficient (HI).

Gene Group\ Metric	**ADR**		**DOMINO**			**GPP**			**VIRLoF**		
	n	**%**	**n**	**%**	***P***	**n**	**%**	***P***	**n**	**%**	***P***
AD	271	60.63	264	59.06	8.27E-01	237	53.02	2.43E-01	242	54.14	3.19E-01
"AD and AR"	84	37.84	95	42.79	5.35E-01	125	56.31	2.31E-02	70	31.53	3.50E-01
AR	114	16.62	118	17.20	8.32E-01	289	42.13	6.23E-15	106	15.45	6.63E-01
MD	525	33.74	536	34.45	7.76E-01	731	46.98	1.01E-06	473	30.40	1.56E-01
ASD, EE or ID de novo	116	73.89	113	71.97	9.31E-01	78	49.68	3.52E-02	115	73.25	1.00E+00
Severe HI	24	92.31	26	100.00	8.46E-01	13	50.00	1.97E-01	24	92.31	1.00E+00
Cell essential	289	49.74	289	49.74	1.00E+00	261	44.92	3.51E-01	273	46.99	6.07E-01
Cell non-essential	66	9.90	72	10.79	6.55E-01	46	6.90	7.66E-02	39	5.85	1.13E-02
Mouse Het Lethal	66	63.46	68	65.38	9.12E-01	43	41.35	7.68E-02	56	53.85	4.95E-01
Mouse Hom Lethal	1480	51.98	1483	52.09	9.82E-01	1355	47.59	5.64E-02	1422	49.95	3.86E-01
Olfactory	0	0.00	1	0.36	1.00E+00	0	0.00	1.00E+00	0	0.00	1.00E+00

The same analysis was carried out on the first decile (n = 1,528/15,278) to evaluate the performance of the metrics at prioritising the most important / likely AD genes (Table 4.2). At this threshold, ADR prioritised fewer AD genes (n = 158, 35.35%) than DOMINO (n = 177, 39.60%), but more than VIRLoF (n = 146, 32.66%). However, the VIRLoF decile contained fewer AR (n = 25, 3.64%) and cell non-essential (n =11, 1.65%) genes than ADR (n = 33, 4.81% and n = 16, 2.40%, respectively) and DOMINO (n = 44, 6.41% and n = 25, 3.75%, respectively). Moreover, VIRLoF also prioritised slightly more genes enriched with *de novo* mutations in individuals with diagnosed ASD, EE and ID disorders (n = 95, 60.51%), than ADR (n = 90, 57.32%) and DOMINO (n = 84, 53.50%). All three metrics prioritised the same number of severe HI genes (n = 21, 80.77%). Note that ADR and DOMINO predicted 20.62% (n = 3151) and 24.92% (n = 3807) of the genes to be AD, respectively. Consequently, all known AD genes should not be expected to be ranked in the first 10% and, therefore, prioritisation of known AD genes might result in deprioritisation of unknown AD genes (i.e. those that have yet to be associated with an AD disease). Moreover, the most important genes might be dominant lethal and, consequently, not associated with documented diagnosed diseases. Therefore, at this threshold, metrics had to be evaluated mainly based on precision, which was higher in VIRLoF (85.38%) than ADR (82.72%) and DOMINO (80.09%). Moreover, VIRLoF also prioritised fewer genes associated with both “AD and AR” (n = 19, 8.56%) diseases than ADR (n = 35, 15.77%) and DOMINO (n = 44, 19.82%). If these genes were counted as false positives together with AR genes, then VIRLoF would be even more precise (76.84%) than ADR (69.91%) and DOMINO (66.79%). Consequently, VIRLoF might be more efficient than ADR and DOMINO at ranking the most important genes. However, it should be noted that the difference between ADR and VIRLoF was not significant in all examined gene groups (Table 4.2), except for genes associated with both “AD and AR” diseases (P = 4.39E-02). Nevertheless, genes predicted to be AD by ADR were subsequently sorted based on their VIRLoF scores in the final DIP gene ranking process (n = 3,212/15,794, 21.02%).

Table 4.2: Number and percentage of genes from various groups in the first decile (n = 1,528) of the evaluation dataset (n = 15,278) sorted by DND, DOMINO, GPP and VIRLoF metrics.

The proportion of genes in the ADR group was compared with other metrics using a two-sided Fisher's exact test (P indicates statistical significance; n = number). Used abbreviations: autosomal dominant (AD), autosomal recessive (AR), Mendelian disease (MD), autism spectrum disorder (ASD), epileptic encephalopathy (EE), intellectual disability (ID), haploinsufficient (HI).

Gene Group\ Metric	**ADR**		**DOMINO**			**GPP**			**VIRLoF**		
	n	**%**	**n**	**%**	***P***	**n**	**%**	***P***	**n**	**%**	***P***
AD	158	35.35	177	39.60	4.05E-01	124	27.74	8.75E-02	146	32.66	5.95E-01
"AD and AR"	35	15.77	44	19.82	3.93E-01	73	32.88	1.19E-03	19	8.56	4.39E-02
AR	33	4.81	44	6.41	2.42E-01	141	20.55	1.23E-15	25	3.64	3.49E-01
MD	251	16.13	294	18.89	9.46E-02	380	24.42	2.84E-06	211	13.56	9.00E-02
ASD, EE or ID de novo	90	57.32	84	53.50	7.77E-01	28	17.83	7.30E-07	95	60.51	7.82E-01
Severe HI	21	80.77	21	80.77	1.00E+00	8	30.77	6.23E-02	21	80.77	1.00E+00
Cell essential	150	25.82	156	26.85	7.97E-01	98	16.87	3.23E-03	149	25.65	1.00E+00
Cell non-essential	16	2.40	25	3.75	2.05E-01	6	0.90	5.07E-02	11	1.65	4.38E-01
Mouse Het Lethal	41	39.42	42	40.38	1.00E+00	24	23.08	8.71E-02	30	28.85	2.74E-01
Mouse Hom Lethal	726	25.50	783	27.50	1.93E-01	641	22.51	4.05E-02	701	24.62	5.74E-01
Olfactory	0	0.00	0	0.00	1.00E+00	0	0.00	1.00E+00	0	0.00	1.00E+00

The performance of the DND model was compared with other metrics on the last quartile of genes (n = 3,820/15,278) from the common evaluation dataset (Table 4.3), which was close to the number of genes predicted to be non-disease by GPP (n = 3,855, 25.23%). In the context of continuous gene ranking, this group of genes was expected to be deficient in known disease genes, and enriched with unlikely disease (e.g. olfactory) or possibly less important genes (e.g. cell non-essential). The DND group contained a smaller or similar number of known or potentially disease genes (e.g. cell essential, mouse knock-out lethal) compared with other metrics (significance was measured with a two-sided Fisher's exact test). Specifically, the number of known Mendelian disease (MD) genes in the DND group (n = 80, 5.14%) was significantly smaller than in GPP (n = 145, 9.32%, P = 3.13E-05), VIRLoF (n = 171, 10.99%, P = 2.73E-08) and DOMINO (n = 308, 19.79%, P = 2.08E-29) groups. Although the GPP group contained the largest number of olfactory (n = 279, 100%) and cell non-essential (n = 423, 63.42%) genes, the difference was not significant compared with the DND group (n = 275, 98.57% and n = 403, 60.42%, respectively). Therefore, the DND model performed the best at deprioritising non-disease and possibly non-important genes.

Note that only the DOMINO group contained a similar percentage of AR (n = 217, 31.63%) and cell non-essential genes (n = 220, 32.98%). Moreover, it contained only 27.24% (n = 76) of olfactory genes, whereas all the other metrics contained 90-100% of them in the last quartile. This shows that models designed to distinguish AD from AR genes, such as DOMINO, may not be able to effectively prioritise AR over non-disease genes, and that this behavior should not be assumed by default.

Table 4.3: Number and percentage of genes from various groups in the last quartile (n = 3,820) of the evaluation dataset (n = 15,278) sorted by DND, DOMINO, GPP and VIRLoF metrics.

The proportion of genes in DND group was compared with other metrics using a two-sided Fisher's exact test (*P* indicates statistical significance; n = number). Used abbreviations: autosomal dominant (AD), autosomal recessive (AR), Mendelian disease (MD), autism spectrum disorder (ASD), epileptic encephalopathy (EE), intellectual disability (ID), haploinsufficient (HI).

Gene Group\ Metric	DND		DOMINO			GPP			VIRLoF		
	n	%	n	%	*P*	n	%	*P*	n	%	*P*
AD	19	4.25	33	7.38	6.43E -02	37	8.28	2.67E -02	38	8.50	1.95E -02
"AD and AR"	5	2.25	18	8.11	9.46E -03	20	9.01	3.59E -03	23	10.36	8.10E -04
AR	39	5.69	217	31.63	4.71E -27	59	8.60	5.95E -02	67	9.77	8.82E -03
MD	80	5.14	308	19.79	2.08E -29	145	9.32	3.13E -05	171	10.99	2.73E -08
ASD, EE or ID de novo	8	5.10	13	8.28	3.69E -01	6	3.82	7.86E -01	9	5.73	1.00E +00
Severe HI	0	0.00	0	0.00	1.00E +00	1	3.85	1.00E +00	0	0.00	1.00E +00
Cell essential	17	2.93	84	14.46	3.26E -11	17	2.93	1.00E +00	29	4.99	9.80E -02
Cell non-essential	403	60.42	220	32.98	1.04E -09	423	63.42	5.95E -01	412	61.77	8.24E -01
Mouse Het Lethal	2	1.92	10	9.62	3.50E -02	10	9.62	3.50E -02	7	6.73	1.72E -01
Mouse Hom Lethal	136	4.78	358	12.57	1.86E -22	137	4.81	1.00E +00	152	5.34	3.65E -01
Olfactory	275	98.57	76	27.24	4.43E -18	279	100.00	9.05E -01	252	90.32	5.04E -01

Finally, to analyse overall trends of gene distribution among DIP ranks and facilitate their interpretation, we used the same approach described in our GeVIR study[121] and calculated fold enrichment (FE) of AD, "AD and AR", AR, and cell non-essential genes in a range up to ±5% for each gene (Appendix Table C.1; statistical significance was measured with a two-sided Fisher's exact test). To compare DIP with other metrics, the same analysis was performed on the evaluation dataset (Figure 4.2), but the trends observed were the same in both cases (Figure 4.2 and Appendix Table C.1).

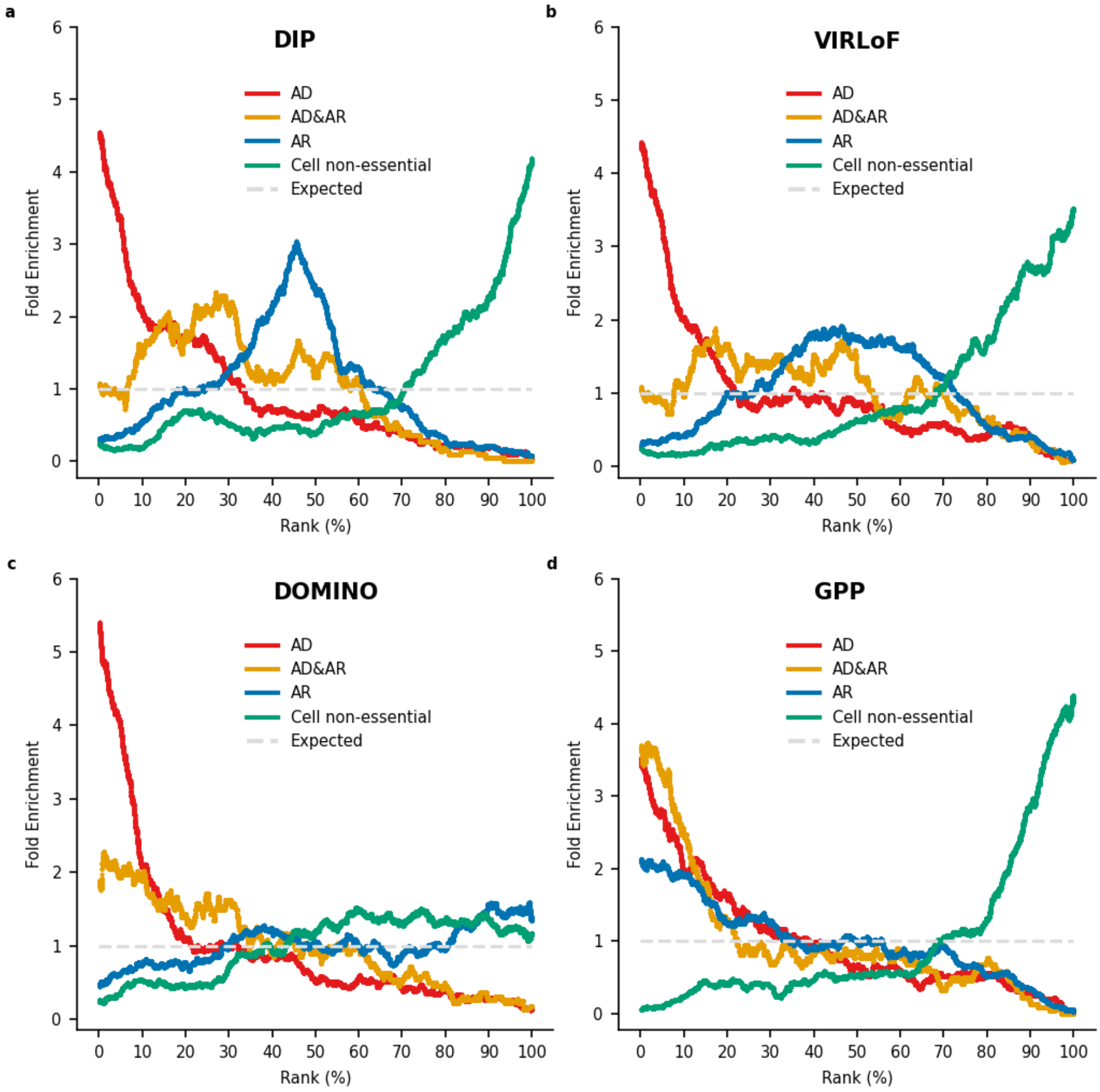


Figure 4.2: Fold enrichment of known disease and cell non-essential genes for each gene in the common evaluation dataset.

Genes were ranked based on (a) DIP, (b) VIRLoF, (c) DOMINO and (d) GPP scores. For each gene (n = 15,278), fold enrichment was calculated by analysing genes with similar ranking scores (up to ±5 percentiles) and comparing the proportion of known AD (n = 447), AD&AR (n = 222), AR (n = 686) and cell non-essential (n = 667) genes.

AD and cell non-essential genes were enriched among the highest and the lowest ~5% of DIP ranks with peaks of FE equal 4.55 (P = 1.21E-30) and 4.19 (P = 1.13E-36) respectively. The AR gene enrichment peak (FE = 3.04, P = 2.27E-34) was in the middle ranks (41-51%), whereas the "AD and AR" gene enrichment peak (FE = 2.34, P = 4.91E-07) was between AD and AR peaks (22-32%, Figure 4.2a). VIRLoF showed similar trends, but had a smaller peak enrichment of AR genes (FE = 1.92, P = 5.98E-10) in the middle ranks (41-51%), mostly due to less effective segregation of AR and cell non-essential genes (Figure 4.2b). However, DOMINO and GPP did not effectively distinguish AR from cell non-essential (Figure 4.2c) and disease genes by mode of inheritance (Figure 4.2d), respectively. Therefore, overall DIP showed the highest trend to continuously rank genes by mode of disease inheritance.

4.5 Discussion

The overall aim of this study was to develop a continuous gene metric that could be used to prioritise genes by Disease Inheritance Patterns (DIP), more effectively than statistical variation intolerance metrics (e.g. GeVIR, LOEUF). To some extent, this was achieved by developing and combining predictions from two supervised machine learning models, which were trained to distinguish disease from non-disease (DND) genes, and AD from AR (ADR) genes. ADR and DND models were trained on the same datasets as DOMINO and GPP respectively, so that their performance could be tested on the maximum number of known disease genes that were unseen by all methods.

We have shown that existing models that were trained to classify genes into two groups produced mostly dichotomous, but not continuous, metrics. DOMINO was able to prioritise AD genes, but could not distinguish AR from non-disease genes, whereas GPP could prioritises disease over non-disease genes, but could not distinguish genes with different inheritance patterns. It should be noted that the GPP study also developed separate metrics for prioritisation of AD and AR genes[3], but these metrics were not evaluated here since the genes used to train the final models were not reported by the authors. The DND model performed significantly better than other metrics at prioritising disease genes, and its combination with ADR in DIP resulted in a peak of AR genes in the middle ranks, not seen in GPP or DOMINO and more pronounced than in VIRLoF.

Overall, ADR performance in the context of AD and AR gene classification was comparable with DOMINO and VIRLoF, but assessment of candidate AD gene ranking (i.e. top ranked genes predicted to be AD) showed ambiguous results. In the first decile, DOMINO had the highest recall, VIRLoF was the most precise (especially if “AD and AR” genes were considered as false positives), whereas ADR was the second best in the both cases. Although the probability for a gene to be AD predicted by ADR and DOMINO tends to correlate with disease severity, similar to VIRLoF, both models were not trained for this. Moreover, considering that the most important genes could be dominant lethal and, therefore, incompatible with life so largely unknown, supervised methods might not be suitable for this task. Although high precision could also be achieved by prioritising unknown non-disease genes over AR genes, VIRLoF had slightly less cell non-essential genes in the first decile and significantly less in the first quartile than ADR or DOMINO. Consequently, VIRLoF might actually perform best at ranking the most important genes despite prioritising less known AD genes, but this could not be assessed since the majority of these genes are still not linked to disease phenotypes. As a compromise, ADR was used to predict candidate AD genes, but VIRLoF was used to rank them.

The quality of the DIP metric is limited by several factors. Firstly, DIP was created using predictions from the supervised machine learning models whose performance depended on the degree to which genes in the training datasets represented all analysed genes[39]. To allow fair comparisons with existing methods, we reused the training datasets from the previous studies[3,17] despite their limitations. Since a dataset of confirmed non-disease genes did not exist, the least important genes, used to train the DND model, were selected based on the presence of deleterious LoF variation in population databases[12,131], however they could still be associated with some unknown mild or late onset disease[3]. Both ADR and DND models were trained on subsets of established known disease genes, which might disproportionally represent the variety of known phenotypes (i.e. selection bias). Nevertheless, distribution of the genes from various examined groups in the evaluation dataset, including those that were not used in the training process (e.g. cell non-essential), has shown that ADR and DND model predictions were generalizable and often better than other metrics (DOMINO and GPP).

Secondly, DIP, as well as DOMINO and GPP, used protein-protein interactions (PPI) networks data, whose usage was criticised since less studied genes were found to have fewer interactions[35]. To reduce bias towards more studied genes, main network-based features of the ADR model (PPI-N1 and PPI-N2) were calculated using proportions instead of absolute numbers of interactions. However, this approach also made network-based predictions less precise, especially in the case of PPI-N1, which had to be compensated by other features (Figure 4.1c) in the ADR model and the DND model was used to filter out unlikely disease genes. We found that 1023/4014 (25.49%) and 474/3315 (14.30%) of genes predicted to be AD (probability >= 0.5) by PPI-N1 and PPI-N2 models had a number of interactions fewer than or equal to the median (n = 9), but only 226/1023 (22.09%) and 147/474 (31.01%) of them were also predicted to be AD by ADR (with DND filtration), respectively. Note that DIP was created using a diverse set of gene properties, and the cumulative weight of network based features in the ADR and DND models were only 38.4% (lower than 47.5% in DOMINO[17]) and 28.6%, respectively.

Finally, since ADR and DND models were trained mostly as dichotomous metrics, usage of their predicted probabilities for ranking might not be optimal, which was highlighted in the evaluation of the highly ranked genes. However, the evaluation also showed that they worked well for majority of the genes, and in the final DIP ranking the top ~21% were resorted based on their VIRLoF scores because it is a more objective statistical metric.

The main advantage of DIP is that it provides a continuous and intuitively understandable metric for ranking genes whereby AD and likely non-disease genes are enriched at the opposite ends of the spectrum, with AR genes residing in the middle. Since the ranking of a gene can also be influenced by disease severity (at least for AD genes) and some genes can be associated with multiple diseases with different modes of inheritance (i.e. “AD and AR”), the degree to which genes can be confidently segregated based on their mode of inheritance is unknown. Consequently, continuity is an important characteristic of a metric because it estimates gene predisposition to all modes of inheritance, which is especially important in ambiguous cases. Moreover, continuous metrics measure the relative importance of genes in general, assuming that loss or disruption of less important genes is less likely to result in disease phenotypes.

Another advantage of DIP is the diverse set of gene properties (e.g. evolutionary, variation intolerance and network-based) that were used as features to build the machine learning models behind it. Although individual features have limitations, to some extent they can be negated when features are used together. Previously, this was demonstrated by combining two constraint metrics, GeVIR and LOEUF, to generate VIRLoF[121], whereas in DIP gene ranking was further improved, albeit with some novel limitations, by combining these metrics with other features and using supervised machine learning methods. Further research of statistical and unsupervised machine learning methods for gene classification or ranking is important since they could be less biased by training data (or not used it at all) and could improve the performance of current supervised methods, such as DIP, which uses them as features. Moreover further development of these metrics can also be useful in supervised machine learning research to predict candidate genes for specific phenotypes and modes of inheritance (e.g. gene causing AD developmental conditions). This direction of research was also suggested in the GPP study[3] and investigated for some disease phenotypes (e.g. neurological disorders[132] and schizophrenia[133]).

In conclusion, multiple supervised machine learning models and a statistical gene variation intolerance metric (VIRLoF) have been used in this study to continuously rank autosomal genes by Disease Inheritance Patterns (DIP). Comparisons were made with two published supervised machine learning methods (GPP and DOMINO) that were developed to classify genes into only two groups (disease/non-disease and AD/AR, respectively), and a statistical method to measure gene intolerance to variation (VIRLoF) that continuously ranked genes from all three groups (AD, AR and non-disease). Although it is hard to say which method performed best at ranking candidate AD genes (i.e. gene order in the first quartile), overall DIP has shown the greatest trend to differentiate between the gene groups and rank genes in the following order: AD, "AD and AR", AR and non-disease. Therefore, when analysing whole genome/exome sequencing data, DIP should be used as a gene level metric alongside variant prioritisation metrics[119,120] to discover novel human disease genes. DIP could be particularly useful for prioritising candidate AR disease genes or to filter out genes unlikely to cause disease.

Chapter 5

Conclusion

5.1 Summary of results and contributions

This thesis is written in journal format style with three main chapters formatted as papers (Chapter 2, 3, and 4). This section summarises the results and contributions of these studies. Statistical significance was measured using two-sided Fisher's exact test for all reported results in this chapter, unless stated otherwise.

5.1.1 Chapter 2

In Chapter 2, we analysed deviations from Hardy-Weinberg Equilibrium (HWE)[82] in gnomAD populations to identify candidate recessive disease-causing and potentially heterozygous advantageous variants (Objective 1). We hypothesised that these variants might deviate from HWE due to heterozygote excess (HetExc) due to natural selection and exclusion of individuals with severe paediatric disorders from the dataset (i.e. affected homozygous variant carriers). Initially, we planned to develop a gene variation intolerance metric that would incorporate HetExc statistics. However, during our analysis, we noticed that many HetExc variants might be deviating from HWE due to genotype calling errors. Therefore, we developed a set of strict rules to understand how many potentially genuine HetExc variants were present in the dataset and identified only 161 variants in 149 genes. Although these variants were significantly enriched in known autosomal recessive genes (FE = ~1.6, *P* = ~0.02), it became clear that this property could not be used to make predictions about disease inheritance patterns of more than 19,000 remaining protein-coding genes.

Nevertheless, our final dataset contained two known disease-causing variants in *HBB* and *CFTR* genes, with evidence of heterozygote advantage in the literature[89,102,103] that showed that our filtering strategy could detect these types of variants. Surprisingly, most of the HetExc variants (~79.5%) were detected in African/African American populations, statistically significantly more than expected (FE = ~1.7, *P* = 3.0E-05) based on the proportion of non HetExc variants in African/African American (~48.1%). This observation is interesting since it might mean that African individuals have more heterozygous advantageous variants or are more prone to genotype calling errors. Therefore, we decided to present the results of our analysis in the context of heterozygote advantage. This type of variant is rare, evidence is often speculative, and even one new potential heterozygote advantageous variant could be considered an exciting finding.

Approximately half a year before we published a preprint of Chapter 2, Wei and Nielsen[134] published an article in *Nature Medicine* where they argued that CCR5-Δ32 could be heterozygous advantageous based on HWE statistics in a variant population database. The CCR5-Δ32 variant in both homozygous and heterozygous states is known to protect against HIV[135], but is also associated with lower resistance to influenza[136]. Wei and Nielsen[134] analysed data from the UK Biobank (409,693 individuals) and found that individuals with homozygous CCR5-Δ32 variants had increased all-cause mortality rate, whereas heterozygous carriers had the same rate as unaffected individuals. However, ~2 months before we published a preprint, Karczewski *et al.*[97] published a preprint where they explained that the deviation from HWE equilibrium of the CCR5-Δ32 variant was caused by a genotype calling error. Specifically, some individuals had CCR5-Δ32 variants with high allele balance (>0.9) but were misclassified as heterozygous instead of homozygous[97]. Around the same time, Wei and Nielsen[137] retracted their publication.

Based on this, we can make the following conclusions. First, using HWE statistics and variant population databases to identify disease and potentially heterozygous advantageous variants was a trending topic in the research community. Second, authors, reviewers, and editors of the article[134] were not aware that deviations from HWE due to HetExc, besides natural selection, could also be caused by technical errors still present in variant population databases. Third, researchers who developed variant population databases were unaware of this type of genotype calling error in their data. After this publication, Karczewski *et al.*[97] identified and flagged variants that might be affected by this error in the gnomAD database. We also recalculated HWE statistics for these variants using a more conservative allele balance threshold (>0.8) than suggested by Karczewski *et al.*[97] (>0.9) to minimise false-positive HetExc variants in our dataset. However, unlike Wei and Nielsen[134], we did not make strong claims that variants with HetExc in our final dataset were deviating due to natural selection. Instead, we presented it as supporting *in silico* evidence for potential future work that might confirm it as having a biological role. Moreover, we highlighted that deviation might also be caused by unknown genotype calling errors or by chance due to insufficient population size in many cases. Therefore, we believe that our work contributed to the research community's understanding of the deviations from HWE in variant population databases, even if, in reality, none of the HetExc variants in our dataset were heterozygous advantageous.

5.1.2 Chapter 3

In Chapter 3, we analysed variant distribution at a gene level in the gnomAD database and developed a novel metric named gene variation intolerance rank (GeVIR), that can be used to prioritise candidate disease-causing genes and, to some extent, distinguish them by inheritance patterns (dominant and recessive) (Objective 2). We hypothesised that some of the gene regions could be so important for their function that they are entirely free of functional variants (missense and loss-of-function), and the longer the region is, the less likely it is variant-free by chance. We named them variation intolerant regions (VIRs). This hypothesis was strongly supported by our analysis that showed that known pathogenic missense variants were extremely statistically significantly ($P < 2.2 \times 10{-}308$) more likely (FE = ~3.7) to be present in the long VIRs (>20 amino acids). To calculate a gene-level metric from VIRs (GeVIR), briefly, we assigned each region a weight based on how frequently VIRs of each length were observed in all genes, then calculated average score of well-covered VIRs for each of the 19,361 protein-coding genes, and ranked them based on these scores. However, we found that some VIRs might be false positives due to variant calling errors (i.e. VIR reports that region is free of variants, but it might not be valid if it contains some non-pass quality variants that are real). We assumed that genuine VIRs should also be evolutionary conserved and recalculated GeVIR scores with VIR weights multiplied by their average evolutionary conservation scores (measured with GERP++). The addition of evolutionary conservation scores slightly improved the ability of GeVIR to prioritise and distinguish dominant and recessive genes (<1% AUC and F1 scores). However, it also greatly improved the overall ability of GeVIR to prioritise disease genes by assigning lower ranks to possibly non-disease genes (by 5.8% and 7.5% AUC for Null and cell non-essential genes, respectively).

When we were working on the GeVIR manuscript, a study of constraint coding regions (CCRs)[33] and a preprint of novel gnomAD gene LoF intolerance scores (LOEUF)[11] were published, both of which had an impact on our work. Despite some methodological differences, CCRs were based on essentially the same principle as VIRs (i.e. the distance between functional variants)[33]. However, Havrilla *et al.*[33] focused on using CCRs to prioritise candidate disease variants, whereas we used VIRs to calculate gene-level variation intolerance scores. Nevertheless, Havrilla *et al.*[33] compared CCRs correlation with gene-level variation intolerance metrics (pLI and missense z-scores) by ranking genes based on the number of the longest CCRs (top 5%). Although we agreed with CCRs usefulness for candidate disease variant prioritisation, we considered their approach to calculating and evaluating gene-level metrics derived from the variant-free region data incorrect and insufficient, respectively.

First, approximately two-thirds of protein-coding genes (n > 11,000) did not contain even one CCR at or above the 95th percentile[33] and, consequently, their metric could not be used to estimate variation intolerance of the majority of the genes. They explicitly stated that CCRs were ill-suited to prioritise recessive disease-causing variants since they were expected to be present in gnomAD and, consequently, disrupt variant-free regions[33]. Second, by default, genes with longer protein-coding sequences could contain more CCRs, and our analysis confirmed that their metric was significantly biased by gene length (Spearman r = -0.29). Finally, by measuring only correlation, they showed that CCRs and gnomAD gene variation intolerance scores prioritised different genes[33], but did not show how they performed on known sets of disease genes.

Our study was superior to Havrilla *et al.*[33] based on all the above points in the context of the gene-level intolerance metric. First, GeVIR scores were calculated for all protein-coding genes that had gnomAD variation intolerance metrics (n = 19,361 versus n = 7,000 genes in their study). Moreover, GeVIR performed surprisingly well for prioritising known autosomal recessive genes that were enriched ~1.8 times at the middle of the ranking (40th–60th percentiles). The enrichment of AR genes at the same range was lower for GeVIR without GERP++ adjustments (~1.5 times), but still statistically significant (*P* = 2.27E-19). Second, GeVIR bias by gene length was minimal (Spearman r = 0.03) based on the analysis of the same amount of top-ranked genes (n = 7,000, different for CCR and GeVIR), although it was still biased when all genes were analysed (Spearman r = -0.26). Finally, GeVIR outperformed the CCR gene-level metric at distinguishing autosomal dominant and recessive genes by a ~5.1% (~4.4% for GeVIR without GERP++) F1 score, and was thoroughly compared with gnomAD gene variation intolerance metrics (LOEUF, MOEUF, missense z-score) on various gene sets. Therefore, although GeVIR is based on the same basic idea as CCRs, it is a substantial step forward from the Havrilla *et al.*[33] work.

GeVIR without GERP++ outperformed gnomAD missense intolerance scores (missense z-score and MOEUF) in prioritising autosomal dominant and distinguishing them from autosomal recessive genes by 4.1% AUC and 1.2% F1 scores, respectively. The differences in AUC and F1 scores were ~0.7% and ~0.9% higher, respectively, when GERP++ evolutionary conservation adjustments were used. Moreover, GeVIR was able to prioritise Autosomal Recessive (AR) genes that were statistically significantly ($P \leq$ 2.90E-16) deficient in the last 30% of the genes (FE = ~0.45 and ~0.6 for GeVIR with and without GERP++ adjustments, respectively). In contrast, gnomAD missense intolerance metrics had an approximately expected number of AR genes in the same range. Therefore, even without evolutionary conservation adjustments, GeVIR is superior to the state-of-the-art gnomAD missense intolerance methods.

GeVIR and LOEUF had a similar performance (56.4% and 56.6% F1 score, respectively) at distinguishing AD and AR genes for approximately the top 15% of the genes (threshold suggested by LOEUF authors for the most intolerant genes). However, LEOUF had a higher peak F1 score (64.2%) than GeVIR (62.34%) at 27.8 and 29.9 gene rank percentiles, respectively, despite GeVIR being slightly better at prioritising AD genes (by ~1.1% AUC). LOEUF, similarly to GeVIR, had enrichment (FE = ~1.72, *P* = 2.99E-31) and deficiency (FE = ~0.53, *P* = 1.79E-22) of AR genes in the middle and last 30% of the ranked genes, respectively. This correlation is interesting because both metrics were calculated using substantially different statistical methods and subsets of gnomAD variants. The ability to rank all genes across the spectrum of LoF intolerance was highlighted as one of the main strengths of LOEUF[11], and our study added additional evidence that such ranking of AR genes might be natural. Although GeVIR and LOEUF showed similar performance in many assays, it is important to note that they prioritised substantially different genes. For example, only 55.9% of the most LoF intolerant genes, according to LOEUF (among approximately the top 15%), were also in the top GeVIR genes. This could be because GeVIR was less biased by gene length than LOEUF (Spearman r = -0.26 and r = -0.54, respectively) and was able to prioritise potentially important short genes (e.g. known disease genes or those performing basic biological functions such as mRNA metabolism), which was a limitation of the LOEUF metric, acknowledged by the authors. Therefore, GeVIR complements the LOEUF metric, and we showed that their combination (VIRLoF) outperformed both metrics in most assays. We recommend using GeVIR to estimate gene missense variation intolerance, especially if they are short, and using LOEUF to estimate gene LoF intolerance, especially if they are large. VIRLoF can be used when a single gene variation intolerance metric is required.

5.1.3 Chapter 4

In Chapter 4, we used supervised machine learning (ML) to develop a metric that can be used to categorise genes into three groups: autosomal dominant (AD), autosomal recessive (AR), and non-disease (Objective 4). Previous studies developed various supervised ML models for gene classification[3,15,17,18,34–38], but all of them were trained to predict only one group of genes (e.g. haploinsufficient (HI), AD, AR, Mendelian disease (MD)), and little effort was made to understand how these models work together. Therefore, we developed a single metric that continuously ranked genes (i.e. similarly to GeVIR and LOEUF) based on their Disease Inheritance Patterns (DIP).

First, we developed two supervised ML models using the k-nearest neighbour (KNN) algorithm to distinguish AD and AR genes based on their first and second interaction partners in protein-protein interaction networks (PPI-N1 and PPI-N2, respectively). Both models were trained on the DOMINO training dataset (AD = 291, AR = 693)[17]. We hypothesised that genes that interact more with AD than AR genes were also more likely to be AD and vice versa (Objective 3). Our models were compared with the DOMINO approach to PPI based features calculation (counting first degree interactions with AD genes) on the DOMINO training set using the subset of PPIs that showed the best performance in the DOMINO study[17]. The maximum achievable performance of their DOMINO approach was a 65.0% F1 score (i.e. the threshold was assigned based on all training data), whereas PPI-N1 and PPI-N2 best 10-fold cross-validation results were 66.5% and 65.5% F1 scores, respectively. Moreover, PPI-N1 and PPI-N2 models showed their best performance on the text-mining PPI subset, where they achieved 67.5% and 66.4% F1 scores, respectively. These results supported our hypothesis and showed that simultaneous analysis of AD and AR interactions could have higher discriminative power than just analysing interactions with positive samples (the approach used by DOMINO).

Second, we developed two supervised ML models using the random forest (RF) algorithm that was trained to distinguish AD from AR genes (ADR), and disease from non-disease (DND) genes. The ADR and DND model used various features (including predictions of PPI-N1 and PPI-N2 models and the GeVIR scores developed in Chapter 3) and were trained on DOMINO and GPP training sets[3,17], respectively. The ADR and DND models gene predicted probabilities were combined into DIP ranking by first ranking all the remaining “unseen” genes (n = 15,794) based on their probability of being disease causing, as predicted by DND, and then re-ranking the first half of them based on their probabilities to be AD, predicted by ADR. The DIP performance was evaluated and compared with DOMINO, GPP and VIRLoF based on the number of genes from various groups (e.g. AD, AR, cell essential/non-essential, olfactory) in the first and last quartiles of the ranked gene list. We used quartiles since DOMINO and GPP predicted similar number of genes to be AD (n = 3,807, 24.9%) and non-disease (n = 3,855, 25.2%), respectively. At this threshold, the ADR model achieved a slightly higher F1 score than DOMINO (65.1% and 63.7%, respectively), but there was no statistically significant difference in the number of genes in any examined groups. However, the DND model had significantly less known disease genes (n = 80 (5.1%)) in the last quartile than GPP (n = 145 (9.3%), *P* = 3.13E-05), DOMINO (n = 308 (19.8%), *P* = 2.08E-29), and VIRLoF (n = 171 (11.0%), *P* = 2.73E-08). Note that GeVIR was the most influential feature used by the DND model (weight = 39.4%). Moreover, the combination of ADR and DND model probabilities resulted in a significant enrichment of AR genes in the middle rank, with a peak of approximately three times more than expected AR genes in the 41-51% ranks (*P* = 2.27E-34). VIRLoF had a smaller peak of AR genes (FE = ~1.92, *P* = 5.98E-10) at the same ranks, whereas DOMINO and GPP could not effectively distinguish AR genes from cell non-essential and AD genes, respectively. Therefore, while the ADR model might be slightly better than DOMINO for prioritising AD genes, the DND model was superior for distinguishing disease and non-disease genes than the other metrics. Moreover, their combination resulted in the development of the best metric for the prioritisation of AR genes.

Finally, comparing DIP (ADR) and VIRLoF performance on the top 10% (n = 1,528) of the ranked gene lists highlighted a possible advantage of variation intolerance methods over supervised ML models. Although DIP (ADR) had slightly more AD genes than VIRLoF (n = 158 (35.4%) and n = 146 (32.7%), *P* = 5.95E-01, respectively), the latter had statistically significantly (*P* = 4.39E-02) less genes associated with both AD and AR diseases (n = 35 (15.8%) and n = 19 (8.6%), respectively), and slightly less AR (n = 33 (4.8%) and n = 25 (3.6%), respectively) and cell non-essential (n = 16 (2.4%) and n = 11 (1.7%), respectively). Considering that top-ranked genes were expected to consist of the most functionally important dominant genes (e.g. lethal), genes from all these categories might be considered false positives. However, the ADR model was trained and optimised to effectively distinguish AD from AR genes and not rank AD genes by their importance. Gene functional importance is hard to estimate, but it is expected to be correlating with variation intolerance[19]. Therefore, to combine the strengths of both metrics, we re-sorted the genes predicted to be AD by ADR with probability ≥ 0.5 (n = 3,212 (21.02%)) based on their VIRLoF scores in the final DIP ranking list.

5.1.4 Recap of objectives and main contributions

This study aimed to develop computational metrics that can be used to prioritise candidate Mendelian disease genes (i.e. classify genes as disease and non-disease) and distinguish them by probable inheritance patterns (i.e. classify disease genes as dominant and recessive). We planned to achieve this by investigating under-researched gene-level properties and developing novel features (Objectives 1, 2, and 3) that can be used to create a novel supervised machine learning model for gene classification (Objective 4). However, at the beginning of the work, it was unknown which of the under-researched gene-level properties would result in novel features and how effective they would be at distinguishing dominant, recessive and non-disease genes on their own or in combination with others.

The analysis of deviations from Hardy-Weinberg equilibrium showed that homozygous deficiency is unlikely to be used as a feature with the current size of variant population databases and the accuracy of variant genotyping methods for prioritisation of candidate recessive disease-causing variants and, consequently, genes (Objective 1). Nevertheless, we decided to publish this research to explain the potential problems of applying Hardy-Weinberg equilibrium for such tasks to future researchers. Considering the recent retraction of Wei and Nielsen's article due to a lack of understanding of genotyping errors that can result in homozygous deficiency in the scientific community[137], a systematic and large scale investigation of this problem was needed. Despite being published approximately two years ago (March 2020), our article was viewed and downloaded more than 97% (n > 25,000) and 92% (n > 2,200) of all articles from the publisher (Frontiers) over more than ten years (accessed April 2022)[a]. Therefore, although the investigation performed for Objective 2 turned out to be not helpful for the aim of this thesis, it contributed to the understanding of existing variant population data

[a] http://loop-impact.frontiersin.org/impact/article/516957#totalviews/views

and can be beneficial for other researchers, as demonstrated by the high number of views of the article.

The analysis of variant distribution patterns within genes in the gnomAD database resulted in the development of GeVIR (Objective 2), a novel gene constraint metric that outperformed existing gnomAD missense constraint metrics (Missense z-scores and MOEUF). GeVIR also complemented the LoF constraint metric (LOEUF), owing to its ability to estimate intolerance of short genes. It can be used to prioritise candidate disease genes or as a feature in machine learning models. The latter was demonstrated in Chapter 4, where GeVIR was the most influential feature (weight = 39.4%) in a model trained to distinguish candidate disease and non-disease genes. To facilitate the usage of GeVIR scores, we developed a website (www.gevirank.org) which, over approximately two years (since March 2020), was visited more than 6,900 times from 49 countries (accessed April 2022)[b]. We noticed a couple of recurrent visitors from Leipzig (Germany) and Hyderabad (India), contributing ~34% and ~24% of the total number of visits, respectively. However, considering that the data on the website is also available in supplementary materials of the published article[121], the actual number of GeVIR users is untraceable and could be higher. Nevertheless, these statistics demonstrate that some users probably used GeVIR in their everyday analyses.

The simultaneous analysis of protein-protein interactions with known genes associated exclusively with AD and AR recessive diseases (Objective 3) was performed as part of development of a novel supervised machine learning model (DIP) used to classify genes as AD, AR, and non-disease (Objective 4). DIP showed comparable performance at distinguishing AD and AR genes and was significantly better at distinguishing disease and non-disease than existing solutions (DOMINO and GPP, respectively). Moreover, unlike existing binary supervised ML models and similarly to GeVIR and LOEUF, DIP provides a single and easily interpretable metric that can be used to distinguish three studied groups of genes (AD, AR, and non-disease). However, we also showed that DIP and other supervised machine learning solutions could be biased by training gene sets, and their actual performance is hard to measure since selected testing gene sets might

[b] https://www.revolvermaps.com/livestats/56szvr48wyh/

not be representative of all genes whose status these models are trying to predict. In a field where performance is a crucial measurement of success, we believe that an open discussion about this problem is also a valuable and long required contribution.

Initially, we expected that the solution based on the supervised machine learning methods (Objective 4) would be the main contribution of this work, and other objectives (1, 2, and 3) will be stepping stones to this goal. However, it turned out that GeVIR (Objective 2) could potentially be the most impactful contribution to the field due to its performance among similar class metrics (gnomAD gene constraint scores) and ability to be used as a feature in future machine learning solutions to this or similar gene classification tasks. Nevertheless, the results of all objectives contributed to the knowledge in the investigated under-researched areas.

5.2 Limitations and future work

The analysis of deviations from Hardy-Weinberg Equilibrium (HWE) presented in Chapter 2 showed that gnomAD populations were not large enough to detect statistically significant deviations due to heterozygous excess of rare variants (e.g. AF < 0.0072 for the largest population - Non-Finish European, n = 64,603). Moreover, considering that the database was not completely free of known homozygous recessive disease-causing variants and genotype errors, this type of research probably should be delayed until a substantially more extensive variant population database becomes available (e.g. consisting of millions of individuals). Meanwhile, variants with heterozygous excess that survived our conservative filtering process, can be examined to identify overlooked types of genotype errors. The literature review of biological function of the genes with these variants might allow the formulation of speculative hypotheses and further narrow down the set of potentially heterozygous advantageous variants. Both of these tasks were not beneficial for developing gene classification methods and, therefore, were out of the scope of our study.

The main limitation of the GeVIR method, described in Chapter 3, is the requirement of variation intolerant regions to be completely free of missense and LoF variants. Although it worked reasonably well, it is naive to assume that functionally important regions in dominant genes cannot have some non-constrained amino acids. Moreover, recessive genes are expected to accumulate lots of rare disease-causing heterozygous variants in functional important regions in genes due to the presence of unaffected carriers as the database population size increases. Note that the CCR study had the same weakness[33]. Therefore, although the GeVIR study showed that variation distribution should be considered in gene variation intolerance estimations, a more sophisticated statistical approach that is robust to rare random variants in functionally important regions has to be developed to measure it. Possibly, this might require a different statistical method, assessment of variant frequencies, and their probabilities to be damaging by various other metrics.

The supervised ML gene classification models, such as ADR and DND developed in Chapter 4, might benefit from future research that improves their features (e.g. novel gene variation intolerance metrics, evolutionary conservation scores, and network analysis methods). However, the major limitation of our and previous supervised ML models is that they are trained on the gene sets that might not be representative samples of all protein-coding genes from the studied groups. We reused training datasets from DOMINO and GPP studies[3,17] to compare the performance of the models at the training stage and simultaneously test them on the maximal amount of genes unseen by all models. However, we found that ADR performance was different on training, validation, and testing gene sets (76.3%, 82.1%, and 65.1% F1 scores, respectively), and that AD genes from these sets had different variation intolerance levels based on VIRLoF (13.4%, 7.7%, and 19.7% median VIRLoF scores). Although it could be because training and validation gene sets were manually curated, it shows that the models were trained and validated on the subsets of AD genes with different profiles than the remaining known disease genes. Variation intolerance metrics are one of the most informative features used by the supervised ML models, and consequently, they could not be used to balance the training and testing gene sets. Therefore, future studies that will use supervised ML for gene classification based on disease inheritance patterns might benefit from manually curated and balanced training and testing gene sets based on associated disease severity and type (e.g. development, skeletal). However, systematic disease severity estimation on a large scale (i.e. across different types of diseases) is a substantial task that requires surveying by multiple experts[23], and actual proportions of genes associated with different types of diseases are unknown (e.g. lethal dominant genes are hard to discover). Therefore, it might be reasonable to narrow the scope and develop supervised ML models for specific diseases and their inheritance modes. This future direction was also suggested by the authors of the GPP model[3]. Alternatively, usage of unsupervised ML methods for this task could be explored.

5.3 Concluding remarks

Human disease gene discovery can be facilitated by computational methods that can prioritise candidate disease genes. The recent development of large variant population databases provided novel data for computational analysis. The scope of this study was metrics that can be used to predict inheritance patterns of candidate disease genes. A review of previous work in the literature, showed several under-researched areas that were explored in this study. In particular, we studied (1) deviations from Hardy-Weinberg equilibrium due to heterozygous excess, (2) variant distribution patterns within genes, (3) simultaneous analysis of gene interaction partners with different inheritance patterns in networks, and (4) combination of predictions from multiple supervised ML models, trained to classify different groups of genes, into a continuous gene ranking metric.

The results of the research objectives turned out to have different usability from the declared aim. However, all the work presented contributed to the understanding of variant population data and the application of supervised ML methods to classify candidate disease genes in the context of disease inheritance patterns. The main outcome of this research was the development of two continuous gene metrics, GeVIR and DIP, available for 19,361 and 15,794 protein-coding genes, respectively. Both metrics can be used to distinguish dominant, recessive and non-disease genes, and both have their strengths, limitations, and applications. Although DIP used GeVIR as one of the features and had a superior performance, GeVIR on its own can be reused in future ML models that might outperform DIP. We anticipate that both metrics will aid clinical researchers in the prioritisation of candidate disease genes.

References

1. Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. *Am J Hum Genet*. 2019;105(3):448-455. doi:10.1016/j.ajhg.2019.07.011

2. Fahrioğlu U. Problems of unknown significance: counseling in the era of next generation sequencing. *Balk J Med Genet BJMG*. 2018;21(1):73-76. doi:10.2478/bjmg-2018-0003

3. He S, Chen W, Liu H, Li S, Lei D, Dang X, Chen Y, Zhang X, Zhang J. Gene pathogenicity prediction of Mendelian diseases via the random forest algorithm. *Hum Genet*. 2019;138(6):673-679. doi:10.1007/s00439-019-02021-9

4. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

5. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet*. 2014;59(1):5-15. doi:10.1038/jhg.2013.114

6. Bromberg Y. Chapter 15: Disease gene prioritization. *PLoS Comput Biol*. 2013;9(4). doi:10.1371/journal.pcbi.1002902

7. Grimm DG, Azencott C, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat*. 2015;36(5):513-523. doi:10.1002/humu.22768

8. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003;31(13):3812-3814.

9. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet Editor Board Jonathan Haines Al*. 2013;Chapter 7:Unit7.20. doi:10.1002/0471142905.hg0720s76

10. Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S, Beutler B, Whittle B, Bertram EM, Enders A, Goodnow CC, Andrews TD. Comparison of predicted and actual consequences of

missense mutations. *Proc Natl Acad Sci*. 2015;112(37):E5189-E5198. doi:10.1073/pnas.1511585112

11. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7

12. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057

13. Alliance G, ScreeningServices TNYMAC for G and N. *Inheritance Patterns*. Genetic Alliance; 2009. Accessed April 6, 2018. https://www.ncbi.nlm.nih.gov/books/NBK115561/

14. Fisher E, Scambler P. Human haploinsufficiency — one for sorrow, two for joy. *Nat Genet*. 1994;7(1):5-7. doi:10.1038/ng0594-5

15. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and predicting haploinsufficiency in the human genome. *PLOS Genet*. 2010;6(10):e1001154. doi:10.1371/journal.pgen.1001154

16. Wilkie AO. The molecular basis of genetic dominance. *J Med Genet*. 1994;31(2):89-98.

17. Quinodoz M, Royer-Bertrand B, Cisarova K, Alessandro Di Gioia S, Superti-Furga A, Rivolta C. DOMINO: using machine learning to predict genes associated with dominant disorders. *Am J Hum Genet*. 2017;101(4):623-629. doi:10.1016/j.ajhg.2017.09.001

18. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang Z, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld J, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335(6070):823-828. doi:10.1126/science.1215040

19. Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. *Nat Rev Genet*. 2018;19(1):51-62. doi:10.1038/nrg.2017.75

20. Hart T, Tong AHY, Chan K, Leeuwen JV, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Tworzyanski L, Lawson K, Sartori MA, Alibeh S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Usaj M, Billmann M, Rahman M, Costanzo M, Myers CL, Andrews BJ, Boone C, Durocher D, Moffat J. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 Genes Genomes Genet*. 2017;7(8):2719-2727. doi:10.1534/g3.117.041277

21. Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera MDC, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol*. 2014;32(3):267-273. doi:10.1038/nbt.2800

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

23. Lazarin GA, Hawthorne F, Collins NS, Platt EA, Evans EA, Haque IS. Systematic classification of disease severity for evaluation of expanded carrier screening panels. *PLoS ONE*. 2014;9(12):e114391. doi:10.1371/journal.pone.0114391

24. Dawes R, Lek M, Cooper ST. Gene discovery informatics toolkit defines candidate genes for unexplained infertility and prenatal or infantile mortality. *Npj Genomic Med*. 2019;4(1):8. doi:10.1038/s41525-019-0081-z

25. 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*. 2019;18(1):23-29. doi:10.1093/bfgp/ely033

26. 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):e1003709. doi:10.1371/journal.pgen.1003709

27. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, DePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD,

Cook Jr EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ. A framework for the interpretation of de novo mutation in human disease. *Nat Genet*. 2014;46(9):944-950. doi:10.1038/ng.3050

28. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, Samocha KE, O'Donnell-Luria A, MacArthur DG, Daly MJ, Beier DR, Sunyaev SR. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet*. 2017;49(5):806-810. doi:10.1038/ng.3831

29. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, Daly MJ. Regional missense constraint improves variant deleteriousness prediction. *bioRxiv*. Published online June 12, 2017:148353. doi:10.1101/148353

30. 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*. 2016;17. doi:10.1186/s13059-016-0869-4

31. Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, Scott E, Ciancanelli MJ, Lafaille FG, Markle JG, Martinez-Barricarte R, de Jong SJ, Kong XF, Nitschke P, Belkadi A, Bustamante J, Puel A, Boisson-Dupuis S, Stenson PD, Gleeson JG, Cooper DN, Quintana-Murci L, Claverie JM, Zhang SY, Abel L, Casanova JL. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci U S A*. 2015;112(44):13615-13620. doi:10.1073/pnas.1518646112

32. 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):471-474. doi:10.1093/bioinformatics/btv602

33. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. *Nat Genet*. Published online December 10, 2018:1. doi:10.1038/s41588-018-0294-6

34. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. *PLOS Comput Biol*. 2013;9(3):e1002886. doi:10.1371/journal.pcbi.1002886

35. Steinberg J, Honti F, Meader S, Webber C. Haploinsufficiency predictions without study bias. *Nucleic Acids Res*. 2015;43(15):e101-e101. doi:10.1093/nar/gkv474

36. Hsu JS, Kwan JSH, Pan Z, Garcia-Barcelo MM, Sham PC, Li M. Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. *Bioinformatics*. 2016;32(20):3065-3071. doi:10.1093/bioinformatics/btw381

37. Shihab HA, Rogers MF, Campbell C, Gaunt TR. HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinforma Oxf Engl Bioinforma*. 2017;33, 33(12, 12):1751, 1751-1757. doi:10.1093/bioinformatics/btx028, 10.1093/bioinformatics/btx028

38. Han X, Chen S, Flynn E, Wu S, Wintner D, Shen Y. Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nat Commun*. 2018;9(1):2138. doi:10.1038/s41467-018-04552-7

39. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332. doi:10.1038/nrg3920

40. Nandi S, Ganguli P, Sarkar RR. Essential gene prediction using limited gene essentiality information – an integrative semi-supervised machine learning strategy. *PLOS ONE*. 2020;15(11):e0242943. doi:10.1371/journal.pone.0242943

41. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database Issue):D514-D517. doi:10.1093/nar/gki033

42. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. *Nat Genet*. Published online April 8, 2019:1. doi:10.1038/s41588-019-0383-1

43. Dimensions - the next evolution in linked scholarly information. Dimensions. Accessed August 2, 2021. https://www.dimensions.ai/

44. The cost of sequencing a human genome. Genome.gov. Accessed August 6, 2021. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

45. Behjati S, Tarpey PS. What is next generation sequencing? *Arch Dis Child Educ Pract Ed*. 2013;98(6):236-238. doi:10.1136/archdischild-2013-304340

46. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med*. 2018;20(10):1122-1130. doi:10.1038/gim.2017.247

47. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15(9):733-747. doi:10.1038/gim.2013.92

48. Barbitoff YA, Polev DE, Glotov AS, Serebryakova EA, Shcherbakova IV, Kiselev AM, Kostareva AA, Glotov OS, Predeus AV. Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci Rep*. 2020;10(1):2057. doi:10.1038/s41598-020-59026-y

49. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci*. 2015;112(17):5473-5478. doi:10.1073/pnas.1418631112

50. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, Barrett JC, Biesecker LG,

Conrad DF, Cooper GM, Cox NJ, Daly MJ, Gerstein MB, Goldstein DB, Hirschhorn JN, Leal SM, Pennacchio LA, Stamatoyannopoulos JA, Sunyaev SR, Valle D, Voight BF, Winckler W, Gunter C. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508(7497):469-476. doi:10.1038/nature13127

51. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181-1186. doi:10.1038/ng1007-1181

52. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Mark J. Rieder, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, NHLBI Exome Sequencing Project, Akey JM. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013;493(7431):216-220. doi:10.1038/nature11690

53. Consortium T 1000 GP. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073. doi:10.1038/nature09534

54. Consortium T 1000 GP. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. doi:10.1038/nature11632

55. gnomAD v3.1 | gnomAD news. Accessed August 5, 2021. https://gnomad.broadinstitute.org/news/2020-10-gnomad-v3-1/

56. Exome variant server. Accessed June 23, 2021. https://evs.gs.washington.edu/EVS/

57. Francioli L. gnomAD v2.1. MacArthur Lab. Published October 17, 2018. Accessed November 13, 2018. https://macarthurlab.org/2018/10/17/gnomad-v2-1/

58. Francioli L. gnomAD v3.0. MacArthur Lab. Published October 16, 2019. Accessed August 5, 2021. https://macarthurlab.org/2019/10/16/gnomad-v3-0/

59. Kobayashi Y, Yang S, Nykamp K, Garcia J, Lincoln SE, Topper SE. Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med*. 2017;9. doi:10.1186/s13073-017-0403-7

60. Tarailo-Graovac M, Zhu JYA, Matthews A, van Karnebeek CDM, Wasserman WW. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet Med*. Published online May 4, 2017. doi:10.1038/gim.2017.50

61. Hall CL, Sutanto H, Dalageorgou C, McKenna WJ, Syrris P, Futema M. Frequency of genetic variants associated with arrhythmogenic right ventricular cardiomyopathy in the genome aggregation database. *Eur J Hum Genet*. 2018;26(9):1312-1318. doi:10.1038/s41431-018-0169-4

62. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet*. 2013;132(10):1077-1130. doi:10.1007/s00439-013-1331-2

63. Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, Zhang J, Zhou H, Tian L, Prakash O, Lemire M, Sleiman P, Cheng W yi, Chen W, Shah H, Shen Y, Fromer M, Omberg L, Deardorff MA, Zackai E, Bobe JR, Levin E, Hudson TJ, Groop L, Wang J, Hakonarson H, Wojcicki A, Diaz GA, Edelmann L, Schadt EE, Friend SH. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol*. 2016;34(5):nbt.3514. doi:10.1038/nbt.3514

64. Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, Barton PJR, Funke B, Cook SA, MacArthur D, Ware JS. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med*. 2017;19(10):1151-1158. doi:10.1038/gim.2017.26

65. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL, Committee on behalf of the ALQA. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-423. doi:10.1038/gim.2015.30

66. Mugal CF, Wolf JBW, Kaj I. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol Biol Evol*. 2014;31(1):212-231. doi:10.1093/molbev/mst192

67. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, Civello D, Adams MD, Cargill M, Clark AG. Natural selection on protein-coding genes in the human genome. *Nature*. 2005;437(7062):1153-1157. doi:10.1038/nature04240

68. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. doi:10.1038/ng.2892

69. ClinGen genome dosage map. Accessed January 31, 2021. https://dosage.clinicalgenome.org/

70. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M. Natural selection on genes that underlie human disease susceptibility. *Curr Biol CB*. 2008;18(12):883-889. doi:10.1016/j.cub.2008.04.074

71. Graffelman J, Jain D, Weir B. A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. *Hum Genet*. 2017;136(6):727-741. doi:10.1007/s00439-017-1786-7

72. Chen B, Cole JW, Grond-Ginsbach C. Departure from Hardy Weinberg equilibrium and genotyping error. *Front Genet*. 2017;8. doi:10.3389/fgene.2017.00167

73. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. *Proc Natl Acad Sci U S A*. 2013;110(24):9851-9855. doi:10.1073/pnas.1302575110

74. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJ, van Karnebeek C, Wasserman WW. FLAGS, frequently mutated genes in public exomes. *BMC Med Genomics*. 2014;7(1):64. doi:10.1186/s12920-014-0064-y

75. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015;43(D1):D447-D452. doi:10.1093/nar/gku1003

76. Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. In: *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. ; 2016:1310-1315.

77. Davydov EV, 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):e1001025. doi:10.1371/journal.pcbi.1001025

78. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110-121. doi:10.1101/gr.097857.109

79. Ripley BD. *Pattern Recognition and Neural Networks*. Cambridge University Press; 1996. doi:10.1017/CBO9780511812651

80. Robinson PN, Webber C. Phenotype ontologies and cross-species analysis for translational research. *PLOS Genet*. 2014;10(4):e1004268. doi:10.1371/journal.pgen.1004268

81. Ronemus M, Iossifov I, Levy D, Wigler M. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet*. 2014;15(2):133-141. doi:10.1038/nrg3585

82. Edwards AWF. G. H. Hardy (1908) and Hardy–Weinberg equilibrium. *Genetics*. 2008;179(3):1143-1150. doi:10.1534/genetics.104.92940

83. Thiessen D, Gregg B. Human assortative mating and genetic equilibrium: An evolutionary perspective. *Ethol Sociobiol*. 1980;1(2):111-140. doi:10.1016/0162-3095(80)90003-5

84. Garnier-Géré P, Chikhi L. Population subdivision, Hardy–Weinberg equilibrium and the wahlund effect. In: *ELS*. American Cancer Society; 2013. doi:10.1002/9780470015902.a0005446.pub3

85. Sinnock P. The Wahlund effect for the two-locus model. *Am Nat*. 1975;109(969):565-570.

86. About dbSNP reference (rs) number. Accessed April 16, 2022. https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/

87. Ashley-Koch A, Yang Q, Olney RS. Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. *Am J Epidemiol*. 2000;151(9):839-845.

88. Withrock IC, Anderson SJ, Jefferson MA, McCormack GR, Mlynarczyk GSA, Nakama A, Lange JK, Berg CA, Acharya S, Stock ML, Lind MS, Luna KC, Kondru NC, Manne S, Patel BB, de la Rosa BM, Huang KP, Sharma S, Hu HZ, Kanuri SH, Carlson SA. Genetic diseases conferring resistance to infectious diseases. *Genes Dis*. 2015;2(3):247-254. doi:10.1016/j.gendis.2015.02.008

89. Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br Med J*. 1954;1(4857):290-294.

90. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*. 2016;44(D1):D862-868. doi:10.1093/nar/gkv1222

91. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux AF, Smith T, Antonarakis SE, Taschner PEM. HGVS recommendations for the description of sequence variants: 2016 update. *Hum Mutat*. 2016;37(6):564-569. doi:10.1002/humu.22981

92. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. *Genome Biol*. 2016;17:122. doi:10.1186/s13059-016-0974-4

93. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47(D1):D766-D773. doi:10.1093/nar/gky955

94. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 2005;76(5):887-893.

95. Graffelman J, Moreno V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat Appl Genet Mol Biol*. 2013;12(4):433-448. doi:10.1515/sagmb-2012-0039

96. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17(3):261-272. doi:10.1038/s41592-019-0686-2

97. Karczewski KJ, Gauthier LD, Daly MJ. Technical artifact drives apparent deviation from Hardy-Weinberg equilibrium at CCR5-Δ32 and other variants in gnomAD. *bioRxiv*. Published online October 2, 2019:784157. doi:10.1101/784157

98. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. *Science*. 2002;297(5583):1003-1007. doi:10.1126/science.1072047

99. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573-580.

100. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Hinrichs AS, Gonzalez JN, Gibson D, Diekhans M, Clawson H, Casper J, Barber GP, Haussler D, Kuhn RM, Kent WJ. The UCSC genome browser database: 2019 update. *Nucleic Acids Res*. 2019;47(Database issue):D853-D858. doi:10.1093/nar/gky1095

101. Muyas F, Bosio M, Puig A, Susak H, Domènech L, Escaramis G, Zapata L, Demidov G, Estivill X, Rabionet R, Ossowski S. Allele balance bias identifies systematic genotyping errors and false disease associations. *Hum Mutat*. 2019;40(1):115-126. doi:10.1002/humu.23674

102. Rodman DM, Zamudio S. The cystic fibrosis heterozygote--advantage in surviving cholera? *Med Hypotheses*. 1991;36(3):253-258.

103. Bosch L, Bosch B, De Boeck K, Nawrot T, Meyts I, Vanneste D, Le Bourlegat CA, Croda J, da Silva Filho LVRF. Cystic fibrosis carriership and tuberculosis: hints toward an evolutionary selective advantage based on data from the Brazilian territory. *BMC Infect Dis*. 2017;17. doi:10.1186/s12879-017-2448-z

104. Gazal S, Sahbatou M, Babron MC, Génin E, Leutenegger AL. High level of inbreeding in final phase of 1000 Genomes Project. *Sci Rep*. 2015;5:17453. doi:10.1038/srep17453

105. Kargapolova Y, Rehimi R, Kayserili H, Brühl J, Sofiadis K, Zirkel A, Palikyras S, Mizi A, Li Y, Yigit G, Hoischen A, Frank S, Russ N, Trautwein J, van Bon B, Gilissen C, Laugsch M, Gusmao EG, Josipovic N, Altmüller J, Nürnberg P, Längst G, Kaiser FJ, Watrin E, Brunner H, Rada-Iglesias A, Kurian L, Wollnik B, Bouazoune K, Papantonis A. Overarching control of autophagy and DNA damage response by CHD6 revealed by modeling a rare human pathology. *Nat Commun*. 2021;12(1):3014. doi:10.1038/s41467-021-23327-1

106. Alfonso Roberto, Lutz Thomas, Rodriguez Ariel, Chavez J. Pablo, Rodriguez Paloma, Gutierrez Silvia, Nieto Amelia. CHD6 chromatin remodeler is a negative modulator of influenza virus replication that relocates to inactive chromatin upon infection. *Cell Microbiol*. 2011;13(12):1894-1906. doi:10.1111/j.1462-5822.2011.01679.x

107. Alfonso R, Rodriguez A, Rodriguez P, Lutz T, Nieto A. CHD6, a cellular repressor of influenza virus replication, is degraded in human alveolar epithelial cells and mice lungs during infection. *J Virol*. 2013;87(8):4534-4544. doi:10.1128/JVI.00554-12

108. Cross NCP, Tolan DR, Cox TM. Catalytic deficiency of human aldolase B in hereditary fructose intolerance caused by a common missense mutation. *Cell*. 1988;53(6):881-885. doi:10.1016/S0092-8674(88)90349-2

109. UK Department of Health and Social Care. Matt Hancock announces ambition to map 5 million genomes. GOV.UK. Published October 2, 2018. Accessed September 20, 2019. https://www.gov.uk/government/news/matt-hancock-announces-ambition-to-map-5-million-genomes

110. Sivley M. Comprehensive analysis of constraint on the spatial distribution of missense variants in human protein structures. *Am J Hum Genet*. 2018;102(3):415-426. doi:10.1016/j.ajhg.2018.01.017

111. Karczewski K. The genome aggregation database (gnomAD). MacArthur Lab. Published February 27, 2017. Accessed November 13, 2018. https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

112. Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E, Sigurdsson GT, Jonasdottir A, Jonasdottir A, Sigurdsson A, Magnusson OT, Kong A, Helgason A, Holm H, Thorsteinsdottir U, Masson G, Gudbjartsson DF, Stefansson K. Identification of a large set of rare complete human knockouts. *Nat Genet*. 2015;47(5):448-452. doi:10.1038/ng.3243

113. Motenko H, Neuhauser SB, O'Keefe M, Richardson JE. MouseMine: a new data warehouse for MGI. *Mamm Genome*. 2015;26(7-8):325-330. doi:10.1007/s00335-015-9573-z

114. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database Group. The mouse genome database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res*. 2015;43(Database issue):D726-736. doi:10.1093/nar/gku967

115. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res*. 2017;45(Database issue):D619-D625. doi:10.1093/nar/gkw1033

116. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825−2830.

117. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

118. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923

119. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24(8):2125-2137. doi:10.1093/hmg/ddu733

120. Walters-Sen LC, Hashimoto S, Thrush DL, Reshmi S, Gastier-Foster JM, Astbury C, Pyatt RE. Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Mol Genet Genomic Med*. 2015;3(2):99-110. doi:10.1002/mgg3.116

121. 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*. 2020;52(1):35-39. doi:10.1038/s41588-019-0560-2

122. Braschi B, Denny P, Gray K, Jones T, Seal R, Tweedie S, Yates B, Bruford E. Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res*. 2019;47(D1):D786-D792. doi:10.1093/nar/gky930

123. Zhao G, Li K, Li B, Wang Z, Fang Z, Wang X, Zhang Y, Luo T, Zhou Q, Wang L, Xie Y, Wang Y, Chen Q, Xia L, Tang Y, Tang B, Xia K, Li J. Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res*. 2020;48(D1):D913-D926. doi:10.1093/nar/gkz923

124. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome Database Group. Mouse genome database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res*. 2018;46(D1):D836-D842. doi:10.1093/nar/gkx1006

125. Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, Meehan TF, Weninger WJ, Westerberg H, Adissu H, Baker CN, Bower L, Brown JM, Caddle LB, Chiani F, Clary D, Cleak J, Daly MJ, Denegre JM, Doe B, Dolan ME, Edie SM, Fuchs H, Gailus-Durner V, Galli A, Gambadoro A, Gallegos J, Guo S, Horner NR, Hsu CW, Johnson SJ, Kalaga S, Keith LC, Lanoue L, Lawson TN, Lek M, Mark M, Marschall S, Mason J, McElwee ML, Newbigging S, Nutter LMJ, Peterson KA, Ramirez-Solis R, Rowland DJ, Ryder E, Samocha KE, Seavitt JR, Selloum M, Szoke-Kovacs Z, Tamura M, Trainor AG, Tudose I, Wakana S, Warren J, Wendling O, West DB, Wong L, Yoshiki A, The International Mouse Phenotyping Consortium, MacArthur DG, Tocchini-Valentini GP, Gao X, Flicek P, Bradley A, Skarnes WC, Justice MJ, Parkinson HE, Moore M, Wells S, Braun RE, Svenson KL, de Angelis MH, Herault Y, Mohun T, Mallon AM, Henkelman RM, Brown SDM, Adams DJ, Lloyd KCK, McKerlie C, Beaudet AL, Bućan M, Murray SA. High-throughput discovery of novel developmental phenotypes. *Nature*. 2016;537(7621):508-514. doi:10.1038/nature19356

126. Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, Blake A, Chen CK, Easty R, Di Fenza A, Fiegel T, Grifiths M, Horne A, Karp NA, Kurbatova N, Mason JC, Matthews P, Oakley DJ, Qazi A, Regnart J, Retha A, Santos LA, Sneddon DJ, Warren J, Westerberg H, Wilson RJ, Melvin DG, Smedley D, Brown SDM, Flicek P, Skarnes WC, Mallon AM, Parkinson H. The international mouse phenotyping consortium web portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res*. 2014;42(Database issue):D802-809. doi:10.1093/nar/gkt977

127. Mainland JD, Li YR, Zhou T, Liu WLL, Matsunami H. Human olfactory receptor responses to odorants. *Sci Data*. 2015;2(1):150002. doi:10.1038/sdata.2015.2

128. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, Watts NA, Solomonson M, O'Donnell-Luria A, Baumann A, Munshi R, Walker M, Whelan CW, Huang Y, Brookings T, Sharpe T, Stone MR, Valkanas E, Fu J, Tiao G, Laricchia KM, Ruano-Rubio V, Stevens C, Gupta N, Cusick C, Margolin L, Taylor KD, Lin HJ, Rich SS, Post WS, Chen YDI, Rotter JI, Nusbaum C, Philippakis A, Lander E, Gabriel S, Neale BM, Kathiresan S, Daly MJ, Banks E, MacArthur DG, Talkowski ME. A structural variation reference for medical and population genetics. *Nature*. 2020;581(7809):444-451. doi:10.1038/s41586-020-2287-8

129. Huang YF. Unified inference of missense variant effects and gene constraints in the human genome. *PLOS Genet*. 2020;16(7):e1008922. doi:10.1371/journal.pgen.1008922

130. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, Billis K, Boddu S, Marugán JC, Cummins C, Davidson C, Dodiya K, Fatima R, Gall A, Giron CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Lavidas I, Le T, Lemos D, Martinez JG, Maurel T, McDowall M, McMahon A, Mohanan S, Moore B, Nuhn M, Oheh DN, Parker A, Parton A, Patricio M, Sakthivel MP, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sycheva M, Szuba M, Taylor K, Thormann A, Threadgold G, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Flint B, Frankish A, Hunt SE, IIsley G, Kostadima M, Langridge N, Loveland JE, Martin FJ, Morales J, Mudge JM, Muffato M, Perry E, Ruffier M, Trevanion SJ, Cunningham F, Howe KL, Zerbino DR, Flicek P. Ensembl 2020. *Nucleic Acids Res*. 2020;48(D1):D682-D688. doi:10.1093/nar/gkz966

131. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, The 1000 Genomes

Project Consortium, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75-81. doi:10.1038/nature15394

132. Botia JA, Guelfi S, Zhang D, D'Sa K, Reynolds R, Onah D, McDonagh EM, Rueda-Martin A, Tucci A, Rendon A, Houlden H, Hardy J, Ryten M. G2P: Using machine learning to understand and predict genes causing rare neurological disorders. *bioRxiv*. Published online March 27, 2018:288845. doi:10.1101/288845

133. Li J feng, Wang L, Dang X, Feng WM, Ma YT, He SJ, Liang L, Yang HM, Liu HK, Zhang JG. Gene prioritization based on systems biology revealed new insight into genetic basis and pathophysiology underlying schizophrenia. *medRxiv*. Published online June 29, 2020:2020.06.26.20140541. doi:10.1101/2020.06.26.20140541

134. Wei X, Nielsen R. CCR5 -Δ32 is deleterious in the homozygous state in humans. *Nat Med*. 2019;25(6):909-910. doi:10.1038/s41591-019-0459-6

135. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, Saragosti S, Lapouméroulie C, Cognaux J, Forceille C, Muyldermans G, Verhofstede C, Burtonboy G, Georges M, Imai T, Rana S, Yi Y, Smyth RJ, Collman RG, Doms RW, Vassart G, Parmentier M. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*. 1996;382(6593):722-725. doi:10.1038/382722a0

136. Falcon A, Cuevas MT, Rodriguez-Frandsen A, Reyes N, Pozo F, Moreno S, Ledesma J, Martínez-Alarcón J, Nieto A, Casas I. CCR5 deficiency predisposes to fatal outcome in influenza virus infection. *J Gen Virol*. 2015;96(8):2074-2078. doi:10.1099/vir.0.000165

137. Wei X, Nielsen R. Retraction Note: CCR5-Δ32 is deleterious in the homozygous state in humans. *Nat Med*. 2019;25(11):1796-1796. doi:10.1038/s41591-019-0637-6

Appendices

A Supplementary material for Chapter 2

A.1 Data availability

Publicly available datasets were analyzed in this study. This data can be found at the following links:

https://console.cloud.google.com/storage/browser/gnomad-public/release/2.1.1/;
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/.

A.2 Code availability

Code for data analysis and figures can be found at https://github.com/niab/hwe.

A.3 Supplementary tables

Table A.1: Dataset of variants deviating from Hardy-Weinberg Equilibrium due to heterozygote excess (HetExc).

Please see supplementary Excel document at https://data.mendeley.com/datasets/f5v5t2kkvm/2 OR Supplementary Table 1 at https://www.frontiersin.org/articles/10.3389/fgene.2020.00210/full#supplementary-material

Table A.2: Statistical comparison of variants deviating from HWE due to HetExc that are located in segmental duplication (a) or tandem repeat (b) regions with the reference (Ref) group (i.e., all other regions except segmental duplications and tandem repeats).

a		
	HetExc	**All**

Segmental Duplication	56	2676
Ref	348	40801
fold-enrichemnt	2.45	
P	2.045E-08	

b		
	HetExc	All
Tandem Repeat	11	1182
Ref	348	40801
fold-enrichemnt	1.09	
P	0.75	

Table A.3: Statistical comparison of variants with Variant Carriers with "Normal" Allele Balance VCNAB < 50% in the whole Ref group and a subset of variants with statistically significant excess of heterozygotes (HetExc) in the Ref group.

	VCNAB < 50%	All
Ref	1181	40801
Ref HetExc	49	348
fold-enrichemnt	4.86	
P	1.92E-17	

Table A.4: Statistical comparison of proportions of variants deviating and not deviating from HWE due to excess of heterozygotes (HetExc and HetExc-, respectively) in 7 ethnic gnomAD populations (a–g).

a		
	NFE	all populations
HetExc	2	161
HetExc-	11499	39430
fold-enrichemnt	0.04	
P	1.54E-15	

b		
	AMR	all populations
HetExc	1	161
HetExc-	2692	39430
fold-enrichemnt	0.09	

P	5.35E-04	

c		
	SAS	**all populations**
HetExc	0	161
HetExc-	5061	39430
fold-enrichemnt	0	
P	5.69E-09	

d		
	FIN	**all populations**
HetExc	7	161
HetExc-	5914	39430
fold-enrichemnt	0.29	
P	1.96E-04	

e		
	AFR	**all populations**
HetExc	128	161
HetExc-	18957	39430
fold-enrichemnt	1.65	
P	3.01E-05	

f		
	EAS	**all populations**
HetExc	18	161
HetExc-	3621	39430
fold-enrichemnt	1.22	
P	0.42	

g		
	ASJ	**all populations**
HetExc	5	161
HetExc-	2621	39430

fold-enrichemnt	0.47	
P	0.10	

Table A.5: Statistical comparison of missense (a), synonymous (b), and other (c) variant proportions in HetExc and HetExc- datasets.

a		
	Missense	**All**
HetExc	84	161
HetExc-	18808	39430
fold-enrichemnt	0.91	
P	0.49	

b		
	Synonymous	**All**
HetExc	63	161
HetExc-	16954	39430
fold-enrichemnt	1.10	
P	0.56	

c		
	Other	**All**
HetExc	14	161
HetExc-	3668	39430
fold-enrichemnt	1.07	
P	1	

Table A.6: Statistical comparison of proportions of “AD” (a), “AR or AR, AD” (b) and all genes with at least one variant in HetExc and HetExc- datasets.

a		
	AD	**All**
HetExc	8	149
HetExc-	617	11842
fold-enrichemnt	1.03	
P	0.85	

b		
	AR or AR,AD	**All**

HetExc	29	149
HetExc-	1418	11842
fold-enrichemnt	1.63	
P	0.02	

Table A.7: Statistical comparison of Allele Frequencies of heterozygote excess (HetExc) variants in *HBB* and *CHD6* genes between African and African American population in the 1000 Genomes database.

Please see supplementary Excel document at
https://data.mendeley.com/datasets/f5v5t2kkvm/2
OR
Supplementary Table 7 at
https://www.frontiersin.org/articles/10.3389/fgene.2020.00210/full#supplementary-material

B Supplementary material for Chapter 3

B.1 Data availability

The GERP++ file can be found at http://mendel.stanford.edu/SidowLab/downloads/gerp/hg19.GERP_scores.tar.gz.

The ClinVar files can be found at ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz and ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/var_citations.txt.

The CCR files can be found at https://s3.us-east-2.amazonaws.com/ccrs/ccrs/ccrs.autosomes.v2.20180420.bed.gz and https://s3.us-east-2.amazonaws.com/ccrs/ccrs/ccrs.xchrom.v2.20180420.bed.gz.

The OMIM genemap2.txt file can be found, after registration, at https://omim.org/downloads.

The gnomAD gene constraint metric file can be found at https://storage.googleapis.com/gnomad-public/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_transcript.txt.bgz.

The gnomAD exomes variants and coverage files can be found at https://storage.googleapis.com/gnomad-public/release/2.0.2/vcf/exomes/gnomad.exomes.r2.0.2.sites.vcf.bgz and https://storage.googleapis.com/gnomad-public/release/2.0.2/coverage/combined_tars/gnomad.exomes.r2.0.2.coverage.all.tar, respectively.

The gnomAD genomes variants files can be found at https://storage.googleapis.com/gnomad-public/release/2.0.2/vcf/genomes/gnomad.genomes.r2.0.2.sites.coding_only.chr1-22.vcf.bgz and https://storage.googleapis.com/gnomad-public/release/2.0.2/vcf/genomes/gnomad.genomes.r2.0.2.sites.coding_only.chrX.vcf.bgz.

The gnomAD genes, transcripts and exons files can be found at http://broadinstitute.org/~konradk/exac_browser/exac_browser.tar.gz.

The Ensembl coding and peptide sequences from build GRCh37/hg19 can be found at https://grch37.ensembl.org/biomart/martview (data set: Human genes (GRCh37.p13); Attributes → Sequences → 'Coding sequence' and 'Peptide').

The homozygous LOF tolerant genes (that is, nulls) can be found at https://github.com/macarthur-lab/gene_lists/blob/master/lists/homozygous_lof_tolerant_twohit.tsv.

The cell essential and non-essential genes from CRISPR–Cas experiments can be found at https://github.com/macarthur-lab/gene_lists/blob/master/lists/CEGv2_subset_universe.tsv and https://github.com/macarthur-lab/gene_lists/blob/master/lists/NEGv1_subset_universe.tsv, respectively.

The mouse heterozygous lethal genes can be obtained from http://www.mousemine.org/ by querying the database with the following search terms: path = "OntologyAnnotation.ontologyTerm" type = "MPTerm"; path = "OntologyAnnotation.subject" type = "SequenceFeature"; path = "OntologyAnnotation.evidence.baseAnnotations.subject" type = "Genotype"; path = "OntologyAnnotation.evidence.baseAnnotations.subject.zygosity" op = "=" value = "ht" code = "B"; path = "OntologyAnnotation.ontologyTerm. name" op = "CONTAINS" value = "lethal".

The human–mouse ortholog mapping file can be found at http://www.informatics.jax.org/downloads/reports/HMD_HumanPhenotype.rpt.

The HGNC approved gene symbols can be found at https://www.genenames.org/download/statistics-and-files.

B.2 Code availability

Code for calculating GeVIR/VIRLoF scores, data analysis and figures can be found at https://github.com/gevirank/gevir. Computed GeVIR/VIRLoF scores are available in Appendix Table B.3.

B.3 Supplementary notes

Evolutionary conservation adjustments

Although GeVIR without evolutionary conservation adjustments (GERP++) performed reasonably well (Table 3.1), they are an important part of the method and improve its performance, especially at deprioritisation of potentially non-important genes (i.e. nulls and cell non-essential). Null and cell non-essential genes probably had shorter VIRs in general due to the presence of a larger number of variants and, consequently, their weights relied more on GERP++ adjustments. We used the GERP++ metric to measure conservation because it had a few important properties which allowed us to easily integrate it into our model.

Firstly, GERP++ conservation of each nucleotide is independent from adjacent nucleotides, which allowed us to precisely measure evolutionary conservation of a region without including data from positions affected by variants on the borders, that might be less conserved.

Secondly, GERP++ scores can be negative, which allowed us to penalise genes with “suspicious” regions, as we assumed that real variant conserved regions should have at least positive evolutionary conservation scores. For example, there were 94 genes with GeVIR ranking >70% and GeVIR without GERP++ ranking <30% (i.e. significantly shifted from variant tolerant to intolerant ranks when VIR evolutionary conservation was not considered). However, none of them belong to the AD group (expected to be enriched with GeVIR ranking <30%) and 8 were cell non-essential (expected to be enriched with GeVIR ranking >70%). Therefore, these genes might contain long VIRs due to variant filtering errors (e.g. real variants were classified as non-pass quality (i.e. “false negatives”) and were not considered when VIRs were calculated).

Finally, we wanted to use evolutionary conservation to support the signal obtained from variant distribution and not vice versa. Although GERP++ scores vary from approximately -12 to 6, 99% of the VIRs had a mean GERP++ score between -3.62 and 5.64 (Appendix Figure B.1). Additionally region conservation scores that were less than 1 or -1 were rounded to 1 or -1 (mean 0 conservation scores remained) to avoid extreme penalties from multiplication with region weights (e.g. evolutionary conservation of 0.001 would reduce its weight 1000 times). This rounding had a minor impact on the gene scores and GeVIR performance is nearly identical without it (data not shown), when the scores were computed using gnomAD data. However, it ensured that, in general, VIR weights were increased up to ~5.6 times or remained unchanged if they had positive evolutionary conservation score. Negative conservation could increase the weights by up to ~3.6 times, but transform it into a penalty. For example, an autosomal VIR with length 5 (weight = 14.90) and maximal evolutionary conservation (mean GERP++ = 6.18), would still have lower adjusted weight (14.90 × 6.18 = 92.082), than any VIR with length 10 (weight = 105.68) and positive evolutionary conservation. Therefore, GeVIR scores were mostly driven by variant distribution within the genes, whereas evolution adjustments performed smooth correction and allowed possible “false positive” long VIRs to be taken into account.

We anticipate that future improvements in variant filtering and sequencing strategies will allow us to rely more on variant free regions, but at the moment excluding additional evolutionary conservation checks from our GeVIR method would likely prioritise ‘false positive’ variant intolerant genes.

Performance comparison of GeVIR, LOEUF and probability of Loss-of-Function Intolerance (pLI)

The latest version of gnomAD (v2.1) contains new gene Loss-of-function (LoF) and missense constraint metrics based on confidence intervals (CI) of observed over expected variants in genes, together with recalculated pLI and missense z-scores. The authors suggest the use of loss-of-function observed/expected upper bound fraction (LOEUF) instead of pLI as a metric for LoF intolerant genes, because LOEUF scores are continuous and easer to interpret[c]. Unlike LOEUF, which allows to rank all genes based on their intolerance to LoF variation (i.e. a continuous metric like GeVIR), pLI is suggested to be used to separate potentially haploinsufficient genes (pLI > 0.9) from the others (i.e. a non-continuous metric). Interpretation of LoF intolerance of genes with low pLI requires the consideration of two accompanying metrics that represent gene probability of being Recessive (pRec) or Null (pNull) (sum of pLI, pRec and pNull scores of each gene equals 1), but they were recommend to be used with caution by the authors.

Ranking all genes based on pLI scores would be a misuse of the metric since the majority of the genes have extremely low pLI (e.g. 10,398 genes have pLI < 0.01) and their classification to Recessive and Null groups was not possible without taking into account the pRec and pNull scores. Therefore, pLI performance was not assessed alongside other metrics in our study due to lack of continuity. However, since pLI is still a widely used metric we compared GeVIR, LOEUF, and pLI performance in the context of Autosomal Dominant (AD) and Autosomal Recessive (AR) gene classification. We selected 3,129 genes with pLI > 0.9 and the same number of genes ranked the most intolerant to variation by GeVIR and LOEUF. Note that we used pLI and LOEUF metrics calculated for gene canonical transcripts analysed in this study (19,361 genes). The pLI gene set was nearly identical to LOEUF (2,964/3,129, ~95%), and although pLI performed slightly better (F1 = 57.5%) than LOEUF (F1 = 57.0%) or GeVIR (56.7%) the difference was minor.

Interpretation of GeVIR, LOEUF and VIRLoF metrics on the website

To help interpret GeVIR percentiles (%) for each gene X (n = 19,361) in a list we calculated fold enrichment of AD (n = 790) and AR (n = 1,585) genes in a range of up to ±5% of gene X (Appendix Figure B.5). Statistical significance of AD or AR gene enrichment was measured with two-sided Fisher's exact test by comparing number of AD or AR genes (separately) in the examined range, to the number of AD or AR genes in all genes with GeVIR scores. Exact group sizes and numbers of AD and AR genes for each gene are available in Appendix Table B.3. For example, GeVIR ranks RBBP5, a

[c] "Constraint" section in MacArthur Lab blog post: https://macarthurlab.org/2018/10/17/gnomad-v2-1/

gene not associated with any disease, in the top 2.13%, so the range examined is from 0 to 7.13% (1,380 genes). Genes in this range are ~4 times more often associated with AD disease (fold-enrichment (FE) = 4.2, two-sided Fisher's exact test *P* = 2.13e-68), and ~4 times less often associated with AR disease (FE = ~0.25, *P* = 1.63e-21). Therefore, based on what is currently known about genes with similar GeVIR %, RBBP5 is more likely to be associated with AD than AR disease. Similarly, we calculated AD and AR gene enrichment metrics for LOEUF and VIRLoF percentiles. Metrics for all analysed genes are available in Appendix Table B.3 and on our GeVIR website (www.gevirank.org) which supports batch queries.

B.4 Supplementary figures

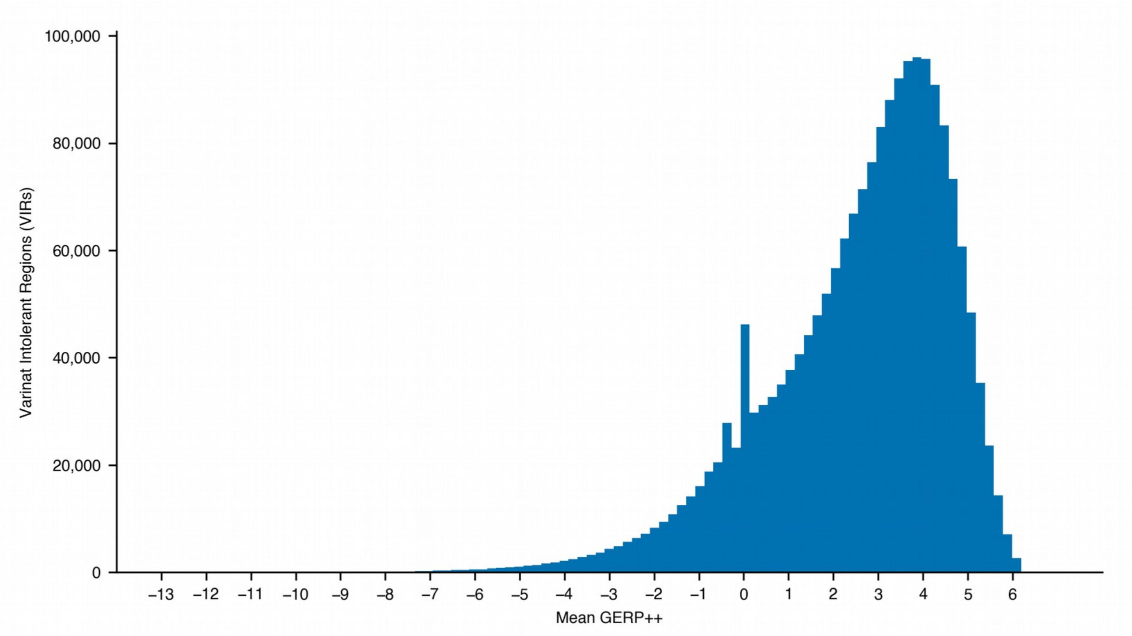


Figure B.1: Evolutionary conservation (GERP++) of Variant Intolerant Regions (VIRs) used in GeVIR score calculation.

N = 1,938,102 VIRs with high coverage and length ≥1 in 18,491 genes, note that 1,570,941 VIRs with 0 length (i.e. regions between adjacent variants) were assigned 0 GERP++ score and are not shown on the figure.

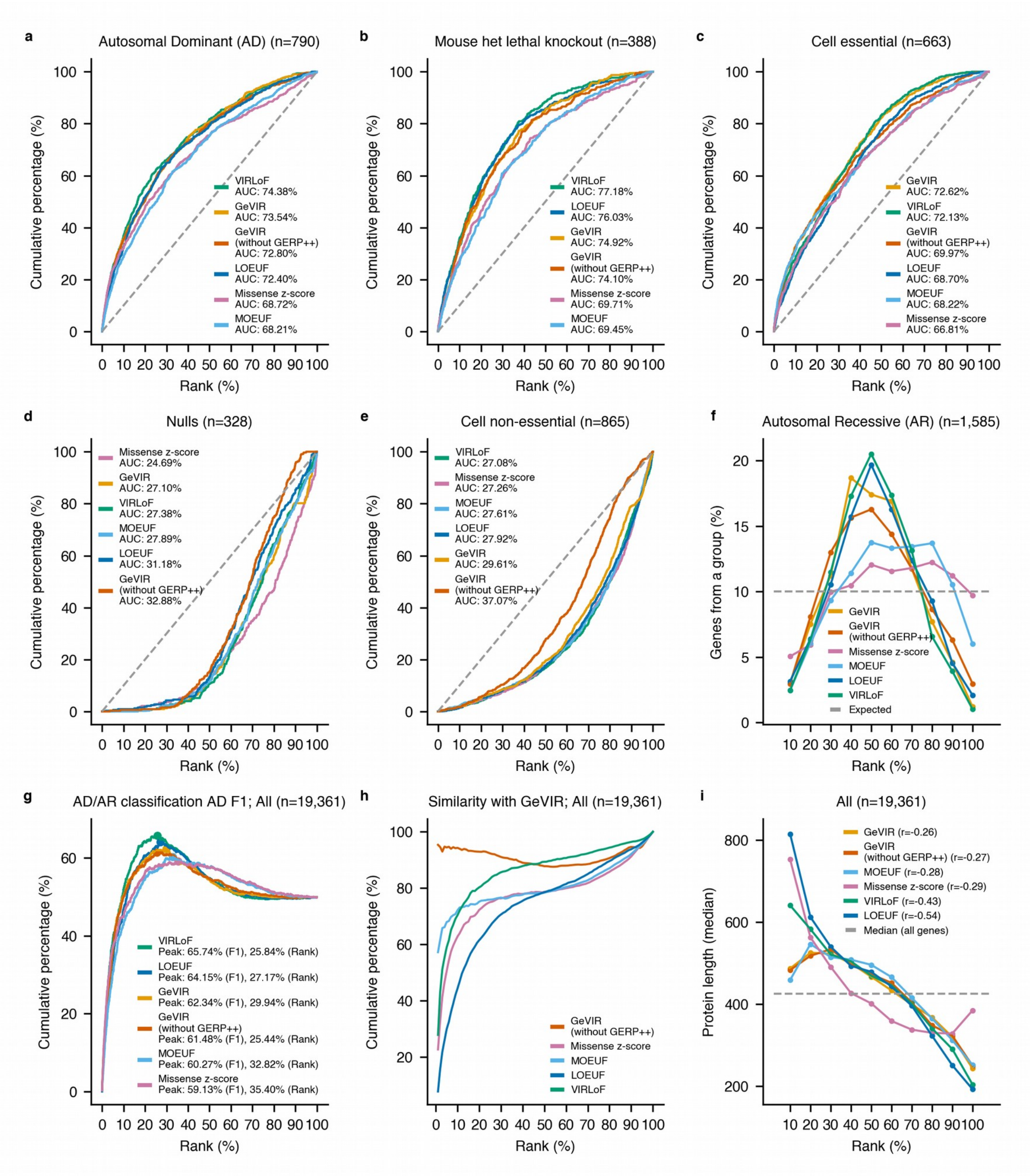


Figure B.2: Comparison of GeVIR gene ranking with gnomAD constraint metrics on 19,361 genes.

Cumulative percentage of: **a)** genes associated exclusively with Autosomal Dominant (AD) diseases in OMIM (n = 790), **b)** heterozygous lethal genes in mouse (n = 388), **c)** human cell essential based on CRISPR/Cas screens (n = 663), **d)** human null genes with at least two different high-confidence homozygous LoF variants observed in healthy populations (n = 328) and **e)** non-essential genes based on CRISPR/Cas screens (n = 865). **f)** Percentage of genes associated exclusively with Autosomal Recessive (AR) diseases out of all AR genes in OMIM (n = 1,585) in each rank decile (non-cumulative). **g)** AD class F1 score calculated at each percentile (cumulative) considering AD genes (n

= 790) as True Positives and AR genes (n = 1,585) as False Positives among all analysed genes (n = 19,361). **h)** Cumulative percentage of genes (n = 19,361) prioritised by other constraint metrics, which are also prioritised by GeVIR at each percentile (ranking similarity). **i)** Median protein length (amino acids) in each rank decile (non-cumulative) among all analysed genes (n = 19,361). Correlation between protein length and gene rank was measured with Spearman's rank correlation coefficient.

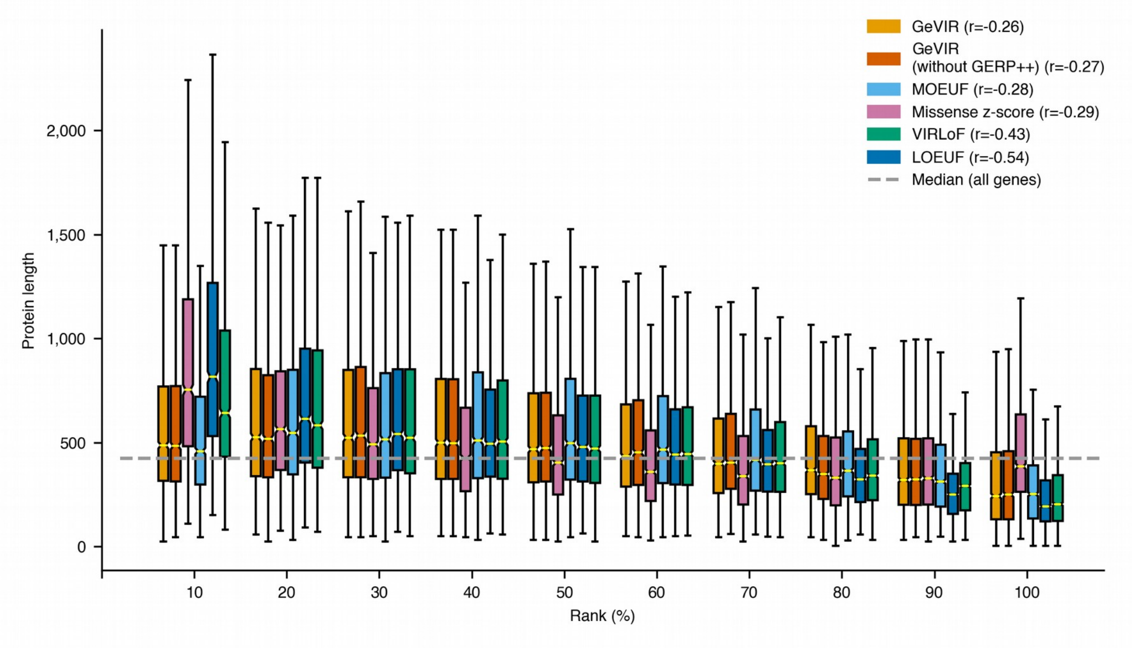


Figure B.3: 19,361 gene protein length (amino acids) grouped in deciles. Standard notations are used for elements of the boxplot (i.e. upper/lower hinges: 75th/25th percentiles; inner-segment: median, notches are calculated using a Gaussian-based asymptotic approximation; and upper/lower whiskers: extension of the hinges to the largest/smallest value at most 1.5 times of interquartile range). Outliers are not shown due to the presence of genes with extreme protein length (e.g. TTN ~36,000 amino acids) in the dataset which would distort the figure. Correlation between protein length and gene rank was measured with Spearman's rank correlation coefficient.

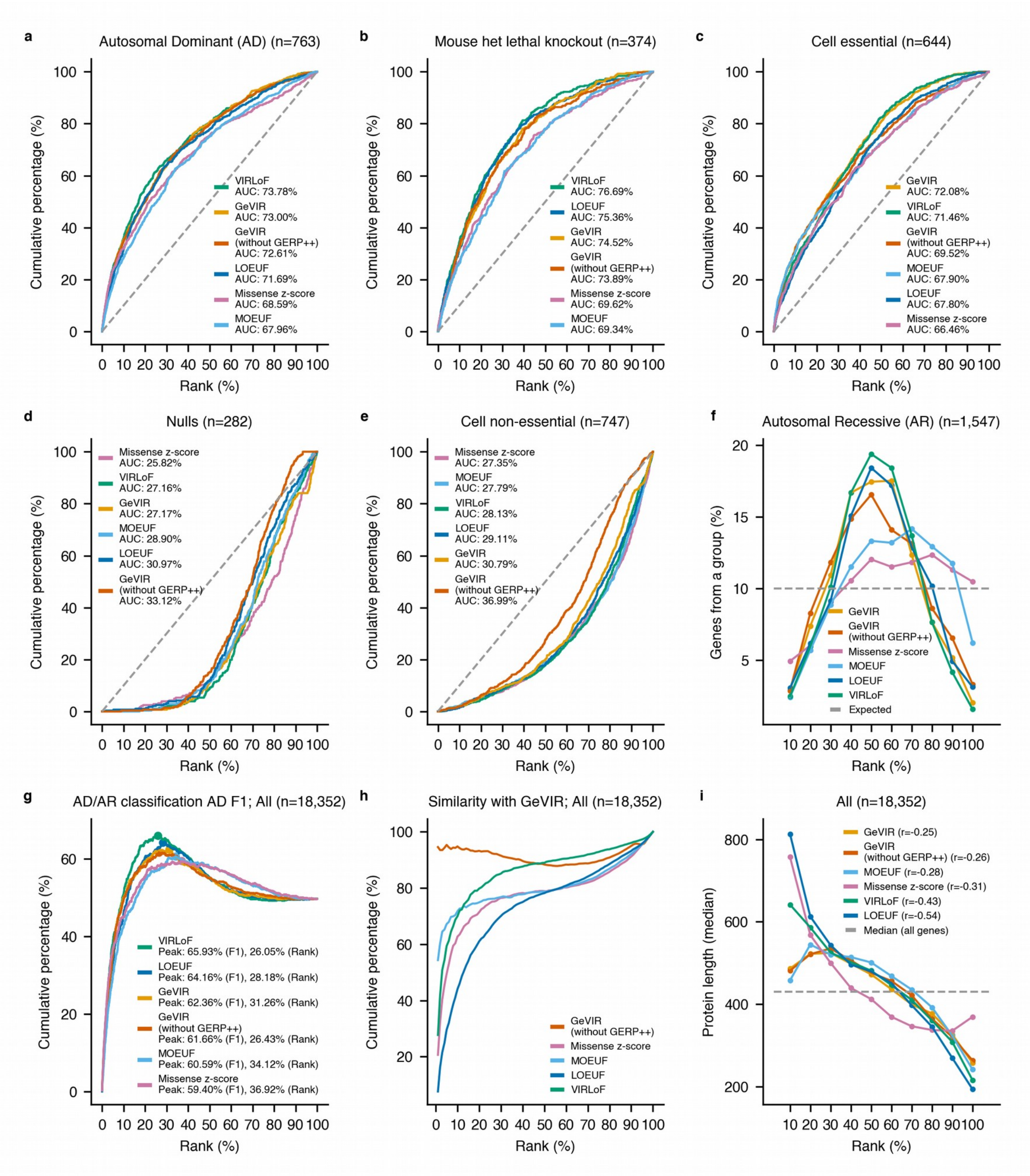


Figure B.4: Comparison of GeVIR gene ranking with gnomAD constraint metrics on 18,352 genes with 1,009 outliers excluded.

Cumulative percentage of: **a)** genes associated exclusively with Autosomal Dominant (AD) diseases in OMIM (n = 763), **b)** heterozygous lethal genes in mouse (n = 374), **c)** human cell essential based on CRISPR/Cas screens (n = 644), **d)** human null genes with at least two different high-confidence homozygous LoF variants which were observed in healthy population study (n = 282) and **e)** non-essential genes based on CRISPR/Cas screens (n = 747). **f)** Percentage of genes associated exclusively with Autosomal Recessive (AR) diseases out of all AR genes in OMIM (n = 1,547) in each rank decile (non-cumulative). **g)** AD class F1 score calculated at each percentile (cumulative) considering AD genes (n = 763) as True Positives and AR genes (n = 1,547) as False Positives among all analysed genes (n = 18,352). **h)** Cumulative percentage of genes (n = 18,352) prioritised by other constraint metrics, which are also prioritised by GeVIR at each percentile (ranking similarity). **i)** Median protein length (amino acids) in each rank decile (non-cumulative) among all analysed genes (n = 18,352). Correlation between protein length and gene rank was measured with Spearman’s rank correlation coefficient.

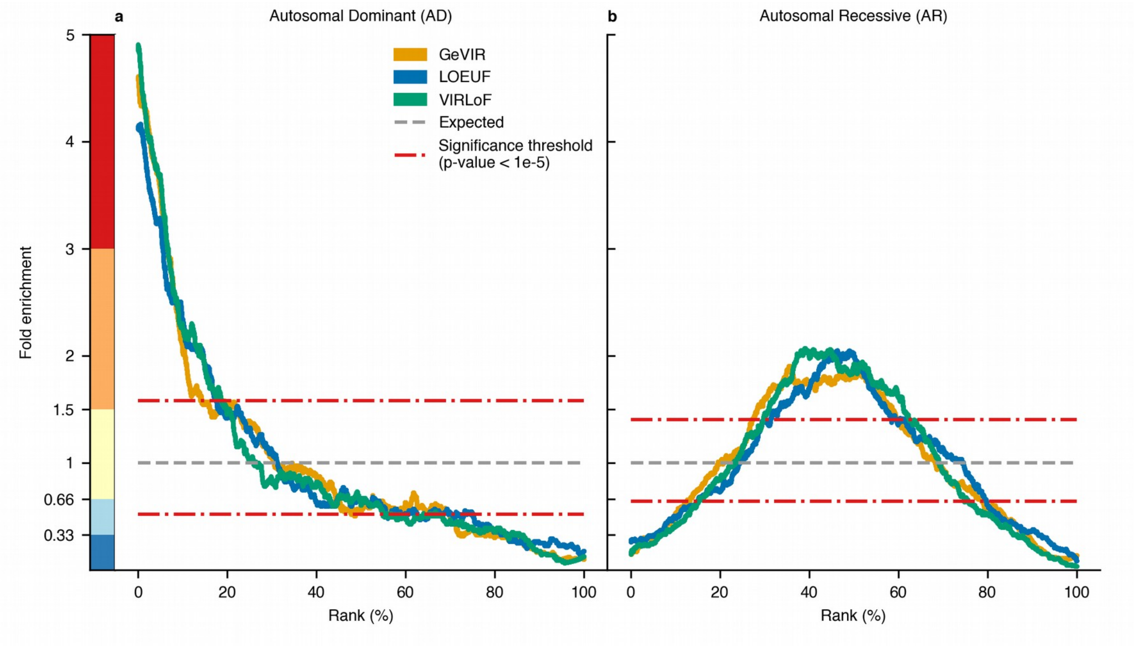


Figure B.5: Fold enrichment of known autosomal dominant and autosomal recessive genes from OMIM for each gene in the ranking lists (GeVIR, LOEUF and VIRLoF).

For each gene (n = 19,361), fold enrichment was calculated by analysing genes with similar ranking scores (up to ±5 percentiles) and comparing the proportion of known AD (n = 790) **(a)** and AR (n = 1,585) **(b)** disease genes. Exact group sizes and numbers of AD and AR gene for each gene are available in Appendix Table B.3. Statistical significance of AD or AR genes enrichment was measured with two-sided Fisher’s exact test.

B.5 Supplementary tables

Table B.1: Autosomal chromosome VIR weights.

Length	Number	Number of VIRs with that length or longer	Frequency	Weight
0	1545129	3427010	1.00E+00	1.00
1	814582	1881881	5.49E-01	1.82
2	446945	1067299	3.11E-01	3.21
3	248659	620354	1.81E-01	5.52
4	141756	371695	1.08E-01	9.22
5	82840	229939	6.71E-02	14.90
6	50855	147099	4.29E-02	23.30
7	31059	96244	2.81E-02	35.61
8	19756	65185	1.90E-02	52.57
9	13000	45429	1.33E-02	75.44
10	8739	32429	9.46E-03	105.68
11	5861	23690	6.91E-03	144.66
12	4122	17829	5.20E-03	192.22
13	3025	13707	4.00E-03	250.02
14	2260	10682	3.12E-03	320.82
15	1633	8422	2.46E-03	406.91
16	1341	6789	1.98E-03	504.79
17	974	5448	1.59E-03	629.04
18	821	4474	1.31E-03	765.98
19	623	3653	1.07E-03	938.14
20	505	3030	8.84E-04	1131.03
21	396	2525	7.37E-04	1357.23
22	311	2129	6.21E-04	1609.68
23	250	1818	5.30E-04	1885.04
24	221	1568	4.58E-04	2185.59
25	185	1347	3.93E-04	2544.18
26	168	1162	3.39E-04	2949.23
27	127	994	2.90E-04	3447.70
28	109	867	2.53E-04	3952.72
29	89	758	2.21E-04	4521.12
30	68	669	1.95E-04	5122.59
31	73	601	1.75E-04	5702.18
32	53	528	1.54E-04	6490.55
33	51	475	1.39E-04	7214.76
34	54	424	1.24E-04	8082.57
35	43	370	1.08E-04	9262.19

36	30	327	9.54E-05	10480.15
37	23	297	8.67E-05	11538.75
38	29	274	8.00E-05	12507.34
39	32	245	7.15E-05	13987.80
40	26	213	6.22E-05	16089.25
41	13	187	5.46E-05	18326.26
42	14	174	5.08E-05	19695.46
43	11	160	4.67E-05	21418.81
44	16	149	4.35E-05	23000.07
45	9	133	3.88E-05	25766.99
46	7	124	3.62E-05	27637.18
47	9	117	3.41E-05	29290.68
48	9	108	3.15E-05	31731.57
49	8	99	2.89E-05	34616.26
50	9	91	2.66E-05	37659.45
51	4	82	2.39E-05	41792.80
52	10	78	2.28E-05	43936.03
53	7	68	1.98E-05	50397.21
54	7	61	1.78E-05	56180.49
55	4	54	1.58E-05	63463.15
56	1	50	1.46E-05	68540.20
57	4	49	1.43E-05	69938.98
58	2	45	1.31E-05	76155.78
59	2	43	1.25E-05	79697.91
60	2	41	1.20E-05	83585.61
61	3	39	1.14E-05	87872.05
62	3	36	1.05E-05	95194.72
63	3	33	9.63E-06	103848.79
64	2	30	8.75E-06	114233.67
67	1	28	8.17E-06	122393.21
70	2	27	7.88E-06	126926.30
71	2	25	7.29E-06	137080.40
73	1	23	6.71E-06	149000.43
74	1	22	6.42E-06	155773.18
75	2	21	6.13E-06	163190.95
76	1	19	5.54E-06	180368.95
77	1	18	5.25E-06	190389.44
79	3	17	4.96E-06	201588.82
80	1	14	4.09E-06	244786.43
84	2	13	3.79E-06	263616.15
85	1	11	3.21E-06	311546.36
92	2	10	2.92E-06	342701.00

100	3	8	2.33E-06	428376.25
105	1	5	1.46E-06	685402.00
110	1	4	1.17E-06	856752.50
112	1	3	8.75E-07	1142336.67
130	1	2	5.84E-07	1713505.00
165	1	1	2.92E-07	3427010.00

Table B.2: Allosomal VIR weights.

Length	Number	Number of VIRs with that length or longer	Frequency	Weight
0	25812	82033	1.00E+00	1.00
1	16274	56221	6.85E-01	1.46
2	11078	39947	4.87E-01	2.05
3	7554	28869	3.52E-01	2.84
4	5257	21315	2.60E-01	3.85
5	3905	16058	1.96E-01	5.11
6	2734	12153	1.48E-01	6.75
7	1999	9419	1.15E-01	8.71
8	1483	7420	9.05E-02	11.06
9	1103	5937	7.24E-02	13.82
10	884	4834	5.89E-02	16.97
11	662	3950	4.82E-02	20.77
12	507	3288	4.01E-02	24.95
13	427	2781	3.39E-02	29.50
14	361	2354	2.87E-02	34.85
15	272	1993	2.43E-02	41.16
16	243	1721	2.10E-02	47.67
17	195	1478	1.80E-02	55.50
18	168	1283	1.56E-02	63.94
19	133	1115	1.36E-02	73.57
20	118	982	1.20E-02	83.54
21	113	864	1.05E-02	94.95
22	82	751	9.15E-03	109.23
23	80	669	8.16E-03	122.62
24	72	589	7.18E-03	139.28
25	68	517	6.30E-03	158.67
26	43	449	5.47E-03	182.70
27	44	406	4.95E-03	202.05
28	35	362	4.41E-03	226.61
29	27	327	3.99E-03	250.87
30	35	300	3.66E-03	273.44

31	32	265	3.23E-03	309.56
32	11	233	2.84E-03	352.07
33	13	222	2.71E-03	369.52
34	20	209	2.55E-03	392.50
35	14	189	2.30E-03	434.04
36	14	175	2.13E-03	468.76
37	16	161	1.96E-03	509.52
38	11	145	1.77E-03	565.74
39	14	134	1.63E-03	612.19
40	13	120	1.46E-03	683.61
41	7	107	1.30E-03	766.66
42	7	100	1.22E-03	820.33
43	9	93	1.13E-03	882.08
44	6	84	1.02E-03	976.58
45	6	78	9.51E-04	1051.71
46	4	72	8.78E-04	1139.35
47	6	68	8.29E-04	1206.37
48	4	62	7.56E-04	1323.11
49	1	58	7.07E-04	1414.36
50	4	57	6.95E-04	1439.18
51	5	53	6.46E-04	1547.79
53	3	48	5.85E-04	1709.02
54	2	45	5.49E-04	1822.96
55	1	43	5.24E-04	1907.74
56	2	42	5.12E-04	1953.17
57	3	40	4.88E-04	2050.83
58	4	37	4.51E-04	2217.11
59	3	33	4.02E-04	2485.85
60	5	30	3.66E-04	2734.43
61	2	25	3.05E-04	3281.32
62	1	23	2.80E-04	3566.65
63	2	22	2.68E-04	3728.77
64	1	20	2.44E-04	4101.65
65	1	19	2.32E-04	4317.53
67	1	18	2.19E-04	4557.39
69	1	17	2.07E-04	4825.47
70	1	16	1.95E-04	5127.06
72	2	15	1.83E-04	5468.87
75	1	13	1.58E-04	6310.23
76	1	12	1.46E-04	6836.08
82	1	11	1.34E-04	7457.55
83	2	10	1.22E-04	8203.30

84	1	8	9.75E-05	10254.13
88	1	7	8.53E-05	11719.00
90	2	6	7.31E-05	13672.17
92	1	4	4.88E-05	20508.25
94	1	3	3.66E-05	27344.33
97	1	2	2.44E-05	41016.50
128	1	1	1.22E-05	82033.00

Table B.3: GeVIR, LOEUF and VIRLoF ranks with AD and AR gene enrichment statistics for 19,361 genes.

Please see supplementary Excel document at
https://data.mendeley.com/datasets/f5v5t2kkvm/2
OR
Supplementary Table 2 at
https://doi.org/10.1038/s41588-019-0560-2

Table B.4: Statistical comparison of pathogenic variant enrichment in short (1-5 amino acids length) and long (>20 amino acids length) regions.

Two-sided Fisher Exact Test was used to compare the number of variants (n = 9,650 missense and n = 16,852 Loss-of-Function) in short regions (presented as summed length in amino acids, n = 846,971) with the number of variants (n = 1,906 missense and n = 576 Loss-of-Function) in long regions (presented as summed length in amino acids n = 45,297).

Group	VIRs length group 1	Variants 1	Total length 1	VIRs length group 2	Variants 2	Total length 2	Fold-enrichment	p-value
Missense	1-5	9650	846971	21+	1906	45297	3.69	0
Loss-of-Function	1-5	16852	846971	21+	576	45297	0.64	2.92E-29

Table B.5: Enrichment of AD and AR genes in mouse heterozygous lethal, cell essential, cell non-essential and null gene groups.

Two-sided Fisher Exact Test was used to compare n = number of genes defined in “AD/AR Genes” columns among n = number of genes in a group (e.g. Mouse het lethal) defined in “Genes” column with overall number of AD (n = 790) and AR genes (n = 1,585) among all analysed genes (n = 19,361).

Group Name	Genes	AD Genes	AD Fold-Enrichment	AD p-value	AR Genes	AR Fold-Enrichment	AR p-value
Mouse het lethal	388	96 (24.7%)	6.06	1.95E-037	24 (6.2%)	0.76	0.22
Cell essential	663	30 (4.5%)	1.11	0.551	69 (10.4%)	1.27	0.0655
Cell non-essential	865	26 (3.0%)	0.74	0.155	54 (6.2%)	0.76	0.0631
Null	328	14 (4.3%)	1.05	0.78	10 (3.0%)	0.37	0.00054

Table B.6: Statistical comparison of AR gene enrichment in GeVIR and gnomAD constraint metrics ranked list deciles, based on data shown on Figure 3.5a.

Two-sided Fisher Exact Test was used to compare n = number of genes defined in “AR” columns in among n = number of genes in deciles or ranges in “genes” column with overall number of AR genes (n = 1,585) among all analysed genes (n = 19,361).

Decile	GeVIR AR	GeVIR genes	GeVIR Fold-enrichment	GeVIR p-value
1	48	1937	0.30	6.56E-22
2	119	1936	0.75	3.07E-03
3	178	1936	1.12	1.69E-01
4	296	1936	1.87	2.84E-18
5	276	1936	1.74	3.60E-14
6	268	1936	1.69	9.68E-13
7	186	1936	1.17	5.40E-02
8	122	1936	0.77	6.24E-03
9	73	1936	0.46	1.96E-12
10	19	1936	0.12	3.18E-40
1, 2 (first 20%)	167	3873	0.53	1.18E-16
4, 5, 6 (mid 30%)	840	5808	1.77	1.61E-34
8, 9, 10 (last 30%)	214	5808	0.45	9.74E-32

Decile	GeVIR (without GERP++) AR	GeVIR (without GERP++) genes	GeVIR (without GERP++) Fold-enrichment	GeVIR (without GERP++) p-value
1	47	1937	0.30	2.41E-22
2	128	1936	0.81	2.49E-02
3	206	1936	1.30	1.08E-03

4	248	1936	1.56	2.96E-09
5	258	1936	1.63	6.28E-11
6	228	1936	1.44	2.45E-06
7	186	1936	1.17	5.40E-02
8	137	1936	0.86	1.15E-01
9	100	1936	0.63	5.24E-06
10	47	1936	0.30	2.39E-22
1, 2 (first 20%)	175	3873	0.55	7.00E-15
4, 5, 6 (mid 30%)	734	5808	1.54	2.27E-19
8, 9, 10 (last 30%)	284	5808	0.60	2.90E-16

Decile	**Missense z-score AR**	**Missense z-score genes**	**Missense z-score Fold-enrichment**	**Missense z-score p-value**
1	80	1937	0.50	1.87E-10
2	94	1936	0.59	3.30E-07
3	158	1936	1.00	1.00E+00
4	166	1936	1.05	5.75E-01
5	191	1936	1.21	2.12E-02
6	183	1936	1.15	7.90E-02
7	187	1936	1.18	4.43E-02
8	194	1936	1.22	1.33E-02
9	178	1936	1.12	1.69E-01
10	154	1936	0.97	7.94E-01
1, 2 (first 20%)	174	3873	0.55	4.04E-15
4, 5, 6 (mid 30%)	540	5808	1.14	1.50E-02
8, 9, 10 (last 30%)	526	5808	1.11	5.67E-02

Decile	**MOEUF AR**	**MOEUF genes**	**MOEUF Fold-enrichment**	**MOEUF p-value**
1	39	1937	0.25	2.12E-26
2	96	1936	0.61	8.57E-07
3	148	1936	0.93	4.59E-01
4	181	1936	1.14	1.12E-01
5	218	1936	1.38	4.97E-05
6	211	1936	1.33	3.16E-04
7	213	1936	1.34	1.94E-04
8	217	1936	1.37	5.98E-05

9	167	1936	1.05	5.46E-01
10	95	1936	0.60	5.34E-07
1, 2 (first 20%)	135	3873	0.43	2.60E-25
4, 5, 6 (mid 30%)	610	5808	1.28	8.86E-07
8, 9, 10 (last 30%)	479	5808	1.01	8.92E-01

Decile	**LOEUF AR**	**LOEUF genes**	**LOEUF Fold-enrichment**	**LOEUF p-value**
1	50	1937	0.32	6.78E-21
2	101	1936	0.64	8.07E-06
3	167	1936	1.05	5.46E-01
4	249	1936	1.57	2.31E-09
5	312	1936	1.97	6.92E-22
6	258	1936	1.63	6.28E-11
7	196	1936	1.24	8.26E-03
8	147	1936	0.93	4.33E-01
9	72	1936	0.45	9.81E-13
10	33	1936	0.21	4.76E-30
1, 2 (first 20%)	151	3873	0.48	1.07E-20
4, 5, 6 (mid 30%)	819	5808	1.72	2.99E-31
8, 9, 10 (last 30%)	252	5808	0.53	1.79E-22

Decile	**VIRLoF AR**	**VIRLoF genes**	**VIRLoF Fold-enrichment**	**VIRLoF p-value**
1	39	1937	0.25	2.12E-26
2	100	1936	0.63	5.24E-06
3	182	1936	1.15	9.45E-02
4	274	1936	1.73	8.96E-14
5	325	1936	2.05	4.47E-25
6	275	1936	1.74	4.90E-14
7	208	1936	1.31	5.95E-04
8	104	1936	0.66	2.33E-05
9	62	1936	0.39	5.39E-16
10	16	1936	0.10	8.11E-43
1, 2 (first 20%)	139	3873	0.44	5.29E-24
4, 5, 6 (mid 30%)	874	5808	1.84	5.11E-40

8, 9, 10 (last 30%)	182	5808	0.38	1.83E-41

Table B.7: Summary of functional enrichment (Gene Ontology and KEGG pathways) of the most variant intolerant genes (~15%) ranked by GeVIR, LOEUF and VIRLoF.

Functional enrichment analysis was performed using DAVID 6.8, the statistical significance was calculated using one-sided Fisher Exact Test modified for gene-enrichment analysis and reported by False Discovery Rate (FDR), Bonferroni and Benjamini adjusted p-values.

Please see supplementary Excel document at
https://data.mendeley.com/datasets/f5v5t2kkvm/2
OR
Supplementary Table 6 at
https://doi.org/10.1038/s41588-019-0560-2

C Supplementary material for Chapter 4

C.1 Data availability

The HGNC gene list can be found at https://www.genenames.org/download/statistics-and-files/

The Gene Discovery Informatics Toolkit can be found in supplementary data at https://www.nature.com/articles/s41525-019-0081-z

The Ensembl build GRCh37/hg19 gene, transcript, and protein ids to HGNC name and id mapping data can be found at https://grch37.ensembl.org/biomart/martview

The STRING data can be found at https://string-db.org/

The GeVIR gene scores can be found in supplementary data at https://www.nature.com/articles/s41588-019-0560-2

The UNEECON gene scores can be found at https://psu.app.box.com/s/wur3td0dawju9qtvu7w8orkxu5ur0oo6/file/517942997406

The DOMINO gene scores and features (final, train, and validation) can be found at https://wwwfbm.unil.ch/domino/download.html

The gnomAD genomes variants files can be found at https://storage.googleapis.com/gnomad-public/release/2.0.2/vcf/genomes/gnomad.genomes.r2.0.2.sites.coding_only.chr1-22.vcf.bgz and https://storage.googleapis.com/gnomad-public/release/2.0.2/vcf/genomes/gnomad.genomes.r2.0.2.sites.coding_only.chrX.vcf.bgz.

The gnomAD genes, transcripts and exons files can be found at http://broadinstitute.org/~konradk/exac_browser/exac_browser.tar.gz.

The gnomAD structural variants data and gene constraint scores can be found at https://gnomad.broadinstitute.org/downloads

The GPP gene lists can be found in supplementary data at https://doi.org/10.1007/s00439-019-02021-9

The Gene4Denovo data can be found at https://academic.oup.com/nar/article/48/D1/D913/5603227

The lists of olfactory, cell essential and cell non-essential gene symbols can be found at https://github.com/macarthur-lab/gene_lists/

The severe haploinsufficient gene list can be found in supplementary data at https://www.biorxiv.org/content/10.1101/148353v1

C.2 Code availability

Code for calculating DIP scores, data analysis and figures can be found at https://github.com/niab/dip.

C.3 Supplementary tables

Table C.1: ADR and DND predicted probabilities for 17,857 genes and DIP ranks for 15,794 genes.

Please see supplementary Excel document: https://data.mendeley.com/datasets/f5v5t2kkvm/2