Imperial College London

Developing methods to assess evolutionary and functional equivalence of single nucleotide variants for improved clinical interpretation of human genetic variation

---

Nicholas Li

2022

Institute of Clinical Sciences

Submitted in part fulfilment of the requirements for the degree of

Doctor of Philosophy of Imperial College London

and the Diploma of Imperial College London

Declaration of Originality

I, Nicholas Li, declare that any work presented in this thesis is my own and has not been duplicated from the work of any other authors, journals, or institutions. Any work that does originate from other sources is acknowledged and referenced accordingly.

Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

With advancements in sequencing technology there has been an unprecedented rise in human single nucleotide variant data in recent years. One of the key challenges within clinical genetics is distinguishing truly pathogenic from rare but benign variants. Many in silico tools have been developed with this aim but they often over predict pathogenicity particularly on novel variants. Here, I demonstrate how a framework designed to identify variants with functional equivalence by using information from variants in known related genes can help pathogenic variant interpretation.

Using sequence alignments of human paralogues, known pathogenic variants within aligned positions can be used to transfer their annotations across to aligned variants. This Paralogue Annotation method is shown to be widely applicable exome-wide, with 71% of disease genes having at least one paralogue. As a classifier it performs more precisely than other contemporary variant predictors, having a precision of 94% or higher depending on the data. This however comes at the cost of limited sensitivity (17% and lower).

But this is rescued when the framework was improved by altering the alignments to protein domains instead of whole gene sequences. The sensitivity was increased by 74% with a marginal 6% precision decrease.

By expanding the framework to explore the usage of structural protein alignments instead of sequence alignments there is potential to further improve sensitivity, but current limited structural data means that predicted protein models must be relied on leading to further

assumptions to be taken. In structural space, pathogenic variants across aligned models are statistically more likely to be closer together than benign and pathogenic variants.

This framework can be used as a precise pathogenic variant classifier in sequence space, but overall, it can be used to search for functionally equivalent variants to variants of interest, which is a line of information not used by many.

Contents

Declaration of Originality **2**

Copyright Statement **3**

Abstract **4**

List of Figures **12**

List of Tables **16**

Contributions of others to work presented **18**

Acknowledgements **19**

1 Introduction **22**

1.1 Variation in the Genome 22

1.1.1 How genetic variation arises in DNA 22

1.1.2 DNA repair pathways that lead to variations 24

1.1.3 Types of genetic variations 25

1.1.4 Molecular basis of point mutations, in particular Single Nucleotide Variants 27

1.1.5 Distribution of variation in human populations 30

1.2 Genetic basis of diseases for monogenic dominant SNVs 31

1.3 Pathogenic classification of variants 35

1.4 Current existing predictors for pathogenic variants 39
1.4.1 Traditional conservation based methods 40
1.4.2 Modern machine learning based methods 41
1.4.3 Inadequate performance of current predictors 42
1.5 Defining variant equivalencies and why it can help predictions 43
1.6 Aims and hypothesis of thesis 45

2 Paralogue Annotation 47

2.1 Introduction 47
2.1.1 Defining Paralogues 47
2.1.2 Paralogue Annotation 49
2.2 Materials and Methods 50
2.2.1 Surveying number of paralogues among disease and non-disease genes in the human genome 50
2.2.2 Paralogue Annotation procedure - use as a classifier and implementation 50
2.2.3 Utilizing conservation stringency to enhance Paralogue Annotation as a classifier 52
2.2.4 Using ClinVar for exome wide study and testing 53
2.2.5 Measuring the statistical performance of Paralogue Annotation and other computational tools as classifiers 55
2.2.6 Calculations of Odds Ratios for Case and Control comparison 56
2.2.7 Independent test sets used 57
2.2.8 Gene Ontolgy and PANTHER 57
2.2.9 Commonly used tools and software 57
2.3 Results 58
2.3.1 Number of Paralogues in the human genome and disease 58

2.3.2 Annotation of ClinVar . . . 59
2.3.3 Correlation with across-family paralogue conservation . . . 62
2.3.4 Addressing possible bias issues with unbalanced datasets, particularly at high conservation stringencies . . . 63
2.3.5 Paralogue Annotation can be applied to multiple protein classes . . . 69
2.3.6 Paralogue annotation is more precise but less sensitive than established methods . . . 70
2.3.7 Paralogue Annotation increases odds of correctly calling pathogenic variants in case-control comparisons . . . 72
2.3.8 General applicability of Paralogue Annotation . . . 74
2.4 Discussion . . . 75
2.4.1 Limitations of Paralogue Annotation . . . 79
2.5 Conclusion . . . 81

3 Protein domain Annotation 82

3.1 Introduction . . . 82
3.1.1 Protein domains . . . 83
3.1.2 Pfam and Pfam metadomains . . . 84
3.2 Materials and Methods . . . 85
3.2.1 Using Pfam metadomain alignments instead of Ensembl Paralogue alignments . . . 86
3.2.2 Surveying number of Pfam protein domains in genes and comparison to Ensembl Paralogues . . . 87
3.2.3 Chi-square test of independence for GO protein classes Gene Ontology pathway analyis . . . 88
3.2.4 Adjusted test set for appropriate comparison to Paralogue Annotation . . . 89

3.3 Results . . . 90
3.3.1 Number of genes containing Pfam domains surpass the number of genes with related paralogues . . . 90
3.3.2 Gene Ontology analysis of PANTHER protein class show differences between genes with Pfam domains and Paralogues . . . 91
3.3.3 Gene Ontology analysis of PANTHER pathways show differences between genes with Pfam domains and Paralogues . . . 95
3.3.4 Pfam Annotation of ClinVar . . . 97
3.3.5 Pfam domains provides substantially more positive predictions than Paralogue Annotation with marginal decrease in Precision on standardised tests . . . 98
3.3.6 Gene Ontology analysis of protein classes and pathways show no significant differences between genes that are annotated by Pfam domain and Paralogue Annotation . . . 101
3.3.7 Pfam metadomains have variable improvements in sensitivity on independent test sets compared to Paralogue Annotation . . . 103
3.3.8 Pfam metadomains do not increase odds of correctly calling pathogenic variants in case-control comparisons anymore than Paralogue Annotation 106
3.4 Discussion . . . 107
3.4.1 Limitations of Pfam metadomains . . . 112
3.5 Conclusion . . . 113

4 Expanding from sequence to structure 115

4.1 Introduction . . . 115
4.1.1 Limitations of sequence alignments . . . 116
4.1.2 Looking beyond sequence . . . 118

4.2 Materials and Methods . . . 121
4.2.1 Protein sequence and structure databases used . . . 121
4.2.2 Structural alignment of protein structures/models . . . 121
4.2.3 Calculation of Root Mean Square Deviation . . . 121
4.2.4 Tools used for visualisation of protein structures and sequence alignments 122
4.2.5 *In silico* prediction tools used for structural models . . . 122
4.2.6 Mapping variants onto structures and search space scanning distances . . 123
4.2.7 Statistical Tests . . . 123
4.3 Results and Discussion . . . 124
4.3.1 Limited volume of structural data when compared to sequential equivalent 124
4.3.2 Alignments of available crystal structures of SCN5A and SCN2A provide limited structural coverage . . . 128
4.3.3 Using Alphafold 2 predicted structures of SCN5A and SCN2A compensates for lack of structural coverage . . . 132
4.3.4 Alphafold 2 models too complex for current structural alignment workflow 136
4.3.5 Using only homology template modelling removes random loops and allows viable structural alignment . . . 138
4.3.6 Translating equivalent variant framework from sequence to structure . . . 140
4.3.7 Future works . . . 146
4.4 Conclusion . . . 148

5 Concluding Remarks **150**

6 Tools, Code, and Publications arising from this PhD **153**

6.1 Paralogue Annotator Website . . . 153
6.2 Github repositories . . . 154

6.2.1 VEP Plugin 154
6.2.2 Paralogue Annotation analysis 154
6.2.3 Pfam metadomain analysis 155
6.2.4 Structural alignment analysis 155
6.3 List of Publications 155
6.4 Conference presentations 156

List of Acronyms **157**

Bibliography **160**

Appendix **180**

Permissions to re-use third-party copyrighted works 190

List of Figures

1.1 Relationship between variant effect size and allele frequency. 33

1.2 American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) evidence framework for benign and pathogenic variant classification. 37

1.3 Evidence framework for computational and predictive data as described by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP). 44

2.1 Difference between orthologues and paralogues in regards to homology. 49

2.2 Paralogue Annotation schematic overview. 52

2.3 Pipeline representation of Paralogue Annotation experimental design. 54

2.4 Histogram showing that most human protein-coding genes, including disease associated genes, have at least one paralogue. 59

2.5 Precision-Sensitivity space plot show that Para Z scores supplement Paralogue Annotation conservation stringencies. 63

2.6 Kernel Density Estimation of the distribution of precisions from a balanced test set of Benign/Likely benign and Pathogenic/Likely pathogenic. 68

2.7 Paralogue Annotation is applicable to genes encoding a diverse range of protein classes. 70

2.8 Precision-Sensitivity space plot of predicting pathogenic variants show Paralogue Annotation is more precise but less sensitive than other variant prediction tools. 71
2.9 Improved Odds Ratios before and after Paralogue Annotation is applied. 73
2.10 Schematic of KCNQ1 shows that variants predicted to be pathogenic by Paralogue Annotation cluster near other pathogenic ClinVar variants. 74

3.1 Schematic concept of using homologous human protein domain alignments instead of paralogue alignments. 86
3.2 More genes incorporate a Pfam domain than have a related paralogue. 90
3.3 PANTHER protein classes vary among genes with Pfam domains and paralogues. 91
3.4 Correlation plot showing distinct protein classes among genes with Pfam domains and paralogues. 93
3.5 Sublevel children PANTHER protein class hits within translational proteins vary. 95
3.6 PANTHER pathway hits vary among genes that contain only Pfam domains or only have paralogues. 96
3.7 Pfam domains provide more pathogenic predictions for variants overall than Paralogue Annotation. 99
3.8 PANTHER protein class hits show little variation between genes annotated using Pfam metadomain alignments and Ensembl paralogue alignments. 102
3.9 PANTHER pathway hits show little variation between genes annotated using Pfam metadomain alignments and Ensembl paralogue alignments. 102
3.10 Precision-Sensitivity space plot of predicting pathogenic variants show Pfam metadomain Annotation has generally improved sensitivity. 105
3.11 Odds Ratios calculated before and after Pfam Annotation is applied compared to Paralogue Annotation. 107

4.1 Hypothetical scenario where structural alignments provide more information about possible functional equivalency than structural alignments. 117
4.2 Schematic of possible adaptation to the current framework. 118
4.3 Number of entries in the UniProtKB/TrEMBL sequence database has increased dramatically since 1998 to 2020. 125
4.4 Number of entries in the Protein Data Bank database has increased exponentially since 1990 to 2020. Total cumulative entries shown in blue and newly released structures per year shown in red. 126
4.5 Sequence independent structural alignment of 4DCK and 4JPZ 130
4.6 Ensembl pairwise sequence alignment of paralogues SCN5A and SCN2A. 130
4.7 Mapped Pathogenic/Likely pathogenic missense variants from ClinVar displayed onto the sequence independent structural alignment of SCN5A and SCN2A. . . 131
4.8 Pfam domain schematic of sodium channel proteins SCN5A/SCN2A show how crystal structures 4DCK and 4JPZ cover only a small region of total protein. . . 131
4.9 Alphafold 2 predicted structure of SCN5A. 134
4.10 Alphafold predicted structure of SCN2A. 134
4.11 Alphafold predicted structure of SCN5A with mapped ClinVar variants. 135
4.12 Alphafold predicted structure of SCN2A with mapped ClinVar variants. 135
4.13 TM-align structural alignment of Phyre2 predicted models of SCN5A and SCN2A.140
4.14 Structurally aligned Phyre2 predicted models of SCN5A and SCN2A. 141
4.15 Pathogenic variants in homologous proteins more likely to be located in similar positions than benign and pathogenic. 142
4.16 Ensembl pairwise sequence alignments of SCN5A and SCN2A between regions where structurally aligned variants are positioned. 146
4.17 Closer view of structurally aligned variants. 146

6.1 Front page of the Paralogue Annotator web application 154

List of Tables

2.1 Number of ClinVar variants predicted to be pathogenic using Paralogue Annotation. 61

2.2 Number of ClinVar variants removed after applying conservation stringenicies that are predicted to be pathogenic by Paralogue Annotation. 62

2.3 Number of ClinVar variants removed after applying alternate amino acid conservation stringenicies that are predicted to be pathogenic by Paralogue Annotation. 62

2.4 Number of aligned paralogous positions at varying paralogue conservations among ClinVar variants. 66

2.5 Number of ClinVar VUS predicted as being pathogenic by Paralogue Annotation. 75

3.1 Number of ClinVar variants predicted to be pathogenic using Pfam metadomains instead of Paralogue Annotation. 97

3.2 Precision decreases marginally when using Pfam metadomain instead of Paralogue Annotation. 99

3.3 Sensitivity increases substantially when using Pfam metadomain instead of Paralogue Annotation. 100

4.1 Table showing SCN5A Pathogenic and/or Likely pathogenic (P/LP) variants that were not previously annotated by Paralogue Annotation or Pfam metadomains with SCN2A P/LP variants in close proximity when using structural alignments. 145

5.1 Table summarising the three classes of the equivalent variant framework - Paralogue Annotation; Pfam metadomains; and using structural alignments - and their unique aspects and properties. 152

Contributions of others to work presented

Erica Mazaika developed the original plugin script written in Perl for Ensembl's VEP tool which identified equivalent paralogous positions across Ensembl's paralogue alignments described in Chapter 2. I later took over and ensured it functioned for both genome build coordinates and later versions of VEP.

Panda Theotokis and I co-developed the Paralogue Annotator Webtool application described in Chapter 6. I started the original development of the website with him mostly working on the front-end while I handled the back-end. As I focused more on data analysis, he contributed more by ensuring and maintaining a working version. He also added more features, such as visually showing protein domains in paralogue alignments (see Chapter 6 for more details).

Acknowledgements

I would like to start by thanking my course administrator Claire Marsh and the wider London Institute of Medical Science (LMS) community for establishing a brilliantly structured environment where students can feel more supported during their PhD. I also want to mention my initial director of postgraduate studies Mark Ungless who, during my early doubtful stages, gave me enough motivation to keep on moving to where I am now.

Next I would like to thank the past and present members of the Cardiovascular Genetics and Genomics research group - mentioning everyone will be difficult as it was a big group split amongst different departments. I would start with the LMS members of the group, in particular Konstantinos Vanezis, who was always the reliable person you would turn to when problems arose. He made sure everything kept on running smoothly and was down to earth. I was always comfortable speaking frankly with him. Konstantinos, Nick Quaife, Katie O'Fee, and other group members at the LMS made working there much more enjoyable.

Moving to members outside of the LMS, I would like to thank Pantazis Theotokis who helped with more technical aspects and was someone relatable who I could have a chat with every now and then. He and I solved many problems together and working with him was a pleasure. Xiaolei Zhang was also someone who I could talk to and as a fellow PhD student, we would often discuss the many issues that would occur during our PhDs. She often felt like a companion who I was sharing the same adventure with.

I also appreciate the in-group IT technical support that Mian Ahmed, and later on Mikyung

Jang, provided to ensure that I had access to specific tools and resources, particularly during my early stages.

At the Royal Brompton Hospital, Will Midwinter, Leanne Felkin, Rachel Buchan and other members there were always around to share a joke with. Will in particular helped me to break out of shell when I first started and I very much enjoyed his company.

I am especially grateful to Erica Mazaika and Roddy Walsh, both of whom were the people I first worked with when I started. Erica was a great caring friend and helped me get settled into the group. She supported me when I needed it. Even after she left, we still connected and she still provided help whenever she could. Likewise, Roddy was also a great friend who I still talk to after he left. He kept things real and honest, and in some ways I looked up to him as a role model. We often shared the same sense of humour. Without these two, my PhD would have undoubtedly ended up being completely different as a lot of the work in this thesis is based on prior work done by them.

Following that I would like to also mention more senior members of the group. Paul Barton was a moral compass and at times, felt like a father figure to the group. When there were things I did not feel comfortable talking about with my supervisor, Paul was always there for me to turn to. Like Roddy, he and I shared a similar sense of humour and made completing the PhD much more enjoyable.

Secondly, I am very thankful to Nicky Whiffin. She was a post-doc when I started and later became my secondary supervisor. She now leads her own research group. She perhaps gave me advice that will stick with me forever - "getting things done is better than getting things perfect". Her guidance allowed me to not fixate too much on the details and kept me moving on to the next task at hand. She and my supervisor saw me at my worst and helped me get back up to complete this PhD.

Lastly, I would like to show my biggest gratitude and appreciation to my supervisor, James

Ware. In the early stages of my PhD and during moments in the middle, we did not always see eye to eye. Arguably we still disagree with some things even now. Compared to other principal investigators (PI) at the institute, he had perhaps not been a PhD supervisor for as long and as one his first PhD students, working together was a learning experience for both of us. However, as I gained more experience I think he did as well. Regardless of his professional and academic role, he is genuinely a kind person who means well. He often had to juggle between his duties as a clinician, lecturer, researcher and being a PI of a research group, yet he still tried to make time for me. In the end, he was the person who stuck with me the longest during this PhD. I would like to think that the rough patches in our student-supervisor relationship offered him great experience that will benefit the many future students I am sure he will have.

Chapter 1

Introduction

1.1 Variation in the Genome

The Deoxyribonucleic Acid (DNA) sequence in an organism's genome contains all the instructions required for how every cell should develop and behave as the organism ages over time (Alberts, 2017). If the DNA is altered in any way then this can cause changes in protein sequence, gene expression and/or regulation which could have a detrimental effect on the well being of the organism depending on the severity of the change.

In this first chapter, I will provide a brief overview of the context of genetic variation in the genome and their role in disease before moving on to introducing the main focus of this thesis.

1.1.1 How genetic variation arises in DNA

Changes to the DNA sequence can occur through both endogenous and exogenous means (Chatterjee and Walker, 2017).

Endogenously, most mutations in the DNA arise from errors during DNA replication (Pray, 2008). DNA polymerase has an approximate error rate of about 1 per every 100,000 nucleotides (Mazouzi et al., 2014). Although the overall error rate is lower due to repair mechanisms within

the cell. However, these mechanisms are not perfect, and if they fail to recognise and correct an erroneous nucleotide then that nucleotide will be included within the following rounds of replication (see 1.1.2). Thereby giving rise to a natural mutation that differs from the initial DNA sequence.

Another endogenous source is DNA damage that comes from DNA chemically reacting with internal reactive oxygen species (ROS), which are naturally produced from the mitochondrial electron transport chain (ETC) during aerobic respiration and are also products of many metabolic reactions (Henle and Linn, 1997). These ROS can oxidize the nitrogenous bases changing their chemical structure, for example inducing ring opening in the purine ring structure of adenine (Juan et al., 2021). They can also cause damage to the sugar phosphate backbone leading to single strand breaks.

Other endogenous sources include events such as spontaneous base deamination and DNA methylation. In base deamination, an amino group is removed from the base via hydrolysis. For cytosine, this would cause a structural transition to uracil, which naturally base pairs with adenine instead of guanine (Duncan and Miller, 1980). In DNA methylation, a methyl group is added to the nitrogenous base. Methylation of cytosines at CpG sites around promoter regions is a typical method of silencing gene expression and is particularly studied in Epigenetics (Moore et al., 2013). However, if this methylation is followed by deamination, then the 5-Methylcytosine transitions into thymine (Walsh and Xu, 2006). Again, this would interfere with normal complementary base pairing.

Exogenous DNA damage comes from environmental factors, such as radiation or chemical mutagens. Ionizing radiation (IR) can ionize local water molecules around DNA to create ROS that damage DNA as described above. Direct damage by IR to the sugar phosphate backbone can also cause single stranded breaks (Mavragani et al., 2019). Ultraviolet (UV) radiation can cause inter- and intrastand covalent crosslinking of bases (Clancy, 2008).

Chemical mutagens that enter the body either through diet or as a gas through respiration can also increase mutation likelihood. Tobacco smoke and bi-products from industrial processing are some of the main alkylating agents that can add a bulky alkyl group to nitrogenous bases or the phosphates in the DNA backbone (Chatterjee and Walker, 2017).

These damages can thereby potentially lead to non-complementary base pair mismatches when the DNA tries to repair itself. All of this alters the DNA sequence and the information encoded.

1.1.2 DNA repair pathways that lead to variations

The cell has many different DNA repair pathways that attempt to correct the damages outlined above. But they are not perfect and alterations to the sequence can still remain in place (Clancy, 2008). At times, imperfect repair can also introduce new mutations that did not exist prior. Below are some brief examples of these repair pathways.

When only the nitrogenous base of the nucleotide has been altered by alkylation/methylation, the cell can directly reverse these alterations via reversal enzymes such as the O^6-alkylguanine-DNA alkyltransferase (AGT/MGMT) class of enzymes. A single AGT/MGMT enzyme chemically reacts with a single alkyl group on a guanine base, covalently transferring the alkyl molecule to a free cysteine residue in the catalytic site of the enzyme (Kaina et al., 2007). This reaction however is permanent and renders the AGT/MGMT enzyme inactive and unable to carry out another reversal. Therefore, continuous production of the enzyme is required for the cell to maintain such reversals. A deficiency of AGT/MGMT enzyme expression has been associated with certain cancers (Mokarram et al., 2013).

Other forms of base damage such as oxidization or deamination, as well as alkylation, is repaired through the Base Excision Repair (BER) pathway in which the damaged base is

removed from the nucleotide resulting in an abasic site. The Phosphate-deoxyribose backbone is nicked and then the pathway goes through either short-patch BER (where a single nucleotide gap is filled in) or long-patch BER (where a 2-10 nucleotide gap is filled in) (Krokan and Bjørås, 2013).

1.1.3 Types of genetic variations

Genetic variants can vary from multiple nucleotide variations, reaching as far as large scale structural rearrangements of the chromosome, to single nucleotide point mutations.

Generally, the more base pairs affected by variation, the larger the effect of the variation is expected to be (Loewe, 2008). Particularly, the probability that the outcome of the effect is deleterious increases.

The largest variations that have been observed are large scale structural variations where vast regions (50 or more base pairs (Mahmoud et al., 2019, Sudmant et al., 2015)) of the chromosome are altered, usually during erroneous DNA replication/repair and recombination (Carvalho and Lupski, 2016). This can be through the removal of a section of chromosome (deletion), from which that section could also be then inserted into a different non-homologous chromosome (insertion). This movement of DNA between two non-homologous chromosomes is also known as chromosomal translocation. Structural variation also includes inversions, where a section of DNA could be inverted reversing the DNA sequence for the section alone. Chromosomal inversions can occur during DNA repair processes (see 1.1.2) such as non-homologous end joining (Puig et al., 2015). Furthermore, a DNA region can be duplicated and repeated after itself. This gives rise to Copy Number Variants (CNV) and in healthy individuals can vary from zero to multiple copies (Jackson et al., 2018). Even greater, aneuploidy can occur where entire chromosomes are duplicated, as is the case with Down Syndrome (trisomy 21), where affected individuals have an extra copy

of chromosome 21. Large scale variations, particularly within protein coding regions of DNA, are highly disruptive as they are more likely to interfere with the normal balance of proteomic stoichiometry.

Smaller variations are less likely to be as disruptive since they affect fewer nucleotides, but they can still lead to disease. Due to them being less disruptive, they face less selection pressure and are therefore more common than larger, highly disruptive, structural variations. Collectively, small insertion and deletion events, or indels (less than 50 base pairs) (Sudmant et al., 2015), can have large effects even if they only affect one single nucleotide. As this would still alter the coding frame of the DNA sequence within protein coding regions - leading to a frameshift (Alberts, 2017). This therefore could alter the entire protein sequence downstream of where the variation occurs. The frameshift can result in a premature termination codon (PTC) leading to not just an altered protein sequence but also a shorter truncated one. On occasion, an indel may cause a PTC at the loci of where the variation occurs (Cirulli et al., 2011).

On the other hand, indels can also be inframe if they are a multiple of 3 base pairs. Though these will not change the amico acid sequence that is being coded for by the unaffected nucleotides, the inframe indels can lead to additional amino acids being inserted into or deleted from the sequence (Baker et al., 2008). Depending on the size and where these inframe indels occur, they can result in functional regions of the protein being disrupted or deleted, or influence protein folding negatively (Zhang et al., 2011).

Even smaller are point mutations that involve a single nucleotide being changed. These are discussed in greater detail in section 1.1.4.

All of the variant types described above can disrupt normal protein function to some certain extent and would be deemed as Loss-of-Function (LOF). If there is a complete LOF then then causal variant is described as a null variant (Housden et al., 2017). On the other hand, a variant can sometimes cause the protein product to gain function, otherwise known as Gain-

of-Function (GOF). This can either mean that protein function has increased in activity, such as in Long QT Syndrome (LQTS) where a GOF variant in *SCN5A* causes more sodium ions to leak into cardiomyocytes than normal (Wilde and Amin, 2018); or where the protein gains an entirely new function, such as when a variant in the disordered region of GLUT1 transmembrane protein causes abnormal binding to adaptor proteins, leading to mislocalization of GLUT1 (Li and Babu, 2018, Meyer et al., 2018). Both LOF and GOF variation can lead to disease depending on the protein function and the extent of change away from normal function the variation causes.

For coding regions of the genome, the direct effect a variation has on the resulting protein product can be logically solved based on the codon to amino acid alterations, but in non-coding regions the phenotypic effect can be harder to deduce (Ward and Kellis, 2012). Still, research has shown that even effects on non-coding regions of the DNA can lead to disease. These non-coding variants can be harder to interpret as their phenotypic effects are harder to predict, but changes to regulatory regions of DNA, such as enhancer or suppressor regions, can affect the expression of upstream or downstream genes associated with such regions (Gordon et al., 2020). For example, non-coding variants found in enhancer/suppressor regions for the *NOS1AP* gene has been associated with cardiac arrhythmias and function (Zhang and Lupski, 2015).

This thesis, however, will mainly focus on protein coding variations, in particular Single Nucleotide Variants.

1.1.4 Molecular basis of point mutations, in particular Single Nucleotide Variants

Of great interest to this thesis are single nucleotide variants (SNVs) created by substitution events. Point mutations are the most common source of genetic variation (Brady, 2005). Most SNVs are not are associated with disease and account for normal variation within a healthy

population. But some rare variants have been empirically linked with pathogenicity (Bomba et al., 2017, Katsonis et al., 2014). They are defined by a single nucleotide base change - substitution of a single nucleotide with a different nucleotide at the same position. The mutation rate per single nucleotide position per haploid genome is estimated to be around 1.1 x 10^{-8} (Roach et al., 2010).

SNVs are where a single nucleotide substitution occurs in the DNA without any frameshifts. Since there is no frameshift, these SNVs only affect the amino acid the nucleotide being altered is partly coding for in the context of its triplet codon (Alberts, 2017). The substitution can have multiple effects depending on what the initial and alternative nucleotide change is, the position the nucleotide is within its codon, and the the initial amino acid the codon is coding for. A synonymous substitution occurs where the nucleotide change does not alter the final amino acid being coded for (Alberts, 2017). This can occur quite frequently if the the nucleotide is in the third position of a codon as for most amino acids that position within a codon is the most variable.

Synonymous substitutions are most likely to be benign as they have no direct effect on the protein product. However, as they still alter the DNA sequence, they may still affect non-coding regulatory elements such as enhancer/suppressor binding regions or affect splicing junctions resulting in alternative splicing (Sauna and Kimchi-Sarfaty, 2011). This could then cause expression of a protein product which leads to a disease phenotype. Such as in the case of melanoma and other cancers (Gotea et al., 2015).

Non-synonymous substitutions (or missense variants) are more likely to be directly involved in disease as they alter the amino acid being coded for by the nucleotide in question (Hughes et al., 2008). This can take two forms - nonsense and missense. Nonsense occurs when the nucleotide change creates a PTC leading to a shortened, truncated protein (Alberts, 2017). Depending on where the nonsense mutation occurs along the protein length and the molecular

function of the protein, the extent of truncation can lead to a disease phenotype (DeBoever et al., 2018). Missense occurs when the nucleotide change results in a different amino acid being coded for than the initial one before the substitution (Alberts, 2017). What the amino acid and where in the protein sequence the change is can result in different outcomes in regards to protein function. If the amino acid changes to another one of similar biochemical properties, such as same charge, hydrophobicity, or size, then the viability of the overall protein may still be functional enough to carry out its original purpose. On the other hand, if the change is to an amino acid with a different biochemical property then this can very disruptive to the protein. Where the change occurs is also important as even changes to similar amino acids within conserved regions of a protein sequence can be quite detrimental. These regions tend to be regions of structural and functional importance, such as active sites of an enzyme, and therefore even the slightest of alterations to amino acids can affect the protein (Stefl et al., 2013, Zhang et al., 2012). Because of this, missense mutations typically result in LOF, but depending on the function of the protein, GOF is also possible. For example, missense mutations within the *SCN5A* gene, which codes for the alpha subunit of the sodium channel protein $Na_v1.5$ responsible for regulating sodium current in the myocardium, can cause both LOF and GOF. LOF mutations results in less sodium influx while GOF leads to an increase of sodium into cardiomyocytes (Wilde and Amin, 2018). Both can cause diseases that have distinct differences but ultimately with adverse effects on the heart. In terms of protein altering mutations missense variants are the most common (Miosge et al., 2015). They are not as selected against on average compared to nonsense or frameshift mutations as they are not as disruptive on the protein molecule. With only one amino acid change as oppose to premature termination or shifting of reading frame.

Of course there are other point mutations not mentioned above that may have a greater effect at the protein level such as start loss mutations that prevent the initiation of translation.

Furthermore, some point mutations greatly affect transcript structure, which clearly would also affect the final protein product. For example, elongation of the transcript/protein can be caused by stop loss mutation (Bánfai et al., 2017), and point mutations at intron-exon boundaries and splice junctions can lead to unintentional alternative splicing events (Anna and Monika, 2018). These mutations however will not discussed any further in this thesis and are excluded from the study.

1.1.5 Distribution of variation in human populations

In subsections 1.1.3 and 1.1.4, I described the different types of genetic variations that can occur within the genome. Here I describe the distribution of these variations within populations. Population datasets, such as the 1000 Genome Project from the 1000 Genomes Project Consortium (2015) and The Genome Aggregation Database (GnomAD) from Karczewski et al. (2020) will be used as examples.

The 1000 Genome Project, which started in 2008, towards the end managed to sequence 2,504 individuals from African, American, European, East and South Asian populations with an average sequencing coverage depth of 7x (Consortium, 2015). In version 3 of GnomAD, a total 76,156 genomes has been sequenced at a greater coverage of 32x (Collins et al., 2020).

The first variations I discussed were large structural variants, which though are not as individually common as SNVs, as they affect a larger number of base pairs, they can be an important source of variation. Pang et al. (2010) estimated that the average genomic variation between any two humans is around 1.5% for structural variants compared to 0.1% for SNVs. The telomeres are particularly affected by structural variants (Levy-Sakin et al., 2019).

According to the 1000 Genome Project, structural variants affect on average around 20 million bases (0.007%) of the human genome. The average genome contains between 2,100 to 2,500 structural variants. Approximately 1000 of these are deletions and 160 are CNV. The

rest being insertions with a small number (~ 10) being inversions (The 1000 Genome Project Consortium (2015)).

On the other hand, Collins et al. (2020) has analyzed structural variants across 14,891 genomes from GnomAD. They showed that the median genome contains 7,439 structural variants. Around 3,500 of which are deletions, 2,600 are insertions, 550 are CNVs, 720 are duplications, and also 14 inversions. The median number of base pairs a single structural variant affects is 331bp.

Neither the 1000 Genome Project nor GnomAD provide much data on aneuploidy, but a review by Nagaoka et al. (2012) says that aneuploidy is estimated to occur between 20–40% of all conceptions. This would include conceptions that result in miscarriages. For common aneuplodies, like trisomy 21, Presson et al. (2013) estimated that within the United States population, 8.27 per 10,000 people suffered from Down Syndrome.

Smaller indels and SNVs are much more common than larger structural variants. More than 99.9% of individual variants in the 1000 Genome Project are indels and SNVs. The average genome in the 1000 Genome Project has 566,000 indels and 3.7 million SNVs.

1.2 Genetic basis of diseases for monogenic dominant SNVs

Unrepaired mutations in somatic cells can affect the individual until the cell in question is removed through one of the many different pathways of cell death. The period of time can vary depending on the cell type (Erickson, 2003).

Non-lethal mutations in the germline can be passed down to offspring and be integrated into the affected family throughout the generations, potentially affecting individuals who carry such variations. Lethal mutations or mutations with deleterious effect will be selected against

during the evolutionary process. Yet many variations associated with disease still appear in populations (Campbell and Eichler, 2013).

The link between genomic variation and disease can be difficult to establish. This can be due to the effect size of variants, that is to say how penetrant a variant is, and/or expressivity of the disease. Penetrance is the proportion of individuals within a population that carry the disease-causing variant and express the disease phenotype (Lobo, 2008). Indeed, though an individual may carry a pathogenic variant they are not guaranteed to develop the disease. In the individuals that do develop disease, the range of the symptoms and the extent of disease progression can also vary. This is known as the expressivity of the disease (Lobo, 2008). As a result, varying penetrance and expressivity among different individuals increase the challenge of establishing the genetic basis of diseases.

With complex diseases where a combination of multiple genes and environmental factors contribute to the disease phenotype, the challenge is even greater. Multiple variants within an individual can contribute risk to developing the disease and the difficulty in detecting these alleles can vary (Craig et al., 2008). Common variants can be detected by genome-wide association studies (GWAS) but due to purifying selection, most of these only have low or modest effect on disease (**Fig. 1.1**) (McCarthy et al., 2008). There are only a few examples of common variants that greatly contribute to disease, such as in the case of the Apolipoprotein E (*APOE*) gene, where the *APOE4* allele (Cys112Arg compared to *APOE3*) contributes to Alzheimer's disease (Sadigh-Eteghad et al., 2012). Rare variants with low to modest effect size are hard to identify but given they do not contribute as much risk for disease they are not focused on as much by the research community.

This thesis will focus primarily on monogenic diseases that predominately follow Mendelian inheritance patterns. Variants that cause these diseases have high effect size and as such, they appear with very rare allele frequencies due to high selection pressure (McCarthy et al., 2008).

Given that a pedigree exists, genetic linkage studies are the preferred way of detecting such variants as opposed to GWAS.

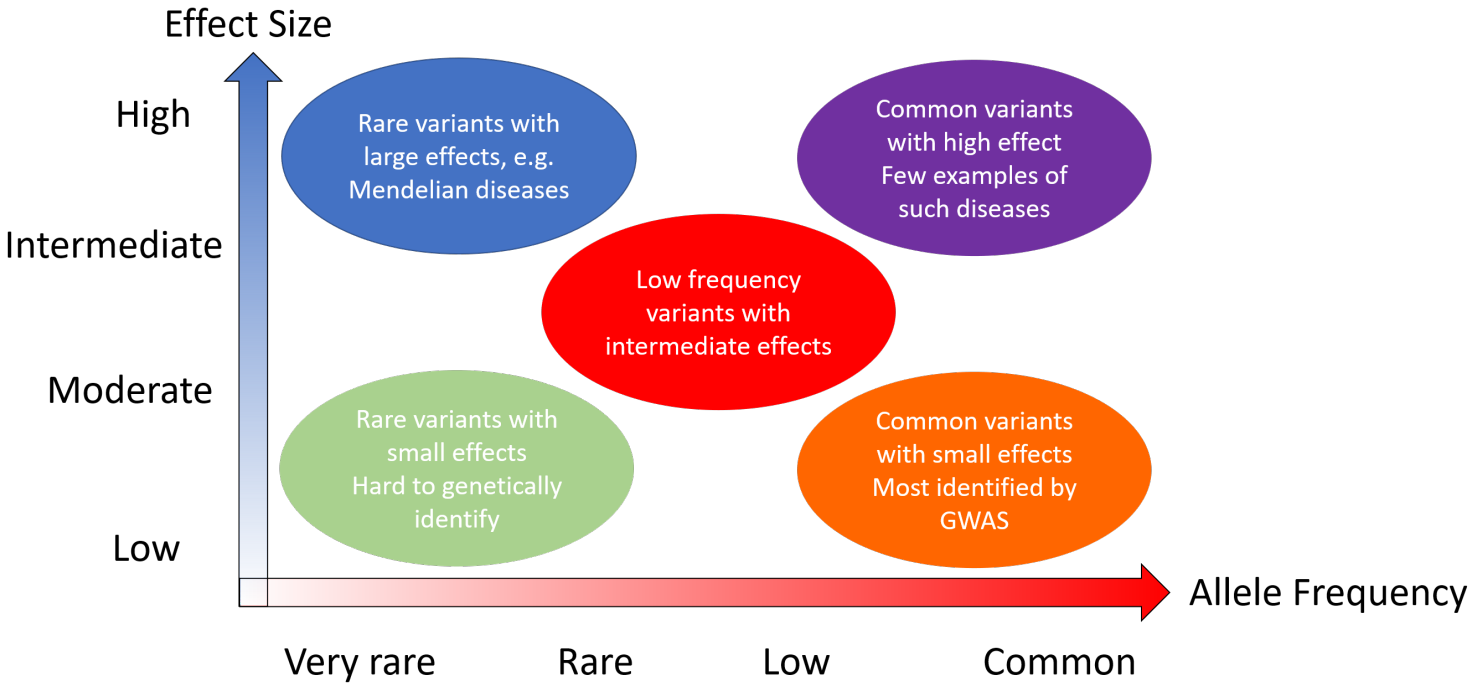


Figure 1.1: Relationship between variant effect size and allele frequency. Reproduced from McCarthy et al. (2008).

Monogenic diseases typically follow one of these Mendelian inheritance patterns (Alliance et al., 2009):

- **Autosomal Dominant** - only 1 copy of the allele (from an autosome) needs to be inherited for disease expression. Characterised by each affected proband having an affected parent and appearing in every generation.

- **Autosomal Recessive** - both copies of the allele (from the same homologous pair of autosomes) need to be inherited for disease expression. Characterised by both parents of proband being at least a carrier of affected allele and therefore not necessarily appearing in every generation.

- **X-linked Dominant** - only 1 copy of the allele (on the X chromosome) needs to be inherited for disease expression. Characterised by more females than males being affected with only daughters of an affected male expressing the disease, since fathers do not pass their X chromosome to their sons.

- **X-linked Recessive** - for females, both copies of the allele (on the X chromosomes) need to be inherited for disease expression, but not for males as they only carry one X chromosome. Characterised by more males than females expressing disease with each male affected in every generation where possible.

- **Y-linked** - only affects males

- **Mitochondrial** - affected allele present in mitochondiral DNA inherited only from the mother. Characterised as affecting both males and females but also showing a disease lineage exclusively following mothers within a family.

Alternatively, the variant associated with the expressing disease may not have observed any of the aforementioned inheritance patterns and can only be detected in the child but neither

parents. In such events, these non-inherited, but newly acquired, variants would be described as *de novo* in the context of that child. If the child passes these variants to their offspring then they would cease being described as *de novo* for those newer generations of family. (Note that *de novo* is not to be confused with *novel* variants, in which the latter describes newly discovered variants never seen before entirely.) *De novo* variants have yet to go through any selection process and therefore ones that appear can be quite deleterious on average compared to inherited variants that have undergone generations of purifying selection (Acuna-Hidalgo et al., 2016).

However, even among autosomal dominant diseases, the penetrance and expressivity of disease may still vary somewhat, with the phenotype sometimes not clear or consistently observed. Some individuals may not be aware that they carry pathogenic variants until there are obvious signs of disease onset. For certain diseases, this can be quite late during the disease progression and at which point the prognosis can be quite poor, for example in the case of Hypertrophic Cardiomyopathy (HCM) (Walsh et al., 2017).

If an individual is aware that they carry a pathogenic variant and therefore have an increased risk of developing such disease, they would be better prepared at anticipating onset and can look out for and deal with symptoms associated.

1.3 Pathogenic classification of variants

Due to the advancements of sequencing technology in the last decade, more and more variants are being identified (Metzker, 2010). However, distinguishing novel pathogenic from rare yet benign variants remains one of the key challenges in clinical genetics. Individual rare variants, particularly singletons that are only seen once in a population dataset, are collectively very common (Lek et al., 2016). Identifying which one of these in a person with disease is the main contributor to said disease is still incredibly difficult (Bomba et al., 2017).

If the variant has already been observed before and documented, then annotating the variant with prior information is trivial. But, if the variant has never been observed then identifying its effects and whether its involved in disease can be challenging.

In order to provide a consensus on the interpretations of sequence variants the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) have introduced standards and guidelines to help with this (Richards et al., 2015). These standards and guidelines attempt to provide researchers a more quantifiable and structured approach to variant interpretation by ordering levels of evidence that meet certain criteria on the basis of how much support they provide a variant for pathogenicity (**Fig. 1.2**).

They recommend standardised terminology for classifying the degree of pathogenicity of a variant using the following:

- Benign
- Likely benign (>90% certainty of being benign)
- variant of uncertain significance (VUS)
- Likely pathogenic (>90% certainty of being disease causing)
- Pathogenic

In order to assign a variant to one of these terms, the variant needs to meet a certain combination of criteria for different data types. Once a number of these criteria are met, the variant can be designated one of the 5 classification tiers.

The framework orders criteria based on the level of support they provide for a benign (Supporting and Strong (with a single special criteria Stand-Alone; see below)) and pathogenic assertion (Supporting, Moderate, Strong, and Very Strong).

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very Strong**
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Figure 1.2: Full evidence framework for benign and pathogenic variant classification as described by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP). Ordered by level of support are the evidence codes and corresponding criteria required to be fulfilled for a benign or pathogenic assertion of a given variant. Note that the numbering of evidence codes does not indicate any weighting for the level of support. Reproduced from Richards et al. (2015) with permission.

The general principles of interpreting variants include identifying if the variant is similar to ones that already exist in curated databases or seeing if the inheritance pattern of the variant fits with the known inheritance pattern of disease as discussed in the previous section.

For example in the case of population data, such as GnomAD (Lek et al., 2016), if a variant has been observed before at high prevalence (MAF $> 5\%$) (evidence code Benign Stand-Alone 1 (BA1)) or its allele frequency is greater than expected for the disorder (Benign Strong 1 (BS1)), then that provides strong evidence for benignity (Richards et al., 2015). On the other hand, if

its not been observed in population datasets (Pathogenic Moderate 2 (PM2)) and its prevalence in disease cohorts is statistically enriched in comparison (Pathogenic Strong 4 (PS4)), then that provides moderate and strong pathogenic evidence respectively.

Computational and predictive data types can most importantly provide insights about how similar a variant is to ones already known to be pathogenic. This and the criteria for such data will be discussed in greater detail in section 1.5

Well-established functional studies that show the variant causes a deleterious effect (PS3), or lie within a mutational hot spot (PM1), provide strong and moderate levels of support for pathogenicity. Likewise, if no deleterious effect is shown (BS3) than this is strong evidence for the variant being benign.

In terms of inheritance, if a variant co-segregates with family members affected by disease within a pedigree (PP1) then this provides supporting evidence for pathogenicity. If there is more segregation data from multiple different families that show similar evidence of segregation than this can increase the support for pathogenicity to moderate or even strong. Similarly for *de novo* variants, if the affected individual with disease harbours variants that are not present in either parent then that is considered moderate (PM6), if there is no confirmation, or strong (PS2) support for pathogenicity, if sequencing confirms both parents do not present said variant.

There are other data types that the framework provides criteria for, such as allelic data, but these do not provide as strong evidence of support for either benignity or pathogenicty in comparison to what has been mentioned.

Once which criteria has been identified and fulfilled by the available data, the ACMG-AMP guidelines then provide rules for combining these criteria in order to reach one of the 5 classification terms.

For example, A variant only needs to fulfil BA1 for it to be classified as Benign. Alternatively, 2 or more strong benign criteria need to be fulfilled (BS1-BS4). If that is not

possible, then for a Likely benign classification, 1 strong and 1 supporting (BP1–BP7) criteria must be fulfilled.

For a Pathogenic classification, 1 very strong (PVS1) and at least 1 strong (PS1–PS4) pathogenic criteria must be fulfilled. Otherwise, 2 or more strong is also acceptable. If not enough strong criteria are present, then a Likely pathogenic classification requires 1 very strong and 1 moderate (PM1–PM6), of if neither very strong nor strong criteria are available then 3 or moderate criteria can be fulfilled.

The rules that Richards et al. (2015) provides include many different combinations allowing many variants to be classified even if they lack some supporting data types. However, when there is a lack of information about a variant and not enough criteria are met, or if there are contradicting pathogenic and benign criteria, then the variant is recommended to be classified as VUS.

Notably the ACMG-AMP guidelines are stated as being intended primarily for Mendelian disorders and not applicable for multigenic complex diseases. As a result, since publication of the guidelines in 2015, there have been several attempts from some research groups and bodies to extend the guidelines to other non-Mendelian diseases and disabilities. Oza et al. (2018) have presented adaptations to the guidelines for sequence variants in hearing loss genes. While Masson et al. (2022) has outlined a possible extension of the guidelines to a general framework that includes multifactorial diseases, with a focus on chronic pancreatitis in mind.

1.4 Current existing predictors for pathogenic variants

In the absence of any direct human data, we may be able to assess the functional consequence of variant. Validation of the variant through biochemical experimentation, such as functional assays, would be desired. But such procedures tend to be time and resource intensive, with no practical means of approaching all newly discovered variants with the same level of scrutiny

(Grimm et al., 2015). As a result, many researchers apply the use of *in silico* tools to help predict which variants are most likely to impact protein function and therefore which variants are worth investigating further. Over the course of the last several years many different prediction tools have been developed with different methods being utilised. All with the intention of predictively differentiating pathogenic from benign variants.

1.4.1 Traditional conservation based methods

Traditionally, Multiple Sequence Alignments (MSAs) of both nucleotide and amino acids have been used extensively throughout bioinformatics as they have proven instrumental in identifying conservation across and within species (Edgar and Batzoglou, 2006, Kemena and Notredame, 2009, Thompson et al., 2011). This thereby identifies regions of significance to both genomes and proteins. In genomes, conserved regions can highlight the presence of genes or regulatory elements, while in proteins they can identify structural or functional domains. Conservation can be used in such ways due to selection pressures faced through evolutionary time constrains regions of importance to only a select few of viable variations. Any variation that occurs in regions that lead to reduced fitness of the organism will be under stronger selection and be selected against. Therefore any variants present in these regions of strong conservation could imply a deleterious effect. Within protein sequences, the physicochemical properties of different amino acids can be used to gauge similarity. With similar amino acids being less likely to have a deleterious effect on protein function than ones that have dissimilar properties, such as opposing charges. Predictions tools such as SIFT (Ng and Henikoff, 2001) and initially PolyPhen (Ramensky et al., 2002) utilize MSAs and such fundamental rules on conservation in their algorithms for predicting variant effect on protein function. Likewise, the use of MSAs and conservation is important to much of the work produced in this thesis.

1.4.2 Modern machine learning based methods

More recently machine learning has grown in popularity, with researchers having taken advantage of the growth of variant data and turning towards more advanced classification methods (Uddin et al., 2019). In these machine learning methods, large amounts of data are analysed in order to learn from past information and gain potentially hidden insights into the patterns within the data (Mitchell and Mitchell, 1997).

The use of supervised learning in particular has been of great importance in prediction algorithms for variant classification (Uddin et al., 2019). Documented variants with known associations to disease are used to train these algorithms in preparation for prediction. The algorithms learn from the features of these variants and also control variants that are not disease associated. The algorithms then distinguishes what features are most related to disease variants compared to controls and uses those during classification. During testing, any test variants that are determined to have similar attributes to the features of the disease variants used in training would be predicted accordingly as also being disease associated.

There are many databases that archive such known variants, one of which is ClinVar. Clinvar is a publicly accessible database that holds information on human variation and their phenotypic relationships, supported with evidence that arise from clinical testing, experimental/functional research, or extracted from literature (Landrum et al., 2015).

Since PolyPhen's initial development, PolyPhen2 has been updated and now includes the additional use of a naive bayes classifier as a machine learning method. This was trained on variants from the HumDiv and HumVar datasets (Adzhubei et al., 2010). From the training variants, the naive bayes classifier utilizes features that are mainly formed from comparisons of properties of the reference allele and the alternate allele. This includes features like how biochemically similar the variant is to the reference amino acid, whether they have a similar sized side chain for example. Other features include whether the variant occurs at the surface of

the protein or within the core, and whether the variant is located at a highly mutable position.

Many other classifiers follow the same workflow during their development. Some classifiers even utilize multiple measures and incorporate pre-existing tools, such as the SIFT and PolPhen scores, as part of their algorithm design. These ensemble methods consider the predicted outputs of other classifiers in order to provide an overall better consensus prediction. Such ensemble predictors include REVEL (Ioannidis et al., 2016) and CADD (Kircher et al., 2014), which uses a random forest and support vector machine respectively.

1.4.3 Inadequate performance of current predictors

Prediction results from traditional conservation based methods were initially of low specificity and could not have been used as sufficient evidence for pathogenicity, especially within the clinical field (Flanagan et al., 2010). Albeit, since then updates to SIFT and PolyPhen2 have improved their performance. With the advent of machine learning techniques, variant prediction tools have become even more accurate. Gunning et al. (2021) showed that REVEL had an area under the curve (AUC) of 0.818 compared to SIFT (AUC of 0.702) when they used both tools on a dataset composed of clinically relevant variants likely to be observed in sequencing pipelines of diagnostic labratories.

Nevertheless, independent studies have shown that current existing methods commonly used in predicting pathogenic missense variants still tend to perform poorly. Particularly on novel variants, or variants appearing in genes not typically involved in disease. Without additional supporting evidence, they are not reliable enough to be used in a clinical setting (Schiemann and Stowell, 2016; Mahmood et al., 2017; Ernst et al., 2018).

Miosge et al. (2015) showed that though many *de novo* or low frequency missense variants found during genome sequencing are predicted to be deleterious by aforementioned methods, many of these variant were in fact nearly neutral and had little influence on the clinical

phenotype. Indeed, a common issue with current predictors is the over prediction of pathogenicity. Grimm et al. (2015) attributed this problem to the fact that during training, the machine learning methods tend to overfit their training data. While they perform effectively on variants that have similar characteristics as to the ones used in training sets, they are unable to generalise onto variants that have dissimilar or uncommon properties. Such as novel or rare variants that are located in genes not well represented in the training datasets.

This lack of coverage in the training, especially when there is limited data, is an inherent problem to machine learning as a whole and not just to variant classification (Hastie et al., 2009). Guidelines on how best to manage these issues have been developed, with additional compensatory methods, such as the use of cross-validation, have been suggested in order to improve separation of training and testing sets (Vihinen, 2013). However, undoubtedly the annotation and experimental validation of novel variants that will allow the expansion of training data still remains as the best solution. But as aforementioned, this is not always practical or even feasible for researchers.

1.5 Defining variant equivalencies and why it can help predictions

As described above, computational predictions on the pathogenicity of novel variants can often be unreliable and difficult to interpret within a clinical context.

Within the ACMG-AMP guidelines, predictive results from computational tools alone are not considered with great significance unless they are used in reference with another variant of the same amino acid position and of known pathogenicity (**Fig. 1.3**, evidence codes **PM5** and **PS1**), if the variant is predicted to cause protein length alterations (**PM4**), or if a null variant

is predicted where LOF is known to cause disease in the gene the variant is in (**PVS1**).

Benign		Pathogenic			
Strong	Supporting	Supporting	Moderate	Strong	Very Strong
NA	BP1 - Missense in gene where only truncating cause disease BP4 - Multiple lines of computational evidence suggest no impact on gene/gene product BP7 – Silent variant with non predicted splice impact	PP3 – Multiple lines of computational evidence support a deleterious effect on the gene/gene product	PM4 – Protein length changing variant PM5 – Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before	PS1 – Same amino acid change as an established pathogenic variant	PVS1 – Predicted null variant in a gene where LoF is a known mechanism of disease

Figure 1.3: Evidence framework for computational and predictive data as described by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP). Ordered by level of support are the evidence codes and corresponding criteria required to be fulfilled for a benign or pathogenic assertion of a given variant. Note that the numbering of evidence codes does not indicate any weighting for the level of support. Adapted from Richards et al. (2015) with permissions.

For **PM5** and **PS1**, these can be difficult to verify when there is limited variant data that can be referred to, such as the case for rare variants or variants occurring at novel positions. However, although a known variant at the same amino acid position may not exist, a functionally equivalent variant within a genomic region that shares the same evolutionary relationship may. That is to say that for example, paralogues - homologous genes within the same genome that have arisen from gene duplication - are evolutionarily related. Indeed, many paralogues that are part of the same family share similar, if not the same, functionality, but are simply expressed in different tissues. For example, the paralogues *SCN5A* and *SCN2A* both code for voltage-dependent sodium ion channel proteins that mediate action potentials across membranes, except that *SCN5A* is expressed in heart muscle (Veerman et al., 2015) and *SCN2A* in brain tissue (Sanders et al., 2018). A known pathogenic variant occurring in

one paralogue may therefore also be deleterious if present in the equivalent position of another paralogue.

Whereas before a newly sequenced variant would either be described as already been seen and documented or as a novel variant with no prior information, using this concept of equivalency provides a middle ground where a variant of similar nature could be used for reference (variants that are part of the same codon can be thought of as using this concept of equivalency albeit to a much lesser extent). These equivalent variants could then provide information about the variant being analyzed and indicate whether they are pathogenic.

Furthermore, by employing this use of equivalency across variants, the range of variant data that could be used to verify **PM5** and **PS1** could be potentially expanded. Thereby increasing the number of variants that can be predicted to be pathogenic. Equivalent variants may also provide additional data that machine learning tools could possibly use during training in order to better generalise their models.

This thesis will examine the definition of functional equivalence for variants that belong to evolutionary related proteins. I will start by defining evolutionary relationships based on paralogues and evaluate their use for identifying equivalent variants and highlight limitations of this approach. Of which I show that by using protein domains - distinct functional and structural units within proteins that evolve and fold independently from the rest of the protein - in the stead of paralogues alleviates some of those limitations. I then further explore the expansion of this functional equivalency framework by moving from the use of sequence data to possibly protein structure data.

1.6 Aims and hypothesis of thesis

The overall hypothesis in this thesis is given a disease associated protein, if an amino acid substitution is observed to alter the protein function in a way that results in disease, then an

amino acid substitution at a functionally equivalent position in a related protein also involved in disease will also be pathogenic.

To test this, the following aims will be tackled:

1. Identifying how two amino acid substitutions can be defined as equivalent across related human gene sequences.

2. Identifying how using equivalency can help in the prediction of pathogenic variants.

3. Exploring the definitions of equivalency and how they can be expanded from:

   (a) protein gene sequences to protein domain sequences.

   (b) protein sequence to protein structure.

4. To utilise these expansions and to explore if they provide improved performance in predicting pathogenic variants.

Chapter 2

Paralogue Annotation

2.1 Introduction

2.1.1 Defining Paralogues

The concept of homology, that is to say shared evolutionary relationships through a common ancestor, has been a fundamental principle used throughout biology.

Homologous genes can be further specified as either orthologues or paralogues (**Figure 2.1**). Orthologues are genes that share an evolutionary relationship (common ancestor) between the genomes of different species. A speciation event lead to the divergence of these different species, each initially having a copy of the same gene, which over time has diverged due to mutational changes. Orthologues have arguably been the most utilized in bioinformatics as direct comparison of universal genes shared between species have been fundamentally informative in evolutionary biology. Paralogues however, have somewhat been underused in comparison.

Paralogues are genes that share an evolutionary relationship within the genome of a single species. A duplication event on a single original gene lead to copies being established. Several models have been proposed explaining the possible outcomes of the gene copies after duplication.

The two main models that have been subfunctionalization and neofunctionalization.

Subfunctionalization exists due to the proposed theory that the common ancestor of a paralogous family originally carried out multiple functions, but as duplications and divergence occurred, these functions were split among the paralogues. Provided that each function that was originally carried out by the ancestral gene is fulfilled by one of the copies, the fitness of the organism would not be affected (Force et al., 1999).

Neofunctionalization proposes that as long as one copy retains the original function of the common ancestor, the remaining copies face less selective pressure to be as equally conserved. Mutations therefore accumulate and overtime give rise to GOF in other copies (Ohno, 2013). Some mutations on the other hand can lead to LOF, and over a long period of evolutionary time, these nonfunctioning paralogues become pseudogenes.

Furthermore, for paralogues of the same family that appear to provide the same function, different paralogues can be found to be expressed in different tissues or at different stages of development, particularly among multicellular eukaryotes (Zallot et al., 2016).

As highlighted in the previous chapter, aligning sequences is an established methodology in bioinformatics. Sequence alignments can highlight regions of similarity and subsequently, aligned positions can be thought of as potentially being functionally equivalent.

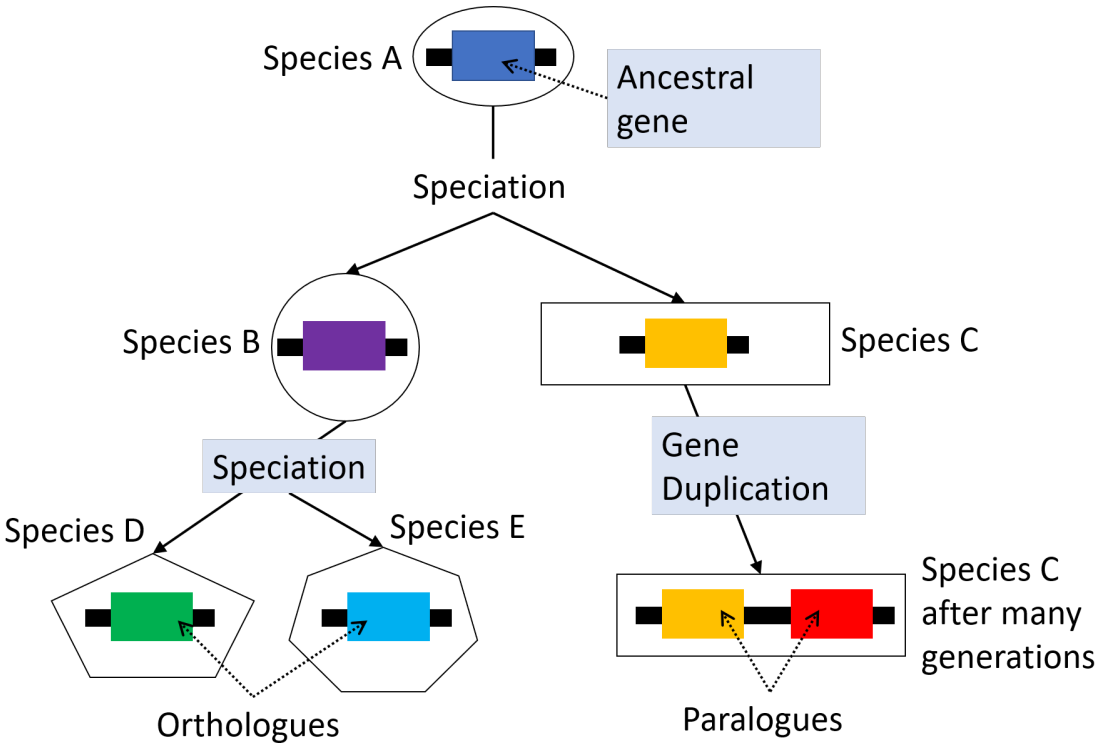


Figure 2.1: Difference between orthologues and paralogues in regards to homology.

2.1.2 Paralogue Annotation

Previously, Ware and Walsh (Ware et al., 2012) developed the concept of Paralogue Annotation as a method for pathogenic variant classification, where variants located in equivalent positions across related paralogues can provide information about pathogenicity. They previously classified pathogenic variants solely on genes involved in LQTS, Brugada Syndrome (BrS) and catecholaminergic polymorphic ventricular tachycardia (CPVT) (Walsh et al., 2014).

The importance of paralogue conservation in pathogenic variant prediction has since been further validated externally across equivalent variants between *SCN4A* and *SCN5A* (two of the main genes involved in LQTS) (Loussouarn et al., 2016), and PA has been shown to perform with higher precision than other computational approaches for disease-causing variants in *SCN5A* and *KCNQ1* (Kapplinger et al. (2015a); Kapplinger et al. (2015b)). Due to this utility, the Clinical Genome (ClinGen) Resource have recently modified the ACMG-AMP criteria for interpreting sequence variants in RASopathies to include the use of

paralogous RASopathy genes when fulfilling the criterias of PS1, PM1 and PM5 (Gelb et al., 2018).

I took this approach and developed it further to enable it to be applied to all possible coding positions in genes that share a evolutionary relationship with a related paralogue.

2.2 Materials and Methods

2.2.1 Surveying number of paralogues among disease and non-disease genes in the human genome

In order to determine how relevant PA is to genes associated with disease, and to genes in general, the number of paralogues each gene had was surveyed. Human genome build GRCh38 data was analyzed from Ensembl Biomart (Ensembl version 93: Jul2018) using R and the BiomaRt package from Bioconductor (Durinck et al., 2005; Kinsella et al., 2011). Only protein coding genes were retrieved. To analyse only disease genes, a subset of genes and their paralogues that contained only P/LP variants from ClinVar was extracted and compared to all genes. A two-sample Kolmogorov–Smirnov test was used to compare distributions.

2.2.2 Paralogue Annotation procedure - use as a classifier and implementation

The general mechanism of PA is illustrated in **Fig. 2.2** steps 1 to 3. Simply, given a query variant or set of query variants, the gene for which these variants belongs to is identified and related paralogues are searched for. The protein sequences of these are aligned together to ascertain equivalent positions across the family to the positions of the query variants. Any known disease-causing variants that are present in these equivalent positions can be searched for and used to infer pathogenicity of the query variants with their annotation being transferred

across the alignment.

Each query variant is handled independently. So if more than one query variant ended up together within the same aligned position with a known pathogenic variant, then that pathogenic variant would annotate all query variants in that aligned position, resulting in the multiple query variants being annotated as pathogenic. If a query variant resulted in an aligned position with more than one pathogenic variant, then this would still only count as a single pathogenic annotation as only a single query variant is being queried. All aligned pathogenic variants to that query are reported and returned, but the additional variants do not add any additional weighting to the annotation. Instead conservation of the variant change was used for increasing confidence of prediction (see section 2.2.3).

For the purpose of distribution to the wider research community, a Perl (Version 5.10.1) plugin for Ensembl's Variant Effect Predictor (VEP) (McLaren et al., 2016) was developed in order to allow users to perform PA on their own set of variants. Query variants are listed in a Variant Call Format (VCF) file and are parsed to VEP. The plugin utilises pre-aligned paralogues from Ensembl's Compara database (Vilella et al., 2009) to return any equivalent amino acids (AA) and positions. A reference set of known disease-causing variants is then used for the transfer of annotations across to the aligned positions of the query variants. Here, ClinVar variants with a clinical significance of P/LP were used as reference as well as being in the test set (see below).

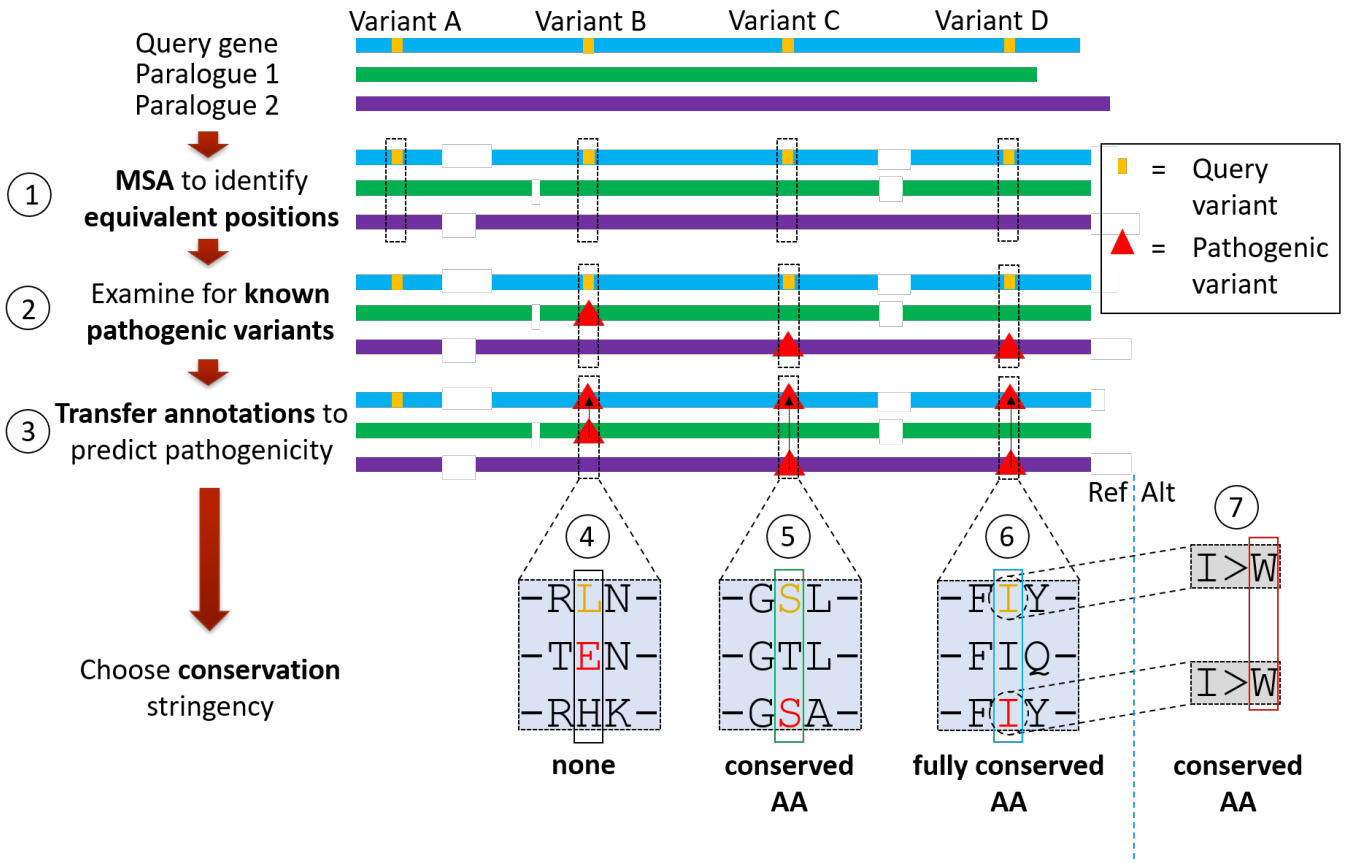


Figure 2.2: Paralogue Annotation schematic overview. (1) Paralogues to the query gene of interest are aligned to identify equivalent positions to the variant of interest (dotted boxes). (2) Any known disease-causing variants in those equivalent positions are identified. Here variant A does not have any aligned disease-causing variants while variants B, C and D do. (3) Their positions are mapped back to the query gene and their annotations transferred, inferring that variants at these sites are likely to also be disease-causing. The conservation between the wild-type reference (Ref) amino acids (AA) in the query gene and reference paralogues can be used to filter annotations. For example, (4) conservation stringency can be ignored and all alignments will be used regardless of reference AA similarity (variant B). (5) Only use alignments where there are conserved AA between the query gene and the paralogue with the annotation (variant C). (6) Full conservation across all genes is required (variant D). (7) In addition, conservation between the alternate (Alt) AA in the query gene and the paralogue can also be used for filtering.

2.2.3 Utilizing conservation stringency to enhance Paralogue Annotation as a classifier

For each query variant and its equivalent positions in aligned paralogues, the reference amino acid conservation across the alignment can be considered via an order of increasingly stringent filters (**Fig. 2.2** Steps 4-6) to increase confidence of true positive annotation transfers. At the lowest stringency (**Fig. 2.2** Step 4) conservation is not evaluated and all aligned positions

are examined (none). At the next level of stringency (**Fig. 2.2** Step 5) only alignments where both the query and reference amino acids are conserved (while not necessarily the whole alignment) are examined (conserved ref AA). Thirdly (**Fig. 2.2** Step 6) only alignments where all equivalent residues across the family in the alignment are conserved are examined (fully conserved ref AA). Finally, in addition to the conservation of the aligned reference amino acids, the conservation of substituting (alternate) amino acids can also be examined (**Fig. 2.2** Step 7); taking only conserved alternate amino acids as well (conserved alt AA). This step can be applied independently following any of the aforementioned steps but has only been done so after conserved ref AA and fully conserved ref AA in this study, seeing as doing so with absolutely no residue conservation did not seem biologically sensible. Furthermore Para Z scores demonstrated by Lal et al. (2020), which quantify paralogue conservation across families were also implemented as cut-off thresholds as a supplementary method for conservation stringency.

2.2.4 Using ClinVar for exome wide study and testing

ClinVar is a free publicly accessible database of known variants associated with genes involved in diseases validated by supporting evidence arising from either clinical testing, research projects or extracted from literature (Landrum et al., 2015). Using version 20190114, ClinVar build GRCh38, provided a gold-standard set of variants that can be used as both reference variants to provide annotations to transfer from and also as test variants to validate the accuracy of PA beyond a couple of genes. The experimental design for this study is shown in **Fig. 2.3**. Variants were downloaded from ClinVar's FTP site (ftp.ncbi.nlm.nih.gov/pub/ClinVar). From 440487 variants, 88528 P/LP and 123567 Benign and/or Likely benign (B/LB) variants with no conflicting interpretations were extracted. 23894 P/LP and 12136 B/LB missense variants in genes with paralogues remained as the test set for analysing the performance of PA. Non-missense variants, such as nonsense, synonymous or indels, and variants with other clinical

classifications, such as VUS, were removed from the study.

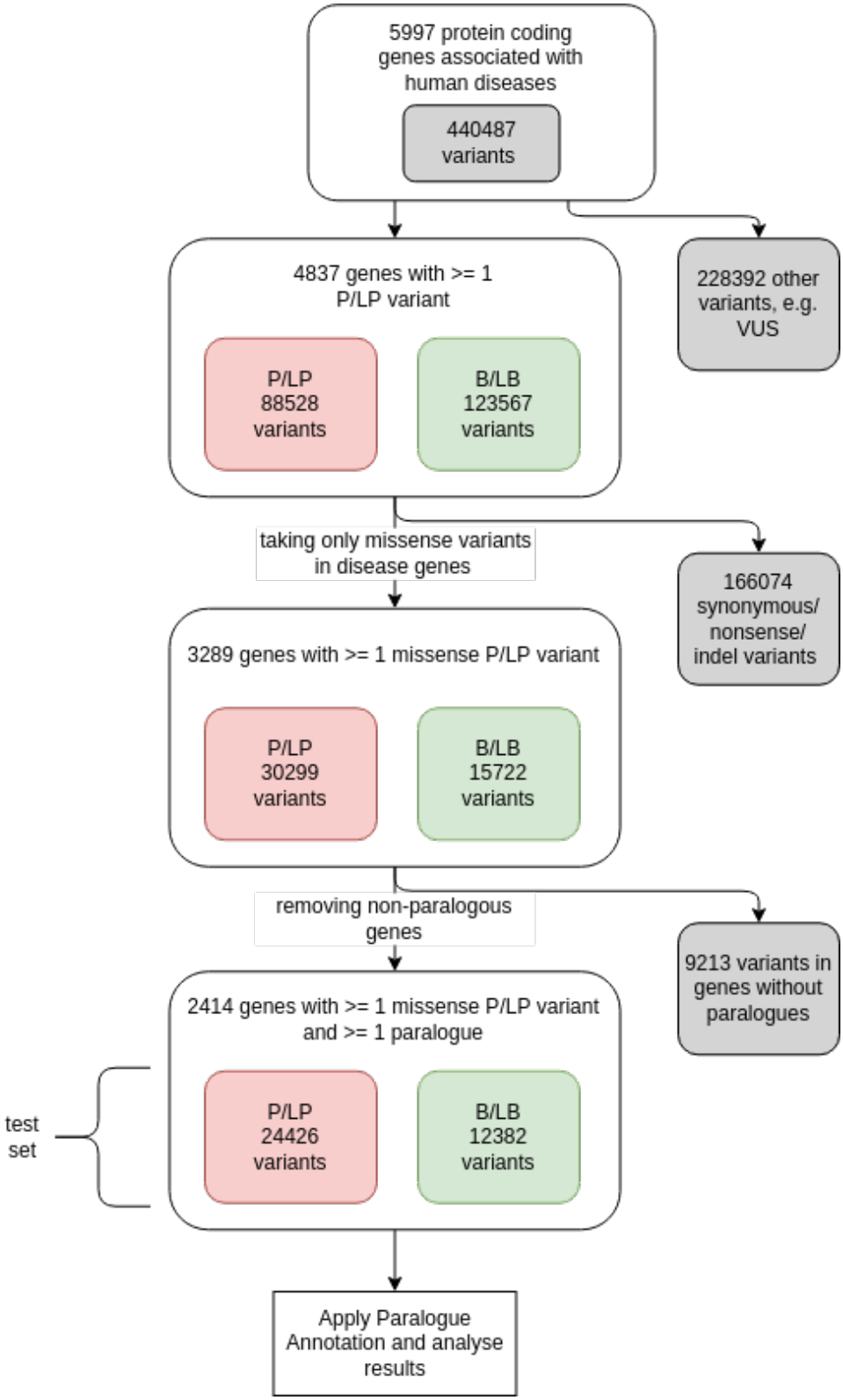


Figure 2.3: Pipeline representation of Paralogue Annotation experimental design. Boxes with straight edges represent experimental process. Boxes with round edges represent numbers of variants that are Pathogenic and Likely pathogenic (P/LP) (red), Benign and Likely benign (B/LB) (green), or variants discarded from the experiment (grey) respectively.

2.2.5 Measuring the statistical performance of Paralogue Annotation and other computational tools as classifiers

The performance of PA and other computational tools were evaluated by 2 statistical measures: 1) The Positive Predictive Value (PPV) (also known as Precision) measures the proportion of variants correctly predicted as being pathogenic out of all variants that were predicted as such; and 2) The Sensitivity (also known as the true positive rate) measures the proportion of actual pathogenic variants being predicted as being pathogenic. In the technical context of PA as a classifier, a True Positive (TP) is when a P/LP query variant aligns with a P/LP reference variant in any one paralogue or more. A False Positive (FP) is when a B/LB query variant aligns with a P/LP variant. Conceptually this means a positive prediction equates to a variant being predicted as pathogenic, while a negative prediction equates to a variant not being pathogenic. The PPV and Sensitivity (also known as Recall) are therefore calculated as follows:

$$PPV = \frac{TP}{TP + FP} \quad ,$$

and

$$Sensitivity = \frac{TP}{P} \quad ,$$

where P is the total number of Positives, that is to say the total number of P/LP test variants used in the analysis.

P values for statistical significance are calculated via a Fisher's exact test on a 2x2 contingency table. The performance of PA as a classifier was compared to REVEL (Ioannidis et al., 2016), SIFT(Ng and Henikoff, 2003), MCAP (Jagadeesh et al., 2016) and CADD (Kircher et al., 2014). All scores were retrieved and annotated using ANNOVAR with dbnsfp35a (Wang et al., 2010; Liu et al., 2016). To ensure comparable statistical analysis, only variants that were able to be annotated by all methods were used during comparisons.

2.2.6 Calculations of Odds Ratios for Case and Control comparison

In addition to using ClinVar as a test set, Odds Ratios (OR) calculated from case and control comparisons were also performed as an orthogonal study to validate PA. Rare variants found in patients diagnosed with LQTS (Kapplinger et al., 2009), BrS (Kapplinger et al., 2010), and Neurodevelopmental Disorder with Epilepsy (NDD+E) (Heyne et al., 2019) were used as cases while equivalently rare variants in GnomAD v2.1.1 (Karczewski et al., 2020) were used as controls. Rare variants in LQTS and BrS were defined as any that had a Minor Allele Frequency (MAF) of less than or equal to 8.2×10^{-6} and 1.0×10^{-5} respectively as determined in methods outlined by Whiffin et al. (2017). Rare variants in NDD+E were defined as any not present in DiscovEHR (Dewey et al., 2016). OR were calculated from the cases and controls harbouring these rare variants and then compared to OR calculated from rare variants that were also predicted pathogenic by PA. OR were calculated by:

$$OR = \frac{(a/b)}{(c/d)}$$

where a is the number of cases with variants that are rare (and predicted to be pathogenic), b is the number of controls with variants that are sufficiently rare (and predicted to be pathogenic), c is the number of cases with variants that are not sufficiently rare (and not predicted to be pathogenic), and d is the number of controls with variants that are not sufficiently rare (and not predicted to be pathogenic). The 95% confidence intervals for OR values were calculated according to Altman (1990) via:

$$95\%\ CI = [e^{\ln(OR) - 1.96 \cdot SE(\ln(OR))}, e^{\ln(OR) + 1.96 \cdot SE(\ln(OR))}]$$

where:

$$SE(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

2.2.7 Independent test sets used

In addition to ClinVar, variants from independent datasets experimentally validated were also tested on. These include variants from Mahmood et al.'s TP53-TA set (functional mutations defined by transactivation assay mutational scanning in the tumour suppressor gene *TP53*) (Mahmood et al., 2017) and Li et al.'s PPARG set (functional classification of all possible missense variants from *PPARG*). An additional dataset containing *de novo* mutations (DNM) identified in patients with autism spectrum disorder (ASD) was also used (Li et al., 2018).

2.2.8 Gene Ontolgy and PANTHER

Functional classification of PANTHER (v16.0) Protein Classes was conducted on http://pantherdb.org/. Tabulated results were downloaded and visual representations was redrawn using the ggplot package in R.

2.2.9 Commonly used tools and software

Pipeline management of workflow and data manipulation was handled primarily in Python 3.7.3. Data and statistical analysis performed on R (v3.6.3). Packages used include tidyverse (v1.3.1) and ggplot2 (v3.3.5).

2.3 Results

2.3.1 Number of Paralogues in the human genome and disease

In Ensembl, 72.0% (14,514/20,158) of all protein coding genes have at least one paralogue (**Fig. 3a**), with a mean of 4.5 paralogues per gene. The largest paralogue family comprised of 49 paralogues, which was the case for a family of Zinc Finger genes. These genes are known to have quite diverse structure and function (Laity et al., 2001).

Taking only genes known to be involved in disease (as defined by Pathogenic/Likely pathogenic variants in ClinVar) gives 5117 genes with a similar proportion of genes having at least 1 paralogue (71%, 3621/5117). But the distribution compared to all genes is significantly different (p = 2.5 x 10-4, Two-sample Kolmogorov–Smirnov test). The variance of disease genes was smaller with a mean of 4.1 and a standard deviation of 4.7, and for the ones that had at least one paralogue a mean of 5.8 ± 4.6 was found. The maximum was still 49, being *ZNF141* from the Zinc Finger family, which is associated with autosomal recessive postaxial polydactyly type A (Klopocki et al., 2013) and Wolf-hirschhorn syndrome (Tommerup et al., 1993). None of the other members of that paralogue family were involved in disease.

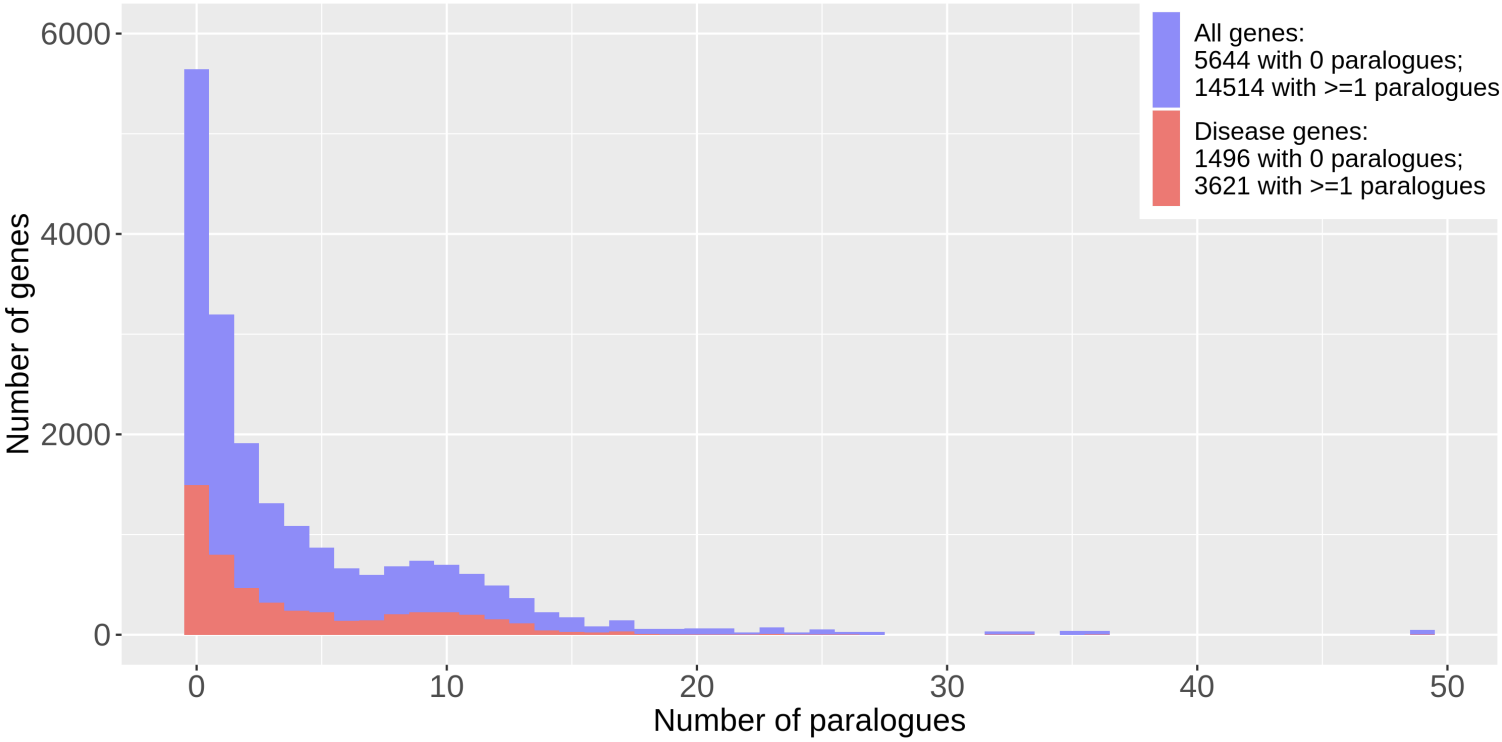


Figure 2.4: Most human protein-coding genes, including those known to be associated with disease, have at least one paralogue. A histogram shows the distribution of the number of paralogues for all human protein-coding genes (blue), and for genes known to be associated with monogenic disease (red), defined as genes with at least one pathogenic or likely pathogenic variant in ClinVar. 71% of disease-associated genes have at least one paralogue, which may be informative for paralogue annotation.

2.3.2 Annotation of ClinVar

When PA is applied to the 23,894 P/LP ClinVar variants, 3977 are correctly predicted to be pathogenic (17%). Conversely, only 235 B/LB variants are incorrectly predicted to be pathogenic out of 12,136 (2%). These predicted variants span across 624 genes (supplementary table A.1). This leads to an initially high precision of 0.94 (compared to 0.66 from a random classifier, see section 2.3.4 for more details), but a sensitivity of only 0.16 (**Table 2.1**).

When the query gene and its paralogue are required to have the same reference amino acid at the position of interest, 3533 P/LP and 91 B/LB variants are predicted to be pathogenic, increasing precision to 0.97, while sensitivity decreases to 0.14. Increasing the conservation stringency further by considering only equivalent positions where the reference amino acids are fully conserved across the whole family additionally decreases the number of P/LP and

B/LB variants being predicted as pathogenic, but the effect is greater for B/LB variants. This therefore increases precision to 0.99, while sensitivity further decreases to 0.079. This trend continues when alternate amino acids are also required to be conserved. The result in an overall precision increase to 1.0 at maximum conservation when considering variants that have fully conserved reference and alternate amino acids. At the same time, sensitivity also drops to a minimum of 0.028.

Important to note that the variants removed as conservation stringency increases should not necessarily be ignored completely as precisions calculated from each removed set of variants are also still quite high with a mean of 0.92 (maximum of 0.99 - variants that have fully conserved reference amino acids but not necessarily conserved alternate amino acids (c - d)) (**Table 2.2**). The ones with non-conserved reference amino acids (a - b) had the lowest precision (0.76) with most benign variants being removed at this stage (144 out of the total 235 B/LB). All further stages had precisions of 0.96 or above. This suggests that these variants have the least amount of information, while at higher conservation stringencies the variants being removed have a higher quality of information. When considering the sole effect of alternate amino acid conservation (b - d; c - e), the variants removed were also of high information with 0.97 and 0.99 precision respectively (**Table 2.3**). Users are therefore recommended to use PA with conserved reference amino acids when classifying variants for optimal performance in both precision and sensitivity as further conservation levels may prove to be too stringent. Though the choice of conservation stringency will ultimately depend on user preference and circumstance.

		Pathogenic variants		Benign variants		Pathogenic to Benign ratio (expected PPV of random classifier)		
Total variants		24426		12382		0.66		
Test		Test Positive (TP)	Test Negative (FN)	Test Positive (FP)	Test Negative (TN)	PPV	Sensitivity	P value
(a) Does variant align to a P/LP paralogue variant?		3977	20449	235	12147	0.94	0.16	<5.0e-324
increasing conservation stringency ⇓⇓	(b) Does variant align to a P/LP paralogue variant with the same conserved ref AA	3533	20893	91	12291	0.97	0.14	<5.0e-324
	(c) Does variant align to a P/LP paralogue variant where ref AA is fully conserved family-wide?	1933	22493	21	12361	0.99	0.079	1.4e-316
	(d) Does variant align to a P/LP paralogue variant with the same conserved ref and alt AA?	1318	23108	15	12367	0.99	0.054	1.3e-212
	(e) Does variant align to a P/LP paralogue variant where ref AA is fully conserved family-wide and with the same conserved alt AA?	675	23751	3	12379	1.0	0.028	8.4e-116

Table 2.1: Shown are number of ClinVar variants with a clinical significance of Pathogenic/Likely pathogenic (P/LP Variants) and Benign/Likely benign (B/LB Variants) that are predicted to be pathogenic using **(a)** Paralogue Annotation. Paralogue Annotation is also subsequently assessed at increasing conservation stringency levels **(b-e)** for the reference (ref AA) and alternate (alt AA) amino acids of the aligned variants (see Methods **Fig. 1**). The Precision, Sensitivity, and Fisher's Exact test p-value are shown to the nearest 2 significant figures where applicable. TP - True Positive, FN - False Negative, FP - False Positive, TN - True Negative. Note: machine performing calculations cannot compute values smaller than 5e-324.

Difference between respective conservation stringencies	Pathogenic variants	Benign variants	PPV
a - b	444	144	0.76
b - c	1600	70	0.96
d - e	643	12	0.98

Table 2.2: Shown are the number of variants predicted to be pathogenic by Paralogue annotation that are removed during filtering between each additional conservation stringency level, where: **a** = Paralogue Annotation with no conservation stringency; **b** = same conserved reference amino acids (ref AA); **c** = fully conserved family-wide ref AA; **d** = same conserved ref AA and conserved alternate amino acids (alt AA); and **e** = fully conserved family-wide ref AA and same conserved alt AA. Note: machine performing calculations cannot compute values smaller than 5e-324

Difference between respective conservation stringencies	Pathogenic variants	Benign variants	PPV
b - d	2215	76	0.97
c - e	1258	18	0.99

Table 2.3: Shown are the number of variants predicted to be pathogenic by Paralogue annotation that are removed during filtering when considering only the conservation of alternate amino acids (alt AA), where **b** = same conserved reference amino acids (ref AA); **c** = fully conserved family-wide ref AA; **d** = same conserved ref AA and conserved alt AA; and **e** = fully conserved family-wide ref AA and same conserved alt AA. Note: machine performing calculations cannot compute values smaller than 5e-324

2.3.3 Correlation with across-family paralogue conservation

Utilizing the Para Z scores from Lal et al. (2020) as additional cut-offs of conservation stringency provides further coverage of precision and sensitivity. This addresses the missing data points (**Fig. 2.5**) between only conserved amino acids in query and reference paralogues to full conservation across alignments. As expected, the trend remains the same. The more conserved paralogues are, the higher the precision, but lower the sensitivity. Here the Para Z scores are utilized with conserved AA but they can be used in addition with any other conservation stringency. The trend would likely remain unchanged. Para Z scores therefore further validate the usage of conservation stringency across paralogues as a method of increasing confidence in predicting pathogenicity.

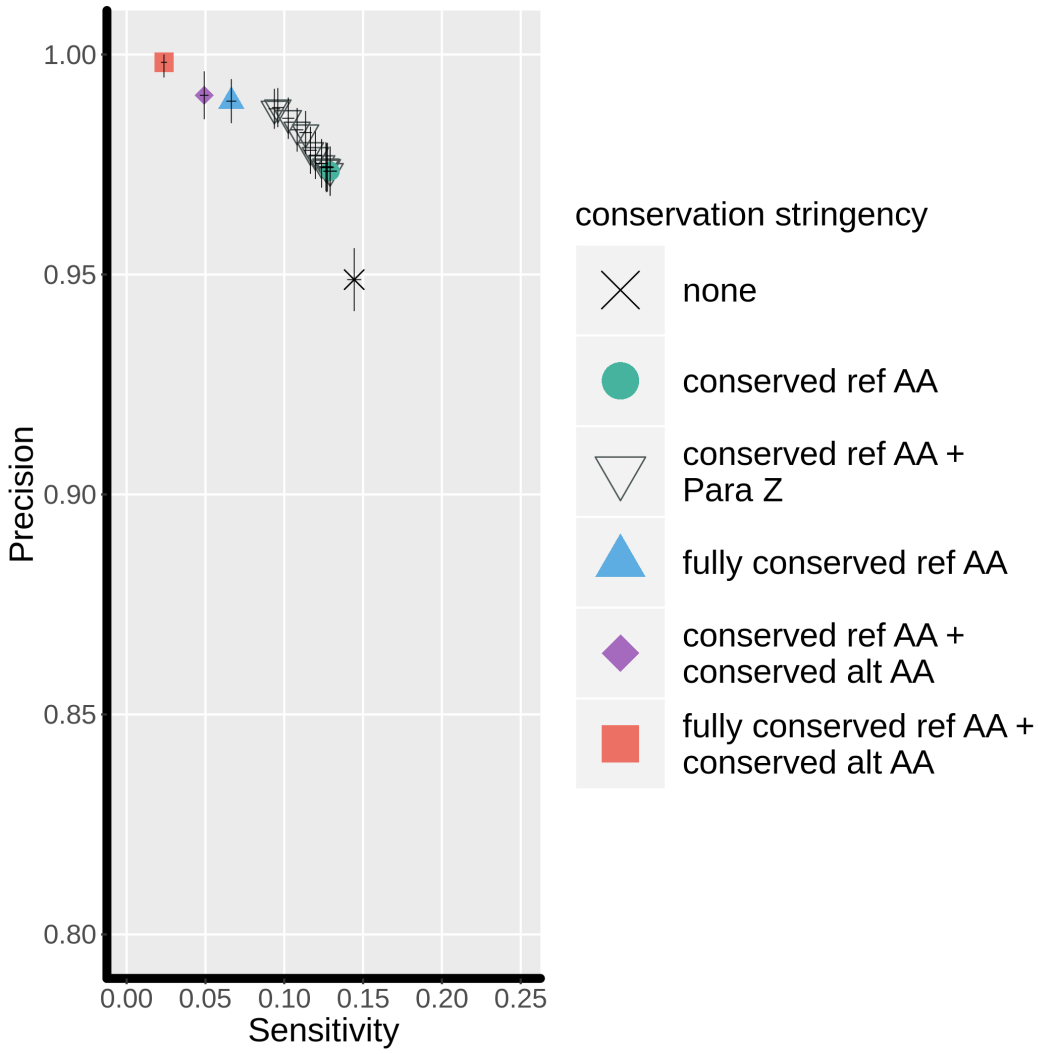


Figure 2.5: Precision-Sensitivity space plot show that predicting pathogenic variants with Paralogue Annotation using Para Z scores as threshold cut-offs supplements conservation stringencies well. The precision and sensitivity of Paralogue Annotation is shown when applied to the whole ClinVar dataset with the use of Para Z scores as cut-offs together with when there are conserved amino acids (AA) between the query gene and the paralogue with the annotation (see Methods 2.2). Other conservation stringencies are also shown.

2.3.4 Addressing possible bias issues with unbalanced datasets, particularly at high conservation stringencies

One concern that arose was that the greater number of P/LP variants being predicted as pathogenic than B/LB variants was simply due to there being an imbalance of more P/LP than B/LB variants in the test set. Particularly at higher conservation stringencies where even more B/LB variants are removed. Therefore, the high precision seen may be due this imbalance and statistical artefact, rather than as a actual result of the Paralogue Annotation mechanism.

This is addressed in three ways. Firstly, the ratio of total pathogenic variants to benign variants is explicitly stated (0.66 in **Tables 2.1** and **2.4**) in order to give a prior expectation of the precision produced by a random classifier. This acts as the baseline precision and helps demonstrate the improvement that PA can provide.

Secondly, the additional conservation stringencies are assessed independently from PA. Since conservation of residues was used as a measure of confidence of annotation, the effects of these two parameters are disentangled in order to evaluate paralogue conservation as a predictor of pathogenicty. This was done by ignoring the number of aligned variants and considering only the aligned paralogous positions. Conservation stringency is then applied on those positions incrementally (**Table 2.4**). Variants at a conserved position are predicted pathogenic, while variants at a non-conserved position are not. The same statistical analyses are then applied as before.

Under this approach, the PPV of whether a variant has an aligned paralogous position would be equal to that of a random classifer at 0.66. But as conservation increases, PPV increases to a maximum of 0.92 if using Lal et al.'s Para Z scores and to 0.91 if looking at aligned paralogue positions where there is full conservation across the family. Conservation alone is comparable but still not as powerful as a predictor as Paralogue Annotation alone - 0.94.

The important point to note is that even when conservation increases to the extent where there is family-wide conservation across an aligned position, there is still a moderately sized number of B/LB variants that have an aligned paralogue position remaining - 822 compared to the 3 remaining B/LB variants with aligned P/LP variants in **Table 2.1e**

Evidently this demonstrates that the issue of low numbers of B/LB variants with aligned P/LP variants at high conservation stringencies is not due to a lack of B/LB variants all together, but rather that the B/LB variants do not align with P/LP variant positions. Thus proving that the high precision is due to Paralogue Annotation mechanism and not due to a

lacking or limited number of B/LB variants in the dataset.

	Pathogenic variants		Benign variants		Pathogenic to Benign ratio (expected PPV of random classifier)		
Total variants	24426		12382		0.66		
Test	Test Positive (TP)	Test Negative (FN)	Test Positive (FP)	Test Negative (TN)	PPV	Sensitivity	P value
Does variant have an aligned paralogue position?	24018	408	12207	175	0.66	0.98	0.06
Does variant have an aligned paralogue position and ParaZ >= 1	16088	8338	4972	7410	0.76	0.66	<5.0e-324
Does variant have an aligned paralogue position and ParaZ >= 3	15729	8697	4549	7833	0.78	0.64	<5.0e-324
Does variant have an aligned paralogue position and ParaZ >= 5	14486	9940	3555	8827	0.80	0.59	<5.0e-324
Does variant have an aligned paralogue position and ParaZ >= 7	13016	11410	2536	9846	0.84	0.53	<5.0e-324
Does variant have an aligned paralogue position and ParaZ >= 9	11017	13409	1448	10934	0.88	0.45	<5.0e-324
Does variant have an aligned paralogue position and ParaZ >= 11	9420	15006	863	11519	0.92	0.39	<5.0e-324
Does variant have an aligned paralogue position that is fully conserved across the family?	8376	16050	822	23604	0.91	0.34	<5.0e-324

Table 2.4: Number of aligned paralogous positions at varying paralogue conservations among ClinVar variants. Shown are the number of aligned paralogue positions among ClinVar Pathogenic/Likely pathogenic and Benign/Likely benign variants at increasing ParaZ score thresholds. The number of aligned positions that are also fully conserved across the family are included for comparison. The Precision, Sensitivity, and Fisher's Exact test p-value are shown to the nearest 2 significant figures where applicable. TP - True Positive, FN - False Negative, FP - False Positive, TN - True Negative. Note: machine performing calculations cannot compute values smaller than 5e-324.

Thirdly, this can be further demonstrated by taking the 822 B/LB variants and randomly sampling an equal number of 822 P/LP variants from the 8376 that are also fully conserved across the family (in **Table 2.4**), thus creating a balanced test set. The same statistical analyses (precision, sensitivity, and p-value from a Fisher's exact test) were then performed as before and this was repeated 1000 times in simulation. A Kernel Density Estimation (KDE) showing the distribution of the precision calculated from the 1000 random samplings demonstrates that precision was still 0.90 on average (**Fig. 2.6**).

As a result, the high precision attained can be attributed to the combined effects of Paralogue Annotation and the use of amino acid/variant conservation, and not due to artefacts of the test set used.

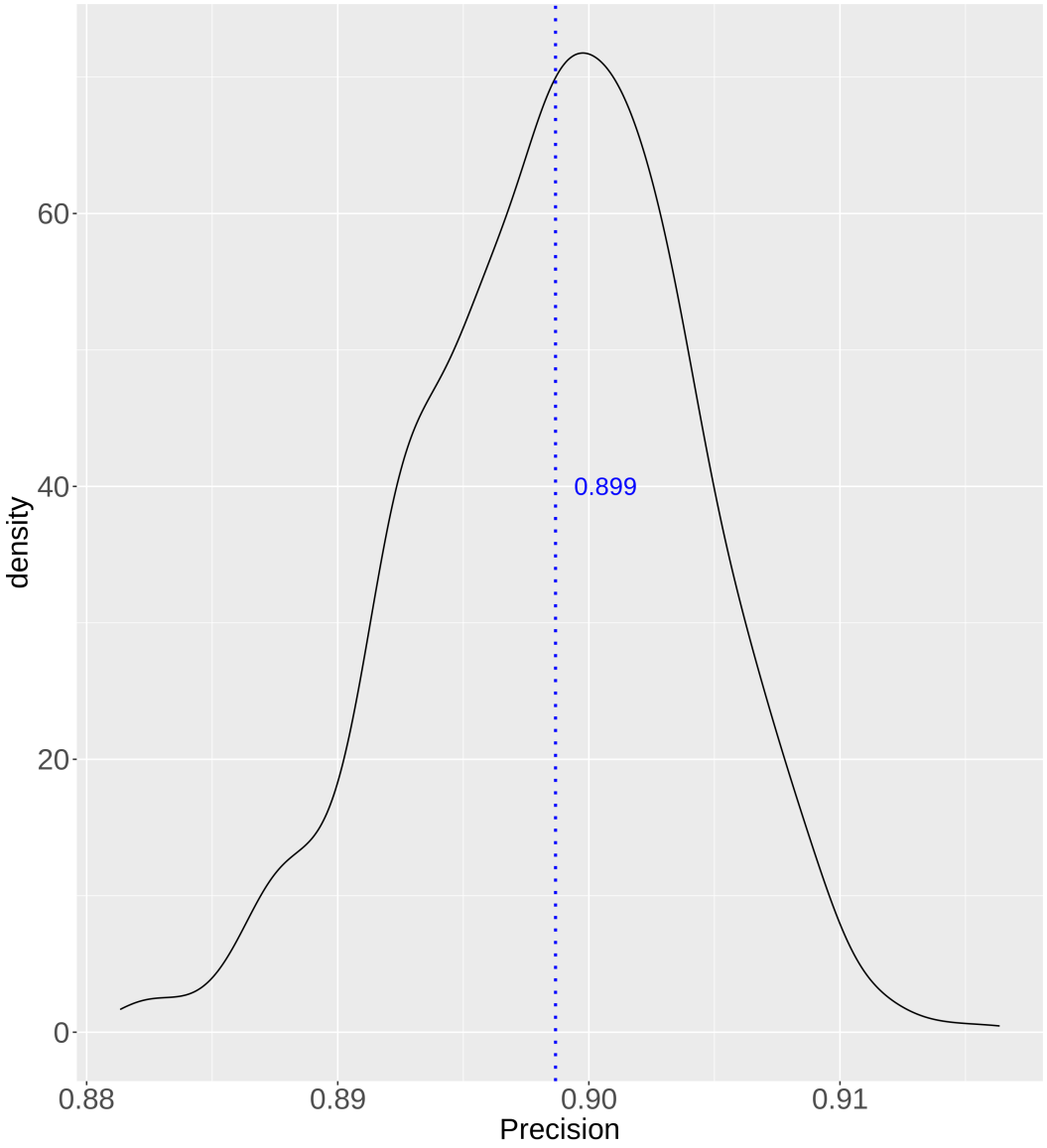


Figure 2.6: Kernel Density Estimation visualising the distribution of precisions calculated from 1000 randomly sampled balanced test sets of Benign/Likely benign (B/LB) and Pathogenic/Likely pathogenic (P/LP) variants that are fully conserved family-wide. The test sets consist of 822 static B/LB variants and an equal number of P/LP variants that are sampled randomly from the 8376 in **Table 2.4** (last row). The mean precision of 0.899 is shown by the blue dotted line.

2.3.5 Paralogue Annotation can be applied to multiple protein classes

Paralogue Annotation was originally only explored among protein ion channels. These genes had many occurring paralogues that allowed Paralogue Annotation to be utilized extensively. Therefore when expanding to other genes, one concern was the ability to translate its utility across to genes of a different nature, where there may not be as many paralogues to work with.

But PANTHER classification of protein classes shows that Paralogue Annotation is not limited to only a specific class of proteins. From the 3621 disease genes with paralogues (**Fig. 2.4**), 1947 of these have paralogues that are also involved in disease. Out of the possible 1947, PANTHER found hits for 1542 genes spread among 22 protein classes (**Fig. 2.7**, blue bars). PANTHER found no hits for 405 genes. The classes with most genes were metabolite interconversion enzymes (277 hits) - this includes enzymes such as hydrolases, isomerases and ligases. This is followed by transporter (220) and gene-specific transcriptional regulator (197) genes. These classes encompass channel proteins and transcription factors respectively. The classes with the fewest genes with paralogues were calcium-binding (6), storage (2) and defense/immunity proteins (2). Among genes with variants predicted to be pathogenic by PA, the distribution is similar though naturally all classes had universally fewer hits 508 (**Fig. 2.7**, red bars) out of 624 genes. Transporter proteins were the class with most genes that had paralogue annotated variants (99), followed by metabolite interconversion enzymes (82) and then transcription factors/gene-specific transcriptional regulators (81). The classes with the fewest genes that had predicted pathogenic variants were scaffold/adaptor proteins (3), calcium-binding proteins (3), and structural proteins (2). Genes classified as translational proteins, chaperones, storage proteins or defense/immunity proteins did not have any annotations. Notably, in the latter two classes the number of genes with paralogues were also near non-existent, whereas for translational proteins and chaperones there were a few

paralogues in those groups.

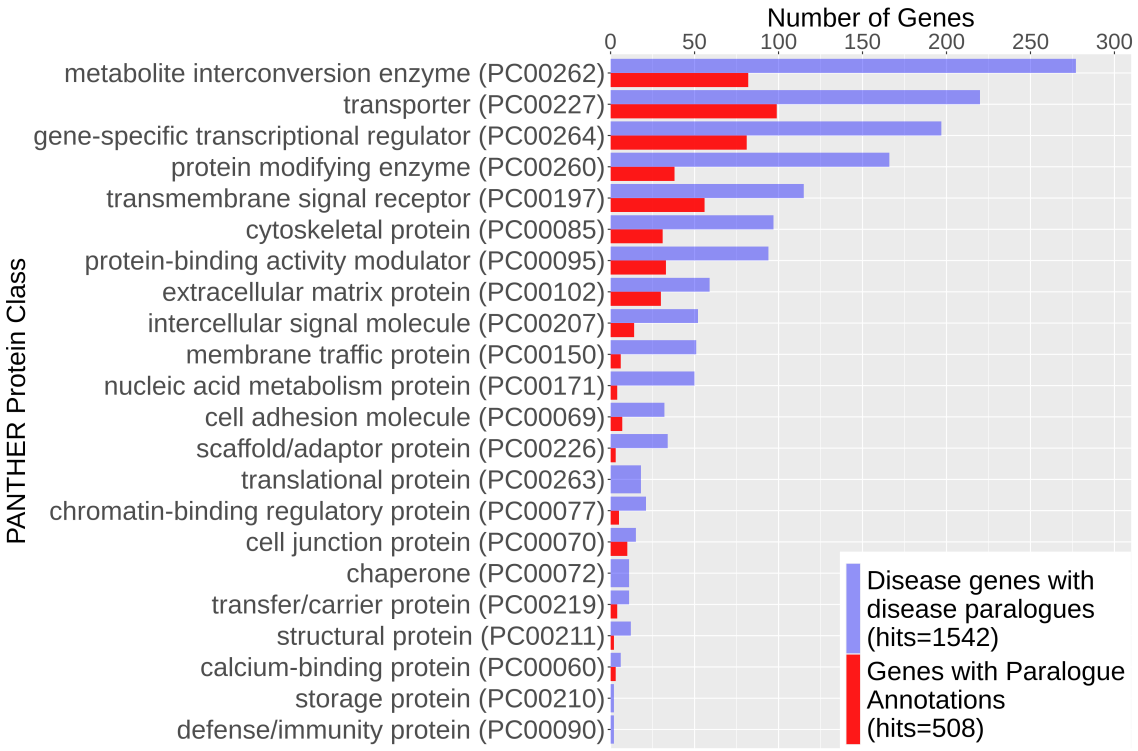


Figure 2.7: Paralogue Annotation is applicable to genes encoding a diverse range of protein classes. The distribution of protein classes encoded by these genes is shown (defined using the PANTHER classification): in blue - ClinVar disease genes that have paralogues which are also involved in disease; in red - genes that had at least one variant predicted to be pathogenic by Paralogue Annotation.

2.3.6 Paralogue annotation is more precise but less sensitive than established methods

To assess whether PA performs better than established methods for predicting the pathogenicity of missense variants I compared its performance to 3 established machine learning based classifiers (MCap, CADD, and Revel) and the traditional homology based method - Sift. Generally, PA has higher precision across multiple genes and test sets, but this is at a trade-off of lower sensitivity however, while the other methods generally had higher sensitivity but at a consistently lower precision than PA. For ClinVar variants, all other methods had a sensitivity of at least 0.89, with Revel having the best precision out of the 4 other methods at 0.95 (**Fig. 2.8a**). PA had a higher precision even at no conservation

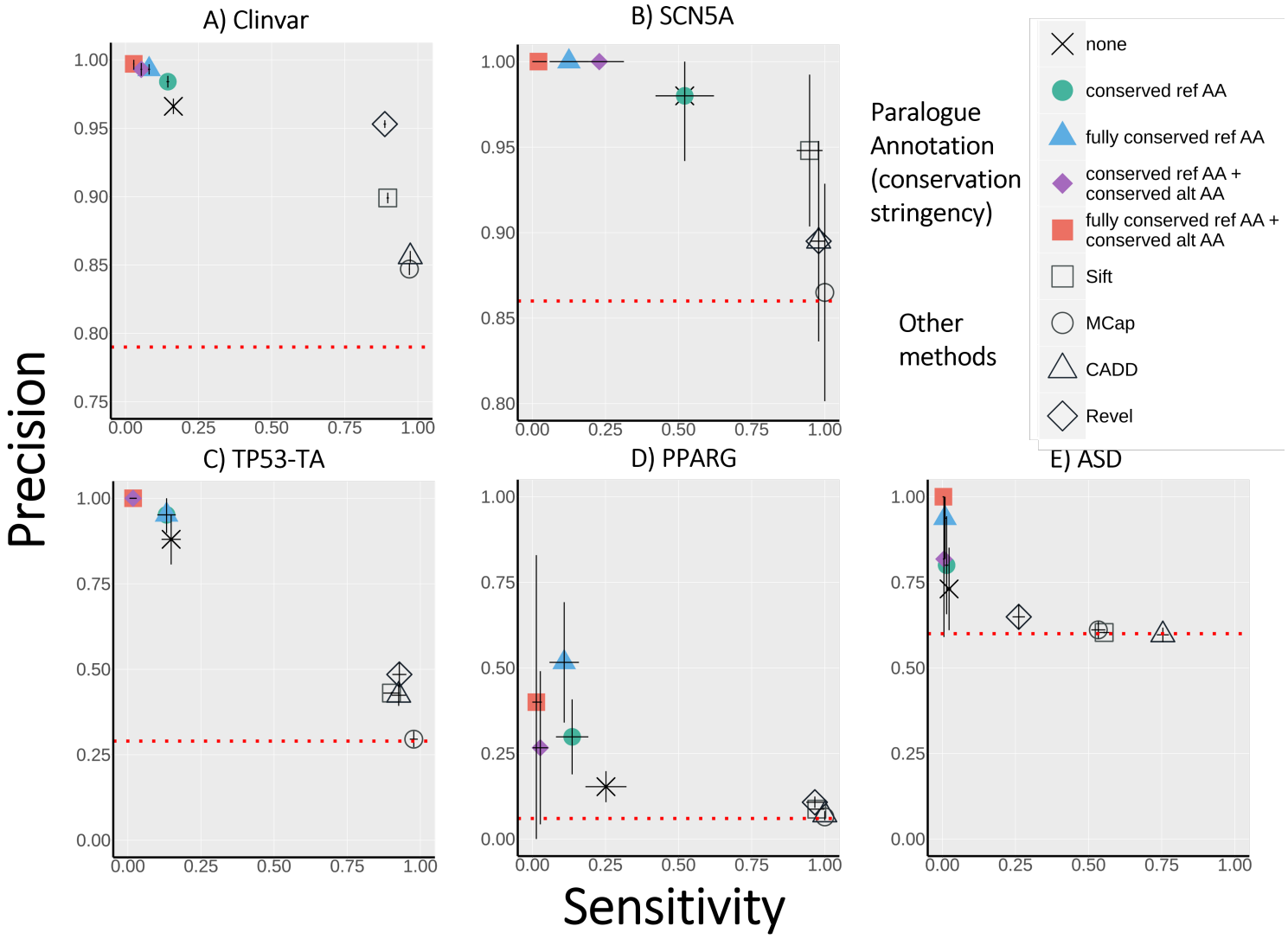


Figure 2.8: Precision-Sensitivity space plot of predicting pathogenic variants show Paralogue Annotation is more precise but less sensitive than other variant prediction tools. The precision and sensitivity of Paralogue Annotation is shown with varying levels of conservation stringency filtering compared with that of Sift, MCap, CADD and Revel applied to a) the whole ClinVar dataset, b) ClinVar variants only in SCN5A , independent test-sets consisting of functional assay-determined variants from c) *TP53*, d) *PPARG* and e) de novo mutations linked to Autism Spectrum Disorder (ASD). 95% confidence intervals shown as error bars.

stringency with 0.97; maximum of 0.997 when at full conservation. But a much lower sensitivity of 0.16 to 0.03 respectively. However, when analysing only variants from *SCN5A*, PA maintained high precision across all conservation stringencies but also had increased overall sensitivity with a maximum of 0.52 (**Fig. 2.8b**). The other methods however did not have any notable changes in sensitivity and although they varied individually in their precision compared to before, they were still within the same relative range.

After performing the same comparison on variants from the TP53-TA test set, all other methods dropped in precision (Revel still outperforming the rest at 0.48) while PA maintained the same relative range (minimum - 0.88; maximum - 1.00) (**Fig. 2.8c**). Sensitivity did not change significantly for any method. A similar effect can be seen for variants from *PPARG*, though for this set all methods had greater reduced precision, with Revel still being the best (0.11) from all other methods and PA still outperforming at all conservation stringencies (minimum of 0.15 at no conservation stringency) (**Fig. 2.8d**). Interestingly, maximum precision was achieved when only AA were fully conserved (0.52) as considering if variants were also conserved caused a decrease in precision. Again, sensitivity remains relatively unchanged for all. Finally, for the DNM among ASD patients, performance dropped significantly for the other methods in both precision and sensitivity compared to their performance on ClinVar variants. Revel still having the highest precision of the group with 0.65 but also with the lowest sensitivity of 0.26. Although PA has slightly worse sensitivity overall for this set of variant as well, it continued to outperform the other methods with higher precision (minimum of 0.73 at no conservation stringency; maximum of 1.00 at full conservation) (**Fig. 2.8e**).

2.3.7 Paralogue Annotation increases odds of correctly calling pathogenic variants in case-control comparisons

An alternative way of validating the ability of Paralogue Annotation to classify pathogenic variants was to see what impact using PA had on the disease burden of rare missense in known disease genes. Specifically, genes known to cause the channelopathies LQTS and BrS, and Neurodevelopmental Disorders with Epilepsy (NDD+E) were used. PA substantially improves the OR in 2 genes involved in LQTS - *KCNQ1* and *KCNH2* (**Fig. 2.9**), from 60.0 and 33.8 to 161.5 and 102.9 respectively. However for *SCN5A*, the increase was negligible. Only for cases

involved in BrS did *SCN5A* have improved OR from 30.4 to 53.7. For 4 genes implicated in NDD+E, *SCN1A*, *SCN8A*, and *KCNQ2* saw small but significant increases in OR from a mean of 1.2 between the 3 genes to 12.7. *SCN2A* saw no significant changes.

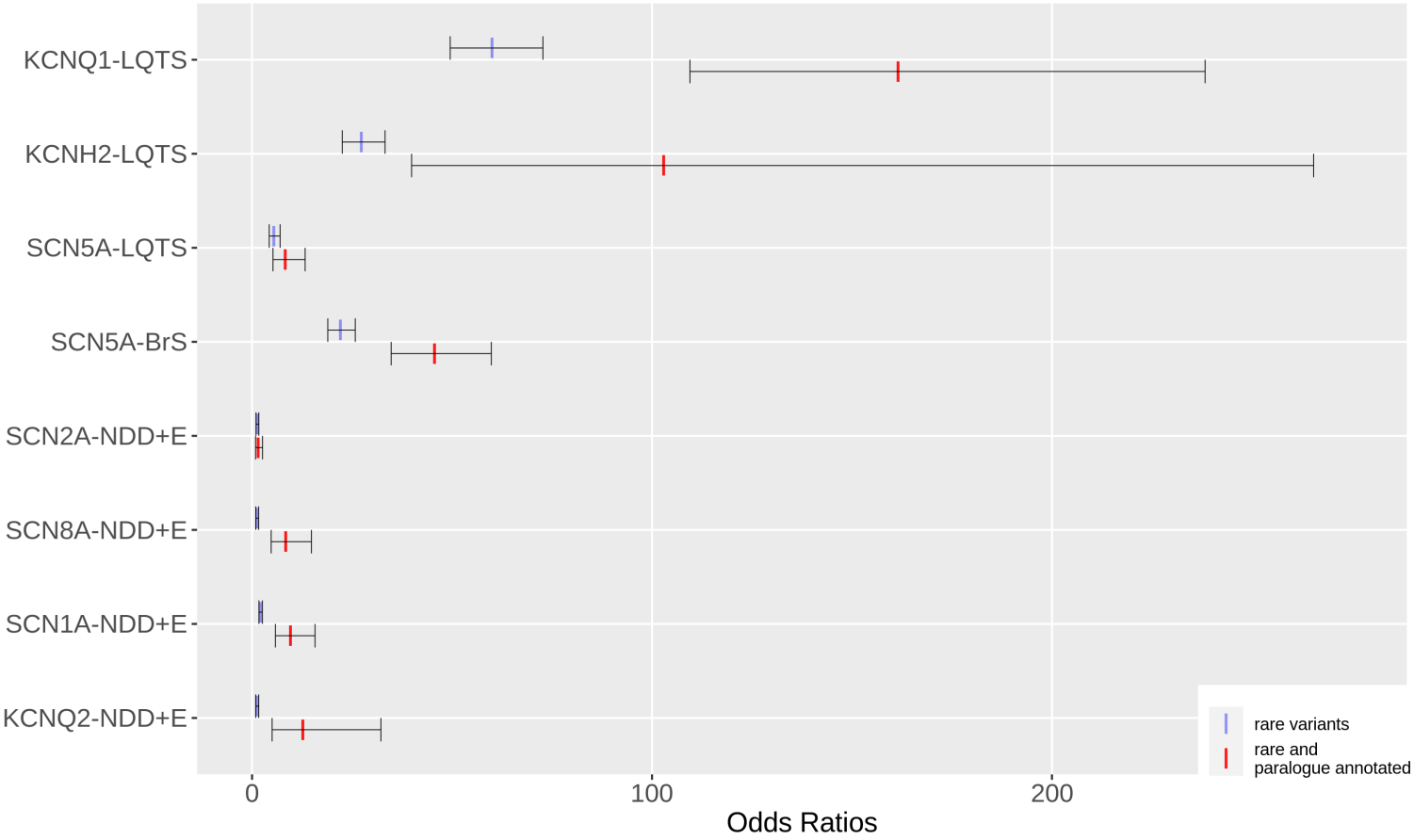


Figure 2.9: Improved Odds Ratios before and after Paralogue Annotation is applied. Shown are Odds Ratios calculated from case and control series for: *KCNQ1* and *KCNH2* in Long QT Syndrome (LQTS); *SCN5A* in LQTS and Brugada Syndrome (BrS); *SCN2A*, *SCN8A*, *SCN1A*, and *KCNQ2* in Neurodevelopmental Disorder with Epilepsy (NDD+E). In blue are variants when filtered only by sufficient rarity and in red are variants that are also additionally paralogue annotated (with conserved AA). Error bars show 95% confidence intervals.

Predicted pathogenic positions coincide with positions of significance. In the case of KCNQ1 protein the distribution of rare variants that are predicted to be pathogenic is not randomly distributed, with variants focused around regions known to be associated with disease (**Fig. 2.10**). *KCNQ1* has 4 paralogues according to Ensembl containing 237 missense variants that are reported as P/LP in ClinVar (as of version 20190520). Mapping those variants back to KCNQ1 annotates 92 amino acid positions, of which 41 have been previously robustly flagged

as sites of P/LP alleles, 0 have been flagged as sites of B/LB alleles, and 51 positions have not been observed or characterised. The majority of sites annotated (82) are located around the vicinity of the 6 transmembrane helices and intramembrane pore-forming loop domain that form the core structure of the channel. The remaining 10 sites are situated around 2 C-terminal domains known to interact with KCNE1 and AKAP9.

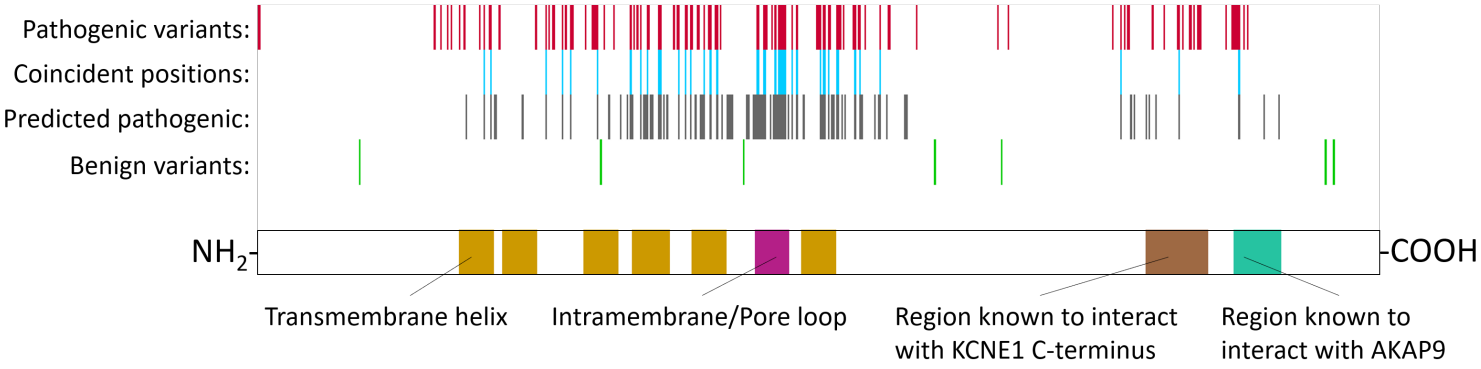


Figure 2.10: Schematic representation of KCNQ1 showing that variants predicted to be disease-causing by Paralogue Annotation cluster in regions densely populated with pathogenic variants from ClinVar. Along the bottom is a schematic diagram of KCNQ1 protein with known Uniprot domains annotated. The 1st and 4th rows show residue positions where known ClinVar pathogenic and known benign variants are located respectively. The 3rd row shows all possible missense variant positions for KCNQ1 that are predicted to be pathogenic by Paralogue Annotation with conserved AA. The 2nd row shows coincident positions of both known pathogenic and predicted variants. There were no coincident positions of both known benign and predicted variants.

2.3.8 General applicability of Paralogue Annotation

From taking all P/LP variants and identifying all possible equivalent positions given by the Ensembl alignments, a total of 178,378 potentially informative annotations can be identified across the 3011 genes. This is an increase from the estimated 150,000 annotations across 1824 genes from the original paper (Ware et al., 2012). Currently, the majority of these annotations are situated in positions where no variant data exists. Another insightful approach is to apply PA to a population dataset. In GnomAD 2.1.1 there are 5,519,773 rare (MAF < 0.001) missense variants out of the reported 17.2 million exome sequenced variants. With conserved AA, 29,123 rare missense are predicted to be pathogenic, of which only 12 are

documented in ClinVar, resulting in a proportion of 0.005 compared to 0.15 and 0.007 for predicted P/LP and B/LB ClinVar variants at the same conservation stringency. Out of the variants that are already reported in ClinVar: 2 are already known as being P/LP; an additional 2 are also reported as B/LB; and the remaining 8 as variants of uncertain significance (VUS). This means that the 29,111 variants that are not described in ClinVar would now have potential evidence to support pathogenicity. Likewise, this would also provide assistance in reclassifying the 8 VUS. In ClinVar, there are 127,080 missense variants classified as VUS. With no conservation stringency, 3,508 of them have at least 1 known P/LP in an equivalent position (**Table 2.5**). This provides evidence in support of pathogenicity for 5% of VUS from disease genes with disease paralogues; 3% of total VUS. At the highest stringency, 185 VUS could be potentially reclassified confidently.

Conservation stringency	Number of VUS
none	3508
conserved ref AA	1871
full conserved ref AA	708
conserved ref AA & conserved alt AA	510
full conserved ref AA & conserved alt AA	185

Table 2.5: Number of ClinVar VUS predicted as being pathogenic by Paralogue Annotation. Shown are missense variants of uncertain significance (VUS) from ClinVar that are predicted to be pathogenic by Paralogue Annotation at varying conservation stringencies.

2.4 Discussion

While traditional homology based methods in bioinformatics, like SIFT, have utilized alignments of conserved sequences before, PA takes a step further by transferring known clinical annotations as well, placing it into a more clinical context than other methods. In its simplest form PA is akin to a search tool like BLAST, providing pathogenicity information about known variants in equivalent positions in related paralogues. This information can then

be used to help predict whether the original query variant is also likely to be pathogenic.

Paralogues are common in most genes irrespective of them being involved in disease or not (**Fig. 2.7**). There is also no obvious enrichment for a particular protein class among paralogues or annotated genes. Though clearly certain classes have more hits than others, this would still demonstrate the wider applicability of PA than previously tested. Therefore outside of current disease genes, if a novel variant is discovered in a gene not known to be diseased, PA could potentially play a role in possible gene discovery, identifying novel disease risk genes, such as if a P/LP variant in *ZNF141* aligned with any novel variants in any one of its paralogues. Current machine learning methods would not have enough training data to provide a reliable prediction for truly novel variants in genes not involved in disease.

In this study, P/LP and B/LB ClinVar variants were used to validate correct predictions and the same P/LP ClinVar variants were used as both the reference set and the test set. One might relate these to typical training and testing data for machine learning algorithms but seeing as PA is not a machine learning based classifier and the reference set is not used for training but instead as a database of known variants to transfer annotations from, this does not infringe on the circularity problems of overlapping training and testing data. Indeed, ClinVar variants were used in this study but if users have their own set of known variants they wished to focus on then those can be used as the reference set instead. However of course the exact performance statistics found in this study would not be guaranteed to apply since that depends on the data being analysed as has been shown (**Fig. 2.8**). For applications in a clinical context however, we would recommend to users to limit reference variants to only those from ClinVar or equivalent databases supported with clinical observations.

With that said, PA in general is shown to have high precision and low sensitivity across different datasets. Even compared to other variant prediction tools, PA is generally more precise but less sensitive. This can be interpreted as though PA would fail to recognise as

many actual pathogenic variants as the other methods might do, the predictions given by PA are more likely to be correct. Together with the clinical context of the reference variants, this is perhaps more important in a clinical application, especially given how established classifiers tend to over predict pathogenicity (Miosge et al., 2015). This can be particularly useful for deciding which variants to focus on for further experimental investigation, such as reclassifying VUS (**Table 2.5**), when resources are limited.

Although the other methods performed with both a high precision and sensitivity on ClinVar variants, suggesting better overall performance, PA maintained a high precision not only on ClinVar variants, buts also independently validated/observed test variants while the other methods performed worse. This is unsurprising as the machine learning based methods would have been likely trained on known variants such as those in ClinVar. As Grimm et al. highlighted, the issue of overlapping training and testing data likely exists here. Some of the variants from ClinVar would've been expected to be used in the training of all the other methods (except Sift) (Grimm et al. (2015);Li et al. (2018)). Using independent test sets that include variants not likely to be present in current databases demonstrate a more realistic representation of performance by the other methods. They therefore do not predict as well on uncommon variants outside of such datasets, while Sift's methodology lacks the additional variant information that PA utilizes. Sift does however manage to perform with comparable levels to the machine learning methods and arguably out performs them in some of the tests (**Fig. 2.8**). This perhaps highlights the simple yet effective nature of using homology and conservation as a framework for predictions.

By increasing how conserved the reference amino acid and substituting variants are, PA's precision and specificity can be typically improved further, while at the cost of even lower sensitivity. In fact, utilizing PA with conserved reference amino acids and fully conserved reference amino acids can be thought of as being analogous to PM5 of the ACMG-AMP general

guidelines and as the same as their modified guidelines for RASopathies as previously mentioned (Richards et al. (2015);Gelb et al. (2018)). Similarly, considering conserved alternate in addition to reference amino acids can be viewed as being equivalent to PS1 of the guidelines. Both of these therefore offer a higher level of evidence for pathogenicity in the ACMG-AMP guidelines compared to relying on multiple computational tools agreeing on prediction outcome, which for some genes can be extremely different (PP3) (Richards et al., 2015).

Although amino acid identities were used when considering conservation stringency, using physicochemical similarities instead would potentially allow a greater number of acceptable changes for conservation, thus increasing precision while not lowering sensitivity as strictly. For example, implementing BLOSUM62 matrix scores as a measure of amino acid similarity.

The trade-off between precision and sensitivity is typical of statistical tests as less stringent parameters typically results in less FN, while also increasing the number of FP. While more stringent parameters decreases FP but allows more FN (Chu, 1999). Here the conservation stringency allows the user to control how sensitive or precise the test can be. If conservation stringency is too high then this may not be practical as sensitivity may be too low resulting in no predictions, or this may even result in reduced precision if the test set contains more benign than pathogenic variants and more TP are removed than FP as stringency increases. This can be seen in the case for *PPARG* (**Fig. 2.8d**), where conserving variants in addition to amino acids becomes too stringent, reducing precision since there were more benign variants than pathogenic variants to begin with. At the highest stringencies, though more FP were removed than TP overall, there remained more FP as too many TP were removed as well. Therefore using conserved reference amino acids is the recommended level of conservation stringency for general usage of PA as a classifier, as that is the most optimal level for high precision yet not lowest sensitivity.

Furthermore, though PA improved most of the ORs, the ORs for *SCN5A* associated LQTS

and *SCN2A* implicated NDD+E did not improve. This is likely due to the fact that variants in *SCN5A* that cause LQTS are gain of function mutations while for BrS they are loss of function (Wilde and Amin, 2018). Gain of function also appears to be the mechanism of mutations observed in *SCN2A* among transgenic mice that experienced seizures (Kearney et al., 2001). Since most mutations are loss-of-function, not many of the annotations transferred would have been appropriate (Griffiths et al., 2005). Therefore ideally the mechanism of disease between paralogues should be assessed on whether they are equivalent before annotation transfer especially when dealing with diseases caused by gain of function mutations. But since this is in the minority, this should not affect PA's wider applicability as much. If different disease mechanisms are known to exist in the paralogues being considered, then one possible alleviation to this would be to utilize all conservation stringency levels before annotation transfer so that the amino acid change is exactly identical across paralogues. This may not solve the issue of different disease mechanisms between different paralogues, but it would likely at least reduce the number of incorrect annotations from such paralogues.

2.4.1 Limitations of Paralogue Annotation

The low sensitivity is undeniably a limitation of PA. This can be explained firstly by not all disease nor non-diseased genes having paralogues, meaning a subset of such genes will never be annotated. Secondly, not all paralogues in a family would be involved in disease necessarily. In other words, many paralogues will not have any known pathogenic variants available as references. Thirdly, even when disease paralogues do exist for a given gene, the alignments may not provide any equivalent positions for annotation transfer. A few P/LP variants would never be able to transfer annotations since they themselves align to gaps in other paralogues - they have no equivalent positions. As a result while PA can be argued to be applicable to multiple

different protein classes of genes, the number of genes and variants PA can ultimately provide predictions for is relatively small.

As mentioned, the paralogue alignments are pre-defined by Ensembl and some of them caused particular issues when transferring annotations across equivalent positions, such as the overlapping genes or read-through transcripts that were self-paralogues (supplementary table A.2). Indeed, as the whole framework of PA depends on how paralogue families are defined and how they align, utilizing a more dynamic measure of relatedness for identifying which positions in sequences are deemed equivalent and a different alignment algorithm for more optimal alignment quality could improve overall predictions. For example, predictions arising from alignments with higher percentage identity would indicate closer relatedness and therefore more reliable predictions compared to alignments with low percentage identity. This might perhaps be useful for paralogues that have arisen through neofunctionalization and therefore may not be as relevant.

As aforementioned, previous homology based methods focus on using orthologues. While utilizing orthologues within this framework may increase sensitivity by increasing the search space for a possible reference variant to align with, the assumption that a pathogenic variant in one species may also cause disease in a different, albeit related, species cannot be adequately justified entirely. Though the genetic code is nearly universal (Koonin and Novozhilov, 2009), meaning that the same variation in two different species will most likely result in the same amino acid change, that does not guarantee disease onset will occur in both as one species may have compensatory mechanisms that the other does not. Without studying the entire biochemical systems of the species being compared, using orthologous sequences within the context of this framework does not seem biologically sensible. Not to mention the lack of variant data among other species besides homo sapiens. Hence, restricting the framework to human paralogues.

Essentially however PA is dependent on and limited by the current extent of available

variant data. A lack of annotation should not be equated to a definitive counter prediction for non-pathogenicity and caution should be used when dealing with such predictions. As more pathogenic variants are discovered and validated over time, this will allow a larger reference set of known variants to be mapped back to alignments and transfer annotations, increasing overall sensitivity and expanding the use of PA.

2.5 Conclusion

To conclude, in this chapter Paralogue Annotation is demonstrated as a high precision pathogenic missense variant prediction framework that is more reliable than current established methods at predicting disease-causing variants. While current methods tend to perform poorly by over-predicting pathogenicity on rare variants or variants that are not well studied, PA does not succumb to such shortcomings. Although it is limited by low sensitivity, which is not necessarily a disadvantage. In a clinical context this is relevant as this predicts variants that are more likely to be truly disease-causing and thus narrow down which ones should be the focus of resource intensive studies. Since 72.0% of genes have paralogues and PA is shown to be applicable to multiple protein classes, the more novel variants that are being discovered the greater the sensitivity will become with time.

In the following chapters, I will examine ways of improving the framework of Paralogue Annotation mainly in regards to the limited sensitivity in predictions while attempting to maintain the overall concept of how the framework functions. Namely the identification of functionally equivalent amino acids and how they can be used to aid pathogenic variant prediction.

Chapter 3

Protein domain Annotation

3.1 Introduction

While Paralogue Annotation has proven to be successful in being able to predict pathogenic variants with high precision, one of the main limitations that was established was its limited sensitivity. In practical terms, this meant that while any positive pathogenic prediction given by Paralogue Annotation could be reliably trusted, many pathogenic variants escaped detection all together. Typically a trade-off between precision and sensitivity exists in many classification algorithms and increasing the performance of both statistics simultaneously in real world scenarios is incredibly challenging. Finding the optimal parameters to improve performance in one statistic without detrimentally affecting the other is one of the main goals in classification.

As discussed in the previous chapter, one way of improving sensitivity is by increasing the volume of reference data that Paralogue Annotation can use to transfer annotations from. This will naturally occur over time as more and more variants are identified and documented by the scientific community. Though the speed at which this occurs is not within the scope of this thesis.

An alternative strategy is to improve the method of the framework itself. More specifically, by altering the alignment of protein sequences to identify a greater number of functionally equivalent positions across related protein families. There are two approaches to this option. One is change alignment algorithms - indeed there are many algorithms that utilise different techniques and perhaps other algorithms are better suited for this task than the one being used. The second option is to change the sequences being aligned. I have opted for the latter approach. Orthologues were introduced in the previous chapters and though their suggested usage in the Paralogue Annotation framework would theoretically improve sensitivity (by increasing the number of possible aligned positions for equivalent variants to be searched in), the same variant in one species cannot be guaranteed to cause disease in a different species without extensive studying. Combined with the inherent lack of up-to-date variant data among other species, orthologues were not determined as an adequate alternative to paralogues. Indeed paralogues and orthologues are the conventional examples of evolutionary relationship, they are however not the only instance of homology present in the genome. Another source can be found in protein domains, which sequences can be aligned instead of paralogues.

3.1.1 Protein domains

Protein domains are conserved regions within proteins that evolve and fold independently to the rest of the protein. Different proteins may consist of multiple different combinations of these domains, with multiple proteins often sharing the same domain. For example, while different transcription factor proteins may be comprised of different activating or regulating domains that allow the dynamic regulation of genes, they can often share the same DNA binding domain.

This modularity of different proteins encompassing the same domain allows the definition of equivalency to also be utilized here. As a result, instead of identifying functionally equivalent variants across paralogues, we can attempt to also identify variants across related protein

domains. As shown in Fig. 2.10, most of the pathogenic variants distributed along KCNQ1 are indeed enriched around domains while depleted outside of them. Variants located in domains of functional and/or structural importance are more likely to be deleterious (Richards et al., 2015). Therefore proteins that share the same functional or structural domains would likely be equally as affected if an equivalent variant were to be present.

By focusing on domains, regions not as consequential to protein function or structure, such as random loops, can be ignored. In fact while the alignments used in Paralogue Annotation encompass the whole protein, these global alignments can lead to many gaps in the alignment, with some aligned positions that may otherwise not occur if a more local alignment were to be used instead. Aligning protein domains only will therefore provide more of these localised alignments by disregarding non-domain regions of the protein. As a result, within these domain alignments there will be more equivalent positions that can be identified as there would be less gaps.

3.1.2 Pfam and Pfam metadomains

The Pfam database consists of protein domain families represented as MSAs generated by hidden Markov models. Fundamentally, Pfam has annotated and classified protein domains from multiple organisms into functional groups allowing evolutionarily-related domains to be identified. As of Pfam version 33.1, the database contains 18,259 different families (Mistry et al., 2021). About 90% of the human protein coding genes classified in Ensembl contain at least one Pfam domain, with 51% of the total proteomic sequence covered by a Pfam domain. The average protein coding gene has 2 Pfam domains.

These Pfam domains can therefore be used as the basis for implementing protein domains into the Paralogue Annotation framework. Initially however, the protein domains from non-human genomes must first be taken account of as their inclusion to the original Pfam MSAs

mean that the alignments are less specific to human sequences. As a result, there would be more gaps and less optimally aligned positions for human variants to encounter. This would mean less functionally equivalent variants can be annotated.

Around the same time of this PhD Wiel et al. (2017) created Pfam metadomain alignments. These metadomains consists of alignments of only homologous human protein domains. In their work they explored the same hypothesis as Paralogue Annotation and mapped population and known pathogenic variants onto these metadomains. They determined that approximately 72% of pathogenic variants from ClinVar occur within a Pfam domain (Wiel et al., 2017). They showed that the metadomains could be used to identify homologous variants across Pfam domains that can help improve variant interpretation similar to what Paralogue Annotation can achieve. However, their work did not go as far as what this PhD has aimed to do, and that is to quantitatively and statistically define how informative functionally equivalent variants are in regards to predicting pathogenicity.

With their permission, I have used their Pfam metadomain alignments within the Paralogue Annotation framework, in the stead of Ensembl's paralogue alignments. This allows the modelling and identification of functionally equivalent variants across protein domains as oppose to paralogues.

3.2 Materials and Methods

As most of the data and methods of analysis used for this chapter are identical to the previous one, only materials and methods that are different will be mentioned.

3.2.1 Using Pfam metadomain alignments instead of Ensembl Paralogue alignments

To improve the sensitivity of the framework, alignment of Pfam domains were used instead of paralogues (**Fig. 3.1**). While generating these alignments personally was possible, with permissions from the author, I utilized the Pfam metadomain alignments generated by Wiel et al. (2017) who had already used their alignments for a similar analysis. Their analysis however did not focus on performing the statistical and quantitative analysis that this PhD has done in order to determine the predictive performance of using such metadomains for clinical variant interpretation.

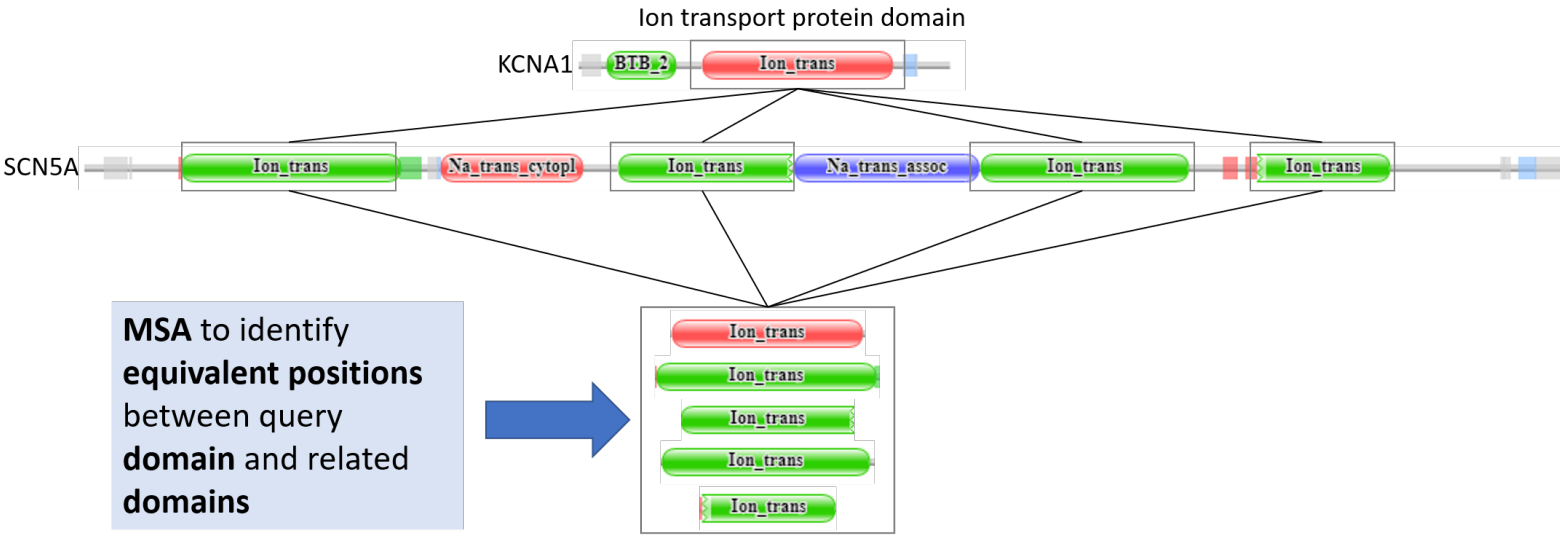


Figure 3.1: Schematic concept of using homologous human protein domain alignments instead of paralogue alignments. Here both KCNA1 and SCN5A contain an Ion transport protein domain (Ion_trans) as defined by Pfam, with the latter containing multiple. These domain sequences can be aligned in a Multiple Sequence Alignment (MSA) similar to paralogue alignments and used in their stead. Protein sequence cartoons extracted from Pfam at pfam.xfam.org (Mistry et al., 2021).

With Paralogue Annotation, the paralogue alignments were leveraged from Ensembl by the usage of VEP and a custom pluggin script. The Pfam metadomains however could not have been used in the same manner and the alignments were instead in a genomic coordinate tabulated format. A new workflow was built to handle this change in the framework. The tabulated genome coordinates were first processed in Python (v3.7.3) to generate two dictionary objects

that mapped the alignments. The first dictionary tracked which Pfam domain and position that a given genomic position was mapped to. The second dictionary tracked the reverse mapping of what other genomic position lies within the same Pfam position in a given domain. Any genomic coordinate can then be queried against these dictionaries to identify any equivalent positions within homologous domains. These equivalent positions are then parsed over and processed in R (v3.6.3), where known reference variants are used to map back to the equivalent positions.

During testing and analysis the Python/R pipeline used a lot of memory due to the nature of the programming languages used and in order to not bottleneck workflow or run out of memory during computing, variant input files were typically split multiple times into smaller files and each were parallel processed simultaneously on separate CPUs. Processed output files were then joined back together at the end to accumulate a coherent result.

However, for extremely large input files the splitting and joining back together of data began to not be practical also and so an alternative mirrored pipeline was constructed. This pipeline was written in C++ with the equivalent R portion being handled by a SQL Database API (SQLite3 v3.30.1) within the C++ program. As a result, mapping reference variants and storing output results could all be done together within a SQL database without the need for splitting any files. Due to the nature of SQL, smaller files that could be split and managed practically would be processed faster in the former pipeline. Results would be gather and read in R using RSQlite (v2.2.8) for statistical analysis.

3.2.2 Surveying number of Pfam protein domains in genes and comparison to Ensembl Paralogues

Statistics and details on Pfam domains and how many are contained in each human protein coding gene were extracted directly from EBI's public FTP database (available at

fhttp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.0/swisspfam.gz). This data was compared to Ensembl's Biomart (Ensembl version 93: Jul2018) data for paralogues in R and visualised using ggplot2.

Gene names were stratified into different groups depending on whether they encompassed a Pfam domain; have any related paralogues; have both Pfam domains and paralogues; or neither. These four groups were then put through Gene Ontology analysis to assess whether there are any significant differences between genes that have Pfam domains or paralogues.

3.2.3 Chi-square test of independence for GO protein classes Gene Ontology pathway analyis

A Chi-square test of independence was conducted in R to better highlight statistically significant differences between PANTHER Protein Classes among the four gene groups mentioned in subsection 3.2.2. The Chi-square statistic is calculated as follows:

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

where o and e are the observed and expected number of gene hits per protein class and gene group in each cell of the contingency table respectively. e for each cell can be calculated by:

$$e = \frac{Mrow \times Mcol}{N}$$

where $Mrow$ and $Mcol$ are the row and column totals for the respective cell, while N is total number of gene hits of all protein classes (McHugh, 2013).

Pearson residuals from the test were extracted and plotted visually using the corrplot package in R to identify which protein class in which gene group was the most influential.

In addition to protein classes, PANTHER (v16.0) Pathways were also analysed. In order to

best visualise the results, cutoff thresholds were used to disregard pathways that were not well represented within the dataset. For the comparison between Pfam domains and paralogues, a pathway hit must have represented at least 1.5% of the total number of pathways present. For the comparison between genes annotated using the Pfam metadomain alignments compared to ones annotated with Paralogue Annotation, a pathway hit cutoff of 1.3% was used.

3.2.4 Adjusted test set for appropriate comparison to Paralogue Annotation

During analyses of Paralogue Annotation results, the variants that were used as the test set were defined as being located in only genes that had paralogous relationships - any variants that existed in non-paralogous genes were removed. If this set were also to be used as the test set for Pfam domains, then there would be a bias of results favouring paralogues. Therefore, a new larger test set that also incorporated variants from non-paralogous genes was generated and used instead to compare the performance between paralogues and Pfam domains.

Furthermore, another reason for the different test set was due to the coordinates for the metadomain alignments being only available for genome build GRCh37, whereas the initial test set for Paralogue Annotation was GRCh38. Hence why there are slight discrepancies in total variant number between the two test sets. Fortunately, Paralogue Annotation can perform on both set of coordinates but given that not all variants have coordinates in both GRCh37 and GRCh38, the impact on performance does vary slightly but not significantly.

This new test set encompasses 30,636 P/LP and 16,724 B/LB variants across 3524 genes, 997 of which do not have paralogues. This set was then tested on both paralogues and pfam domains.

3.3 Results

3.3.1 Number of genes containing Pfam domains surpass the number of genes with related paralogues

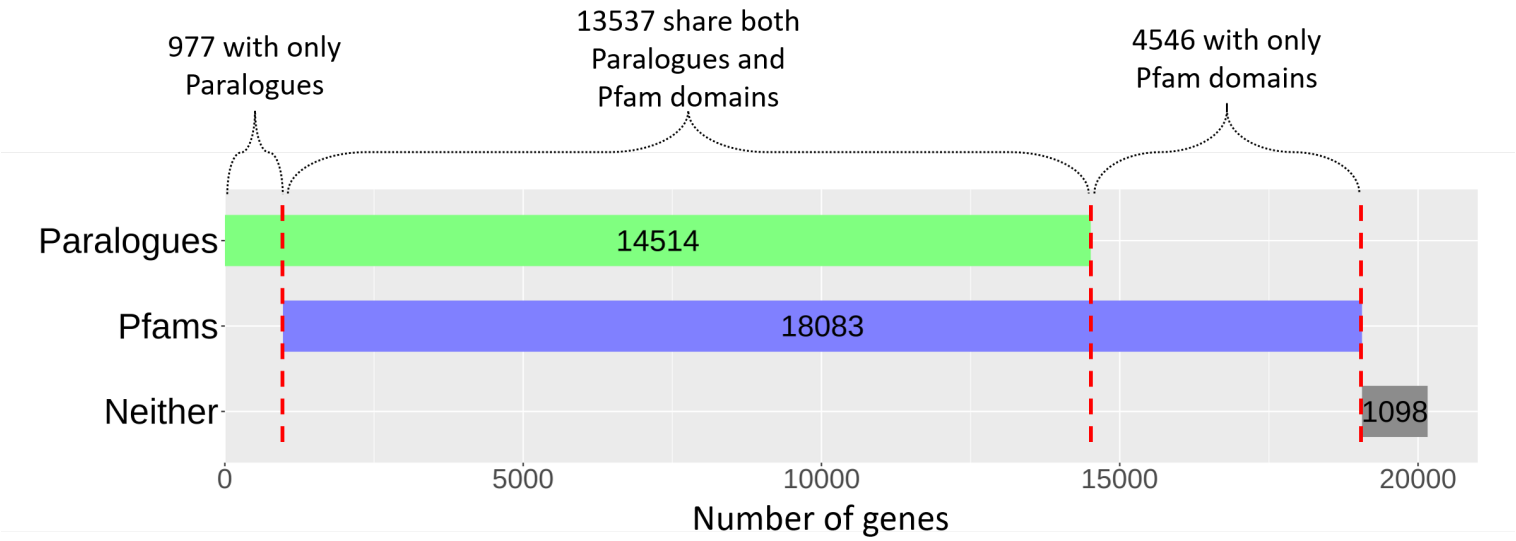


Figure 3.2: More genes incorporate a Pfam domain than have a related paralogue. Plot showing the number of genes that have at least one paralogue (green) and/or incorporate at least one Pfam domain (blue). Genes that have neither are shown in grey.

While 14514 protein coding genes contain a paralogue (**Fig 3.2**), 18083 genes incorporate at least one Pfam domain, with an overlap of 13537 genes sharing both. More importantly, 4546 Pfam containing genes do not have related paralogues. These set of genes would not have been identified by Paralogue Annotation and ultimately overlooked variants that are located in these genes contribute to the limited sensitivity. By using Pfam domains instead of paralogues, the additional variants will theoretically provide more possible aligned positions for the identification of functionally equivalent variants. 977 do not contain any pfam domains but do have paralogous relationships. These would be annotated by Paralogue Annotation only and would not be identified when using pfam domains in the framework. Another subset of genes, 1098, do not contain either pfam domains or have any related paralogues.

3.3.2 Gene Ontology analysis of PANTHER protein class show differences between genes with Pfam domains and Paralogues

Performing a Gene Ontology (GO) analysis on the four different groups of genes can identify if there is any enrichment for particular types of genes within their groups. That is to say (1) genes with only paralogous relationships and no pfam domains; (2) genes that both have paralogues and contain pfam domains; (3) genes that only contain pfam domains and have no paralogues; (4) and genes that have neither paralogues nor pfam domains. GO can therefore elucidate if there is any segregation aspect brought about by genes having related paralogues, containing pfam domains, or both.

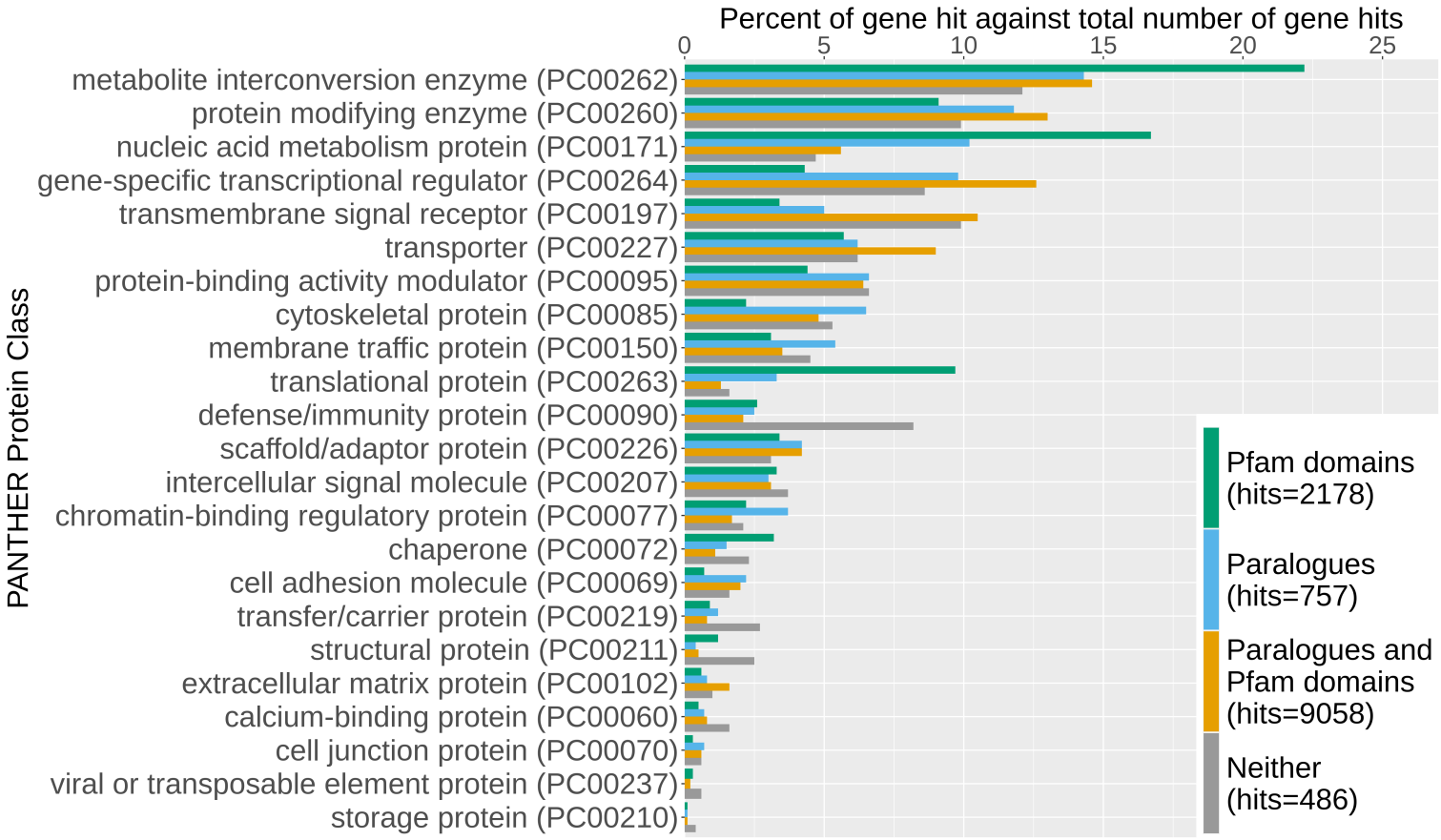


Figure 3.3: PANTHER protein class hits (genes recognised by PANTHER) vary among genes that only contain Pfam domains (green), only have paralogues (blue), both contain pfams domains and have paralogues (yellow), or neither (grey). Along the y-axis are top PANTHER protein class hits for each of the four groups and along the x-axis is the proportional percentage of gene hits against the total number of gene his for each group.

Fig. 3.3 shows a protein class GO analysis using the PANTHER classification system for the four groups of genes. For paralogue only genes, 77% (757/977) of genes found hits. For genes with only Pfam domains 48% (2178/4546) of genes found hits. Whereas for genes that have both paralogues and Pfam domains and genes that had neither, 67% (9058/13537) and 44% (486/1098) of genes found hits respectively.

Generally, for all four gene groups the top protein class was metabolite interconversion enzymes, which occupied: 22.2% of all hits for genes with only Pfam domains; 14.3% for genes with only related paralogues; 14.6% for genes that share both pfam domains and paralogues; and 12.1% for genes that have neither pfam domains nor paralogues. This protein class includes all enzymes involved in converting one molecule into another except those that act on DNA, RNA, or protein. Enzymes such as hydrolases, isomerases, ligases, and more, fall into this class.

After that, genes with only Pfam domains start diverging with the other three gene groups having a second largest protein class of protein modifying enzymes (genes with both pfam domains and paralogues (13%); genes with only paralogues (11.8%); genes with neither (9.9%)). This class includes any proteins that covalently modifies other proteins, such as kinases and proteases. Instead, the second largest protein class for Pfam only genes were nucleic acid metabolism proteins at 16.7%. This was also the third largest protein class for paralogue only genes at 10.2%.

Protein classes, such as cell junction, viral/transposable element, or storage proteins saw very few gene hits irrespective of the group. All had less than 0.8% of total gene hits from the respective groups.

Conducting a Chi-square test of independence on the four gene groups and their associated protein classes allows us to determine if genes that have pfam domains and/or related paralogues are more likely to belong to any particular protein class. The resulting p-value of 4.1×10^{-249} - far less than the standard significance level of 0.05 - allows us to reject the null hypothesis



Figure 3.4: Correlation plot of the Pearson residuals from a Chi-square test of independence show that there are distinct protein classes among the four groups of genes particularly genes with only Pfam domains. Shown are PANTHER protein class definitions against the four associated gene groups - 1) genes that only contain Pfam domains; 2) genes with only have paralogues; 3) genes that contain both pfams domains and paralogues; or 4) neither. The range of residuals are displayed on the scale on the right.

that the is no association between the four groups the genes are in and the number of genes in each of the protein classes. Furthermore, a correlation plot of the residuals (**Fig. 3.4**) can help show which gene group is the more enriched or depleted than expected in each protein class.

Each of the four gene groups appears to have unique profiles of protein class enrichment/depletion. But of particular note for certain protein classes, genes with only Pfam

domains seem to have a reciprocal relationship with genes that have both Pfam domains and paralogues. For example where there is enrichment, that is to say a positive residual, in translational proteins (18.5) and nucleic acid metabolism proteins (14.9) in genes with only Pfam domains, there is depletion, or a negative residual, for such classes in genes that have both Pfam domains and paralogues; -8.9 and -7.5 respectively. This effect is also seen to a lesser extent for transmembrane signal receptors and gene-specific transcriptional regulator. The effect however, is reversed for transmembrane signal receptors and gene-specific transcriptional regulator with Pfam only genes having residuals of -8.6 and -9.3, compared to genes with both Pfam domains and paralogues with residuals of 5.1 for both classes.

The relationship between Pfam domain only and paralogue only genes do not appear to be as antagonistic, with the majority of paralogue only genes being close to expected values and only a few with mild fluctuations. Interestingly, genes classified as defense/immunity protein classes were more abundant than expected in genes that had neither any Pfam domains nor related paralogues, with a residual of 8.3.

The protein classes shown so far are only the overall parent level of the PANTHER Protein Class classification lineage. By looking deeper into each of the respective classes can provide insight into the potential reasons for the variations in enrichment and depletion. For example, **Fig. 3.5** shows the sublevel children protein classes within the translational protein (PC00263) class for genes that only have Pfam domains and genes that have both Pfam domains and Paralogues.

For genes with only Pfam domains, ribosomal proteins are particularly enriched with 145 hits compared to only 30 for genes with both Paralogues and Pfam domains. Aminoacyl-tRNA synthetase was approximately the same for both groups at 22 and 19 hits respecively, while translation factors were moderately lower at 44 hits for Pfam domain only genes compared to 69 hits for genes with both Paralogues and Pfam domains.

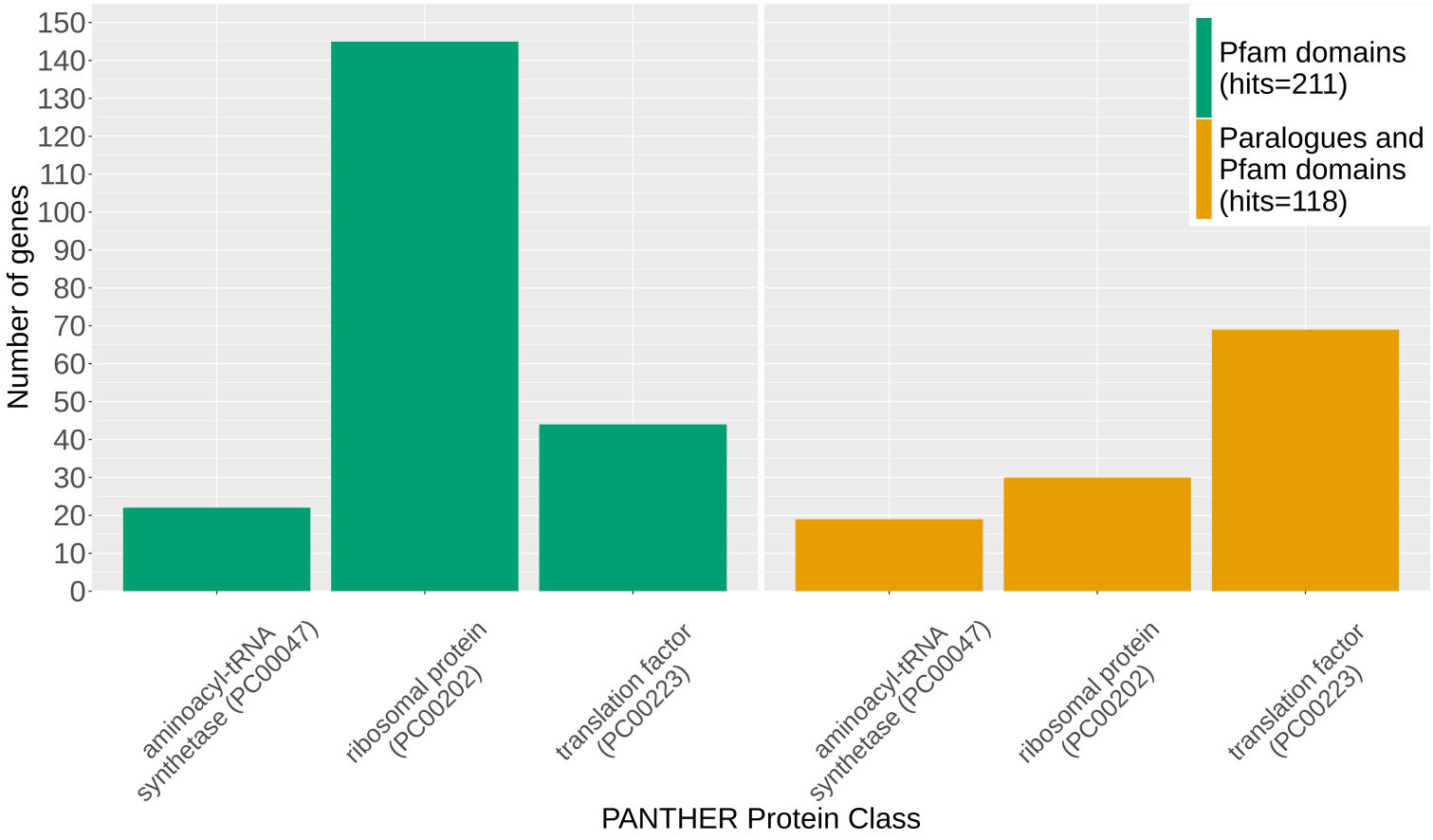


Figure 3.5: Sublevel children PANTHER protein class hits within translational protein (PC00263) vary among genes that only contain Pfam domains (green) and both contain pfams domains and have paralogues (yellow). Along the x-axis are the three children PANTHER protein classes within the translational protein (PC00263) parent class: aminoacyl-tRNA synthetase (PC00047); ribosomal protein (PC00202); and translation factor (PC00223). Along the y-axis are the number of gene hits against the total number of gene hits within the translational protein (PC00263) sublevel for these two groups of genes.

3.3.3 Gene Ontology analysis of PANTHER pathways show differences between genes with Pfam domains and Paralogues

Apart from protein class, a GO analysis of PANTHER pathways can also be conducted. **Fig. 3.6** shows the top pathways (pathways that contain at least 1.5% of total genes hits in either group) that genes containing only Pfam domains are involved in compared to genes with only paralogues. The distribution of hits is spread somewhat evenly, with roughly half of the pathways - 16 - having more paralogue only genes than Pfam domains. 15 pathways had the

opposite effect, with three identifying no hits for paralogue only genes at all. Heme biosynthesis (P02746), Cholesterol biosynthesis (P00014), and the general transcription by RNA polymerase I (P00022) pathways only found hits with genes that only contain Pfam domains. Apart from these three, the biggest difference of hits was within the Transcription regulation by bZIP transcription factor (P00055) pathway with Pfam only genes having 25 hits to 5 for paralogue only genes. The T cell activation pathway (P00053) had an equal number of hits for both groups.

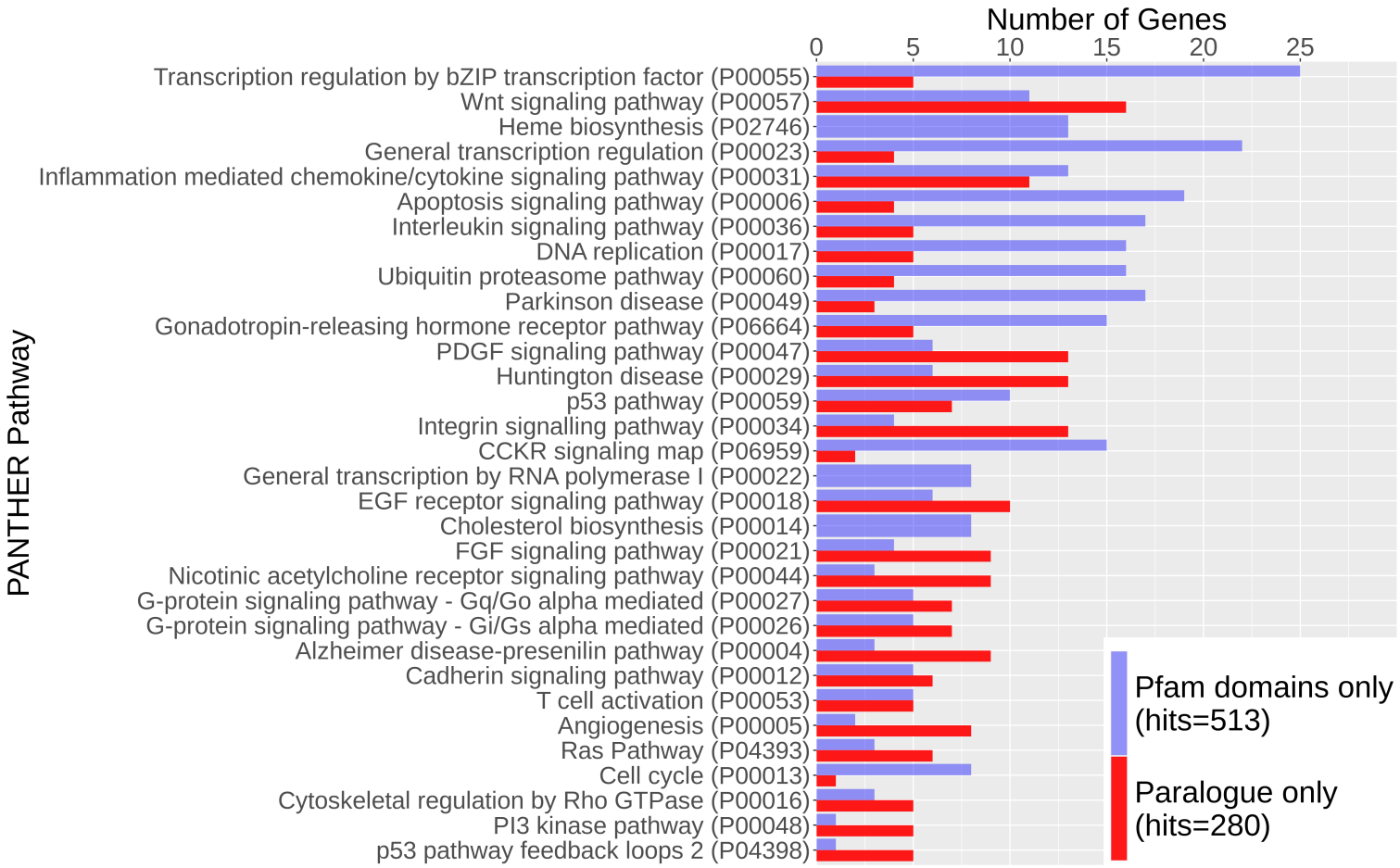


Figure 3.6: PANTHER pathway hits vary among genes that contain only Pfam domains (blue) or only have paralogues (red). Along the y-axis are top PANTHER pathways (pathway must contain at least 1.5% of total genes hits in either Pfam domain only or paralogue only genes) and along the x-axis is the number of gene hits against the total number of gene his for each group.

3.3.4 Pfam Annotation of ClinVar

Following Paralogue Annotation, I applied the framework, now using alignments of Pfam metadomains, to ClinVar. Using the same workflow pipeline, variants that lie outside of domains or within domains that did not align to any other related domain were initially removed. Out of the remaining 18,979 P/LP and 5909 B/LB variants, 7543 P/LP were correctly annotated as pathogenic while 928 B/LB were determined to be false positives (**Table 3.1**). This generated a precision of 0.89 and a sensitivity of 0.40. Similar to Paralogue Annotation, when increasing conservation stringency to remove variants that do not have conserved reference amino acids, this reduces the overall number of annotations to 5,867 P/LP and 281 B/LB. Just as before, the effect is an increase in precision to 0.95 and a decrease in sensitivity to 0.31. The precision and sensitivity among the variants removed in this filtering step was 0.72 and 0.088 respectively.

	Pathogenic variants	Benign variants	PPV	Sensitivity	Specificity	P value
Total variants	18979	5909	-	-	-	-
Pfam Annotation	7543	928	0.89	0.40	0.84	<5.0e-324
Pfam Annotation with conserved ref AA	5867	281	0.95	0.31	0.95	<5.0e-324
non-conserved ref AA variants removed	1676	647	0.72	0.088	0.89	1.5e-06

Table 3.1: Shown are number of ClinVar variants with a clinical significance of Pathogenic/Likely pathogenic (P/LP Variants) and Benign/Likely benign (B/LB Variants) that are predicted to be pathogenic using Pfam metadomain alignments instead of paralogue alignments with the Paralogue Annotation framework and the remaining number of variants after filtering for equivalent variants with conserved reference amino acids (ref AA). Shown along the bottom row are the number of variants removed by the filtering. The Precision (Positive Predictive Value - PPV), Sensitivity, Specificity and Fisher's Exact test p-value are shown to the nearest 2 significant figures where applicable. Note: machine performing calculations cannot compute values smaller than 5e-324.

3.3.5 Pfam domains provides substantially more positive predictions than Paralogue Annotation with marginal decrease in Precision on standardised tests

For a fair comparison between the performances between Pfam domains and paralogues a different testset with a standardised set of variants and a shared genome coordinate system was used. Using this new testset, I can verify that by using the metadomains in our classification framework the number of variants that can be annotated increases from 4328 to 7542 for P/LP variants, while the number of B/LB variants incorrectly annotated also increase from 245 to 928 (**Fig 3.7**). This does lead to a mild decrease of 5.91% in precision (0.95 to 0.89) (**Table 3.2**) but crucially sensitivity increases 74.28% (0.14 to 0.25) (**Table 3.3**). Precision can again be rescued by only considering variants at positions with conserved amino acids. This minimises the precision decrease to just 2.06% while maintaining a high sensitivity increase of 51.06%.

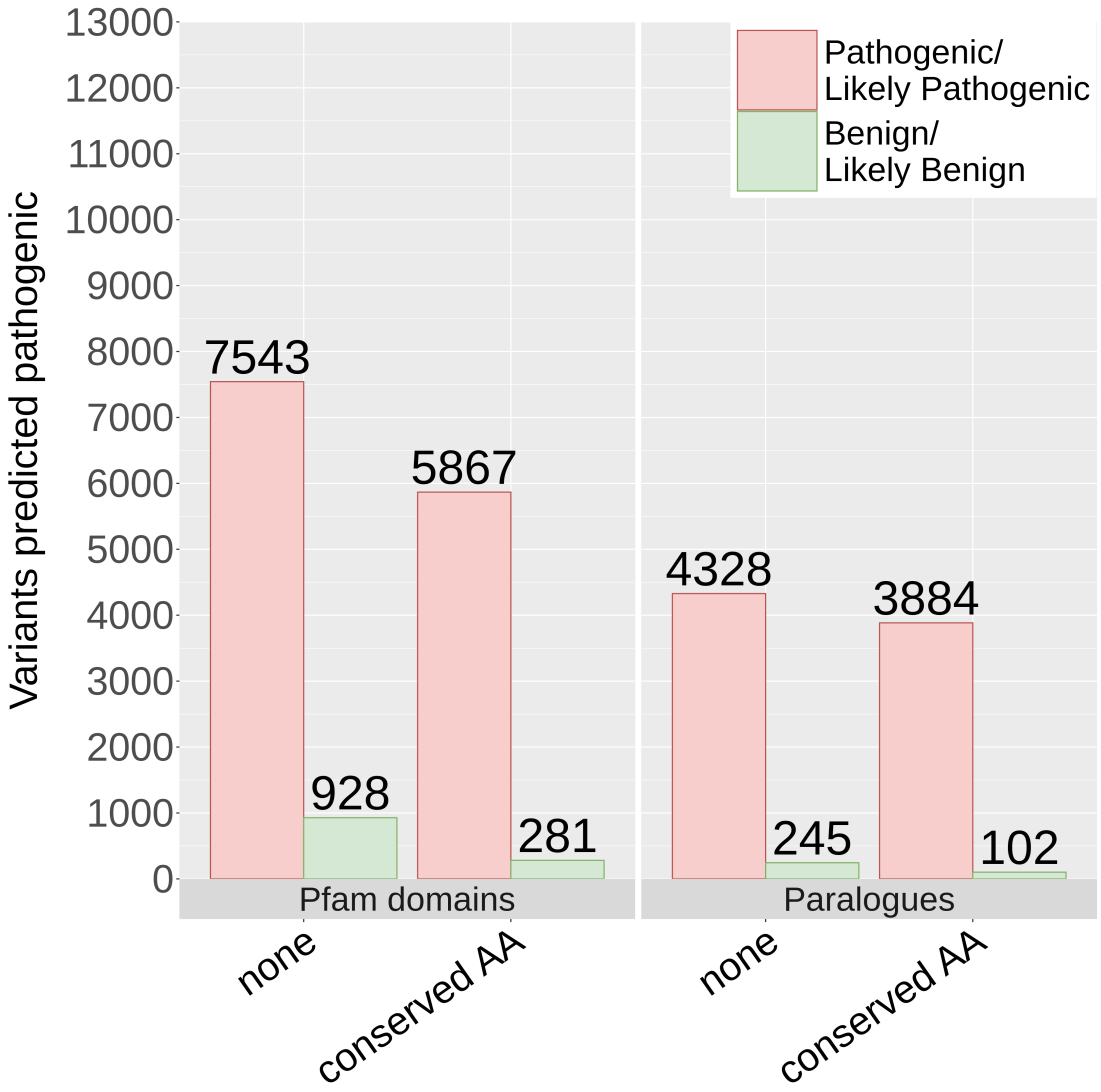


Figure 3.7: Pfam domains provide more pathogenic predictions for variants overall than Paralogue Annotation. Shown on the left are the number of Pathogenic/Likely Pathogenic (red) and Benign/Likely Benign (green) variants that are predicted as being pathogenic when using Pfam metadomain alignments within the Paralogue Annotation framework instead of paralogues. On the right are the respective number of predictions for Paralogue Annotation when using the same standardised testset and genome coordinate system. Secondary bars for when predictions are filtered for variants that have conserved reference amino acids are also shown.

	PPV	PPV % change	PPV with conserved ref AA	PPV with conserved ref AA % change
Pfam	0.89	-5.91%	0.95	-2.06%
Paralogues	0.95		0.97	

Table 3.2: Precision (Positive Predictive Value - PPV) decreases marginally when using Pfam metadomain alignments instead of paralogue alignments within the Paralogue Annotation framework. Shown are the precision of Pfam domains compared to paralogues and their percentage change with and without filtering for conserved reference amino acids (ref AA).

	Sensitivity	Sensitivity % change	Sensitivity with conserved ref AA	Sensitivity with conserved ref AA % change
Pfam	0.25	+74.28%	0.19	+51.06%
Paralogues	0.14		0.13	

Table 3.3: Sensitivity increases substantially when using Pfam metadomain alignments instead of paralogue alignments within the Paralogue Annotation framework. Shown are the sensitivity of Pfam domains compared to paralogues and their percentage change with and without filtering for conserved reference amino acids (ref AA).

3.3.6 Gene Ontology analysis of protein classes and pathways show no significant differences between genes that are annotated by Pfam domain and Paralogue Annotation

Taking a look at the genes that contained variants annotated can provide insight into any crucial differences in genes that are covered by Paralogue Annotation and not by Pfam metadomains, or vice versa. Shown in **Fig. 3.8** are the different protein classes present in genes that are annotated using both paralogues and Pfam metadomains. Out of all the PANTHER protein classes that found at least one hit, the majority of protein classes are relatively similar in the number of genes for both paralogues and Pfam metadomains, but with Pfam metadomains finding more hits than paralogues universally. No genes annotated by Paralogue Annotation had more hits found than their Pfam metadomain counterpart. The biggest difference being protein modifying enzymes where genes with Pfam metadomain annotations found 142 hits (17.6% of the total 808 hits) compared to the 99 hits (19.5% of the total 508) for genes with paralogue annotations.

Paralogue Annotation did not annotate any genes belonging to four protein classes. Translational proteins, defense/immunity proteins, chaperones, and storage proteins all only had genes that could be annotated with Pfam metadomains. Albeit the number of genes in these classes were very low - 6, 4, 4, and 1 respectively.

Among PANTHER pathways between genes that were annotated with Paralogue Annotation and those annotated with Pfam metadomains (**Fig. 3.9**), both returned similar distributions of pathway hits as well, with Pfam metadomains generally identifying more hits than paralogues in each pathway except for the signalling pathways of Integrin, TGF-beta, and Interleukin. In these three pathways, there were more hits for genes annotated by Paralogue Annotation.

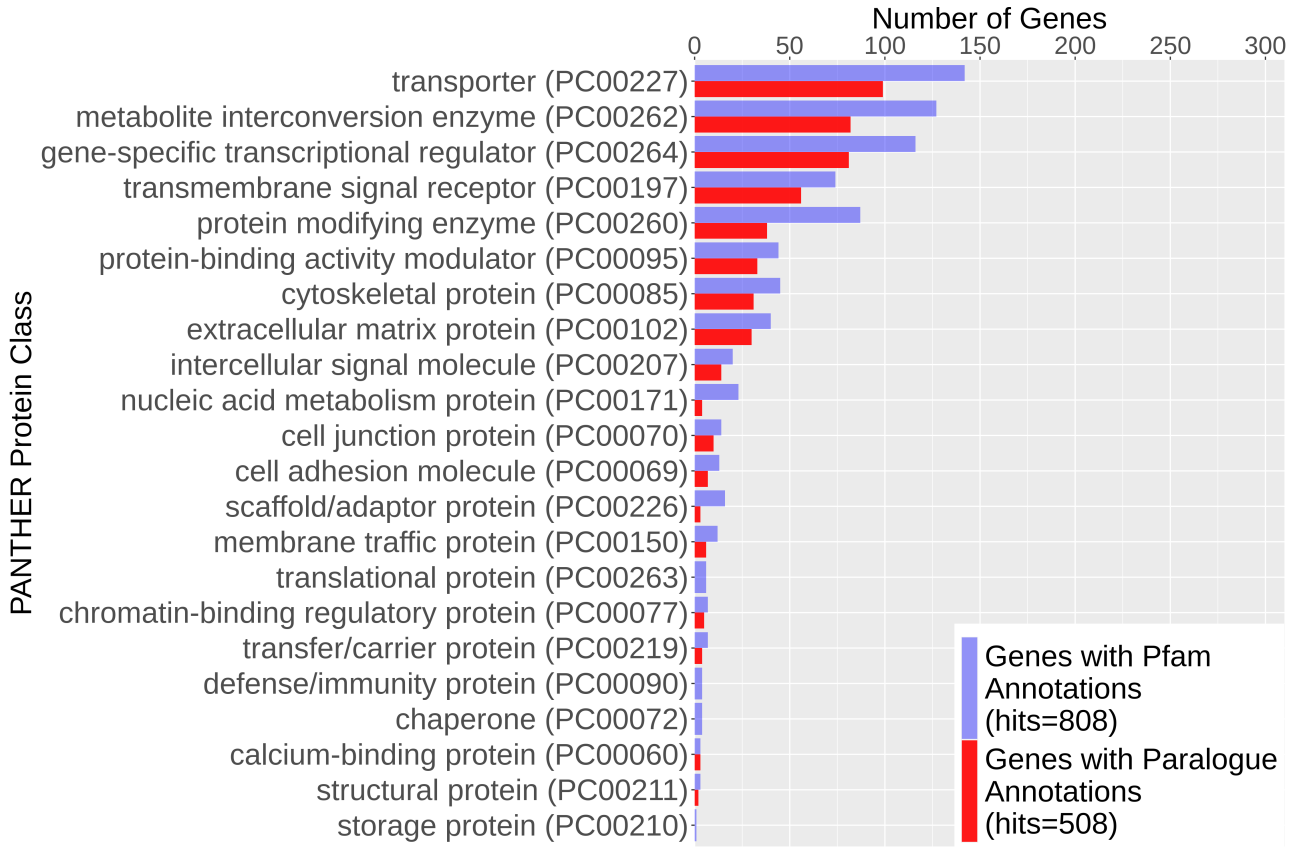


Figure 3.8: PANTHER protein class hits show little variation between genes annotated using Pfam metadomain alignments (blue) and Ensembl paralogue alignments (red). Along the top are the number of genes per protein class. PANTHER protein classes are shown along the left.

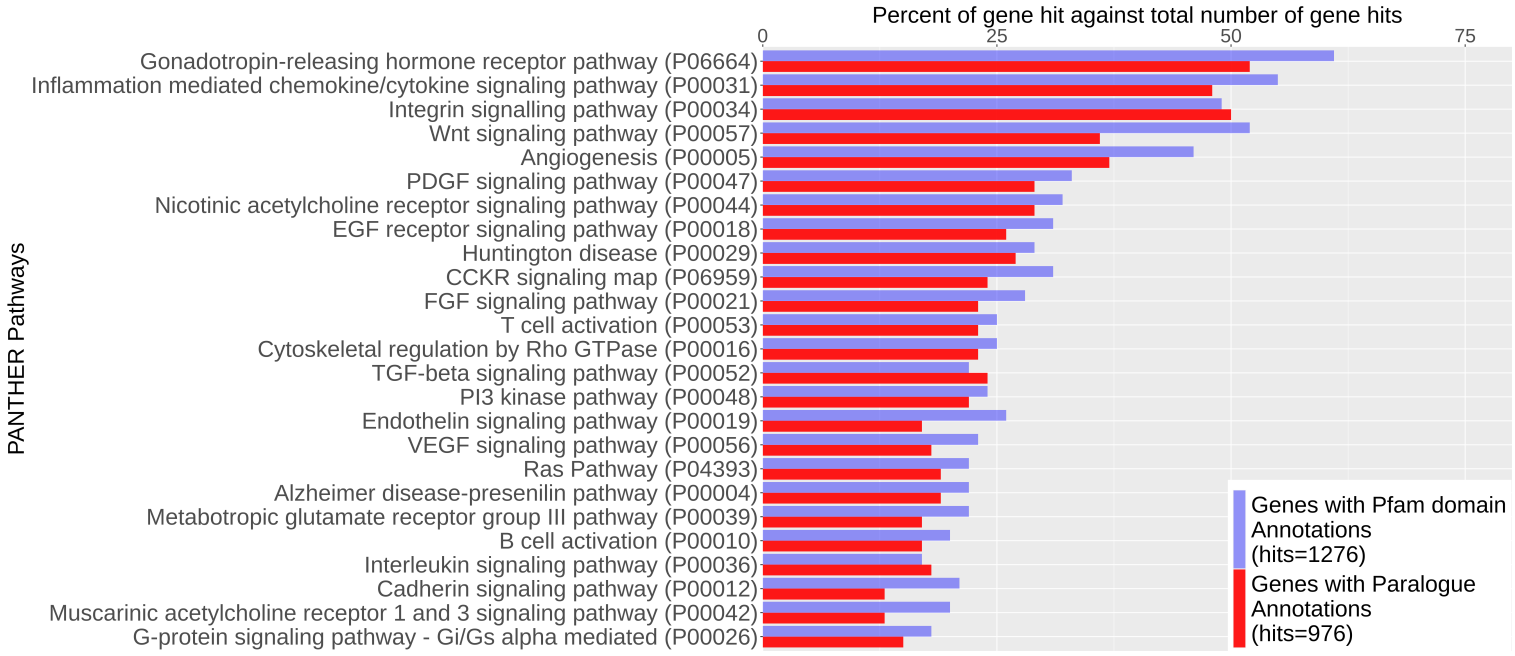


Figure 3.9: PANTHER pathway hits show little variation between genes annotated using Pfam metadomain alignments (blue) and Ensembl paralogue alignments (red). Along the top are the proportional percentages of gene hits against the total number of gene hits for both gene sets per pathway. PANTHER pathways are shown along the left with only pathways that had at least 1.3% of total gene hits in either gene sets being present.

3.3.7 Pfam metadomains have variable improvements in sensitivity on independent test sets compared to Paralogue Annotation

Just as Paralogue Annotation was tested against established methods for predicting pathogenicity, the Pfam metadomains were also tested in the same manner. Using the same independent functionally assayed test sets from Mahmood et al. (2017) and Li et al. (2018), Pfam metadomains had somewhat of a varied performance, though all still had lower sensitivities when compared to the established methods.

For *TP53*, Pfam metadomains performed equally to Paralogue Annotation (**Fig. 3.10a**) with near identical statistics - outperforming the other established methods in precision but not in sensitivity. Precision and sensitivity is slightly higher when no conservation is considered for Pfam metadomains (0.90 and 0.16) compared to Paralogues (0.88 and 0.15). Conserving for reference amino acids increases precision to 0.95 and decreases sensitivity to 0.14, comparable to 0.95 and 0.13 for Paralogue Annotation previously.

For *PPARG*, precision decreased slightly while sensitivity more than doubled, though this is still less than the sensitivity of any of the established methods (**Fig. 3.10b**). With no conservation stringency, precision decreased from 0.15 for paralogues to 0.13 for pfam metadomains, while sensitivity increased from 0.25 to 0.60. The established methods all had a sensitivity of at least 0.96 or more. With conserved reference amino acids, the precision of the Pfam metadomains was 0.24, less than the 0.30 of Paralogue Annotation. While sensitivity was at 0.39 compared to 0.14.

For genes harbouring de novo variants in people with ASD, precision decreased while sensitivity increased both with moderate amounts (**Fig. 3.10c**). While Paralogue Annotation had a precision and sensitivity of 0.73 and 0.02 with no conservation stringency, Pfam metadomains had 0.67 and 0.18 respectively. The precision is similar to the established methods, the closest being Revel with 0.65, but sensitivity is still lagging behind. Revel again

being the closest with a sensitivity of 0.26. Interestingly, when conserving for reference amino acids, precision does not increase but rather decreases minimally for Pfam metadomains to 0.66, while sensitivity decreases as usual to 0.09.

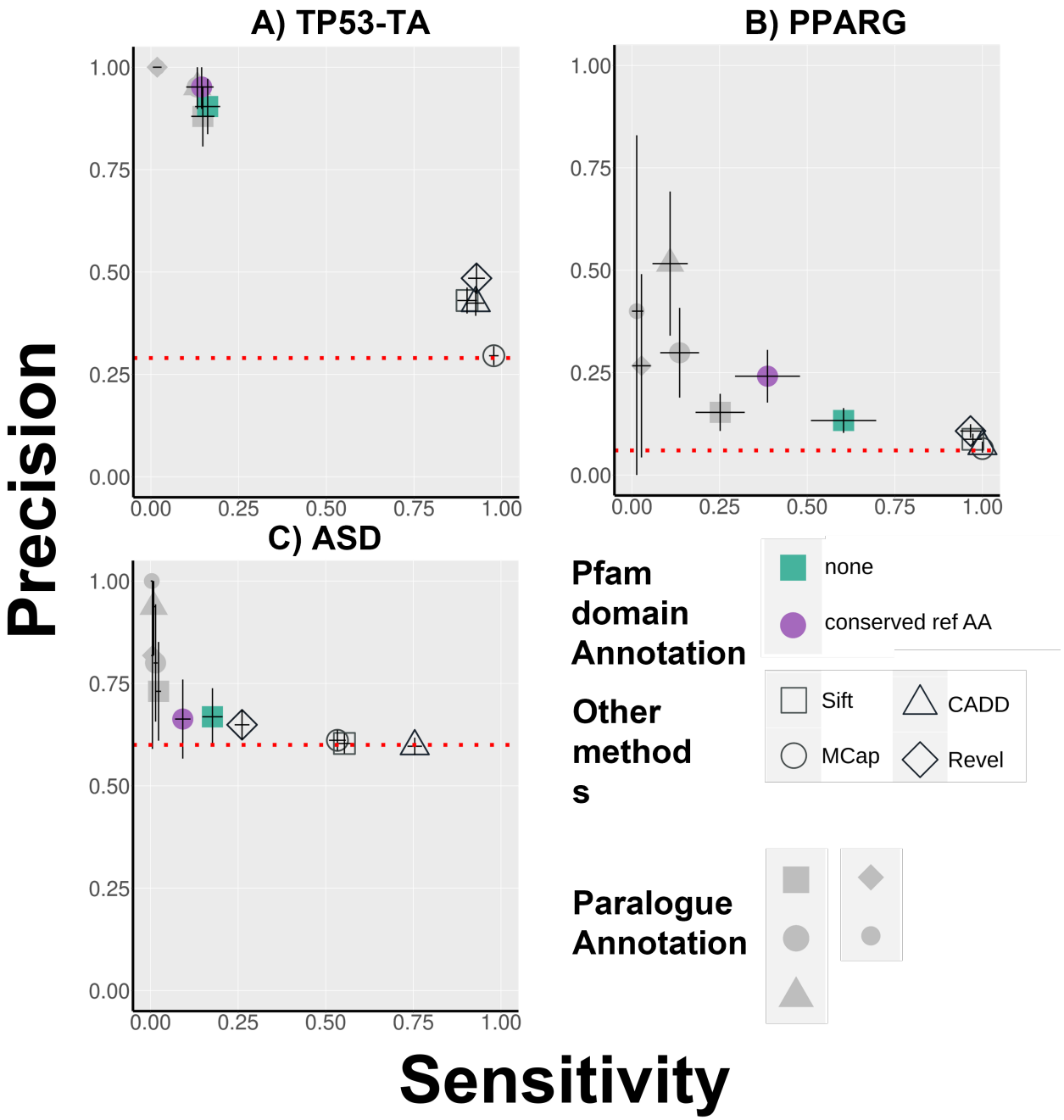


Figure 3.10: Precision-Sensitivity space plot of predicting pathogenic variants show Pfam metadomain Annotation has generally improved sensitivity and comparable precision compared to Paralogue Annotation, but still less sensitivity than other variant prediction tools. The precision and sensitivity of using Pfam metadomains in the Paralogue Annotation framework is shown with and without filtering for conserved reference amino acids (ref AA). This is compared with that of Sift, MCap, CADD, Revel and also Paralogue Annotation (grey; shapes vary with conservation stringency - refer to original Figure 2.8 for details) when applied to independent test-sets consisting of functional assay-determined variants from a) *TP53*, b) *PPARG* and c) de novo mutations linked to Autism Spectrum Disorder. 95% confidence intervals shown as error bars.

3.3.8 Pfam metadomains do not increase odds of correctly calling pathogenic variants in case-control comparisons anymore than Paralogue Annotation

Pfam metadomains did not alter Odds Ratios significantly for any of the genes analysed, with comparable performances to Paralogue Annotation. That is to say, for almost all genes and respective diseases analysed - *KCNQ1* and *KCNH2* for LQTS; *SCN5A* for both LQTS and BrS; and *KCNQ2* for NDD+E - using Pfam metadomains on sufficiently rare variants produced Odds Ratios that were better than just using rare variants alone. But none were any more significant than the Odds Ratios calculated by filtering for rare variants followed by Paralogue Annotation. All had heavily overlapping confidence intervals.

In fact for both *SCN1A* and *SCN8A* implicated NDD+E, Pfam metadomains produced worse Odds Ratios than Paralogue Annotations. 2.4 (upper bound 2.8) and 3.2 (upper bound 4.3) respectively, compared to 9.6 (lower bound 5.8) and 8.4 (lower bound 4.8). For SCN1A, filtering for only rare variants alone produced just as similar Odds Ratios at 2.1 with an upper bound of 2.6.

Only *SCN2A* saw any significant, though small, improvement from 1.5 when using Paralogue Annotation to 3.4 when using Pfam metadomains. The upper and lower bounds of both confidence intervals meet at 2.6.

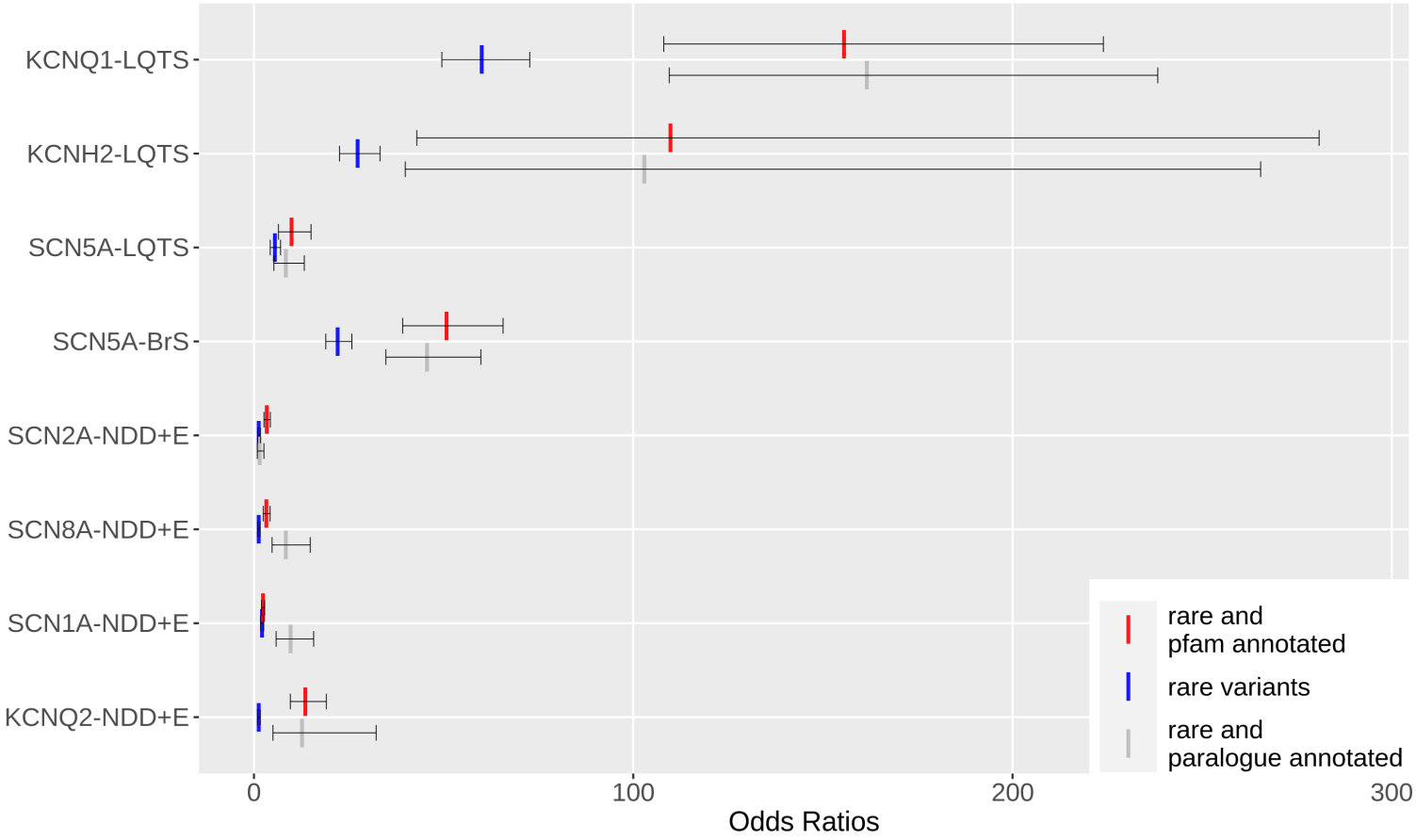


Figure 3.11: Odds Ratios calculated before and after Pfam Annotation is applied compared to Paralogue Annotation. Shown are Odds Ratios calculated from case and control series for: *KCNQ1* and *KCNH2* in Long QT Syndrome (LQTS); *SCN5A* in LQTS and Brugada Syndrome (BrS); *SCN2A*, *SCN8A*, *SCN1A*, and *KCNQ2* in Neurodevelopmental Disorder with Epilepsy (NDD+E). In blue are variants when filtered only by sufficient rarity and in red are variants that are also additionally annotated using Pfam metadomains in the Paralogue Annotation framework (with conserved AA). Previous Odds Ratios generated by Paralogue Annotation are also shown in grey (see original Figure 2.9 for details). Error bars show 95% confidence intervals.

3.4 Discussion

By utilizing the pfam metadomain alignments from Wiel et al. (2017) in our framework, we showed that sensitivity can be improved significantly while the effect on precision, though negative, is minimal. Expanding our framework from paralogy to domain homology seemed logical due to the conserved nature of protein domain evolution within proteins. For example, two proteins involved in transcriptional regulation of two different genes could share the same domain responsible for binding to DNA. As a result, the same application of the ACMG-AMP guidelines could also be applied here. If a variant in the domain of one of the proteins has been

shown to disrupt DNA binding, then a variant with the same or similar resulting amino acid change in the same domain of the other protein could also be theorized as having the same effect.

The increased number of genes covered by pfam domains compared to paralogues, as was shown in **Fig. 3.2**, would also help in compensating for the lack of coverage in certain genes that did not have any related paralogues and thereby hopefully improve the sensitivity issues Paralogue Annotation encountered as demonstrated in Chapter 2. Albeit, a small relative number of genes, 1098, were not covered by paralogues or pfam domains at all. Any variants in these genes would therefore escape annotation by either method. Within this set of genes, perhaps the most interesting feature is that for genes classified with a protein class of defense/immunity protein, genes with neither pfam domains nor related paralogues were enriched (**Fig. 3.3** and **Fig. 3.4**). This protein class includes any proteins involved in the various defenses and immune responses of the body, such as immunoglobulins and antimicrobial peptides, like defensin. These genes are crucial in preventing infectious diseases and so the lessened coverage of this protein class by both paralogues and pfam domains may be a point of concern. However, that is not to say there is no coverage at all, but rather users of the Paralogue Annotation and Pfam metadomain framework should be aware of the possible deficiency within this protein class.

Another interesting aspect of the protein classes classified, was the relationship between genes containing only pfam domains and genes that contained both pfam domains and paralogues. Seeing as the latter group included genes that also contained pfam domains, the antagonistic nature of the two groups seemed surprising. In fact, this sort of relationship was initially expected to occur between genes with only pfam domains and genes with only paralogues, again as the stated desire was to improve paralogue annotation coverage. In certain protein classes, such as metabolite interconversion enzymes and chaperones, this is

indeed the case. For other class, like cytoskeletal proteins, the relationship is even in the opposite direction. But overall the extent of which this occurs is less.

PANTHER defines their protein classes hierarchically (Mi et al., 2021). The parent protein class is based on the overall function of a protein family. Proteins can be further subdivided into more specific children protein classes within the parent. So different children classes can be present in the same parent if they belong to the same family, and looking closer into these classes where the antagonistic relationship occurs can highlight why these differences exist. Translational proteins was the example selected (**Fig. 3.5**), and ribosomal proteins was highly enriched in Pfam domain only genes. Ribosomal proteins constitute the ribosomal subunits of the ribosome and are highly conserved across organisms. As expected there are many different orthologues of these proteins, but not as many paralogues - at least not when defined by Ensembl.

Wiel et al. (2017) implied that homology between domains could be used to identify equivalent variants and thereby help variant interpretation, but here I have shown that just as with Paralogue Annotation, the domains can indeed be used in our framework where a one to one relationship between equivalent pathogenic variants can be used as a prediction for pathogenicity. With ClinVar again acting as the reference set of variants to transfer annotations from and to, Pfam metadomains provided many more annotations than Paralogue Annotations. This of course also meant that there were more false positive benign variants being annotated as being pathogenic as well. So overall, though the results still produced a relatively high precision, it is less than that of Paralogue Annotation. Crucially however, sensitivity vastly improved, fulfilling the original objective of this chapter.

Interestingly, though there were 4,546 and 977 genes with either only Pfam domains and paralogues, GO analysis of the genes that were annotated by both Pfam metadomains or Paralogue Annotation showed that there is not much difference in the distribution of

PANTHER proteins classes or pathways between them (**Fig. 3.8** and **Fig. 3.9**). This seemed to contrast the GO profiles of when looking at Pfam domain only and paralogue only genes before annotation (**Fig. 3.3** and **Fig. 3.6**), where there are more striking differences in the protein classes and pathways between the two groups.

Therefore, Pfam metadomains and Paralogue Annotation most likely annotate approximately the same groups of genes, that is to say genes containing both Pfam domains and paralogues. This is of course not too unexpected as the majority of protein coding genes contain both (13,537 from **Fig. 3.2**).

However, one key aspect and advantage that Pfam metadomains provided was the increased number of hits per protein class or pathway. Even more so for certain protein classes, such as translational proteins, where there were no hits at all when annotating with Paralogue Annotation. The increase in sensitivity has therefore allowed more variants from different classes of genes to be annotated that would otherwise be missed by Paralogue Annotation.

Ultimately however, this is dependent on the current state of reference variants available in ClinVar. With updates to their database in the future or using a different database altogether, the GO profile of genes annotated could differ.

Indeed, in regards to being a classifier the trend of sensitivity increase and precision decrease generally remained. But as describe in Chapter 2 for Paralogue Annotation, the magnitude of this effect again depend entirely on the data being considered and what variants were being annotated from and to. The independent functionally-assayed variants in **Fig. 3.10** demonstrated this.

For *TP53*, Pfam metadomains performed no worse or better than Paralogue Annotation. Most likely due to the fact that in ensembl, *TP53* is defined as having only two paralogues, *TP63* and *TP73*, with Paralogue Annotation unable to annotate with any variants from the latter. While according to Pfam, there are four domains present in TP53 - P53 transactivation

motif, transactivation domain 2, P53 DNA-binding domain, and P53 tetramerisation motif. Only the P53 DNA-binding domain shared any alignments within the metadomain alignments and this was only with one other gene - *TP63*. In this case, both Paralogue Annotation and Pfam metadomains were essentially aligning the same sequences.

For *PPARG*, Paralogue Annotation utilised four paralogues for annotation - *VDR*, *RORA*, *THRB*, and *RARB*. All the while, Pfam metadomains identified thirteen different genes (three of which were the same - *VDR*, *RORA*, and *THRB*) from two Pfam domains - C4 type Zinc finger and ligand-binding domain of nuclear hormone receptor. As a result, sensitivity increased dramatically while precision decreased slightly. In this scenario, Pfam metadomains could be considered more useful than Paralogue Annotation. At the very least, it can act as the middle ground between the lower precision, but higher sensitivity of the current established methods, and the higher precision, but lower sensitivity of Paralogue Annotation.

For the ASD genes, the sensitivity increase was not as substantial and the precision decrease was more noticeable. In fact, REVEL had similar precision statistics yet greater sensitivity. So as a classifier, one might not consider Pfam metadomains as being any more informative than current established methods. If high precision is desired then Paralogue Annotation should be chosen. If sensitivity is the goal, then REVEL could be used.

The informative value of Pfam metadomains is further questioned when considering the if they provide any additional confidence in calling pathogenic variants in case-control studies. The ORs calculated were not significantly better than the ones for Paralogue Annotation. The case numbers for the disease cohorts increased, but somewhat surprisingly so did the control cohorts. This lead to no significant changes in the overall ORs. The increase in number of cases was expected, but the extent of increase in the number of controls was not.

This increase in control numbers could be due to incomplete penetrance for certain variants. Variants from ClinVar can be considered as a gold standard source of data that are relatively

well verified in terms of their pathogenicity or lack thereof. But submitters of variants to ClinVar can be unintentionally bias in the sense that their roles often involve the active search for disease. Therefore, the penetrance of variants in ClinVar as a dataset is likely to be higher than the average variant in a population dataset, where penetrance and pathogenicity are not as distinct. The control cohort is from such a population dataset - GnomAD, and therefore one would expect general disease incident rates to be applicable in such cohorts. The prevalence rates for LQTS is about 1 in 2,000 (Schwartz et al., 2009) and for BrS about 3-5 in 10,000 (Sarquella-Brugada et al., 2016). For NDD+E in particular, prevalence rates are difficult to discern but development disorders caused by de novo variants are estimated to be between 1 in 213 to 1 in 448 births (Study, 2017). This is higher than both LQTS and BrS, and may explain why Pfam metadomains performed relatively worse in *SCN8A* and *SCN1A* as more pathogenic variants were annotated in the control cohorts than by Paralogue Annotation.

When encountering variants that more realistically matches those faced at the forefront of clinical genetics and not ones that are well defined, such as from ClinVar, the classification power of Pfam metadomains is on par with Paralogue Annotation. But more importantly, perhaps as a search tool for functionally equivalent variants, Pfam metadomains offers more possible and different results. This is where the utility of Pfam metadomains could provide beneficial assistance to Paralogue Annotation. Not to say that the latter is obsolete as there will undoubtedly be circumstances where the Pfam metadomains cannot annotate a variant but Paralogue Annotation can. Using the two in conjunction will provide optimal results.

3.4.1 Limitations of Pfam metadomains

The utilisation of Pfam metadomains in the Paralogue Annotation framework in its current state does lack the same intensive conservation stringency filters as Paralogue Annotation. This is of course except for the first stringency level of conserving for the same reference amino acid

in query variant and the annotating variant. Due to the increased technical challenges posed by the Pfam metadomains in comparison to Ensembl's VEP, primarily being the vast increase in number of aligned positions that need to be considered per variant, more development and testing time would be required to implement the remaining conservation filters. In doing so, the precision could potentially be improved further for certain variants. However, even with these features implemented they will unlikely surpass Paralogue Annotation due to the natural trade-off between sensitivity and precision.

Overall where there are differences in performance, sensitivity can be said to have improved while precision has lessened. However, the sensitivity increase is still not enough to match that of current established methods. Though to reiterate points made previously, a higher precision than sensitivity is likely more desirable within the clinical context of predicting pathogenic variants. Nevertheless in the interest of exploring this framework of identifying functionally equivalent variants, sensitivity is once again an issue.

The limitations of Pfam metadomains are similar to the limitations experience by Paralogue Annotation in that regard. That is to say though the Pfam metadomains generate more aligned positions than Ensembl's paralogues, the variant of interest must still be located within one of these positions in the domains. That domain must also be involved in disease. As with Paralogue Annotation, as more pathogenic variants are discovered and documented in time, there will be more functionally equivalent variants available for the alignments and transfer of annotations. This will gradually lead to improvements in sensitivity.

3.5 Conclusion

In conclusion, this chapter has shown how to develop further the framework of Paralogue Annotation by switching the paralogue alignments used by Ensembl to Pfam protein domain alignments, also known as Pfam metadomains. This has mainly lead to the improvement of

providing more functionally equivalent variants to search for and transfer annotations from. Particularly for variants that would have otherwise been ignored by Paralogue Annotation. In general, Pfam metadomains can be thought as being more sensitive than its predecessor but with a marginal cost to precision. The increased sensitivity is still however not as extensive as those from current established methods, which act as more contemporary classifiers for predicting pathogenic variants. Overall, the suggested preference of usage to users of this framework would be not to simply rely on Paralogue Annotation or Pfam metadomains alone, but to use both in conjunction.

With that said, being unable to annotate a variant due to no aligned positions in sequence existing is the universal problem that affects both Paralogue Annotation and Pfam metadomains. Both depend on the variant of interest either being located in a gene that has related paralogues or a domain that exists in multiple instances, whether within the same gene or different genes. In the following chapter, I will examine if moving beyond the limitations of sequence alignments is possible in an attempt to further extend the definition of functionally equivalent variants.

Chapter 4

Expanding from sequence to structure

4.1 Introduction

In the previous chapters I have shown through both Paralogue Annotation and the use of Pfam metadomains that functionally equivalent variants can be identified across either aligned paralogues or homologous protein domains. These variants can then be used to inform on whether other variants in the same aligned position are likely to be pathogenic. I have quantified this framework as having relatively high precision compared to current established methods for predicting pathogenic variants, but the trade-off for higher precision is limited sensitivity in return. This means that many true positive pathogenic variants would never be identified and annotated.

Part of this is due to the current availability of data. As stated before, the more variants discovered and documented over time, the more that will be available to be used in this framework. Sensitivity will naturally increase as variant research progresses.

Another reason for the limited sensitivity is due to the inherent aspects of sequence alignments, regardless of whether that involves aligning the sequences of paralogues or protein domains.

4.1.1 Limitations of sequence alignments

For sequence alignments to work they require a discrete binary selective criteria. In other words, amino acid positions are discrete. When two sequences are aligned in a sequence alignment, the positions of two sequences must be in frame with each other. Positions in one sequence cannot be half (or any other non-integer value) a position out of frame with the other for example, otherwise the whole concept of sequence alignments breakdown. Therefore, for any given position in one sequence, they exist in a binary state of being either aligned with a position in the other sequence, or they are not aligned at all. In the latter, they may be aligned to a gap in the other sequence, or be represented as a gap themselves where they do not exist in the alignment. Aligning homologous protein domain sequences, as oppose to paralogue sequences, would result in less gaps appearing in the alignment as the less conserved regions between domains are removed. But even so, eliminating all gaps in the alignment would be impossible.

Furthermore, even within the most optimised alignment algorithms, there will still be instances where due to the nature of the framework, as long as two variants are not within the same alignment column, they would not be deemed equivalent (**Fig. 4.1a**) and therefore the information from one variant would not be used to inform on the other. This notion implies that neighbouring amino acid positions do not affect the ability of adjacent amino acids to perform their role in the overall protein, whatever that functional or structural role may be. We know this to not be true. Though these nearby positions may not have as a detrimental effect on the fitness of the protein, they can still affect the structural and functional viability of other adjacent positions up and downstream of the protein in the local vicinity of the sequence. This could lead to potential changes to disease or fitness.

Therefore, within the sequence alignment, one could look at positions that are adjacent to the aligned amino acids but not directly aligned themselves in the sequence with the reference variant (**Fig. 4.2**).

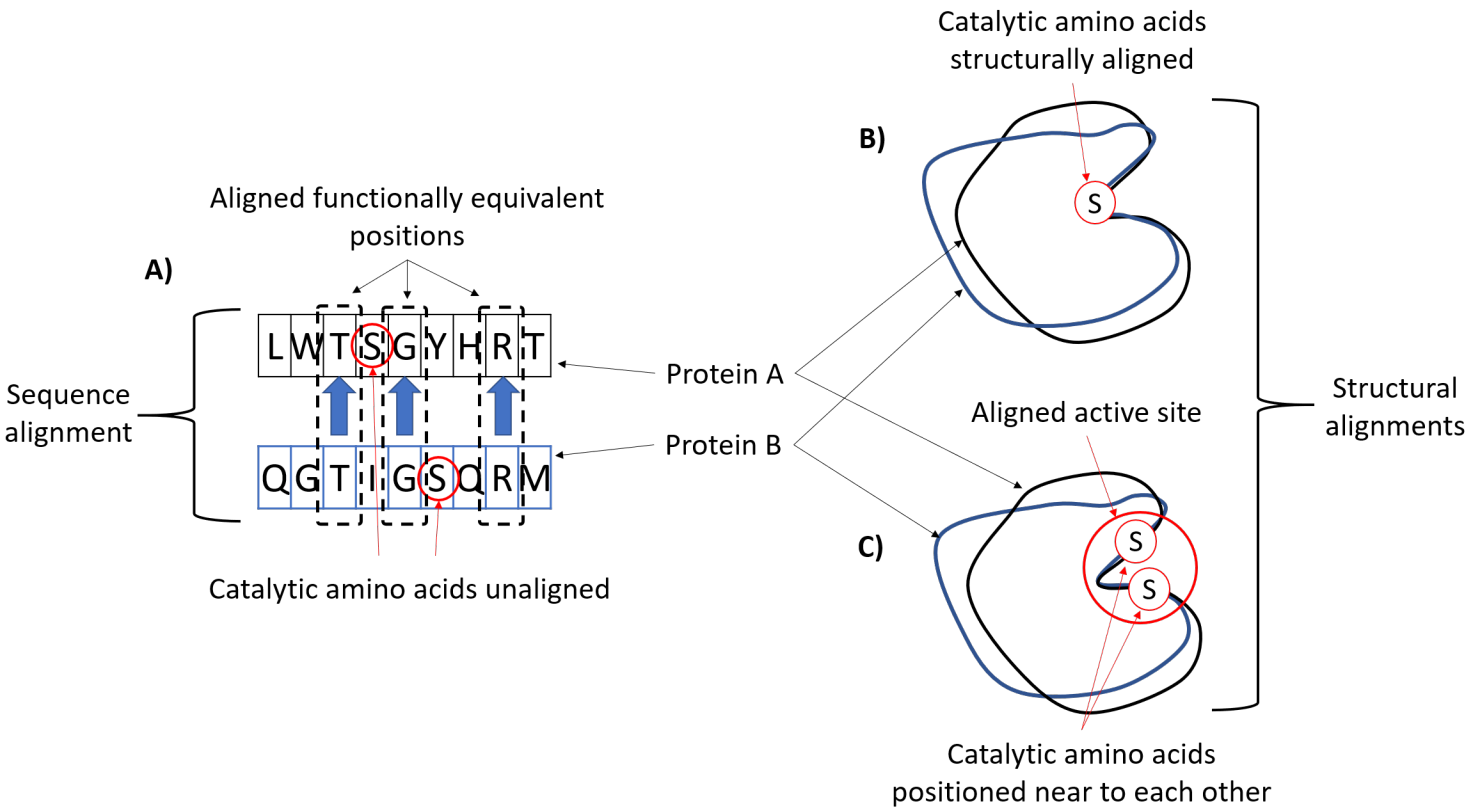


Figure 4.1: Hypothetical scenario where structural alignments provide more information about possible functional equivalency than structural alignments. a) Schematic of two protein sequences where their catalytic amino acids are unaligned in a sequence alignment due to alignment algorithm optimising for better alignment of surrounding conserved amino acids. b) But the structural alignment of the same two proteins could show that the same two catalytic amino acids actually do occupy the same 3D space. c) Alternatively, the structural alignment may not align the catalytic amino acids exactly, but may align the active sites of the proteins together showing that amino acids are at least located within close proximity.

However how far up or downstream from the aligned position is searched requires thorough examination. The same window search size may not be equally appropriate for all protein positions, particularly if you consider conservation patterns. Blocks of aligned positions that are well conserved would better fit this approach, for example when looking at pfam/protein domains. But high fidelity regions where there are fewer aligned positions or positions where the immediate surrounding segments are not very well conserved may provide inaccurate information, such as the case in certain parts of paralogue alignments. In addition, gaps in alignments must be taken into consideration as they misrepresent the real sequential distance the amino acid positions are in and could potentially mislead the additional search locations.

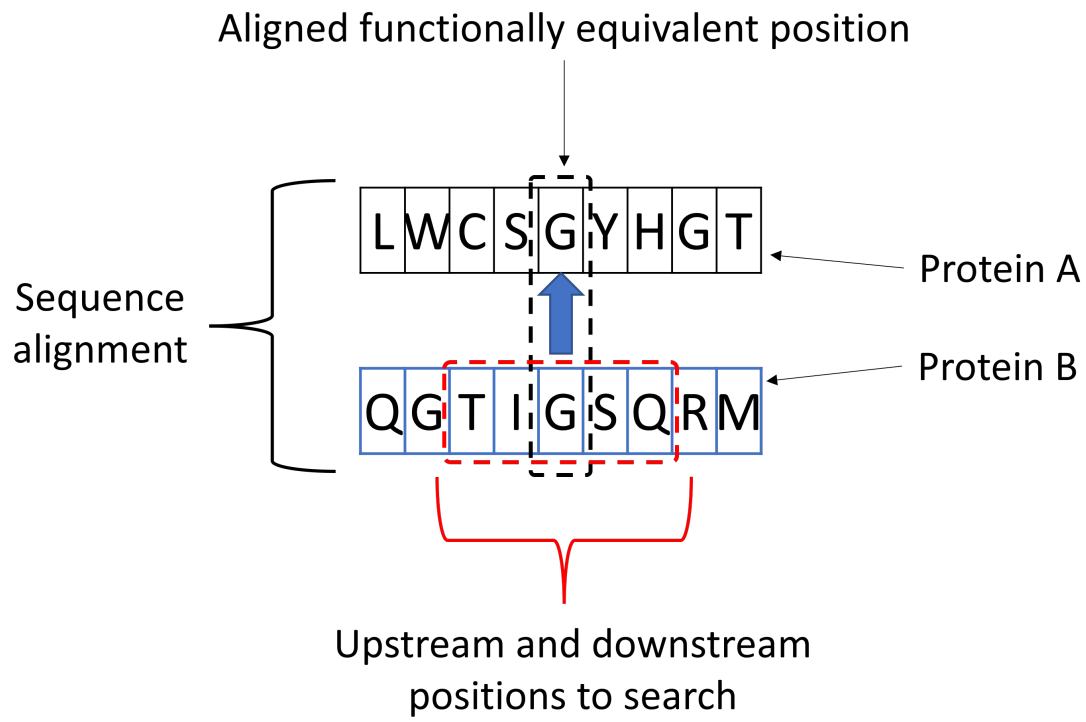


Figure 4.2: Schematic of possible adaptation to the current framework. Protein A is the query sequence and is aligned in a pairwise sequence alignment with Protein B, which is the reference. Two functionally equivalent amino acids are aligned in the middle and these positions could have variants. Highlighted in red are the potential positions where additional reference variants could be searched for although the search window could vary.

4.1.2 Looking beyond sequence

A better approach is perhaps to look beyond sequence alignments entirely. When considering proteins in their native state, they do not exist purely as sequences but instead are folded into complex three dimensional structures. Though sequential information and representation is easier to deduce and process, looking at protein structures could provide information that more closely resembles the reality that proteins exist in. Therefore, the natural step forward from protein sequence would be to look at protein structure. That is to say, instead of using sequence alignments, would using structural alignments of protein structures within the same framework be possible and would it provide more information?

The structure of a protein gives more indication of function than sequence alone, as indeed a protein's structure dictates its function (Alberts et al., 2003). Therefore, by looking at structure, the definition of functional equivalence of a variant could be expanded. While using a sequence

alignment, two catalytic amino acids that may not be aligned in sequence (as in **Fig. 4.1a**) may in fact be aligned in structure (**Fig. 4.1b**). Indeed, two proteins only need approximately 20-25% amino acid conservation to have high structural similarity (Krissinel, 2007). High structural similarity being defined as any alignments with a Root Mean Square Deviation (RMSD) ≤ 2 Å (See Material and Methods for more details on how RMSD is calculated).

Furthermore, even though structural alignments may identify the same equivalent residue positions as the ones identified by sequence alignments, where sequence alignments may fail to align certain segments of amino acids due to inadequate sequence similarity, aligning based on structure may allow more flexible and dynamic annotations of equivalence. Therefore, providing a greater number of proteins with a possible annotation of some kind.

Even if the two catalytic amino acids are not aligned exactly in structure, they may be aligned within close proximity or the immediate local surrounding structure of the amino acids may align similarly, and this could still provide information about possible pathogenicity. For example in **Fig. 4.1c**, two catalytic amino acids may not align in sequence, but aligning the structures of their proteins together may show that the active sites of these two proteins are structurally quite similar and the catalytic amino acids are both located within the same sites with similar 3D coordinates.

The concept mentioned above can be expanded further. Looking at protein structure rather than sequence can also provide additional information about far reaching interactions between distant amino acids. That is to say two residues that may not be close together in sequence can be in structure space due to the folding of the protein (Klein-Seetharaman et al., 2002). Take for example, the active site of serine proteases - an enzyme that cleaves peptide bonds in other proteins. The cleavage reaction requires the use of a catalytic triad - Ser195, Asp102, and His57 (Hedstrom, 2002). Though they are more than 50 amino acids apart from each other in sequence, looking at the protein structure reveals that these three amino acids are

located right next to each other in the active site and act coordinately together to catalyse the breaking of peptide bonds. A variant at either of those three positions would undoubtedly affect the catalytic mechanism. But a variant that is close enough to those positions could also potentially affect the positioning of the catalytic amino acids in the active site or change the shape of the active site. This could lead to an inefficient or incomplete reaction mechanism as the catalytic triad or the substrate are blocked from reacting or binding appropriately.

In fact, this observation of multiple variants affecting a localised region when looking at the structure of a protein is well characterised, particularly for deleterious variants that cluster together in tumour suppressor genes (Gao et al., 2017). Considering variants in isolation can indeed provide useful pathogenic information, but perhaps using multiple variants that affect a localised region together can impart more information about the surrounding vicinity and any other variants within the same region. That is to say that other variants nearby to a variant of interest can be used to help interpret that variant's pathogenicity. This could be extended across structurally aligned proteins where multiple reference variants within a localised region of one reference structure could potentially help annotate a query variant in the query structure.

Hence, by using structural alignments instead of sequence alignments the possible search space for functionally equivalent variants is no longer limited to a discrete 1 dimensional system, but a continuous 3 dimensional one. Instead of looking for equivalent single positions and transferring annotations across sequences, we can now look for equivalent structural spaces and use variants within a local proximity to inform pathogenicty across structures.

The ultimate question however, is how feasible this new approach of the framework is. Resources such as the CATH database of structurally aligned protein domains already exist (Sillitoe et al., 2021). Mapping variants onto these structures and their alignments may provide additional insight for inferring variant pathogenicity. This chapter will differ to the previous chapters and will focus more on some of the exploratory work I have done and resources I have

used to investigate this new direction.

4.2 Materials and Methods

4.2.1 Protein sequence and structure databases used

For representing overall publicly available protein sequence and structure data, UniProtKB/TrEMBL database statistics and Protein Data Bank (PDB) statistics were retrieved and reproduced where possible from https://www.ebi.ac.uk/uniprot/TrEMBLstats and https://www.rcsb.org/stats/growth/growth-released-structures respectively.

Experimentally derived protein structures were also manually queried and retrieved from the PDB webservice. Structures with the best combination of highest resolution and protein length were chosen as the basis of study for the exploratory work in this chapter. This was typically structures solved by X-ray crystallography. The .pdb files for the chosen structures were downloaded for processing.

4.2.2 Structural alignment of protein structures/models

Sequence independent structural alignment of protein structures was performed by TM-align using their online service available at https://zhanggroup.org/TM-align (Zhang and Skolnick, 2005). Two structures that were chosen to be aligned would have their .pdb files submitted online and the alignment file would be returned also in .pdb format.

4.2.3 Calculation of Root Mean Square Deviation

Though RMSD is not perfect for measuring protein structural similarities, for example side chains are not typically considered in the calculation, it is quantitatively the simplest way of representing how close proteins are to each other in terms of their structure (Li, 2013). Just

as how sequence alignments have percentage identity scores that describe the quality of the sequence alignment, RMSD can be used to score the quality of structural alignments.

The RMSD, measured in Å and is calculated by:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$$

where n is the number of pairs of equivalent (α carbon) atoms between two protein structures being considered and d_i is the distance between the i^{th} pair of atoms (Kufareva and Abagyan, 2011).

4.2.4 Tools used for visualisation of protein structures and sequence alignments

Visualisation and manipulation of 3 dimensional protein structures and models executed using PyMol, available at https://pymol.org (Schrödinger, LLC, 2021). .pdb files were loaded in with peptide chains being typically represented in ribbon style while specific residues of interest as atomic spheres. Related images presented in this thesis were produced directly from Pymol as PNGs. Visualisation of associated protein sequence alignments were produced using MView, available at https://www.ebi.ac.uk/Tools/msa/mview (Madeira et al., 2019). Sequence alignments in .msf file format were submitted to MView under default parameters except for "consensus" which was set to off.

4.2.5 *In silico* prediction tools used for structural models

Alphafold 2 protein structural models were queried and retrieved manually from https://alphafold.ebi.ac.uk (Tunyasuvunakool et al., 2021, Varadi et al., 2022). .pdb files and images were downloaded and processed as necessary.

Phyre2 protein structural models were retrieved by submitting the Ensembl

canonical/MANE Select v0.95 protein sequences of interest to http://www.sbg.bio.ic.ac.uk/phyre2 (Kelley et al., 2015) under default parameters and downloading the returned .pdb files.

4.2.6 Mapping variants onto structures and search space scanning distances

Custom Python (v3.7.3) scripts were developed to handle the mapping back of variants onto aligned protein structures. The script would query a pre-downloaded tabulated .vcf file with the gene of choice and subsequently extract the unique protein positions of where these variants were located on the chosen gene. ClinVar (build GRCh37 v20201114) P/LP, B/LB, and GnomAD (r2.1.1) vcf files were used. The script was ran directly in PyMol for automated mapping onto structures.

Another custom Python script was developed for computing the distance between a query variant and the closest reference variant on the equivalently aligned structure. The 3 dimensional coordinates of every α carbon atom in the aligned .pdb file was iterated through and parsed into a k-dimensional tree object using the SciPy package (class: `scipy.spatial.KDTree`), which allowed multidimensional nearest neighbour searches to be conducted. The distances between each α carbon on the query structure and its nearest neighbouring α carbon on the reference structure was recorded. These distances were then parsed into R for graphical representations and statistcal testing.

4.2.7 Statistical Tests

Statistical tests were performed using base R commands but the formula for tests used are as follows.

Welch's T test statistic is calculated by:

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

where $\bar{X_1}$, s_1 , and N_1 are the sample mean, standard deviation, and sample size of the first set of distances between structurally aligned variants; and $\bar{X_2}$, s_2 , and N_2 are the respective sample mean, standard deviation, and sample size of the second set of distances being considered (Welch, 1947). In R, the command `t.test()` was used.

Two sample Kolmogorov–Smirnov statistic is calculated by:

$$D_{n,m} = sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

where sup_x is the supremum of the set of distances, $F_{1,n}(x)$ and $F_{2,m}(x)$ are the empirical distribution functions of the first and second set of distances being considered (Berger and Zhou, 2014). The empirical distribution functions can be calculated by:

$$F_n(x) = \frac{number\ of\ distances\ in\ the\ set \leq x}{n}$$

where n is the total number of distances in the set and x is the distance chosen. In R, the command `ks.test()` was used.

4.3 Results and Discussion

4.3.1 Limited volume of structural data when compared to sequential equivalent

The first immediate question that arises is how much structural data is available. Due to the advancement of sequencing technology over the past decades, acquiring sequence data is mostly trivial at current times with the growth of sequence databases, such as Uniprot (**Fig. 4.3**),

experiencing exponential-like increases in the last decade or so. More importantly, the data in these databases are not raw, they have been annotated and curation has been maintained to a high standard. Therefore, finding and using a specific sequence desired is not challenging in the context of today for most researchers.

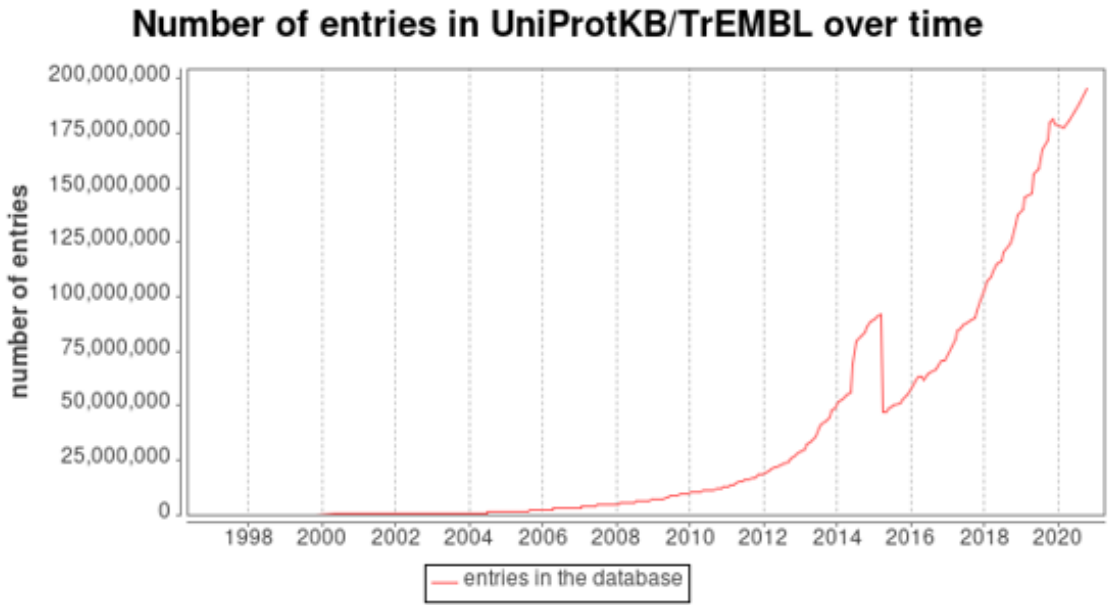


Figure 4.3: Number of entries in the UniProtKB/TrEMBL sequence database has increased dramatically since 1998 to 2020. Image downloaded from the UniProt Consortium (2021). The drop in entries at around 2015 is due to a procedure carried out by the consortium in order to reduce proteome redundancy.

Compared to sequence data, the situation that protein structure data is in is much more complicated. Firstly, the total number of entries in protein structure databases, such as the PDB, is significantly less than that of sequence databases (**Fig. 4.4**). Even though the trend in growth has also been exponential, there has been far less structures available than there has been sequences. Where there has been over 175,000,000 sequence entries for UniProt in 2020, there has been just over 171,000 total structure entries in PDB in the same year.

The main reason for the huge contrast in numbers is due to the increased challenges that faces protein structure determination. Though sequencing peptides directly is still significantly more challenging than sequencing nucleotides, there are still more avenues that protein sequencing can use to supplement its shortcomings compared to protein structure. The obvious being indirect deduction of sequence through translation of nucleotide sequences

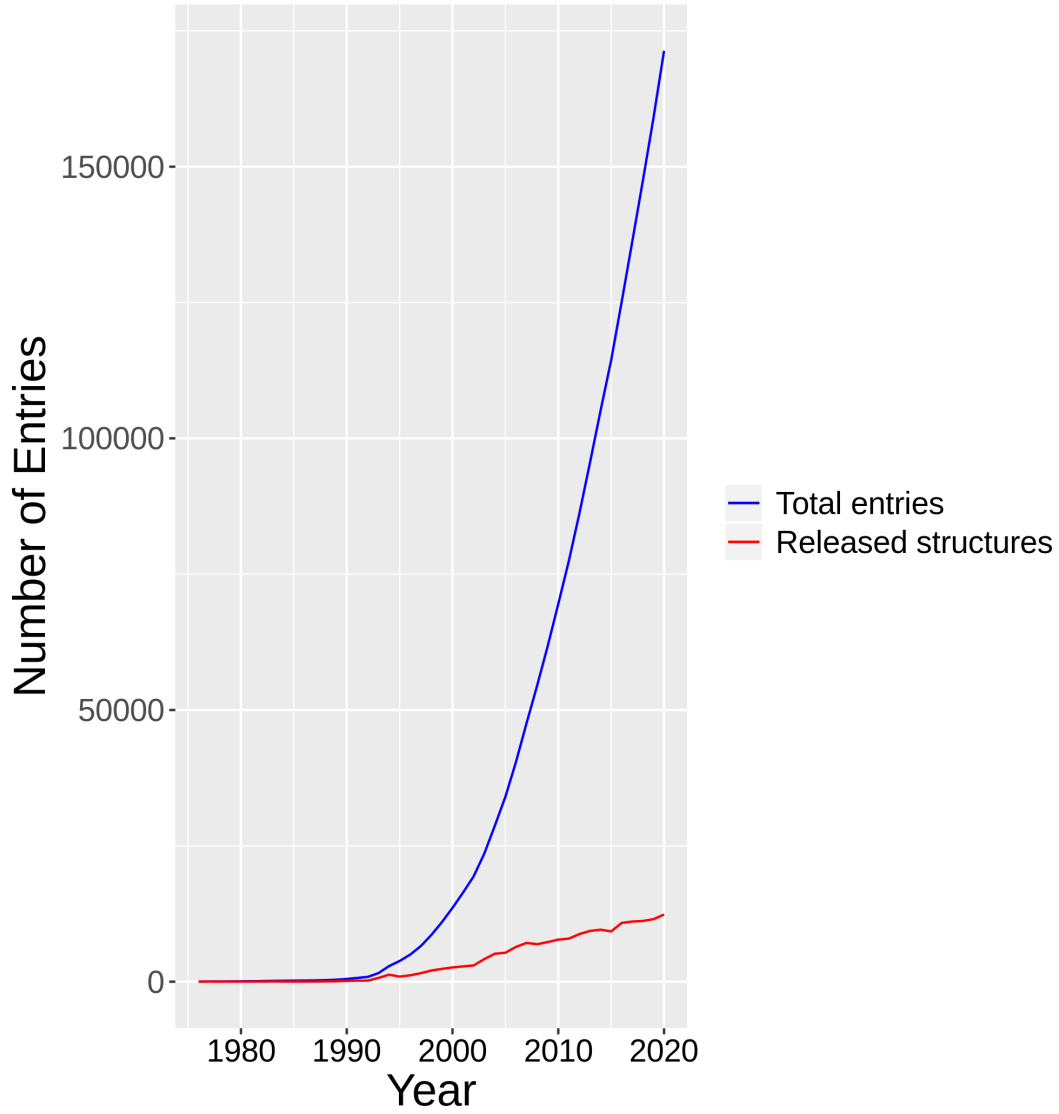


Figure 4.4: Number of entries in the Protein Data Bank database has increased exponentially since 1990 to 2020. Total cumulative entries shown in blue and newly released structures per year shown in red.

gained from transcriptome analysis. For protein structure, there is no equivalent alternative.

The current technological limitations mean that determining structure for most proteins can be incredibly difficult and even impossible for some. There are currently 3 main experimental techniques for determining atomic structures: X-ray crystallography; Electron Microscopy (EM); and Nuclear Magnetic Resonance (NMR).

X-ray crystallography is the most commonly used with 88% of structures in the PDB being derived from this method (Burley et al., 2021). Mainly due to the method obtaining the highest resolution structures out of the three with an average resolution of 2 Å. However, this technique

requires the ability for the proteins of interest to form crystal latices before being exposed to x-rays in order for their to be a x-ray diffraction pattern. This pattern is what is used in order for a model of the structure to be calculated. Proteins that are insoluble, disordered, or highly flexible are very difficult to crystallise and therefore x-ray crystallography is not suitable for many proteins. But for ones that can be, the resulting model structures are often detailed at high resolution.

5% of structures in PDB are derived by EM, specifically cryo-EM. Though proteins no longer need to be crystallised, the method produces on average 3.5 Å resolution structures and is limited to proteins with a molecular weight bigger than 50 kDa (Liu et al., 2019). This is due to the method freezing multiple copies of the protein in place while they are in different orientations and then computatively adding and calculating what the overall model of the protein structure looks like. Cryo-EM is useful for large molecules, but for high fidelity research, such as what is proposed in this chapter, cryo-EM structures are unlikely to be viable.

The remaining 7% of structures in PDB are models derived from NMR. In this technique, proteins in solution are exposed to magnetic fields which induces their atoms to resonate when radio waves are introduced. The resonating frequencies differ depending on their surrounding atoms and are represented as signal peaks on a NMR spectrograph. An overall model of the protein structure can be computed from these peaks. However, proteins must be water soluble and suspended in solution for NMR to function. As a result, NMR is ideal for studying protein dynamics, but in order to generate a single structural model, an ensemble of multiple models, each representing the protein in different dynamical states, is used to generate an average model structure. Because of these dynamics proteins are limited to a molecular weight of less than 30 kDa (Sugiki et al., 2017). Any proteins larger than that become difficult to determine as large proteins start to show overlapping signals in the NMR spectra.

Due to these limitations, finding complete structural models that are high resolution can be

incredibly challenging. Currently, only x-ray crystallography can provide structural information that does not require computing an average model from multiple different representations of the same protein. In that sense, it is the only technique that can determine the real atomic distances of the whole protein being measured. Often though, the whole protein cannot be crystallised and only a segment of the protein that can be is characterised leaving large chunks of missing protein.

4.3.2 Alignments of available crystal structures of SCN5A and SCN2A provide limited structural coverage

The choice of which protein structures to align is an important one. The more structurally similar two proteins are the easier the alignment. There is also the biological significance of choosing similar structured proteins. As mentioned prior, structure dictates function and so the more structurally similar two proteins are the more likely they carry out a similar function. In this framework of annotation transfer across functionally equivalent variants, it is vital for the proteins being aligned to have similar function. Otherwise, the possibility of transferring biologically irrelevant annotations could lead to false positive predictions or provide misleading information about the variant in question. Of course a safeguard against the possible risk of transferring annotations across biologically dissimilar proteins would be the quality of the alignment. If two proteins have a high RMSD between them than the alignment should not be used.

For the purpose of this exploratory work, the paralogues SCN5A and SCN2A were chosen as two example proteins to be used in this structural adaption of the Paralogue Annotation framework. SCN5A and SCN2A are transmembrane proteins and so are very challenging to crystallise as they are insoluble. Their hydrophobic surfaces that come into contact with the lipid bi-layer of membranes aggregate with each other preventing crystallisation and studies

in NMR. As a result, most crystal structures of these genes have limited length as they only represent the parts of the protein that could be crystallised.

The canonical transcripts of SCN5A and SCN2A produces proteins that are 2015 and 2005 residues long respectively. However, for SCN5A there are only 6 entries within PDB all around 150 residues long or less. The highest resolution entry is 4DJC with 1.35 Å but only positions 1491-1522 (32 residues) of the entire protein are present making it rather unsuitable for the intentions of this chapter. Considering both resolution and protein length, 4DCK was chosen as the structure that best represented SCN5A. With a resolution of 2.2 Å and a length of 168 (positions 1773-1940) it was determined as the most optimal entry.

For SCN2A there are only 2 crystal structures in PDB: 4RLY, which is only 19 residues long (positions 1102-1120) with a resolution of 2.5 Å; and 4JPZ, which is longer at 161 residues (positions 1777-1937) but with a lower resolution of 3.02 Å. The lack of other options meant that 4JPZ had to be used for aligning with 4DCK.

Fortunately the structural alignment of both structures gives a relatively decent sized alignment of length 132 residues with high similarity as denoted with a RMSD of 2.67 Å (**Fig. 4.5**). Both structures overlap each other well with 4 β strands and 4 α helices aligning near perfectly apart from the 2 α helices. One at the C-terminal end that spans about 30 residues for both (bottom left of **Fig. 4.5**), and the other helix approximately 15 residues upstream that also spans 15 residues (upper right of **Fig. 4.5**).

The strong structural alignment is not too surprising seeing as SCN2A has 64% sequence identity and coverage to SCN5A. As both entries represent roughly the same region of protein, the region in question is highly conserved both structurally and sequentially (**Fig. 4.6**). Though in the sequence alignment there is a small patch of unconserved residues between positions 1880-1895. This region partly overlaps with the aforementioned 15 residue sized α helices that did not structurally align as well.

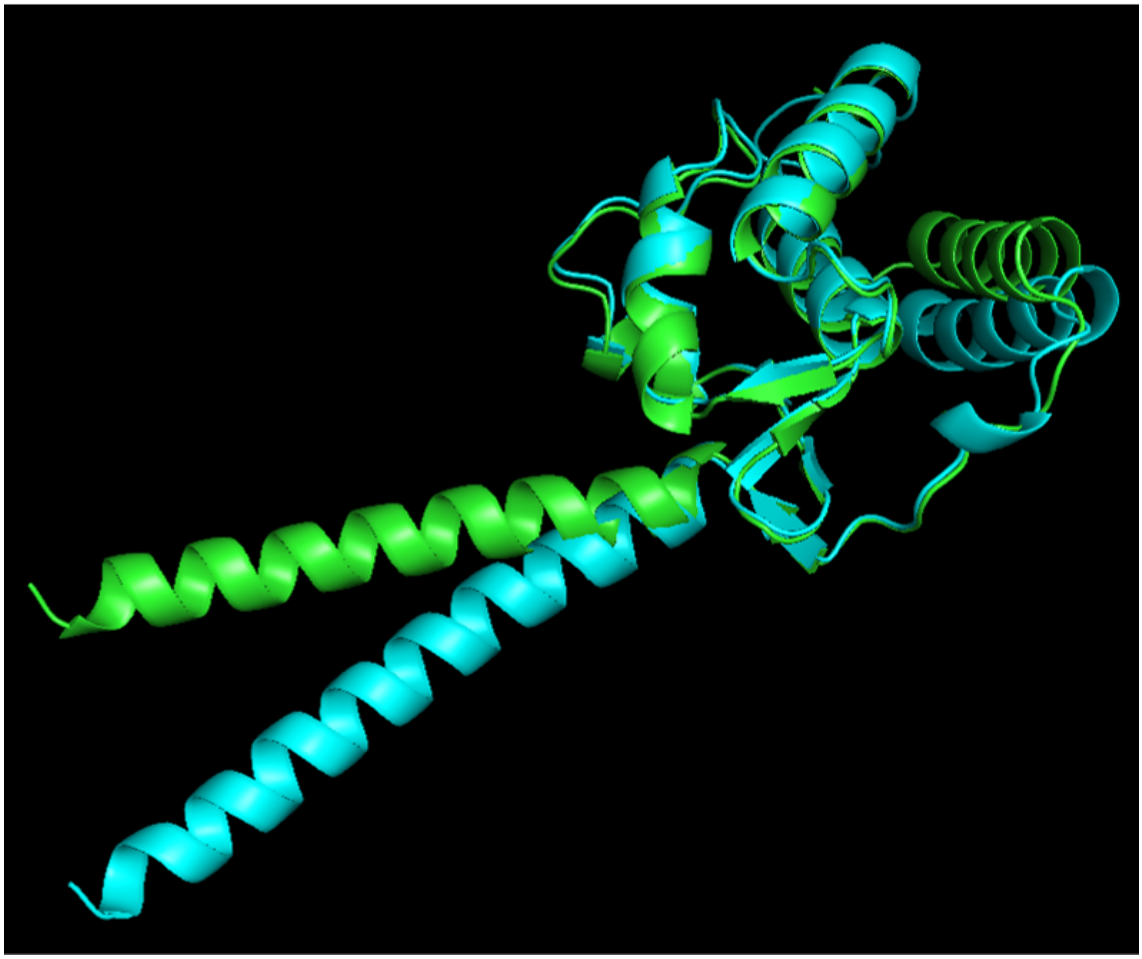


Figure 4.5: Sequence independent structural alignment of 4DCK (SCN5A positions 1773-1940) in cyan and 4JPZ (SCN2A positions 1777-1937) in green using TMalign. Length = 132 residues. RMSD = 2.67 Å.

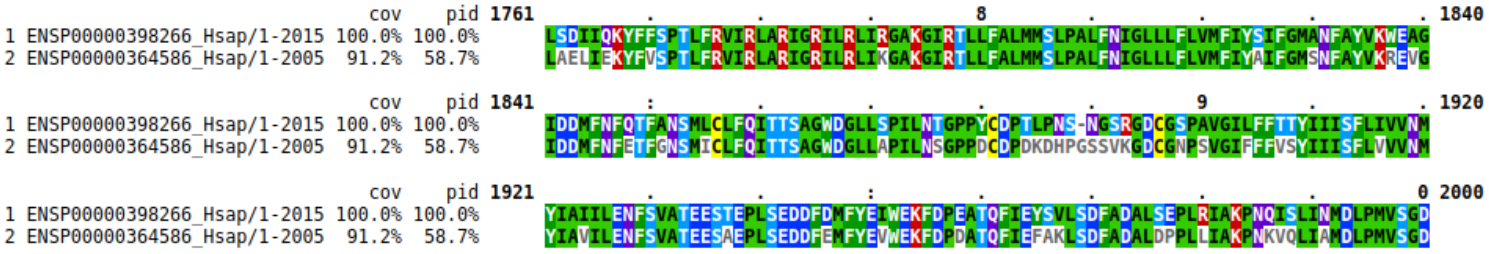


Figure 4.6: Ensembl pairwise sequence alignment of paralogues SCN5A (top sequence) and SCN2A (bottom sequence) between positions 1761 and 2000. Canonical protein sequences ENSP00000398266 and ENSP00000364586 used respectively. Whole protein target (SCN2A) %identity at 63.84% and %coverage at 63.52%. Positions that are conserved have columns that are coloured in top and bottom.

Taking all P/LP missense variants from ClinVar and mapping them onto the structural alignment results in only 2 and 6 variants being able to do so for SCN5A and SCN2A respectively (**Fig. 4.7**). This is out of 97 and 115 total missense variants in ClinVar respectively. Clearly there are not enough variants available that can be mapped back onto the structural alignment

for there to be any reasonable analysis as the region aligned does not cover enough of the whole protein (**Fig. 4.8**).

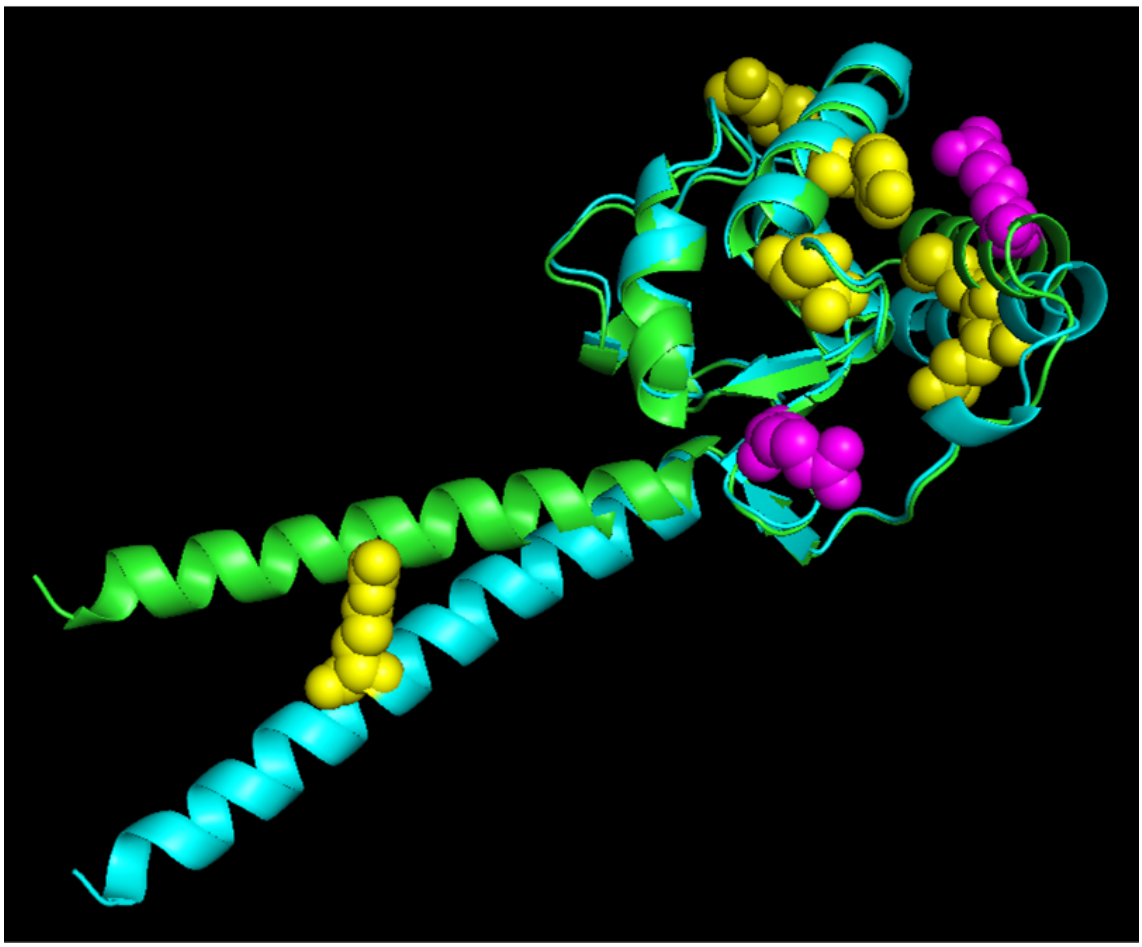


Figure 4.7: Mapped Pathogenic/Likely pathogenic missense variants from ClinVar displayed onto the sequence independent structural alignment of SCN5A and SCN2A. SCN5A variants (two) shown in magenta and SCN2A variants (six) shown in yellow.

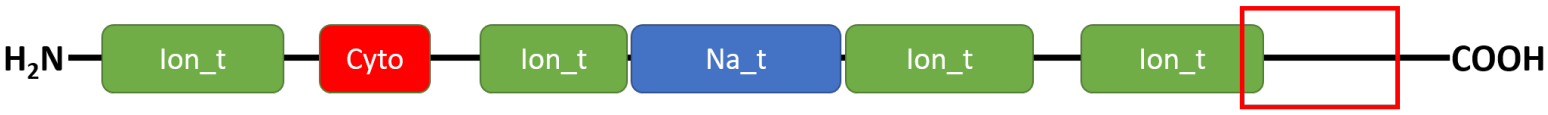


Figure 4.8: Pfam domain schematic of sodium channel proteins SCN5A/SCN2A show how crystal structures 4DCK and 4JPZ cover only a small region of total protein. Ion transport protein domain (**Ion_t**), Cytoplasmic domain of voltage-gated Na+ ion channel (**Cyto**), and Sodium ion transport-associated (**Na_t**) shown as green, red, and blue boxes respectively. SCN5A and SCN2A, being well conserved paralogues, share an identical layout of Pfam domains. The red square highlights the approximate combined coverage of both 4DCK (SCN5A positions 1773-1940) and 4JPZ (SCN2A positions 1777-1937) as they cover approximately the same region in sequence.

The issues highlighted above are common for most proteins. Though any two structures could be theoretically aligned, if the structures are of low quality then the overall alignment will

not be useful or reliable. That is to say if the structures do not adequately represent enough of the whole protein then not enough variants will be mapped back onto the structure. If the structures have poor resolution then atomic positions of residues may be imprecise, leading to possible inaccurate alignments and incorrect transfer of annotations across alignments. Finding complementing structures to align is currently quite difficult even though parts of the whole protein have crystal structures, there simply is not enough structural coverage. In the future, empirically quantifying how many structures that are of high resolution and represent most of the complete protein would be desired.

4.3.3 Using Alphafold 2 predicted structures of SCN5A and SCN2A compensates for lack of structural coverage

Due to the challenges of determining protein structure, many *in silico* tools that predict the structure of proteins have been developed. These tools could be a potential solution to the lack of structures currently available. Though using experimentally derived structures is ideal, if few are practically viable then using predicted models is perhaps a reasonable alternative.

Quite recently, the people at DeepMind developed Alphafold, a machine learning based prediction algorithm that utilises neural networks for deep learning. In 2018, Alphafold finished in first place in the Free Modeling category at the 13th Critical Assessment of protein Structure Prediction (CASP) (AlQuraishi, 2019). In 2020, DeepMind released Alphafold 2, which came first again at CASP14 with further improved results (Jumper et al., 2021). Alphafold 2 was shown as being able to predict novel protein folds with high accuracy and reliability, more so than any of its competitors by a large margin.

At the start of the Alphfold 2 structure prediction pipeline, the input amino acid sequence is queried through multiple protein sequence and structure databases to construct a MSA and a preliminary representation of the possible contact interactions between amino acids in the

sequence. This is done by extracting co-evolution information between the aligned sequences and by using template structures from similar protein sequences with known structures.

After which at the core of the Alphfold 2 pipeline lies two deep learning neural networks that can be thought of as operating separately one after the other. The first neural network takes the MSA and templates, and extracts any possible geometric information that could help indicate residue positioning and contact points. The second neural network takes all the information extracted from the first and uses it to construct the final three dimensional model of what the structure of the protein could be.

Using protein structure prediction tools, like Alphafold 2, may be an intermediate solution to the current lack of structural coverage until there are more structures solved experimentally. **Fig. 4.9** and **4.10** shows Alphafold 2 predicted structures of SCN5A and SCN2A respectively. Both are represented in their full 2016 and 2005 residue protein length, with the front and core of the proteins - the pore forming α helices - predicted very confidently. This side of the proteins face outwards into the extracellular space. Moving back and towards the outer regions, the confidence of prediction lessens. The α helices that form the tetrameric-like transmembrane region are predicted with slightly less confidence. But in particular, towards the back of protein is where residues are predicted with very low confidence. This is to be expected as this region consists of mostly disordered random loops that face the cytoplasmic side of the membrane towards the centre of the cell. For SCN5A, this would be cardiac muscle cells (cardiomyocytes). For SCN2A, neuron and oligodendrocyte precursor cells.

In ClinVar there are 97 P/LP variants documented for SCN5A. Mapping them back onto the predicted structure from Alphafold 2 and taking only unique residue locations results in 86 possible variant positions. For SCN2A, there are an equivalent 115 P/LP variants with 100 unique mapped variant positions on the predicted structure (**Fig. 4.11** and **4.12**). Doing the same for B/LB variants results in 22 and 6 unique variant positions respectively. No

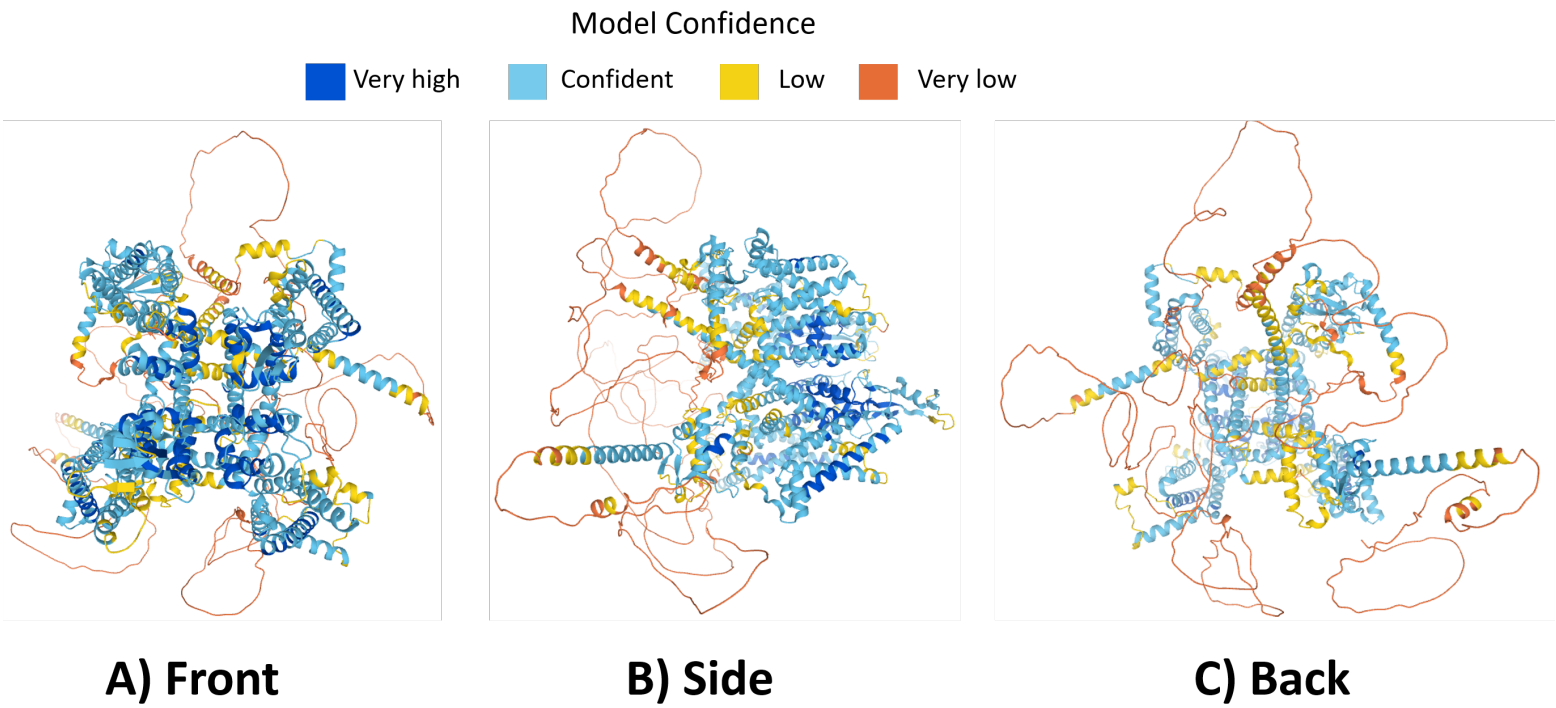


Figure 4.9: Alphafold 2 predicted structure of SCN5A (2016 residues) from a) the front, b) the side, and c) the back points of view shown in ribbon form. Residue confidence scores range from 0 to 100 with the following levels: Very low (<50)(orange); Low (>50 and <70)(yellow); Confident (>70 and <90)(light blue); and Very high (>90)(dark blue).

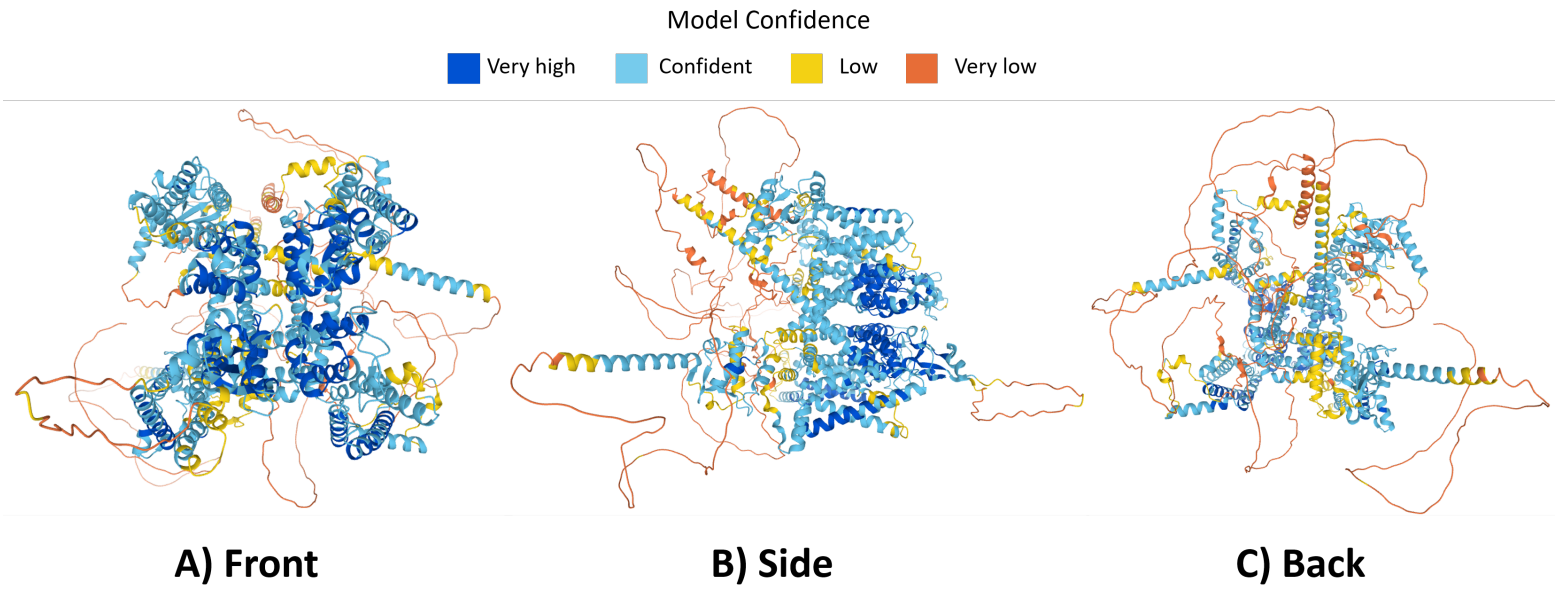


Figure 4.10: Alphafold predicted structure of SCN2A (2005 residues) from a) the front, b) the side, and c) the back points of view shown in ribbon form. Residue confidence scores range from 0 to 100 with the following levels: Very low (<50)(orange); Low (>50 and <70)(yellow); Confident (>70 and <90)(light blue); and Very high (>90)(dark blue).

benign variants were documented as being located in the same position as another. The large discrepancy between pathogenic and benign variants in ClinVar is due to the nature of scientific

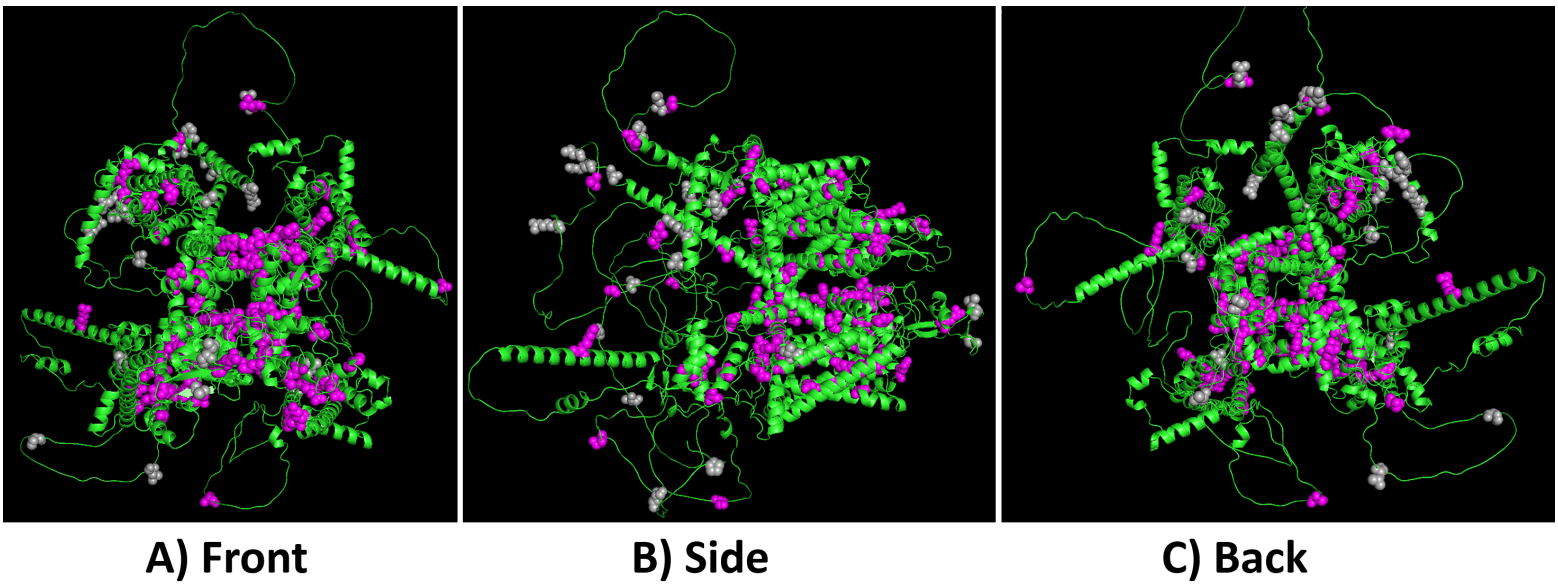


Figure 4.11: Alphafold predicted structure of SCN5A (2016 residues) with mapped ClinVar variants from a) the front, b) the side, and c) the back points of view shown in ribbon form. Unique positions where a ClinVar Pathogenic and/or Likely pathogenic variant is present shown as magenta atoms while a Benign and/or Likely benign variant is shown as grey atoms.

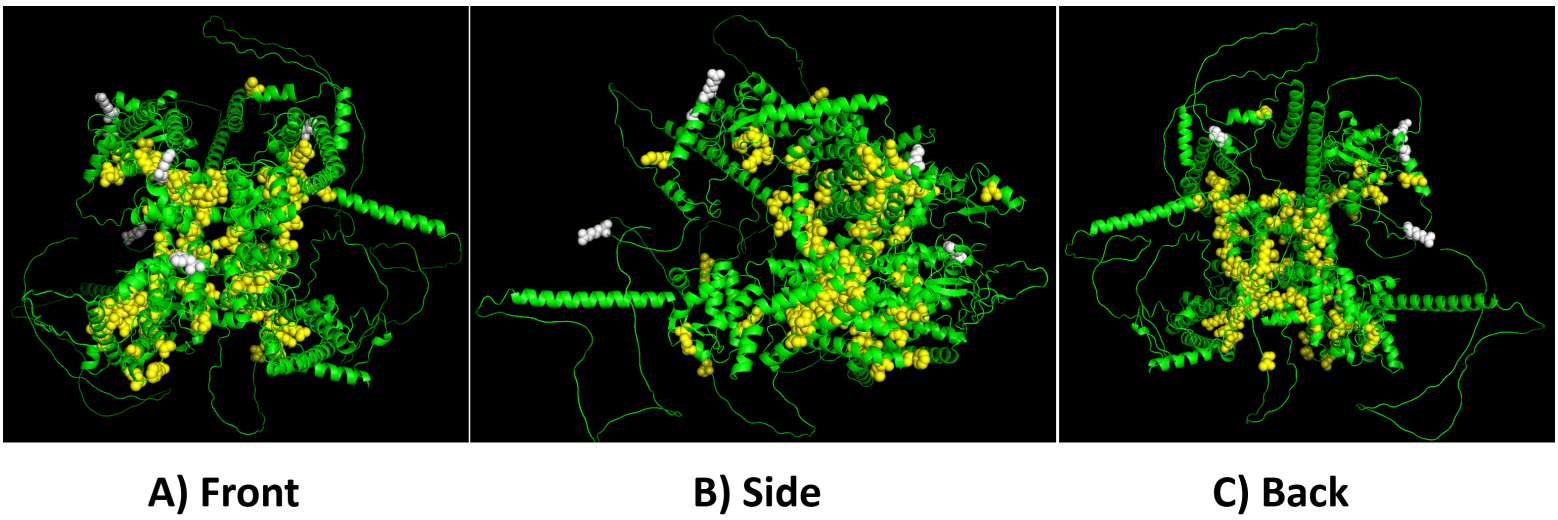


Figure 4.12: Alphafold predicted structure of SCN2A (2005 residues) with mapped ClinVar variants from a) the front, b) the side, and c) the back points of view shown in ribbon form. Unique positions where a ClinVar Pathogenic and/or Likely pathogenic variant is present shown as yellow atoms while a Benign and/or Likely benign variant is shown as white atoms.

reporting where most researchers and studies actively look for disease and therefore most report only positive results, but rarely any negative results.

The mapped pathogenic variants are mainly positioned around and within the α helices that form the pore of the sodium channel at the core of proteins, with some also located in the transmembrane α helices. This is to be expected as any disruption to these regions could

interrupt proper ion transport or integration into the cellular membrane.

The benign variants on the other hand are mostly positioned on the surface of the protein. Many are located within random loops and turns that face either the cytoplasmic or extracellular side of the membrane. These are expected positions where variants are least likely to deal structural damage to the protein.

In the case of SCN5A, a few pathogenic variants however were also located in the middle of random loops, particularly at the back of the protein that faces the cytoplasmic side. Considering Alphafold 2's low predicted confidence scores within these regions, discerning the significance of these variants in these positions within the proposed structural annotation framework is difficult. The model may not be accurate enough here for any informative interpretations.

However, a few studies have shown that within the cytosol, other proteins do indeed bind and interact with SCN5A (and SCN2A) at some of these random loop regions (Remme (2013), Shy et al. (2013), Sanders et al. (2018)). Though Alphafold 2 may not be confident in its prediction of these loops, the lack of secondary structures is plausible and variants here could potentially lead to altered interactions with binding proteins for similar proteins.

4.3.4 Alphafold 2 models too complex for current structural alignment workflow

But the real concern arises when considering the structural alignment of these random loop segments. Due to the nature of a lack of secondary structure or any regular structural pattern, apart from small motifs, aligning random loops will always be a somewhat futile task as there is essentially no inherent structure. This issue is demonstrated when attempting to use the structures within the current framework. Aligning the Alphafold 2 predicted structures together required intensive computational demand and due to limited resources an alignment was unable

to be computed.

Albeit, the first reason for this outcome is likely to be due to the sheer size of the predicted structures; both over 2000 residues in length which is among the longest proteins in eukaryotes. The longest protein TM-align was originally tested on was 1058 residues long (Zhang and Skolnick, 2005). Though prior knowledge dictates that the structures are similar due to sharing a similar sequence, TM-align does not utilise sequence information and attempts to align *ab initio* based on structural geometry alone. Nevertheless however, the amount of disorderly random loops likely provided a lot of additional noise as well and were incapable of being aligned as they inherently have no structural similarity.

This raises the fact that within this structural alignment framework, variants within random loops, or indeed regions with low structural order, that cannot be reliably aligned structurally will not be applicable and cannot be annotated in this way.

Concerning theses random loop variants, as these regions lie outside of regular protein domains and in disordered regions, Pfam metadomains are also not applicable. For Paralogue Annotation, though a paralogous positions may be possibly identified no known P/LP variant would likely occupy these positions considering how most variants for SCN5A and SCN2A are not positioned in random loops. Indeed, predicting the structure and function of random loops, let alone variants that inhabit them, has always been more challenging when compared to α helices or β sheets. Generally, loops have a greater degree of conservation variability and sequence composition (Papaleo et al., 2016).

Moving forward in the framework, as these random loops do not harbour the majority of pathogenic variants, and due to the generally low confidence scores in these predicted regions, they could be removed from the structures prior to aligning. This would require specific *ad hoc* processing for each different predicted protein structure and is something that could be worked on in the future.

Another heuristic approach that is potentially worth investigating given more time is to install and use a local version of TM-align instead of using their web service. Particularly for larger, more complex, proteins, having a local version where there is more direct control over resource and memory allocation can perhaps provide more efficient computation time.

4.3.5 Using only homology template modelling removes random loops and allows viable structural alignment

An alternate less resource intensive prediction technique is to use Phyre2. Phyre2 uses homology template based modelling analogous to the first stage of Alphafold 2's modelling pipeline that excludes the neural networks. In the same way, Phyre2 uses the input amino acid sequence to search through protein databases to generate MSAs with the end goal of identifying similar sequences with known structures. These known structures are then also used as templates to predict the backbone of the query sequence (Kelley et al., 2015). Afterwards, loops and side chains are factored into the model and the final prediction is given. This is where Phyre2's pipeline stops, whereas with Alphafold 2 the information provided in the MSAs and templates are used as inputs for the second and third stages of their pipeline - the neural networks.

Fig. 4.13 shows the structural alignment of the Phyre2 predicted structures of SCN5A and SCN2A. The figure appears to show only one structure but it is in fact two structures aligned together. This is because Phyre2 used the same templates for both SCN5A and SCN2A resulting in both structures completely overlapping with each other. This makes sense seeing as they are strongly conserved paralogues and have high sequence similarity. Alphafold 2 also likely uses the initial same templates for both SCN5A and SCN2A. Comparing the models from the two different techniques, the core of the proteins are ultimately very similar, with the pore forming α helices and tetrameric-like transmembrane helices nearly identical. But with the additional work provided from Alphafold 2's neural networks, the differences between helices in SCN5A

and SCN2A are more nuanced. Looking further outward, Phyre2 is unable to predict the outer regions of proteins. In particular, there are far fewer random loops that are modelled, as there are unlikely to be any viable templates for these segments, and though not clearly shown in the figure, there are gaps in the model.

With less amino acids present in the models (both SCN5A and SCN2A Phyre2 models contain 1320 residues each) and the two models being identical due to the same templates being used, structural alignment of SCN5A and SCN2A was easily achieved.

With that said, the circumstances surrounding the usage of the same templates should be addressed. If the two proteins in question are strongly related to each other, that is to say they have high sequence similarity and conservation, then their structures in real terms should also be highly similar. In these scenarios, using the same template would be acceptable. In fact in the case of Phyre2, using the same template would actually nullify the necessity of aligning the structures together afterwards. This is due to the fact that both structurally aligning and using the same templates for structural prediction essentially utilises the same biological assumptions; that the structures being considered are highly similar.

The issue arises when the two proteins are not as strongly related but the same templates are used due to the current limited availability of experimentally solved protein structures. In such cases, the assumption that the structures are highly similar would be incorrect. The structural predictions would be less confident and the structural alignment would also suffer. Therefore the positions of variants in 3D space cannot be assumed to be accurate and cannot be used to inform on the other variants in the other structure. This issue perhaps does not affect Alphafold 2 as much as Phyre2 since the additional neural networks compensate somewhat. As more structures are defined experimentally, the number of such that can be used as templates increases and more information will be available for more accurate structural predictions. For now, structures predicted from the same templates should be carefully observed for possible

inaccuracies.

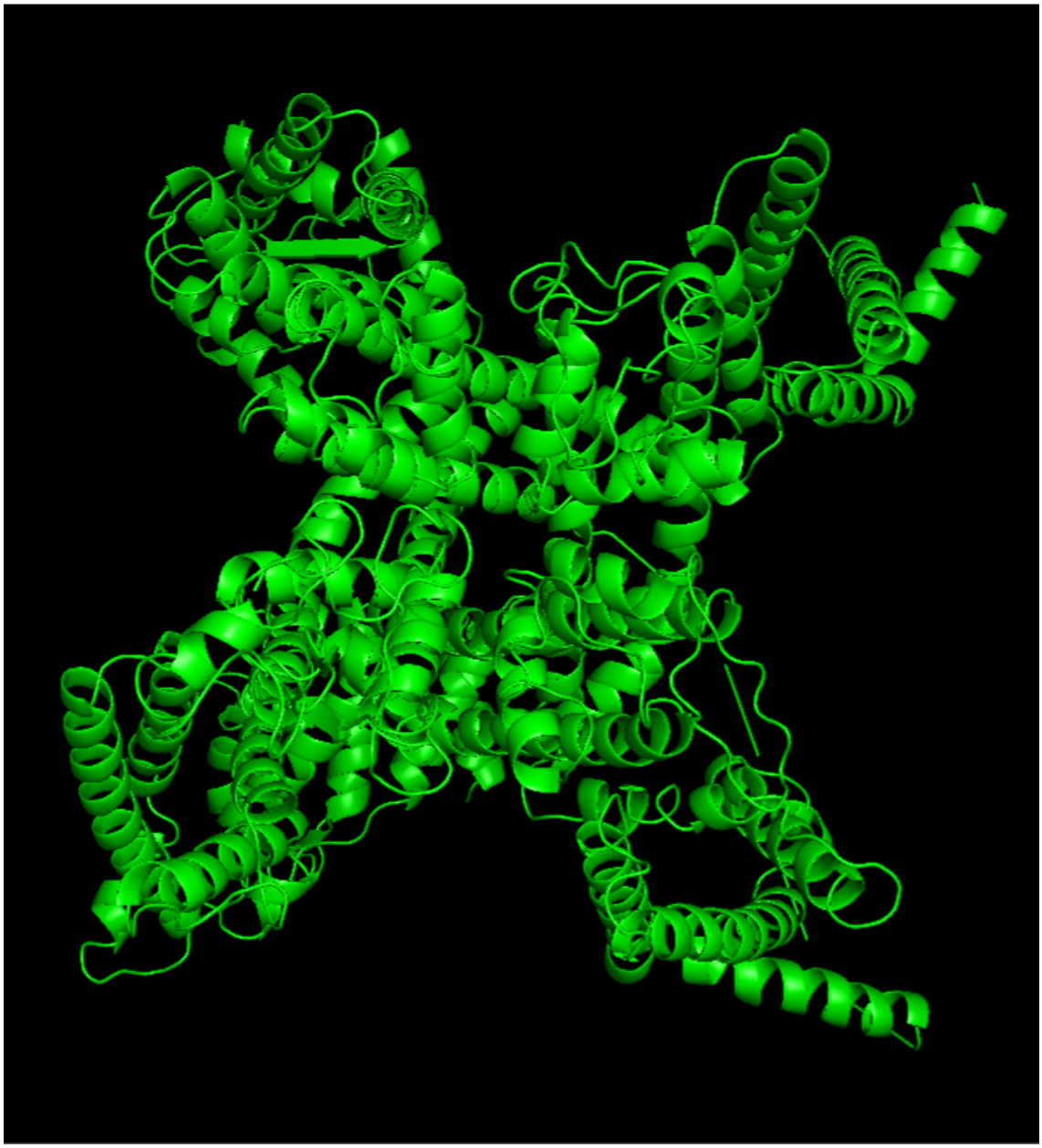


Figure 4.13: TM-align structural alignment of Phyre2 predicted models of SCN5A and SCN2A (both 1320 residues each). Identical templates being used by Phyre2 for the prediction of both SCN5A and SCN2A results in complete overlap of one model over the other during alignment, resulting in the model appearing as if only one protein is present.

4.3.6 Translating equivalent variant framework from sequence to structure

After the structural alignment was achieved the next step was to identify if there was enough P/LP and B/LB variants available and how they would be used in this framework. Using only

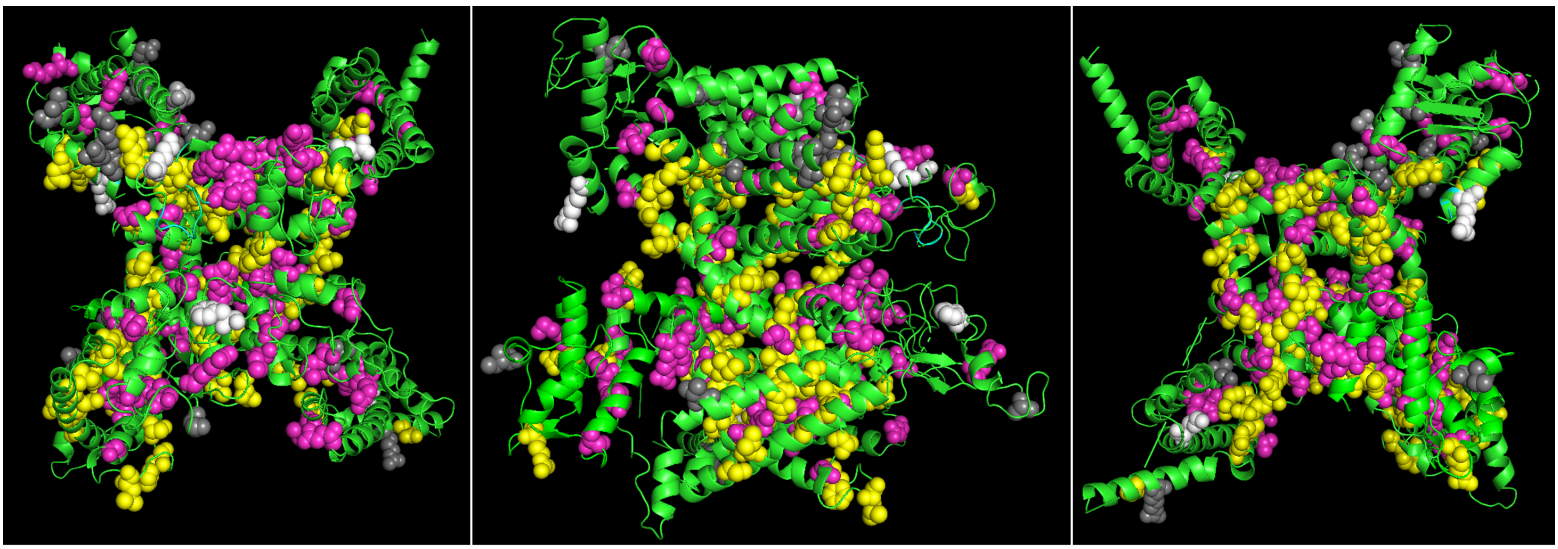


Figure 4.14: Structurally aligned Phyre2 predicted models of SCN5A and SCN2A, shown from a) the front, b) the side, and c) the back points of view in ribbon form. Unique positions where a ClinVar Pathogenic and/or Likely pathogenic variant for SCN5A and SCN2A is present shown as magenta and yellow atoms respectively. Benign and/or Likely benign variants are also shown as grey and white atoms. GnomAD variants are not shown in order to prevent visual obscurity.

B/LB variants from ClinVar led to a small sample size. For SCN5A only 22 B/LB variants were documented and only 6 for SCN2A. Due to gaps in the phyre model, even fewer variants were capable of being mapped back - only 5 for SCN5A and 2 for SCN2A. This did not allow for any meaningful analysis.

Therefore in addition to the ClinVar variants, population variants from GnomAD were also mapped back in order to increase the sample size of non-pathogenic variants, or at least variants that were less likely to be disease causing than confirmed P/LP variants in ClinVar. In GnomAD, there are 1065 and 516 variants for *SCN5A* and *SCN2A* respectively. Out of those, 429 and 215 were mapped back, giving a total of 434 and 217 variants that could be used as control reference variants compared to confirmed P/LP ClinVar variants.

Again due to the gaps in the Phyre2 model, a handful of P/LP variants were not able to be mapped back to the structure. 22 and 21 P/LP variants were missing for SCN5A and SCN2A respectively, out of a total of 97 and 115 P/LP variants. This led to 75 and 94 variants

remaining that could be mapped (**Fig. 4.14**).



Figure 4.15: Pathogenic variants in homologous proteins more likely to be located in similar positions than benign and pathogenic. Shown are measured distances (in Å) between a query variant on SCN5A and the closest Pathogenic/Likely pathogenic (P/LP) ClinVar variant on SCN2A after structural alignment of the Phyre2 predicted models by TM-align. The SCN5A query variants are: Benign/Likely benign (B/LB) ClinVar or GnomAD population variants (green); Pathogenic/Likely pathogenic ClinVar variants (red); and Pathogenic/Likely pathogenic ClinVar variants not already annotated by Paralogue Annotation or Pfam metadomains (blue). For the latter, there appears to be two groups among the distribution: 1) closely located variants that are likely to be informative to the query variant; 2) distant variants that likely do not provide reliable information for the query. * indicates statistical significance ($p < 0.05$).

After examining the distributions of P/LP variants compared to B/LB variants on the Alphafold 2 structures, an initial systematic approach I took was to look at if the differences between the distributions could be quantified on a global scale for the whole protein. Using an analogous approach to the calculation of the statistical performance of Paralogue Annotation, I considered the position of each B/LB or GnomAD variant mapped on to SCN5A and measured the distance of the closest P/LP SCN2A variant on the other structure in the alignment. I

did the same for P/LP SCN5A variants and their closest P/LP SCN2A variants. The distance distributions of SCN5A B/LB/GnomAD to SCN2A P/LP variants and SCN5A P/LP to SCN2A P/LP variants are shown in **Fig. 4.15**. 602 distances were measured for B/LB/GnomAD to P/LP and 91 for P/LP to P/LP.

In doing so, one can see that across the structural alignment, pathogenic variants are more likely to be located closer to each other, that is to say that different P/LP variants from the two proteins are located around similar positions in the proteins. Looking at the distances between SCN5A B/LB/GnomAD variants to SCN2A P/LP, the distribution is more spread out (11.1 $\pm$ 6.9Å) than distances between SCN5A P/LP to SCN2A P/LP (7.7 $\pm$ 5.0Å). Perfoming both a Welch t-test and Kolmogorov-Smirnov (KS) test showed that the differences in distributions are statistically significant (p-values of 9.4 x 10^{-7} and 7.4 x 10^{-5} respectively). This suggests that P/LP variants are more likely to be enriched at closer distances than B/LB/GnomAD and that even distance alone could be an indicative factor when considering what variants can be used as references across the alignment. Clearly however, distance should not be the only factor as there is still a large overlap with many B/LB/GnomAD variants also being quite close to P/LP variants.

The KS test was performed in addition to the t-test as I was not confident that the underlying parametric of the distributions were Gaussian. Firstly, the distributions were positively skewed, 0.71 for B/LB/GnomAD to P/LP and 0.43 for P/LP to P/LP. Secondly, there is a clear range of distances between approximately $0 < x \leq 3.8$Å that have no measurements. This indicates that variants were either exactly on top of each other (aligning perfectly in the structural alignment), or they were at an immediate adjacent residue position in sequence, seeing as α carbons in peptide bonds that are mostly in *trans* configuration are 3.8Å apart (Liwo et al., 1993). 6 B/LB/GnomAD to P/LP variants had distances that were smaller than 3.8Å, the smallest being 3.3Å. Upon closer inspection of the structures, these variants were adjacent

to each other as well in sequence but their peptide bond was in *cis* conformation, which are known to cause the α carbons to be closer to each other in space (Chakraborty et al., 2013). Furthermore, as distances are being measured across aligned structures and not within the same protein, there may be minor deviations away from the standard α carbon peptide bond length.

One key objective was to investigate whether structural space could provide any additional information on variants that were not annotated by either Paralogue Annotation or Pfam metadomains, hence helping to improve the sensitivity of the framework. The blue data points in the third column of **Fig. 4.15** shows query variants that were not identified by previous methods. In this specific example between SCN5A and SCN2A, no variants that were perfectly structurally aligned were missed by Paralogue Annotation or Pfam metadomains. Three SCN5A variants had reference variants that were in close proximity around 5 Å (**Table 4.1**); two of which shared the same reference variant. The remaining variants had reference variants that were too far away and these would unlikely provide any informative information regarding pathogenicity for their query variants.

Looking at the three SCN5A variants and their neighbouring SCN2A variants in a sequence alignment shows that though they are not in alignment they are extremely close to each other. SCN5A Ala425 is 2 residues downstream of SCN2A Thr435 in the alignment (**Fig. 4.16a**). The upstream region of where these 2 variants are located is well conserved but conservation starts to drop at the variants' positions. SCN5A Tyr1495 is situated 3 residues upstream of SCN2A Met1501, and SCN5A Leu1501 being a further 3 residues downstream from the position of SCN2A Met1501 (**Fig. 4.16b**). The surrounding residues in this region are well conserved throughout both sequences.

As the variants are not in alignment, neither Paralogue Annotation nor Pfam metadomains would have identified them unless the window for alignment was expanded as

SCN5A query variant	SCN2A reference variant	Distance between query and reference (Å)
Tyr 1495 Ser	Met 1501 Thr	4.94
Leu 1501 Val		5.08
Ala 425 Thr	Thr 435 Ile	5.39

Table 4.1: Table showing three SCN5A P/LP variants that were not previously annotated by either Paralogue Annotation or Pfam metadomains as having SCN2A P/LP variants in close proximity when using structural alignments. These SCN2A variants could provide indication for the pathogenicity of the SCN5A variants.

described previously in **section 4.1.1** and in **Fig. 4.2**, but again this would have additional issues.

A benefit of using structural alignments as oppose to sequence alignments is that the structural surroundings of these three variants and their aligned reference variants can now be investigated further. This may then reveal some possible details into the mechanism of pathogenicity.

For Ala425Thr and Thr435Ile **Fig. 4.17a**, both are positioned on the same aligned α helix with their side chains facing away. Apart from a random loop there does not appear to be any other secondary structures nearby that may interact with the helix.

For Tyr1495 and Leu1501, they can be seen positioned in the same aligned α helix as Met1501, with Tyr1495 slightly above Met1501 in **Fig. 4.17b** and Leu1501 below. The side chains of Tyr1495 and Met1501 face in a similar direction towards an adjacent helix. Leu1501's side chain faces more towards the right of Tyr1495 and Met1501, but also seems to interact with another adjacent helix. Apart from Leu1501Val, both Tyr1495Ser and Met1501Thr are variants that change from a hydrophobic residue to ones that are polar. Evidently these changes can affect the interactions with the neighbouring α helices and this may play a role in the disease mechanism.

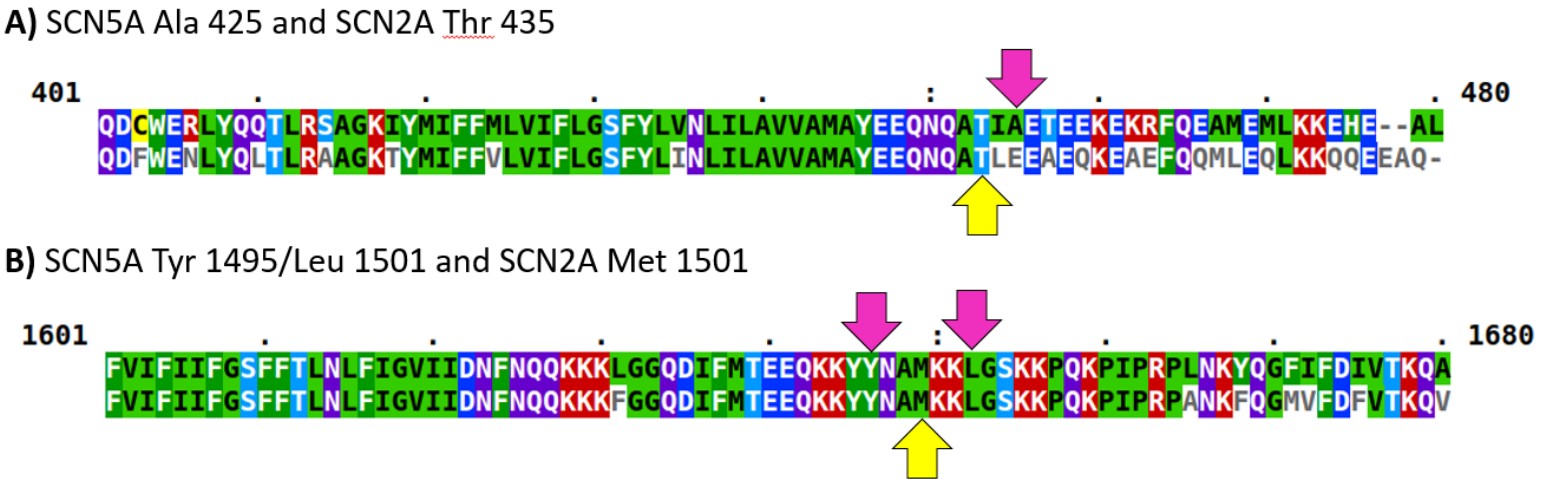


Figure 4.16: Ensembl pairwise sequence alignments of SCN5A (top sequence) and SCN2A (bottom sequence) between regions where A) SCN5A Ala425 with SCN2A Thr435 and B) SCN5A Tyr1495/Leu1501 with SCN2A Met1501 are positioned. SCN5A variants are shown as magenta arrows while SCN2A variants are shown as yellow arrows.

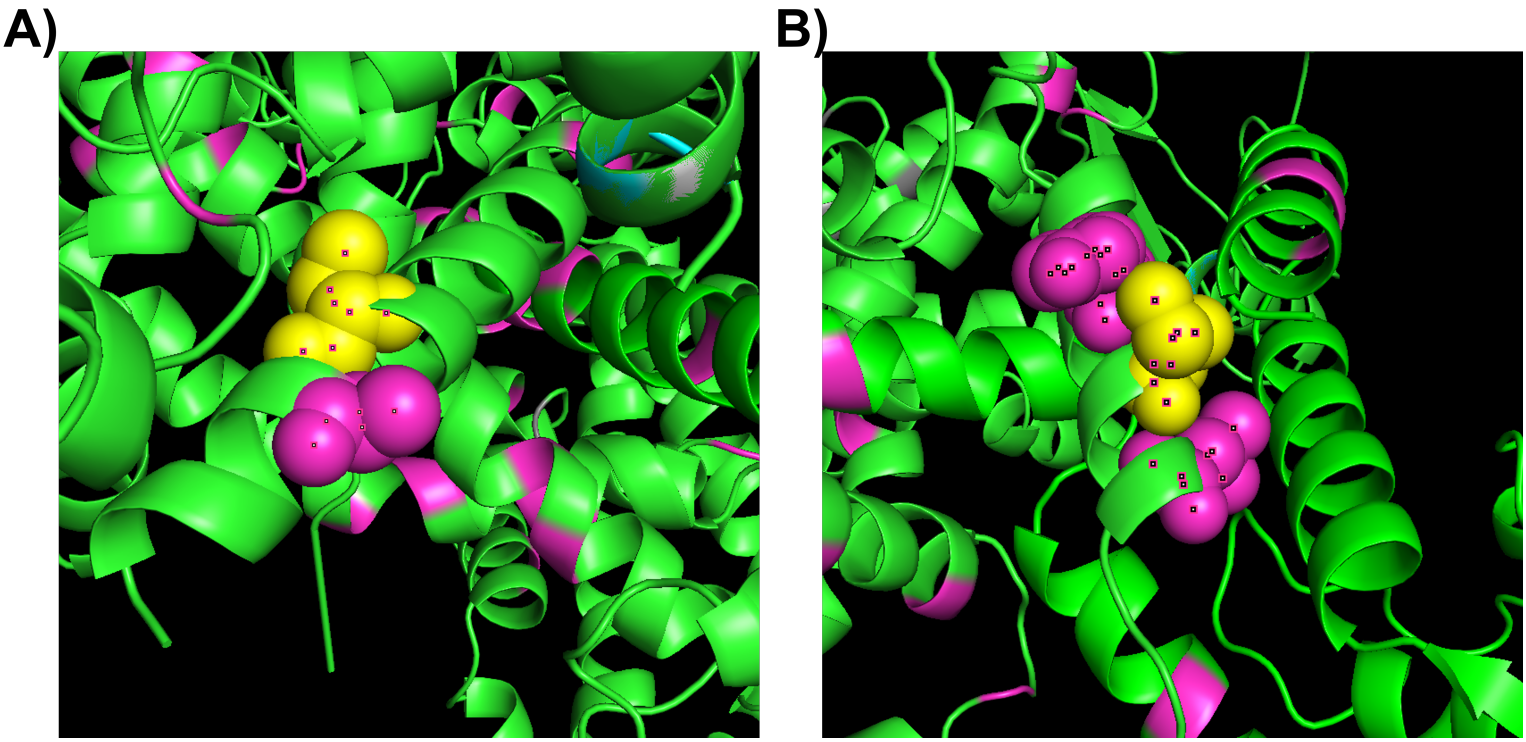


Figure 4.17: Zoomed-in view of the structural positions of A) SCN5A Ala425 with SCN2A Thr435; and B) SCN5A Tyr1495(upper part of helix)/Leu1501(lower part of helix) with SCN2A Met1501. SCN5A variants are shown in magenta while SCN2A variants are shown in yellow.

4.3.7 Future works

At the moment I have only described how using structural alignments could potentially provide more information that could aid in interpreting variants beyond sequence alignment methods like Paralogue Annotation or Pfam metadomains. But in order to be used as a predictive

classifier, more quantifiable aspects need to be formulated.

The idea of exploring the local structural proximity around a residue of interest has already been well established prior. For example, many researchers, such as Homburger et al. (2016), have already used the method of expanding 3D search spaces, or "spheres", around a structural point to look for pathogenic hot spots. Their research was primarily focused on surface-exposed residues, but that does raise the interesting idea of possibly grouping variants initially on whether they are located on the surface of the protein or within the core. As depending on the protein, surface variants and variants at the core can have differentiating effects. From there, a similar 3D search space for neighbouring pathogenic variants can be initiated. For this framework however, the novelty comes with searching for variants across the structural alignment within a different protein to the one the query variant is in. Further statistical methods can then also be employed to improve the reliability of the predictive call, such as using $k > 1$ search parameters during the k-nearest neighbour search.

Additionally, I have only looked at a specific case study between SCN5A and SCN2A. More thorough testing en masse on further proteins will be required to investigate whether this framework is viable and can provide information that can help benefit variant interpretation overall, or if it can only function for a select group of proteins. Further testing is particularly needed for variants that are potentially close to each other in structure but distant in sequence, as none of the SCN5A query variants that had nearby SCN2A reference variants fell into this category.

In regards to structures to be used for alignments, the predicted models provided by Alphafold 2 in addition to the modifications to the current workflow pipeline as described previously in section 4.3.4 would be ideal. Alphafold 2 currently covers about 98.5% of human proteome (Tunyasuvunakool et al., 2021). Proteins that do not have predicted structures are ones that exist in extreme cases, such as extremely short (<16 residues) or long (>2700

residues) proteins, and proteins incorporating non-standard amino acids. Likely, however, Alphafold will eventually provide structural predictions for these proteins as well sometime in the future.

Therefore, as time progresses the lack of structural coverage of the human proteome shrinks. Whether it be by advances in experimental techniques that derive real structures or by more model structures predicted *in silico*. The ultimate goal being that all proteins having experimentally derived structures. We are not there yet and so have to rely on predictions such as those provided by Alphafold 2.

Ultimately, being able to classify and provide a reliable predictive pathogenic interpretation on a query variant would be the desired outcome, but given that this framework will rely on the quality of either empirically solved structures or predicted structures, and also the quality of alignment, there are possibly too many assumptions to be made for this to function as a high fidelity predictor for the clinical outcome of a variant. Instead considering it as an additional tool that other researchers can use to search and identify possible variants of interest that can be considered similar or equivalent to their own variants under investigation may be more practical. This would be useful in the scenario where no other documented variants can be used as a referential comparison, particularly in the case of many novel variants.

4.4 Conclusion

To conclude, in this chapter I have explored how to potentially expand the Paralogue Annotation/Pfam metadomain framework from a sequence alignment based method to one that instead uses structural alignments in hopes of overcoming some of the sensitivity limitations that the prior methods encounter. In doing so, more reference variants can be searched for and thereby theoretically enhancing sensitivity. I have shown what possible resources and data currently available can be used to make such a concept viable.

By expanding the dimensionality of the framework, from a discrete sequential search space to a continuous structural one, the sensitivity may indeed increase but that would naturally come at a cost of precision. Currently, the structural framework is unlikely to be usable as a reliable predictive classifier, but it can still offer useful information regarding reference variants in close proximity.

More work needs to be done to thoroughly examine whether the underlying assumptions, that being the structural alignment framework and the use of predicted protein models, can be used reliably as a pathogenic variant classifier akin to Paralogue Annotation/Pfam metadomains.

Chapter 5

Concluding Remarks

Advances in the field of sequencing technology have meant that over the last several decades there has been a surge of genetic and genomic data at tremendous volumes. This will undoubtedly continue for the years to come. With that data, comes a lot of questions that are naturally asked by the scientific community. In particular, the one key question being asked within clinical genetics, and the whole premise that this PhD was set on, is deciphering which single nucleotide variants are disease causing, or at least likely to be contributing to disease, and distinguishing those from variants that are rare but not pathogenic. As more variants are continuously being discovered, this question will continue to be asked.

In Chapter 1, I described some of the *in silico* tools used for predicting the clinical consequences of variants when experimental assays are not the desired choice of action. I also went through some of problems these tools encounter, such as over predicting pathogenicity due to over fitting of data, particularly for novel variants with little to no prior information.

In that respect, what the framework that was initially started by Paralogue Annotation (Chapter 2) does is twofold. Firstly, it tries to mirror what the other tools do in being a classifier for pathogenic variants, and statistically it performs quite well in that when providing a pathogenic prediction, users can be very confident that the prediction for the variant in

question is indeed correct. In the clinical context, this is probably the most desirable outcome rather than having to filter out a surplus of false positives.

By not being a contemporary machine learning algorithm, Paralogue Annotation does not fall into the same pitfalls although there are some similarities. The reference variants being used to transfer annotations from could be considered analogous to the training data set but the fact that the prediction of results are not solely based on the reference variants, but also on the alignments, help offset any over reliance or over fitting of the reference variant dataset.

The second aspect of the framework is to search for and provide functionally equivalent variants that may be used to improve the training of the machine learnt prediction tools, or researchers could use these variants themselves in other ways, such as in comparative studies. Though statistically tested as a classifier, at the very least this framework can be considered as a search tool, and it is this second usage that is likely to be the most useful for researchers.

The thought process behind considering ways of improving Paralogue Annotation while maintaining the core concept of the framework contemplated on the native state of proteins and how they function *in vivo*. The most apparent solution was to alter the sequence alignments to focus more on the regions responsible for function, that is to say protein domains and thereby the use of the Pfam metadomains (Chapter 3), instead of the entire protein sequence. The transition into structure space from sequence (Chapter 4) was again an attempt to more naturally reflect the way in which proteins function and exist in the cell. The idea is that the closer the framework is to modelling *in vivo* protein comparison the more informative and reliable variants deemed as being functionally equivalent will be. **Table 5.1** summarises the three ways in which I have defined equivalency for variants of interest throughout this thesis.

Clearly however, there is still much more that can be explored and as more data arises, our understanding of the variant landscape, including the variety of diseases and genes involved, will further improve. For instance, a lot of focus on the genetic basis of diseases has been on protein

coding variants. Rightfully so as their consequences are more observable due to the direct effects on protein products. But the importance of variants affecting non-coding regions are becoming more apparent, particularly ones positioned at regulatory/enhancer regions (French and Edwards, 2020). Applying a similar framework to the one described in this thesis may be possible but whether it would function as a high precision classifier would require thorough examination. There would clearly be no structural alignments and the sequence alignments would have to rely on nucleotide sequences instead of protein. Transfer of pathogenicity or disease annotations would have to be replaced by more functional information instead, such as changes to gene expression levels.

Either way I believe the core idea of the framework, that is to say the concept of finding functional equivalency as a way of comparing aspects being studied, is a simple one and intuitive. I am sure in time the progress within other scientific endeavours will produce greater advances that can be made to the application of theories outlined in this thesis and elsewhere.

	Paralogue Annotation	Pfam Metadomains	Structural alignments
type of data	paralogue sequences	protein domain sequences	protein structures
current data availability	sufficient empirical data	sufficient empirical data	limited empirical data; predicted models required
alignment mechanism	discrete binary	discrete binary	continuous
search space dimensions	1 dimensional	1 dimensional	3 dimensional
statistical performance as a classifier	high precision; low sensitivity	marginally lower precision; higher sensitivity	untested; hypothetically higher sensitivity
unique capabilities	can potentially offer alignments for inter-domain linker/random loop regions that have no apparent structure	limited to intra-domain regions but for positions within domains can offer greater sensitivity	can theorectically detect long range interacting variants that are sequentially distant but structurally close

Table 5.1: Table summarising the three classes of the equivalent variant framework - Paralogue Annotation; Pfam metadomains; and using structural alignments - and their unique aspects and properties.

Chapter 6

Tools, Code, and Publications arising from this PhD

6.1 Paralogue Annotator Website

The Paralogue Annotator web application, available at https://www.cardiodb.org/paralog_app, allows any researcher a quick and easy way of using Paralogue Annotation for their own variants of interest (**Fig. 6.1**).

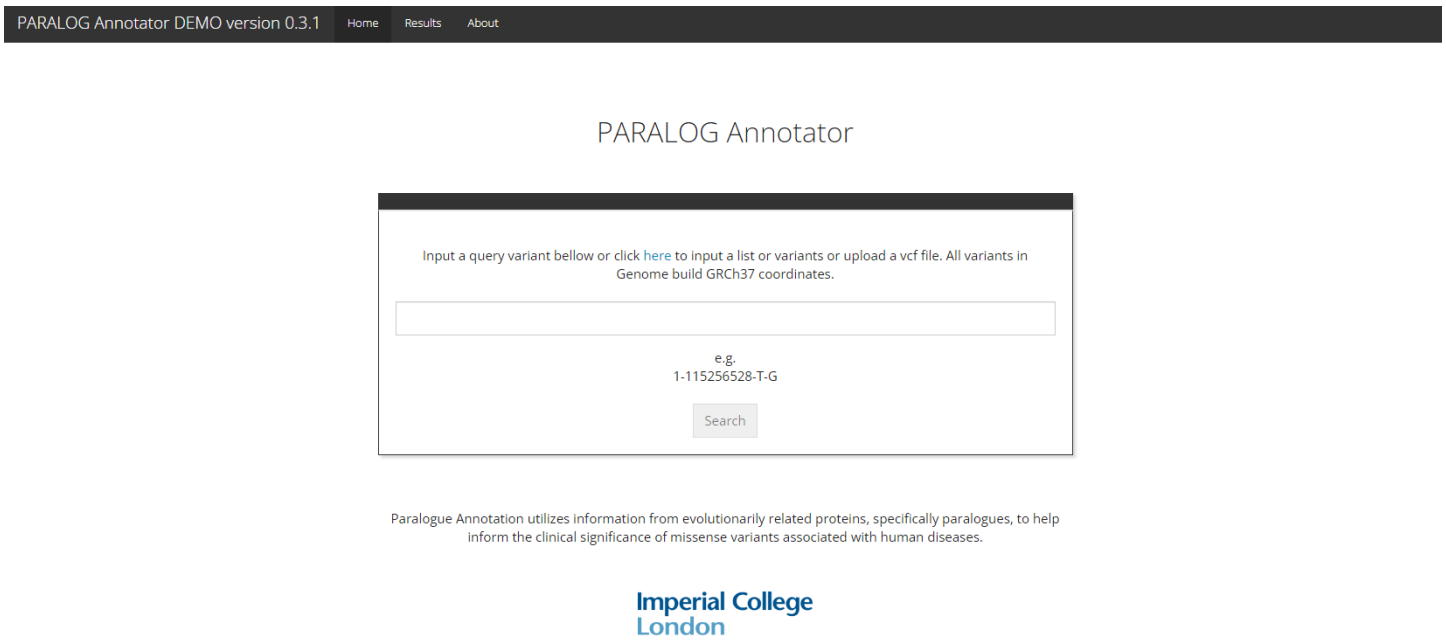


Figure 6.1: Front page of the Paralogue Annotator web application available at https://www.cardiodb.org/paralog_app. Users can input multiple variants of their choice by providing the chromosome, position, reference allele and alternate allele.

6.2 Github repositories

6.2.1 VEP Plugin

For researchers who would like more direct access to code behind identifying equivalent paralogous positions described in Chapter 2, the VEP plugin script offers more control over VEP parameters and results generated. The plugin can be downloaded at https://github.com/ImperialCardioGenetics/paralogueAnnotator.

6.2.2 Paralogue Annotation analysis

Apart from the plugin, all other code and data used in the analysis shown in Chapter 2 can be accessed at https://github.com/ImperialCardioGenetics/ParalogueAnnotation_personal.

6.2.3 Pfam metadomain analysis

All code and data used in the analysis shown in Chapter 3 can be accessed at https://github.com/ImperialCardioGenetics/Pfam_meta_domains

6.2.4 Structural alignment analysis

All code and data used in the analysis shown in Chapter 4 can be accessed at https://github.com/nli8888/3D_Paralogue_Annotation.

6.3 List of Publications

1. **Nicholas Li**, Erica Mazaika, Pantazis Theotokis, Mikyung Jang, Xiaolei Zhang, Mian Ahmad, Paul Barton, Roddy Walsh, Nicola Whiffin, and James S. Ware. Paralogue Annotation identifies pathogenic missense variants with high precision on an Exome-wide scale. [**In preparations for submission to *Genome Medicine***], 2022

2. Erica Mazaika, **Nicholas Li**, Roddy Walsh, Nicola Whiffin, Mian Ahmed, Fiona Cunningham, Paul JR Barton, and James S. Ware. ParalogueAnnotator: a plug-in for the Ensembl Variant Effect Predictor to identify disease-associated variation across protein-families. [**In preparations for submission to *Genome Medicine***], 2022

3. Xiaolei Zhang, Pantazis Theotokis, **Nicholas Li**, Caroline Wright, Kaitlin Samocha, Nicola Whiffin, and James S Ware. Genetic constraint at single amino acid resolution improves missense variant prioritisation and gene discovery. [**Under Review at *Nature Communications***], 2022

4. Xiaolei Zhang, Roddy Walsh, Nicola Whiffin, Rachel Buchan, William Midwinter, Alicja Wilk, Risha Govind, **Nicholas Li**, Mian Ahmad, Francesco Mazzarotto, et al. Disease-

specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. Genetics in Medicine, 23(1):69–79, 2021

5. Roddy Walsh, Francesco Mazzarotto, Nicola Whiffin, Rachel Buchan, William Midwinter, Alicja Wilk, **Nicholas Li**, Leanne Felkin, Nathan Ingold, Risha Govind, et al. Quantitative approaches to variant classification increase the yield and precision of genetic testing in mendelian diseases: the case of hypertrophic cardiomyopathy. Genome medicine, 11(1):1–18, 2019

6.4 Conference presentations

1. ePoster presentation at the European Society of Human Genetics (ESHG) virtual conference, 2020: *High precision exome-wide identification of pathogenic variants by Paralogue Annotation*

2. Poster presentation at the conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB), Basel, 2019: *Exome wide identification of pathogenic variants by Paralogue Annotation*

List of Acronyms

ACMG-AMP American College of Medical Genetics and Genomics and the Association for Molecular Pathology

AGT/MGMT O^6-alkylguanine-DNA alkyltransferase

APOE Apolipoprotein E

AUC area under the curve

B/LB Benign and/or Likely benign

BER Base Excision Repair

BrS Brugada Syndrome

CASP Critical Assessment of protein Structure Prediction

CNV Copy Number Variants

CPVT Catecholaminergic Polymorphic Ventricular Tachycardia

DNA Deoxyribonucleic Acid

DNM *de novo* mutations

EM Electron Microscopy

ETC electron transport chain

FN False Negative

FP False Positive

GnomAD The Genome Aggregation Database

GO Gene Ontology

GOF Gain-of-Function

GWAS genome-wide association studies

HCM Hypertrophic Cardiomyopathy

IR ionizing radiation

KDE Kernel Density Estimation

KS Kolmogorov-Smirnov

LOF Loss-of-Function

LQTS Long QT Syndrome

MSA Multiple Sequence Alignment

NMR Nuclear Magnetic Resonance

P/LP Pathogenic and/or Likely pathogenic

PDB Protein Data Bank

PPV Positive Predictive Value

PTC premature termination codon

ROS reactive oxygen species

RMSD Root Mean Square Deviation

SNV Single Nucleotide Variant

TN True Negative

TP True Positive

UV ultraviolet

VCF Variant Call Format

VEP Variant Effect Predictor

VUS variant of uncertain significance

Bibliography

Rocio Acuna-Hidalgo, Joris A Veltman, and Alexander Hoischen. New insights into the generation and role of de novo mutations in health and disease. *Genome biology*, 17(1): 1–19, 2016.

Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248, 2010.

Bruce Alberts. *Molecular biology of the cell*. WW Norton & Company, 2017.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter, et al. Molecular biology of the cell. *SCANDINAVIAN JOURNAL OF RHEUMATOLOGY*, 32(2):125–125, 2003.

Genetic Alliance, New York-Mid-Atlantic Consortium for Genetic, Newborn Screening Services, et al. Understanding genetics: a new york, mid-atlantic guide for patients and health professionals. 2009.

Mohammed AlQuraishi. Alphafold at casp13. *Bioinformatics*, 35(22):4862–4865, 2019.

Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990.

Abramowicz Anna and Gos Monika. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *Journal of applied genetics*, 59(3):253–268, 2018.

Tania A Baker, Stephen P Bell, Alexander Gann, Michael Levine, Richard Losick, and CHSLP Inglis. *Molecular biology of the gene*. San Francisco, CA, USA:: Pearson/Benjamin Cummings, 2008.

Zsolt Bánfai, Kinga Hadzsiev, Endre Pál, Katalin Komlósi, Márton Melegh, László Balikó, and Béla Melegh. Novel phenotypic variant in the myh7 spectrum due to a stop-loss mutation in the c-terminal region: a case report. *BMC medical genetics*, 18(1):1–7, 2017.

Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

Lorenzo Bomba, Klaudia Walter, and Nicole Soranzo. The impact of rare and low-frequency genetic variants in common disease. *Genome biology*, 18(1):1–17, 2017.

Scott Brady. *Basic neurochemistry: molecular, cellular and medical aspects*. Elsevier, 2005.

Stephen K Burley, Charmi Bhikadiya, Chunxiao Bi, Sebastian Bittrich, Li Chen, Gregg V Crichlow, Cole H Christie, Kenneth Dalenberg, Luigi Di Costanzo, Jose M Duarte, et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49 (D1):D437–D451, 2021.

Catarina D Campbell and Evan E Eichler. Properties and rates of germline mutations in humans. *Trends in Genetics*, 29(10):575–584, 2013.

Claudia Carvalho and James R Lupski. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238, 2016.

Sandeep Chakraborty, Ravindra Venkatramani, Basuthkar J Rao, Bjarni Asgeirsson, and

Abhaya M Dandekar. Protein structure quality assessment based on the distance profiles of consecutive backbone cα atoms. *F1000Research*, 2, 2013.

Nimrat Chatterjee and Graham C Walker. Mechanisms of dna damage, repair, and mutagenesis. *Environmental and molecular mutagenesis*, 58(5):235–263, 2017.

Kevin Chu. An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine*, 11(3):175–181, 1999.

Elizabeth T Cirulli, Erin L Heinzen, Fred S Dietrich, Kevin V Shianna, Abanish Singh, Jessica M Maia, James J Goedert, and David B Goldstein. A whole-genome analysis of premature termination codons. *Genomics*, 98(5):337–342, 2011.

Suzanne Clancy. Dna damage & repair: mechanisms for maintaining dna integrity. *Nature Education*, 1(1):103, 2008.

Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, et al. A structural variation reference for medical and population genetics. *Nature*, 581 (7809):444–451, 2020.

1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Johanna Craig et al. Complex diseases: research and applications. *Nature Education*, 1(1):184, 2008.

Christopher DeBoever, Yosuke Tanigawa, Malene E Lindholm, Greg McInnes, Adam Lavertu, Erik Ingelsson, Chris Chang, Euan A Ashley, Carlos D Bustamante, Mark J Daly, et al. Medical relevance of protein-truncating variants across 337,205 individuals in the uk biobank study. *Nature communications*, 9(1):1–10, 2018.

Frederick E Dewey, Michael F Murray, John D Overton, Lukas Habegger, Joseph B Leader, Samantha N Fetterolf, Colm O'Dushlaine, Cristopher V Van Hout, Jeffrey Staples, Claudia Gonzaga-Jauregui, et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the discovehr study. *Science*, 354(6319), 2016.

Bruce K Duncan and Jeffrey H Miller. Mutagenic deamination of cytosine residues in dna. *Nature*, 287(5782):560–561, 1980.

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

Robert C Edgar and Serafim Batzoglou. Multiple sequence alignment. *Current opinion in structural biology*, 16(3):368–373, 2006.

Robert P Erickson. Somatic gene mutation and human disease other than cancer. *Mutation Research/Reviews in Mutation Research*, 543(2):125–136, 2003.

Corinna Ernst, Eric Hahnen, Christoph Engel, Michael Nothnagel, Jonas Weber, Rita K. Schmutzler, and Jan Hauke. Performance of in silico prediction tools for the classification of rare brca1/2 missense variants in clinical diagnostics. *BMC medical genomics*, 11(1):35, 2018.

Sarah E Flanagan, Ann-Marie Patch, and Sian Ellard. Using sift and polyphen to predict loss-of-function and gain-of-function mutations. *Genetic testing and molecular biomarkers*, 14(4):533–537, 2010.

Allan Force, Michael Lynch, F Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151 (4):1531–1545, 1999.

JD French and SL Edwards. The role of noncoding variants in heritable disease. *Trends in Genetics*, 36(11):880–891, 2020.

Jianjiong Gao, Matthew T Chang, Hannah C Johnsen, Sizhi Paul Gao, Brooke E Sylvester, Selcuk Onur Sumer, Hongxin Zhang, David B Solit, Barry S Taylor, Nikolaus Schultz, et al. 3d clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome medicine*, 9(1):4, 2017.

Bruce D Gelb, Hélène Cavé, Mitchell W Dillon, Karen W Gripp, Jennifer A Lee, Heather Mason-Suares, Katherine A Rauen, Bradley Williams, Martin Zenker, and Lisa M Vincent. Clingen's rasopathy expert panel consensus methods for variant interpretation. *Genetics in Medicine*, 20(11):1334–1345, 2018.

M Grace Gordon, Fumitaka Inoue, Beth Martin, Max Schubach, Vikram Agarwal, Sean Whalen, Shiyun Feng, Jingjing Zhao, Tal Ashuach, Ryan Ziffra, et al. lentimpra and mpraflow for high-throughput functional characterization of gene regulatory elements. *Nature protocols*, 15(8):2387–2412, 2020.

Valer Gotea, Jared J Gartner, Nouar Qutob, Laura Elnitski, and Yardena Samuels. The functional relevance of somatic synonymous mutations in melanoma and other cancers. *Pigment cell & melanoma research*, 28(6):673–684, 2015.

John F Griffiths, Anthony JF Griffiths, Susan R Wessler, Richard C Lewontin, William M Gelbart, David T Suzuki, Jeffrey H Miller, et al. *An introduction to genetic analysis*. Macmillan, 2005.

Dominik G Grimm, Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G MacArthur, Kaitlin E Samocha, David N Cooper, Peter D Stenson, Mark J Daly, Jordan W Smoller, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human mutation*, 36(5):513–523, 2015.

Adam C Gunning, Verity Fryer, James Fasham, Andrew H Crosby, Sian Ellard, Emma L Baple, and Caroline F Wright. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *Journal of medical genetics*, 58(8):547–555, 2021.

Tibshirani Hastie, R Tibshirani, and J Friedman. The elements of statistical learning new york. *NY: Springer*, 2009.

Lizbeth Hedstrom. Serine protease mechanism and specificity. *Chemical reviews*, 102(12): 4501–4524, 2002.

Ernst S Henle and Stuart Linn. Formation, prevention, and repair of dna damage by iron/hydrogen peroxide. *Journal of Biological Chemistry*, 272(31):19095–19098, 1997.

Henrike O Heyne, Mykyta Artomov, Florian Battke, Claudia Bianchini, Douglas R Smith, Nora Liebmann, Vasisht Tadigotla, Christine M Stanley, Dennis Lal, Heidi Rehm, et al. Targeted gene sequencing in 6994 individuals with neurodevelopmental disorder with epilepsy. *Genetics in Medicine*, 21(11):2496–2503, 2019.

Julian R Homburger, Eric M Green, Colleen Caleshu, Margaret S Sunitha, Rebecca E Taylor, Kathleen M Ruppel, Raghu Prasad Rao Metpally, Steven D Colan, Michelle Michels, Sharlene M Day, et al. Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proceedings of the National Academy of Sciences*, 113(24):6701–6706, 2016.

Benjamin E Housden, Matthias Muhar, Matthew Gemberling, Charles A Gersbach, Didier YR Stainier, Geraldine Seydoux, Stephanie E Mohr, Johannes Zuber, and Norbert Perrimon. Loss-of-function genetic tools for animal models: cross-species and cross-platform differences. *Nature Reviews Genetics*, 18(1):24–40, 2017.

Austin L Hughes, Robert Friedman, Pierre Rivailler, and Jeffrey O French. Synonymous and

nonsynonymous polymorphisms versus divergences in bacterial genomes. *Molecular biology and evolution*, 25(10):2199–2209, 2008.

Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, et al. Revel: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016.

Maria Jackson, Leah Marks, Gerhard HW May, and Joanna B Wilson. The genetic basis of disease. *Essays in biochemistry*, 62(5):643–723, 2018.

Karthik A Jagadeesh, Aaron M Wenger, Mark J Berger, Harendra Guturu, Peter D Stenson, David N Cooper, Jonathan A Bernstein, and Gill Bejerano. M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics*, 48 (12):1581, 2016.

Celia Andrés Juan, José Manuel Pérez de la Lastra, Francisco J Plou, and Eduardo Pérez-Lebeña. The chemistry of reactive oxygen species (ros) revisited: outlining their role in biological macromolecules (dna, lipids and proteins) and induced pathologies. *International Journal of Molecular Sciences*, 22(9):4642, 2021.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Bernd Kaina, Markus Christmann, Steffen Naumann, and Wynand P Roos. Mgmt: key node in the battle against genotoxicity, carcinogenicity and apoptosis induced by alkylating agents. *DNA repair*, 6(8):1079–1099, 2007.

Jamie D Kapplinger, David J Tester, Benjamin A Salisbury, Janet L Carr, Carole Harris-Kerr, Guido D Pollevick, Arthur AM Wilde, and Michael J Ackerman. Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the familion® long qt syndrome genetic test. *Heart rhythm*, 6(9):1297–1303, 2009.

Jamie D Kapplinger, David J Tester, Marielle Alders, Begoña Benito, Myriam Berthet, Josep Brugada, Pedro Brugada, Véronique Fressart, Alejandra Guerchicoff, Carole Harris-Kerr, et al. An international compendium of mutations in the scn5a-encoded cardiac sodium channel in patients referred for brugada syndrome genetic testing. *Heart rhythm*, 7(1):33–46, 2010.

Jamie D Kapplinger, John R Giudicessi, Dan Ye, David J Tester, Thomas E Callis, Carmen R Valdivia, Jonathan C Makielski, Arthur A Wilde, and Michael J Ackerman. Enhanced classification of brugada syndrome–associated and long-qt syndrome–associated genetic variants in the scn5a-encoded nav1. 5 cardiac sodium channel. *Circulation: Cardiovascular Genetics*, 8(4):582–595, 2015a.

Jamie D Kapplinger, Andrew S Tseng, Benjamin A Salisbury, David J Tester, Thomas E Callis, Marielle Alders, Arthur AM Wilde, and Michael J Ackerman. Enhancing the predictive power of mutations in the c-terminus of the kcnq1-encoded kv7. 1 voltage-gated potassium channel. *Journal of cardiovascular translational research*, 8(3):187–197, 2015b.

Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

Panagiotis Katsonis, Amanda Koire, Stephen Joseph Wilson, Teng-Kuei Hsu, Rhonald C Lua,

Angela Dawn Wilkins, and Olivier Lichtarge. Single nucleotide variations: biological impact and theoretical interpretation. *Protein Science*, 23(12):1650–1666, 2014.

JA Kearney, NW Plummer, MR Smith, J Kapur, TR Cummins, SG Waxman, AL Goldin, and MH Meisler. A gain-of-function mutation in the sodium channel gene scn2a results in seizures and behavioral abnormalities. *Neuroscience*, 102(2):307–317, 2001.

Lawrence A Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael JE Sternberg. The phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6):845–858, 2015.

Carsten Kemena and Cedric Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19):2455–2465, 2009.

Rhoda J Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, et al. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, 2011, 2011.

Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014.

Judith Klein-Seetharaman, Maki Oikawa, Shaun B Grimshaw, Julia Wirmer, Elke Duchardt, Tadashi Ueda, Taiji Imoto, Lorna J Smith, Christopher M Dobson, and Harald Schwalbe. Long-range interactions within a nonnative protein. *Science*, 295(5560):1719–1722, 2002.

Eva Klopocki, Naveed Wasif, Muhammad Tariq, Saadullah Khan, Jochen Hecht, Peter Krawitz, Stefan Mundlos, Wasim Ahmad, et al. Whole exome sequencing identified a novel zinc-finger gene znf141 associated with autosomal recessive postaxial polydactyly type a. *Journal of medical genetics*, 50(1):47–53, 2013.

Eugene V Koonin and Artem S Novozhilov. Origin and evolution of the genetic code: the universal enigma. *IUBMB life*, 61(2):99–111, 2009.

Evgeny Krissinel. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics*, 23(6):717–723, 2007.

Hans E Krokan and Magnar Bjørås. Base excision repair. *Cold Spring Harbor perspectives in biology*, 5(4):a012583, 2013.

Irina Kufareva and Ruben Abagyan. Methods of protein structure comparison. In *Homology Modeling*, pages 231–257. Springer, 2011.

John H Laity, Brian M Lee, and Peter E Wright. Zinc finger proteins: new insights into structural and functional diversity. *Current opinion in structural biology*, 11(1):39–46, 2001.

Dennis Lal, Patrick May, Eduardo Perez-Palma, Kaitlin E Samocha, Jack A Kosmicki, Elise B Robinson, Rikke S Møller, Roland Krause, Peter Nürnberg, Sarah Weckhuysen, et al. Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome medicine*, 12(1):1–12, 2020.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44 (D1):D862–D868, 2015.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

Michal Levy-Sakin, Steven Pastor, Yulia Mostovoy, Le Li, Alden KY Leung, Jennifer McCaffrey, Eleanor Young, Ernest T Lam, Alex R Hastie, Karen HY Wong, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature communications*, 10(1):1–14, 2019.

Jinchen Li, Tingting Zhao, Yi Zhang, Kun Zhang, Leisheng Shi, Yun Chen, Xingxing Wang, and Zhongsheng Sun. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic acids research*, 46(15):7793–7804, 2018.

Shuai Cheng Li. The difficulty of protein structure alignment under the rmsd. *Algorithms for Molecular Biology*, 8(1):1–9, 2013.

Xiao-Han Li and M Madan Babu. Human diseases from gain-of-function mutations in disordered protein regions. *Cell*, 175(1):40–42, 2018.

Xiaoming Liu, Chunlei Wu, Chang Li, and Eric Boerwinkle. dbnsfp v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs. *Human mutation*, 37(3):235–241, 2016.

Yuxi Liu, Duc T Huynh, and Todd O Yeates. A 3.8 å resolution cryo-em structure of a small protein bound to an imaging scaffold. *Nature communications*, 10(1):1–7, 2019.

A Liwo, RJ Wawak, HA Scheraga, MR Pincus, and S Rackovsky. Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment. *Protein Science*, 2(10):1697–1714, 1993.

Ingrid Lobo. Same genetic mutation, different genetic disease phenotype. *Nature Education*, 1 (1):64, 2008.

Laurence Loewe. Genetic mutation. *Nature education*, 1(1):113, 2008.

Gildas Loussouarn, Damien Sternberg, Sophie Nicole, Céline Marionneau, Francoise Le Bouffant, Gilles Toumaniantz, Julien Barc, Olfat A Malak, Véronique Fressart, Yann Péréon, et al. Physiological and pathophysiological insights of nav1. 4 and nav1. 5 comparison. *Frontiers in pharmacology*, 6:314, 2016.

Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian RN Tivey, Simon C Potter, Robert D Finn, et al. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636–W641, 2019.

Khalid Mahmood, Chol-hee Jung, Gayle Philip, Peter Georgeson, Jessica Chung, Bernard J. Pope, and Daniel J. Park. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Human genomics*, 11(1):10, 2017.

Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck. Structural variant calling: the long and the short of it. *Genome biology*, 20(1):1–14, 2019.

Emmanuelle Masson, Wen-Bin Zou, Emmanuelle Génin, David N Cooper, Gerald Le Gac, Yann Fichou, Na Pu, Vinciane Rebours, Claude Férec, Zhuan Liao, et al. Expanding acmg variant classification guidelines into a general framework. *Human Genomics*, 16(1):1–15, 2022.

Ifigeneia V Mavragani, Zacharenia Nikitaki, Spyridon A Kalospyros, and Alexandros G Georgakilas. Ionizing radiation and complex dna damage: from prediction to detection challenges and biological significance. *Cancers*, 11(11):1789, 2019.

Abdelghani Mazouzi, Georgia Velimezi, and Joanna I Loizou. Dna replication stress: causes, resolution and disease. *Experimental cell research*, 329(1):85–93, 2014.

Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):1–14, 2016.

Michael L Metzker. Sequencing technologies—the next generation. *Nature reviews genetics*, 11 (1):31–46, 2010.

Katrina Meyer, Marieluise Kirchner, Bora Uyar, Jing-Yuan Cheng, Giulia Russo, Luis R Hernandez-Miranda, Anna Szymborska, Henrik Zauber, Ina-Maria Rudolph, Thomas E Willnow, et al. Mutations in disordered regions can cause disease by creating dileucine motifs. *Cell*, 175(1):239–253, 2018.

Huaiyu Mi, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent-Philippe Albou, Tremayne Mushayamaha, and Paul D Thomas. Panther version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive api. *Nucleic acids research*, 49(D1):D394–D403, 2021.

Lisa A Miosge, Matthew A Field, Yovina Sontani, Vicky Cho, Simon Johnson, Anna Palkova, Bhavani Balakishnan, Rong Liang, Yafei Zhang, Stephen Lyon, et al. Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences*, 112(37):E5189–E5198, 2015.

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al.

Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021.

Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

P Mokarram, M Zamani, S Kavousipour, F Naghibalhossaini, C Irajie, M Moradi Sarabi, and SV Hosseini. Different patterns of dna methylation of the two distinct o6-methylguanine-dna methyltransferase (o 6-mgmt) promoter regions in colorectal cancer. *Molecular biology reports*, 40(5):3851–3857, 2013.

Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

So I Nagaoka, Terry J Hassold, and Patricia A Hunt. Human aneuploidy: mechanisms and new insights into an age-old problem. *Nature Reviews Genetics*, 13(7):493–504, 2012.

Pauline C Ng and Steven Henikoff. Predicting deleterious amino acid substitutions. *Genome research*, 11(5):863–874, 2001.

Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 2013.

Andrea M Oza, Marina T DiStefano, Sarah E Hemphill, Brandon J Cushman, Andrew R Grant, Rebecca K Siegert, Jun Shen, Alex Chapin, Nicole J Boczek, Lisa A Schimmenti, et al. Expert specification of the acmg/amp variant interpretation guidelines for genetic hearing loss. *Human mutation*, 39(11):1593–1613, 2018.

Andy W Pang, Jeffrey R MacDonald, Dalila Pinto, John Wei, Muhammad A Rafiq, Donald F Conrad, Hansoo Park, Matthew E Hurles, Charles Lee, J Craig Venter, et al. Towards a

comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5):1–14, 2010.

Elena Papaleo, Giorgio Saladino, Matteo Lambrughi, Kresten Lindorff-Larsen, Francesco Luigi Gervasio, and Ruth Nussinov. The role of protein loops and linkers in conformational dynamics and allostery. *Chemical reviews*, 116(11):6391–6423, 2016.

Leslie Pray. Dna replication and causes of mutation. *Nature education*, 1(1):214, 2008.

Angela P Presson, Ginger Partyka, Kristin M Jensen, Owen J Devine, Sonja A Rasmussen, Linda L McCabe, and Edward RB McCabe. Current estimate of down syndrome population prevalence in the united states. *The Journal of pediatrics*, 163(4):1163–1168, 2013.

Marta Puig, Sonia Casillas, Sergi Villatoro, and Mario Cáceres. Human inversions and their functional consequences. *Briefings in functional genomics*, 14(5):369–379, 2015.

Vasily Ramensky, Peer Bork, and Shamil Sunyaev. Human non-synonymous snps: server and survey. *Nucleic acids research*, 30(17):3894–3900, 2002.

Carol Ann Remme. Cardiac sodium channelopathy associated with scn5a mutations: electrophysiological, molecular and genetic aspects. *The Journal of physiology*, 591(17):4099–4116, 2013.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, and Elaine Spector. 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. *Genetics in medicine*, 17(5):405, 2015.

Jared C Roach, Gustavo Glusman, Arian FA Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, et al. Analysis

of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978): 636–639, 2010.

Saeed Sadigh-Eteghad, Mahnaz Talebi, and Mehdi Farhoudi. Association of apolipoprotein e epsilon 4 allele with sporadic late onset alzheimer's disease. *A meta-analysis. Neurosciences (Riyadh)*, 17(4):321–326, 2012.

Stephan J Sanders, Arthur J Campbell, Jeffrey R Cottrell, Rikke S Moller, Florence F Wagner, Angie L Auldridge, Raphael A Bernier, William A Catterall, Wendy K Chung, James R Empfield, et al. Progress in understanding and treating scn2a-mediated disorders. *Trends in neurosciences*, 41(7):442–456, 2018.

Georgia Sarquella-Brugada, Oscar Campuzano, Elena Arbelo, Josep Brugada, and Ramon Brugada. Brugada syndrome: clinical and genetic findings. *Genetics in Medicine*, 18(1): 3–12, 2016.

Zuben E Sauna and Chava Kimchi-Sarfaty. Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10):683–691, 2011.

AH Schiemann and KM Stowell. Comparison of pathogenicity prediction tools on missense variants in ryr1 and cacna1s associated with malignant hyperthermia. *BJA: British Journal of Anaesthesia*, 117(1):124–128, 2016.

Schrödinger, LLC. The PyMOL molecular graphics system, version 2.0. August 2021.

Peter J Schwartz, Marco Stramba-Badiale, Lia Crotti, Matteo Pedrazzini, Alessandra Besana, Giuliano Bosi, Fulvio Gabbarini, Karine Goulene, Roberto Insolia, Savina Mannarino, et al. Prevalence of the congenital long-qt syndrome. *Circulation*, 120(18):1761–1767, 2009.

Diana Shy, Ludovic Gillet, and Hugues Abriel. Cardiac sodium channel nav1. 5 distribution in

myocytes via interacting proteins: the multiple pool model. *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research*, 1833(4):886–894, 2013.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

Shannon Stefl, Hafumi Nishi, Marharyta Petukh, Anna R Panchenko, and Emil Alexov. Molecular mechanisms of disease-causing missense mutations. *Journal of molecular biology*, 425(21):3919–3936, 2013.

The Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542(7642):433, 2017.

Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

Toshihiko Sugiki, Naohiro Kobayashi, and Toshimichi Fujiwara. Modern technologies of solution nuclear magnetic resonance spectroscopy for three-dimensional structure determination of proteins open avenues for life scientists. *Computational and structural biotechnology journal*, 15:328–339, 2017.

Julie D Thompson, Benjamin Linard, Odile Lecompte, and Olivier Poch. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, 6(3):e18093, 2011.

Niels Tommerup, Lissi Aagaard, Connie L Lund, Esper Boel, Sarah Baxendale, Gillian P Bates, Hans Lehrach, and Henrik Vissing. A zinc-finger gene znf141 mapping at 4p16. 3/d4s90 is

a candidate gene for the wolf-hirschhorn (4p-) syndrome. *Human molecular genetics*, 2(10): 1571–1575, 1993.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.

Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Christiaan C Veerman, Arthur AM Wilde, and Elisabeth M Lodder. The cardiac sodium channel gene scn5a and its gene product nav1. 5: Role in physiology and pathophysiology. *Gene*, 573(2):177–187, 2015.

Mauno Vihinen. Guidelines for reporting and using prediction tools for genetic variation analysis. *Human mutation*, 34(2):275–282, 2013.

Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research*, 19(2):327–335, 2009.

CP Walsh and GL Xu. Cytosine methylation and dna repair. *DNA Methylation: Basic Mechanisms*, pages 283–315, 2006.

Roddy Walsh, Nicholas S Peters, Stuart A Cook, and James S Ware. Paralogue annotation identifies novel pathogenic variants in patients with brugada syndrome and catecholaminergic polymorphic ventricular tachycardia. *Journal of medical genetics*, 51(1):35–44, 2014.

Roddy Walsh, Rachel Buchan, Alicja Wilk, Shibu John, Leanne E Felkin, Kate L Thomson, Tang Hak Chiaw, Calvin Chin Woon Loong, Chee Jian Pua, Claire Raphael, et al. Defining the genetic architecture of hypertrophic cardiomyopathy: re-evaluating the role of non-sarcomeric genes. *European heart journal*, 38(46):3461–3468, 2017.

Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012.

James S Ware, Roddy Walsh, Fiona Cunningham, Ewan Birney, and Stuart A Cook. Paralogous annotation of disease-causing variants in long qt syndrome genes. *Human mutation*, 33(8): 1188–1191, 2012.

Bernard L Welch. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

Nicola Whiffin, Eric Minikel, Roddy Walsh, Anne H O'Donnell-Luria, Konrad Karczewski, Alexander Y Ing, Paul JR Barton, Birgit Funke, Stuart A Cook, Daniel MacArthur, et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine*, 19(10):1151–1158, 2017.

Laurens Wiel, Hanka Venselaar, Joris A Veltman, Gert Vriend, and Christian Gilissen.

Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Human mutation*, 38(11):1454–1463, 2017.

Arthur AM Wilde and Ahmad S Amin. Clinical spectrum of scn5a mutations: long qt syndrome, brugada syndrome, and cardiomyopathy. *JACC: Clinical Electrophysiology*, 4(5):569–579, 2018.

Rémi Zallot, Katherine J Harrison, Bryan Kolaczkowski, De Crécy-Lagard, et al. Functional annotations of paralogs: a blessing and a curse. *Life*, 6(3):39, 2016.

Feng Zhang and James R Lupski. Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–R110, 2015.

Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

Zhe Zhang, Maria A Miteva, Lin Wang, and Emil Alexov. Analyzing effects of naturally occurring missense mutations. *Computational and mathematical methods in medicine*, 2012, 2012.

Zheng Zhang, Jie Huang, Zengfang Wang, Lushan Wang, and Peiji Gao. Impact of indels on the flanking regions in structural domains. *Molecular biology and evolution*, 28(1):291–301, 2011.

Appendix

A.1: Table showing the 624 genes and the number of pathogenic annotations provided by Paralogue Annotation when annotating ClinVar variants with themselves.

Var1	Freq
COL3A1	226
COL4A5	116
SCN1A	110
COL1A1	105
COL1A2	97
KCNQ1	76
KCNQ2	75
COL2A1	58
SCN2A	53
SCN5A	51
SCN8A	49
SCN4A	45
TP53	42
ATP1A3	41
BRAF	39
GFAP	39
ACVRL1	38
BMPR2	37
COL4A3	35
FBN1	35
HRAS	35
KRAS	35
GJB2	34
KIT	34
COL4A4	33
LDLR	33
LMNA	31
F8	30
COL4A1	27

Var1	Freq
FGFR2	27
GJB1	27
ABCC6	25
MAP2K1	25
NRAS	25
FGFR1	24
FLT3	24
TGFBR2	23
COL6A2	22
HBB	22
TP63	22
TUBB4A	22
RYR1	21
FGFR3	20
ABCC8	19
KCNJ2	19
RYR2	19
FBN2	18
KCNQ4	17
PIK3CA	16
RAF1	16
SCN9A	16
COL6A1	15
MAP2K2	15
ACTA1	14
ATP7B	14
HBA1	14
KCNJ11	14
ACADM	13

Var1	Freq
CSF1R	13
GALNS	13
GCDH	13
GRIN2B	13
KRT16	13
KRT6A	13
TGFBR1	13
ARSA	12
ATP7A	12
EGFR	12
F9	12
GRIN1	12
GRIN2A	12
KCNQ3	12
KRT17	12
KRT9	12
LRP5	12
AR	11
ARSB	11
CRB1	11
HBA2	11
PAX6	11
RET	11
ACADVL	10
ACTG2	10
CYP27A1	10
DES	10
KRT14	10
NOTCH3	10

Var1	Freq
ACADS	9
ACTA2	9
ACTB	9
ACTG1	9
CLCN1	9
COL5A1	9
GABRB3	9
GJA1	9
KRT10	9
KRT12	9
LHCGR	9
RS1	9
TGM1	9
ABCA4	8
ATP1A2	8
BMPR1A	8
COL5A2	8
GJA8	8
GLRA1	8
KRT1	8
KRT5	8
THRB	8
COL11A1	7
CP	7
CYP11B1	7
CYP27B1	7
FOXC1	7
GABRB2	7
GATA2	7

Var1	Freq
IVD	7
JAK2	7
KRT2	7
MAFB	7
MYH3	7
PAX3	7
PDGFRB	7
PROC	7
TUBB2B	7
APC	6
AVPR2	6
CACNA1A	6
CLCN4	6
CLCN5	6
COL4A2	6
DHCR7	6
ERBB2	6
FSHR	6
GATA4	6
IDH1	6
INSR	6
ITPR1	6
MYO7A	6
NEFL	6
PIK3R1	6
SERPINC1	6
SLC26A4	6
SOX9	6
STXBP1	6

Var1	Freq
TRPS1	6
TSHR	6
WT1	6
ABCA1	5
ABCA12	5
ABCA3	5
ARX	5
CACNA1G	5
GATA6	5
GJA3	5
GJB3	5
GJB4	5
GNAQ	5
IDH2	5
KCNJ10	5
KCNJ5	5
KRT86	5
LBR	5
MYH7	5
MYO15A	5
PAH	5
PAX2	5
PAX8	5
PDHA1	5
SCN11A	5
SERPING1	5
SLC26A2	5
SLC26A3	5
SMAD4	5
SRY	5
TUBB2A	5
VDR	5
ABCC2	4
ACAD8	4
ACAD9	4
ATP2A2	4
BMPR1B	4
CALM2	4
CHRNE	4

Var1	Freq
CLCN2	4
CLN3	4
COL11A2	4
CRYAA	4
CYP17A1	4
DDR2	4
DNM1	4
DNM1L	4
F7	4
FBLN5	4
FGFR4	4
FLNB	4
FLT4	4
GATA1	4
GJC2	4
GNA11	4
GNAO1	4
GRIA3	4
GRIA4	4
HEXA	4
HSPB8	4
IL2RG	4
KIF1A	4
KMT2D	4
LRP4	4
MAP3K7	4
MYH9	4
NR2F1	4
NR3C1	4
OTX2	4
PAX9	4
PIEZO2	4
PIK3CD	4
PSEN1	4
PSEN2	4
PYCR1	4
SCN3A	4
SMARCA2	4
SMARCA4	4

Var1	Freq
SOS2	4
SPTA1	4
STS	4
TH	4
TUBA4A	4
TUBB8	4
ABCC9	3
ACTC1	3
ALK	3
ALX3	3
ARAF	3
ATP2C1	3
BTK	3
CACNA1C	3
CACNA1E	3
CACNA1F	3
CALM3	3
CAMK2A	3
CLCN7	3
COL7A1	3
CRB2	3
CRX	3
CRYAB	3
DGUOK	3
DNM2	3
ESR1	3
F13A1	3
FGD4	3
FLNA	3
GNAI2	3
GNRHR	3
HNF1A	3
HNRNPDL	3
HSD11B2	3
HSPB1	3
KCNC1	3
KCNQ5	3
KRT3	3
LPL	3

Var1	Freq
LTBP2	3
MAF	3
NFIX	3
NKX2-1	3
NOTCH2	3
NPC1	3
NSD1	3
PDGFRA	3
PHEX	3
PIEZO1	3
PIK3R2	3
PROP1	3
RARB	3
RASA1	3
RDH5	3
RUNX2	3
SHOX	3
SLC2A1	3
SLC4A11	3
SOS1	3
SOX10	3
SOX2	3
STAT1	3
STXBP2	3
TFAP2B	3
TGFB3	3
TK2	3
TRPV3	3
TUBA1A	3
TUBB	3
TWIST2	3
ABL1	2
ACTN2	2
ACVR1	2
AKT3	2
ALOX12B	2
ALOXE3	2
ALX4	2
AQP2	2

Var1	Freq
ARSE	2
ATP1A1	2
C1R	2
CACNA1D	2
CACNA1H	2
CACNA1S	2
CAMK2B	2
CHRND	2
CLN8	2
CNGA3	2
COL17A1	2
COL27A1	2
CYP21A2	2
DLL4	2
EFEMP1	2
EFEMP2	2
EGR2	2
EPB42	2
ERF	2
F5	2
FGD1	2
FLI1	2
FLNC	2
FN1	2
FOXC2	2
FOXE1	2
FOXE3	2
FOXL2	2
GABRA1	2
GABRG2	2
GALE	2
GATA3	2
GATA5	2
GDF5	2
GJA5	2
GJB6	2
GLA	2
GUCY2D	2
HBD	2

Var1	Freq
HESX1	2
HIST1H3I	2
HNF1B	2
HNRNPA1	2
INPP5E	2
ITGB3	2
JAG1	2
JAK3	2
KCNH2	2
KCNJ1	2
KCNJ6	2
KIF1C	2
KRT13	2
KRT6B	2
KRT81	2
KRT83	2
LIPH	2
MC2R	2
MC4R	2
MUSK	2
MYH14	2
MYH2	2
NAGA	2
NKX2-5	2
NPR2	2
NR2E3	2
NRL	2
OCRL	2
POU3F4	2
POU4F3	2
PTCH1	2
PYCR2	2
RUNX1	2
SDHD	2
SETD2	2
SLC4A1	2
SLC52A2	2
SLC6A1	2
SOX11	2

Var1	Freq
STAT3	2
TCF4	2
TFAP2A	2
TGDS	2
TGM5	2
TMPRSS3	2
TNNT2	2
TNNT3	2
TUBB4B	2
TWIST1	2
VWF	2
ZAP70	2
ACADSB	1
ACTN1	1
AFG3L2	1
AKT1	1
AKT2	1
AMPD2	1
AMPD3	1
ANTXR1	1
ANTXR2	1
AQP5	1
ARID1A	1
ARID1B	1
ARSG	1
ATL1	1
ATL3	1
ATN1	1
ATRX	1
AUH	1
BCHE	1
BMP15	1
BMP4	1
C1QC	1
C1S	1
CALM1	1
CD40	1
CDC42	1
CHAT	1

Var1	Freq
CHD2	1
CHD3	1
CHD4	1
CHD7	1
CHRNA1	1
CHRNB2	1
CHRNG	1
CHST3	1
CHST6	1
CLCNKA	1
CLCNKB	1
CNGA1	1
COL10A1	1
COL12A1	1
COL6A3	1
COQ8A	1
COQ8B	1
CRAT	1
CREBBP	1
CRYBB2	1
CRYGC	1
CYP1B1	1
CYP24A1	1
DHH	1
DLX3	1
DMBX1	1
DTNA	1
ECEL1	1
ECHS1	1
EHMT1	1
EIF2AK3	1
EIF2AK4	1
ELOVL4	1
ELOVL5	1
ERBB3	1
ERBB4	1
ESRRB	1
EZH2	1
FGF17	1

Var1	Freq
FGF8	1
FOXF1	1
FOXG1	1
FZD6	1
GAA	1
GABRB1	1
GANAB	1
GDF6	1
GHRHR	1
GLDN	1
GLUD1	1
GLUD2	1
GNA14	1
GNAI3	1
GNAT1	1
GNB1	1
GNB4	1
GRIK4	1
GRIN2D	1
GUCY2C	1
H3F3A	1
HECW2	1
HEXB	1
HNF4A	1
HNRNPA2B1	1
HSD3B2	1
HSD3B7	1
IHH	1
IKBKG	1
INPPL1	1
ITK	1
JAK1	1
KCNA1	1
KCNA2	1
KCNB1	1
KCNC3	1
KCND3	1
KCNH1	1
KCNJ18	1

Var1	Freq
KCNV2	1
KIF11	1
KLKB1	1
KMT2A	1
KRT4	1
KRT6C	1
KRT74	1
LHB	1
LHX3	1
LHX4	1
LRP2	1
LRP6	1
LTBP3	1
LTBP4	1
MAFA	1
MASP1	1
MRAS	1
MSX1	1
MSX2	1
MYC	1
MYCN	1
MYH11	1
MYH6	1
MYH8	1
MYO5B	1
MYOC	1
NALCN	1
NCF1	1
NDUFA9	1
NEDD4L	1
NEUROD1	1
NEUROG3	1
NFIA	1
NFIB	1
NLGN3	1
NLGN4X	1
NOTCH1	1
NR1H3	1
NR3C2	1

Var1	Freq
OPN1LW	1
OPN1MW	1
OPN1SW	1
OPTN	1
PAX4	1
PHKA1	1
PHKA2	1
PHOX2B	1
PITX1	1
PITX2	1
POMT1	1
POMT2	1
PRKAG2	1
PRKAG3	1
PRRX1	1
PTH1R	1
PTPN11	1
PTPN12	1
PYGL	1
PYGM	1
RAC1	1
RAD54L	1
RERE	1
RHO	1
ROS1	1
RPGRIP1	1
RPGRIP1L	1
RPS6KA3	1
RRAS2	1
RSPO2	1
RSPO4	1
SAMD9	1
SAMD9L	1
SCN10A	1
SCO1	1
SERPINF2	1
SGCD	1
SGCG	1
SH3PXD2B	1

Var1	Freq
SLC1A1	1
SLC1A2	1
SLC1A3	1
SLC1A4	1
SLC22A12	1
SLC22A5	1
SLC25A13	1
SLC26A1	1
SLC2A9	1
SLC52A3	1
SLC5A2	1
SLC6A8	1
SMAD2	1
SMAD3	1
SMAD6	1
SMO	1
SOX3	1
SPG7	1
SPTAN1	1
SPTB	1
SPTBN2	1
STAT5B	1
SYNGAP1	1
TARDBP	1
TBX22	1
TBX5	1
TBXAS1	1
TCF12	1
TECTA	1
TGFB1	1
TGFB2	1
THRA	1
TMPRSS6	1
TNFRSF11B	1
TNNC1	1
TNNI2	1
TNNI3	1
TPM1	1
TPM3	1

Var1	Freq
TRPV4	1
TSHB	1
TUBA3E	1
TUBB6	1
VIM	1
VLDLR	1
VSX1	1
WNT4	1
WNT7A	1
ZBTB18	1
ZBTB42	1

A.2: Table showing ClinVar variants with their chromosome, position, reference allele, alternate allele, ClinVar ID, and the overlapping or read-through genes that they are defined as being located in by Ensembl.

CHROM	POS	REF	ALT	ID	GENES
10	97584912	G	C	33	ANKRD2, C10orf62, HOGA1, AL355315.1
10	97601925	T	G	34	HOGA1, AL355315.1
16	2496587	G	C	48	AC093525.2, TBC1D24
16	2496899	T	C	50	AC093525.2, TBC1D24
11	61437820	G	A	401	SDHAF2, AP003108.2
11	72106094	G	A	543	LAMTOR1, LRTOMT, AP000812.4, ANAPC15, AP000812.5
11	72106165	T	C	544	LAMTOR1, LRTOMT, AP000812.4, ANAPC15, AP000812.5
11	72106180	G	A	545	LAMTOR1, LRTOMT, AP000812.4, ANAPC15, AP000812.5
19	19516253	G	C	1693	NDUFA13, AC011448.1, TSSK6
16	31094645	C	A	2207	VKORC1, AC135050.2
16	31094596	A	G	2208	VKORC1, AC135050.2
16	31094558	T	C	2209	VKORC1, AC135050.2
19	41422643	G	A	2377	BCKDHA, B3GNT8, AC011462.1
19	41424496	T	G	2378	BCKDHA, B3GNT8, AC011462.1
19	41422262	G	A	2380	BCKDHA, B3GNT8, AC011462.1
19	41422704	C	G	2381	BCKDHA, B3GNT8, AC011462.1
19	41422309	C	G	2382	BCKDHA, B3GNT8, AC011462.1
7	66089114	A	G	2399	ASL, AC068533.4
7	66092566	C	T	2401	ASL, AC068533.4
7	66092078	C	T	2403	ASL, AC068533.4
13	76995119	G	A	2569	CLN5, AC001226.2
9	69216367	T	C	2907	TJP2, AL358113.1
11	111836179	C	T	3748	ALG9, AP001781.2
11	111853415	T	C	3749	ALG9, AP001781.2
19	38290215	A	G	5205	SPINT2, AC011479.1
8	94391640	T	C	5638	RAD54B, FSBP
8	94399540	C	A	5639	RAD54B, FSBP
22	50523738	G	A	5679	SCO2, NCAPH2, TYMP, SCO2
22	50523901	G	A	5680	SCO2, NCAPH2, TYMP, SCO2
22	50523994	C	T	5681	SCO2, NCAPH2, TYMP, SCO2
22	50524014	C	T	5684	SCO2, NCAPH2, TYMP, SCO2
2	169487811	G	A	6161	BBS5, AC093899.2
2	169493765	A	G	6162	BBS5, AC093899.2
5	77349456	A	C	6390	PDE8B, AC022414.1
19	18868916	C	T	6748	GDF1, UPF1, CERS1
16	89935235	C	T	6963	TUBB3, AC092143.1
16	89935355	G	A	6964	TUBB3, AC092143.1
16	89935700	G	C	6965	TUBB3, AC092143.1
16	89935700	G	A	6966	TUBB3, AC092143.1

CHROM	POS	REF	ALT	ID	GENES
16	89935679	G	A	6967	TUBB3, AC092143.1
11	117822424	C	T	7684	FXYD2, FXYD6-FXYD2
X	149490322	G	A	10487	IDS, AC244197.3
X	149482894	C	G	10488	IDS, AC244197.3
X	149483135	A	C	10493	IDS, AC244197.3
X	149482997	G	A	10497	IDS, AC244197.3
X	149482996	C	T	10498	IDS, AC244197.3
X	149482996	C	A	10500	IDS, AC244197.3
4	109746398	T	A	12118	CFI, AC126283.2
4	109760567	C	T	12124	CFI, AC126283.2
11	66526181	T	G	12143	BBS1, AP002748.4
2	233768259	C	T	12267	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
2	233767161	A	G	12270	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
2	233767092	G	A	12272	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
2	233772413	T	G	12281	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3, MROH2A
2	233767922	A	G	12283	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
2	233768333	A	G	12286	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
15	40718818	G	A	13127	RAD51, AC012476.1
11	2160825	G	C	13377	TH, INS, INS-IGF2
11	2160829	A	G	13378	TH, INS, INS-IGF2
11	2160872	G	C	13379	TH, INS, INS-IGF2
11	2160901	G	T	13388	TH, INS, INS-IGF2
11	2160829	A	C	13389	TH, INS, INS-IGF2
11	2160956	G	A	13390	TH, INS, INS-IGF2
11	2160835	C	T	13391	TH, INS, INS-IGF2
11	2160809	G	A	13392	TH, INS, INS-IGF2
4	54285926	A	T	13543	PDGFRA, AC058822.1
4	54278380	C	T	13550	PDGFRA, AC058822.1
4	54285937	G	T	13551	PDGFRA, AC058822.1
5	139307669	C	G	14002	MATR3, MATR3
11	5254417	G	A	14981	HBG1, HBG2, AC104389.5
11	5254330	G	A	14989	HBG1, HBG2, AC104389.5
11	5254482	A	G	14996	HBG1, HBG2, AC104389.5
11	1757343	A	T	17573	CTSD, AC068580.4
7	30912022	C	T	17847	AQP1, AC004691.2
3	51987188	A	C	18111	ABHD14B, ACY1, ABHD14A-ACY1
3	51986993	C	T	18113	ABHD14B, ACY1, ABHD14A-ACY1
16	30068824	A	G	18177	ALDOA, AC093512.2
16	30069384	G	A	18178	ALDOA, AC093512.2

CHROM	POS	REF	ALT	ID	GENES
11	93796509	T	C	18441	MED17, AP001273.2
11	2160845	A	C	21114	TH, INS, INS-IGF2
11	2160832	C	A	21115	TH, INS, INS-IGF2
11	2160878	C	T	21122	TH, INS, INS-IGF2
11	2160878	C	G	21123	TH, INS, INS-IGF2
11	5254405	C	T	29753	HBG1, HBG2, AC104389.5
16	89935418	A	G	30272	TUBB3, AC092143.1
16	89935064	G	A	30274	TUBB3, AC092143.1
16	89935356	C	T	30275	TUBB3, AC092143.1
19	38916306	T	C	30982	CCER2, AC011455.2, SARS2
11	2160901	G	A	36401	TH, INS, INS-IGF2
22	50524071	C	T	50905	SCO2, NCAPH2, TYMP, SCO2
13	76995077	G	C	56533	CLN5, AC001226.2
13	76995990	A	G	56539	CLN5, AC001226.2
13	76996008	T	C	56540	CLN5, AC001226.2
13	76996028	C	T	56541	CLN5, AC001226.2
13	76996035	G	C	56542	CLN5, AC001226.2
16	2496834	T	C	56846	AC093525.2, TBC1D24
2	169492900	G	A	68066	BBS5, AC093899.2
6	116431086	C	T	88848	DSE, Z84488.2
16	2496872	C	T	91395	AC093525.2, TBC1D24
16	2496266	C	T	91396	AC093525.2, TBC1D24
16	2496206	C	G	91397	AC093525.2, TBC1D24
X	149482996	C	G	92615	IDS, AC244197.3
X	149482891	A	T	92616	IDS, AC244197.3
19	41423038	C	T	93335	BCKDHA, B3GNT8, AC011462.1
19	41424504	G	A	93344	BCKDHA, B3GNT8, AC011462.1
19	41411004	C	T	93356	BCKDHA, AC011462.1
19	41419282	C	T	93365	BCKDHA, AC011462.1
19	41422176	C	T	93367	BCKDHA, B3GNT8, AC011462.1
19	41422370	G	C	93375	BCKDHA, B3GNT8, AC011462.1
19	41422680	A	C	93377	BCKDHA, B3GNT8, AC011462.1
19	41422754	G	A	93385	BCKDHA, B3GNT8, AC011462.1
19	41424582	T	A	100009	BCKDHA, B3GNT8, AC011462.1
16	2496356	G	T	100677	AC093525.2, TBC1D24
16	2497026	G	C	100678	AC093525.2, TBC1D24
11	1758970	G	A	100732	CTSD, AC068580.4
5	139307759	T	G	126561	MATR3, MATR3
5	139322683	A	G	126562	MATR3, MATR3

CHROM	POS	REF	ALT	ID	GENES
5	139307875	C	T	126563	MATR3, MATR3
11	93790727	C	T	129605	MED17, AP001273.2
19	19526257	G	A	132643	NDUFA13, YJEFN3, AC011448.1
16	2496681	C	T	133246	AC093525.2, TBC1D24
1	160237145	G	A	133348	DCAF8, AL139011.2
19	38290229	G	A	157607	SPINT2, AC011479.1
1	74369495	G	A	161447	TNNI3K, FPGT-TNNI3K
1	211019164	C	G	162523	KCNH1, AL590132.1
16	2496267	G	T	183152	AC093525.2, TBC1D24
16	2496461	T	C	183153	AC093525.2, TBC1D24
2	169514701	C	T	183245	KLHL41, AC093899.2
1	210919979	C	T	183415	KCNH1, AL590132.1
1	210919967	G	C	183416	KCNH1, AL590132.1
1	210919955	C	G	183417	KCNH1, AL590132.1
1	210920047	G	T	183419	KCNH1, AL590132.1
16	4340985	T	C	187811	GLIS2, PAM16, CORO7-PAM16
2	169493750	G	A	191306	BBS5, AC093899.2
1	225921254	G	A	192394	PYCR2, AL117348.2
10	97584810	C	T	204265	ANKRD2, C10orf62, HOGA1, AL355315.1
10	97584837	C	T	204266	ANKRD2, C10orf62, HOGA1, AL355315.1
10	97601884	C	A	204274	HOGA1, AL355315.1
10	97601889	G	A	204275	HOGA1, AL355315.1
19	41422361	G	C	204377	BCKDHA, B3GNT8, AC011462.1
19	41414149	G	A	204378	BCKDHA, AC011462.1
19	41414143	A	C	204380	BCKDHA, AC011462.1
11	1759606	G	C	205350	CTSD, AC068580.4
11	1759569	G	A	205363	CTSD, AC068580.4
11	1758994	C	A	208848	CTSD, AC068580.4
2	233768226	C	T	212543	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
22	50523877	G	A	215130	SCO2, NCAPH2, TYMP, SCO2
16	89933486	G	A	219254	TUBB3, AC092143.1
16	89933512	G	A	219255	TUBB3, AC092143.1
16	89935236	G	A	219256	TUBB3, AC092143.1
16	89935589	C	T	219257	TUBB3, AC092143.1
X	149487089	A	G	221200	IDS, AC244197.3
X	149487072	A	T	221201	IDS, AC244197.3
X	149482966	T	C	221204	IDS, AC244197.3
X	149498132	G	T	221211	IDS, AC244197.3
X	149490436	T	A	221212	IDS, AC244197.3

CHROM	POS	REF	ALT	ID	GENES
X	149490385	C	T	221213	IDS, AC244197.3
X	149490317	G	A	221971	IDS, AC244197.3
X	149487080	T	G	221972	IDS, AC244197.3
7	66088850	C	A	224976	ASL, AC068533.4
7	66089146	C	T	224977	ASL, AC068533.4
1	210919966	A	G	235383	KCNH1, AL590132.1
19	41422353	A	G	235750	BCKDHA, B3GNT8, AC011462.1
16	2496342	G	T	236046	AC093525.2, TBC1D24
16	89935613	A	G	236243	TUBB3, AC092143.1
19	38917929	A	G	242874	AC011455.2, SARS2
11	2160847	A	G	253331	TH, INS, INS-IGF2
1	225921232	A	G	254248	PYCR2, AL117348.2
1	225921590	G	A	254249	PYCR2, AL117348.2
1	225921311	A	C	254250	PYCR2, AL117348.2
16	89935623	G	T	265354	TUBB3, AC092143.1
1	210920032	C	T	279981	KCNH1, AL590132.1
16	89935140	C	T	280212	TUBB3, AC092143.1
X	149483134	C	T	284860	IDS, AC244197.3
11	66515586	G	A	370228	BBS1, AP002748.4
1	235427143	T	A	372201	TBCE, TBCE
16	89935214	G	A	372654	TUBB3, AC092143.1
4	109764567	T	C	372807	CFI, AC126283.2
16	89935065	A	C	373761	TUBB3, AC092143.1
16	89935433	G	A	374309	TUBB3, AC092143.1
16	2496552	C	T	375708	AC093525.2, TBC1D24
4	54277981	C	A	375901	PDGFRA, AC058822.1
4	54277981	C	G	375902	PDGFRA, AC058822.1
4	54285925	GA	AT	376055	PDGFRA, AC058822.1
4	54277977	T	C	376182	PDGFRA, AC058822.1
4	54285923	G	A	376183	PDGFRA, AC058822.1
4	54285934	C	T	376184	PDGFRA, AC058822.1
4	54285959	G	A	376185	PDGFRA, AC058822.1
4	54285925	G	T	376250	PDGFRA, AC058822.1
X	149483006	G	T	383181	IDS, AC244197.3
16	89935313	G	A	384329	TUBB3, AC092143.1
16	2496469	T	A	391688	AC093525.2, TBC1D24
19	18869345	C	A	392938	GDF1, UPF1, CERS1
9	69226175	G	A	397505	TJP2, AL358113.1
19	7444215	A	G	417754	ARHGEF18, AC008878.3

CHROM	POS	REF	ALT	ID	GENES
19	18869315	G	C	418231	GDF1, UPF1, CERS1
19	18868764	C	T	418233	GDF1, UPF1, CERS1
1	210920040	T	G	420958	KCNH1, AL590132.1
16	89935532	C	A	421147	TUBB3, AC092143.1
16	89935296	G	C	421587	TUBB3, AC092143.1
X	149487059	C	T	422490	IDS, AC244197.3
16	2497067	A	C	425556	AC093525.2, TBC1D24
7	66089335	G	C	426324	ASL, AC068533.4
X	149496419	T	A	427181	IDS, AC244197.3
16	89935285	C	G	427208	TUBB3, AC092143.1
16	89935660	G	C	429849	TUBB3, AC092143.1
16	89934875	G	A	430312	TUBB3, AC092143.1
X	149487078	C	T	432040	IDS, AC244197.3
7	66092084	G	C	432076	ASL, AC068533.4
19	38919827	G	A	432690	AC011455.2, SARS2
15	92924407	G	A	433117	CHD2, AC013394.1
16	2496957	G	A	436947	AC093525.2, TBC1D24
16	2497014	C	T	436948	AC093525.2, TBC1D24
X	149487006	T	A	437444	IDS, AC244197.3
2	233767858	C	T	437450	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3
2	169492899	C	T	437993	BBS5, AC093899.2
16	89932649	C	T	438582	TUBB3, AC092143.1
7	99388183	C	T	444878	ARPC1B, AC004922.1
6	116431082	A	G	446171	DSE, Z84488.2
X	149498127	T	A	449348	IDS, AC244197.3
X	149490314	C	A	457353	IDS, AC244197.3
15	40729609	G	A	471141	RAD51, AC012476.1
19	41414140	G	A	493261	BCKDHA, AC011462.1
16	89934974	G	C	520851	TUBB3, AC092143.1
4	109766572	C	T	522480	CFI, AC126283.2
19	18868625	A	G	522571	GDF1, UPF1, CERS1
19	41422718	C	T	522678	BCKDHA, B3GNT8, AC011462.1
11	93812006	G	A	522706	MED17, AP001273.2
1	235414579	T	G	522753	TBCE, TBCE
15	40729632	G	A	522859	RAD51, AC012476.1
16	89919770	C	G	548659	TUBB3, MC1R, AC092143.1
19	41422274	G	A	558191	BCKDHA, B3GNT8, AC011462.1
12	120446808	T	G	559417	TRIAP1, GATC, AL021546.1
4	109766720	G	C	560185	CFI, AC126283.2
1	211019219	T	C	560720	KCNH1, AL590132.1
11	66526137	C	G	585170	BBS1, AP002748.4
11	66531741	A	G	585173	ZDHHC24, BBS1, AP002748.4
2	169487092	A	G	585188	BBS5, AC093899.2
X	149483134	C	A	592185	IDS, AC244197.3
2	233768295	CC	GT	597250	UGT1A6, UGT1A1, UGT1A10, UGT1A9, UGT1A4, UGT1A5, UGT1A7, UGT1A8, UGT1A3

Permissions to re-use third-party copyrighted works

Figure 1.1

License Number	5235200370988
License date	Jan 24, 2022
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Genome-wide association studies for complex traits: consensus, uncertainty and challenges
Licensed Content Author	Mark I. McCarthy et al
Licensed Content Date	Dec 31, 1969
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Will you be translating?	no
Circulation/distribution	1-29
Author of this Springer Nature content	no

Figure 1.2 and 1.3

Unique reference number	220124-008522
Submission ID	1231442
Date	24 Jan 2022 7:14am
Type of Publication	Journal
Title	Genetics in Medicine
Authors	Sue Richards et al
Year	2015
From page	405
To page	424
ISSN	1530-0366
Volume	17
Issue	5
Article title	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
I would like to use	Figure(s)
Quantity of material	1
I am the author of the Elsevier material	No
Elsevier author is involved in my project	No
In what format will you use the material	Print and Electronic
Translation	No
Proposed use	Reuse in a thesis/dissertation
Material can be extracted	No