View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by The University of Utah: J. Willard Marriott Digital Library

COMBINING PHENOTYPE AND GENOTYPE FOR DISCOVERY AND DIAGNOSIS OF GENETIC DISEASE

by

Marc Singleton

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

August 2015

Copyright © Marc Singleton 2015

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Marc Singleton**

has been approved by the following supervisory committee members:

Mark Yandell	, Chair	**06/03/2014** Date Approved
Karen Eilbeck	, Member	**06/03/2014** Date Approved
Lynn B Jorde	, Member	**06/03/2014** Date Approved
James E Metherall	, Member	**06/03/2014** Date Approved
Sean Vahram Tavtigian	, Member	**06/03/2014** Date Approved

and by **Lynn B Jorde** , Chair of

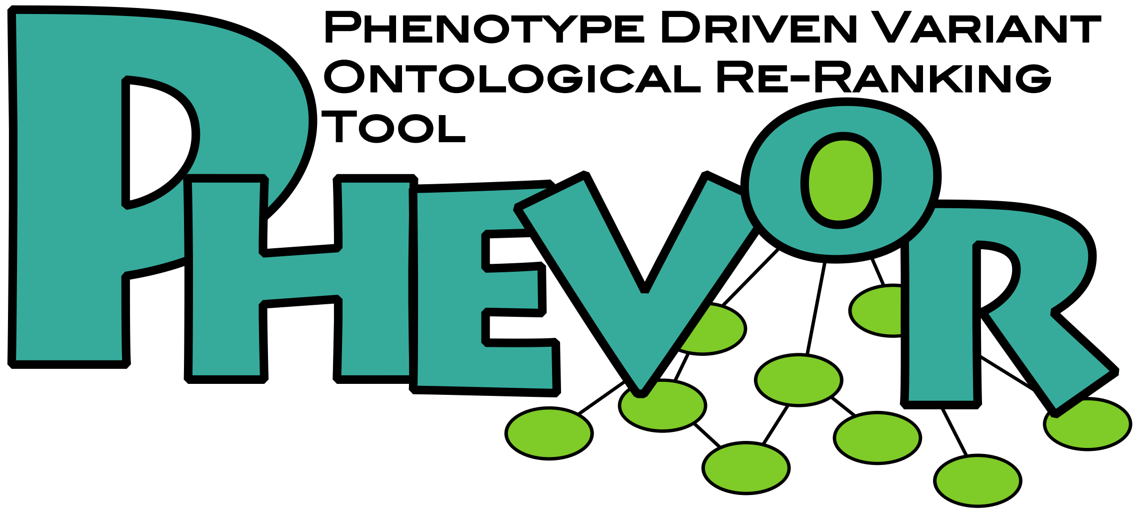
the Department of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Successful molecular diagnosis using an exome sequence hinges on accurate association of damaging variants to the patient’s phenotype. Unfortunately, many clinical scenarios (e.g., single affected or small nuclear families) have little power to confidently identify damaging alleles using sequence data alone. Today’s diagnostic tools are simply underpowered for accurate diagnosis in these situations, limiting successful diagnoses. In response, clinical genetics relies on candidate-gene and variant lists to limit the search space. Despite their practical utility, these lists suffer from inherent and significant limitations. The impact of false negatives on diagnostic accuracy is considerable because candidate-genes and variants lists are assembled *ad hoc*, choosing alleles based upon expert knowledge. Alleles not in the list are not considered—ending hope for novel discoveries. Rational alternatives to *ad hoc* assemblages of candidate lists are thus badly needed. In response, I created Phevor, the Phenotype Driven Variant Ontological Re-ranking tool. Phevor works by combining knowledge resident in biomedical ontologies, like the human phenotype and gene ontologies, with the outputs of variant-interpretation tools such as SIFT, GERP+, Annovar and VAAST. Phevor can then accurately to prioritize candidates identified by third-party variant-interpretation tools in light of knowledge found in the ontologies, effectively bypassing the need for candidate-gene and variant lists.

Phevor differs from tools such as Phenomizer and Exomiser, as it does not postulate a set of fixed associations between genes and phenotypes. Rather, Phevor dynamically integrates knowledge resident in multiple bio-ontologies into the prioritization process. This enables Phevor to improve diagnostic accuracy for established diseases and previously undescribed or atypical phenotypes. Inserting known disease-alleles into otherwise healthy exomes benchmarked Phevor. Using the phenotype of the known disease, and the variant interpretation tool VAAST (Variant Annotation, Analysis and Search Tool), Phevor can rank 100% of the known alleles in the top 10 and 80% as the top candidate. Phevor is currently part of the pipeline used to diagnose cases as part the Utah Genome Project. Successful diagnoses of several phenotypes have proven Phevor to be a reliable diagnostic tool that can improve the analysis of any disease-gene search.



Phenotype Driven Variant
Ontological Re-Ranking
Tool
PHEVOR

“Volumes of history written in the ancient alphabet of G and C, A and T”

- Sy Montgomery, *Search for the Golden Moon Bear: Science and Adventure in Southeast Asia*

TABLE OF CONTENTS

ABSTRACT iii

LIST OF FIGURES x

LIST OF TABLES xiii

ACKNOWLEDGEMENTS xiv

Chapters

1. INTRODUCTION TO NEXT GENERATION SEQUENCING AND CLINICAL GENOMICS 1

    1.1. Where to Start? 2
    1.2. Locus Specific Testing 2
    1.3. Next Generation Sequencing 3
    1.4. Next Generation Sequencing for Clinical Diagnosis 4
    1.5. Exome Sequencing 5
    1.6. Is Exome Sequencing Worth It? 8
    1.7. Transforming the Future of Medical Genetics 9

2. GENOMIC VARIANT INTERPRETATION FOR CLINICAL DIAGNOSIS.. 11

    2.1. What Is in an Exome? 13
    2.2. Clinical Interpretation of Exome Sequence Variants 19
    2.3. Is This Variant Tolerated or Damaging? 20
        2.3.1. Amino Acid Substitutions 20
        2.3.2. Phylogenetic Conservation 27
        2.3.3. Limiting the Variant Search Space 31
        2.3.4. Comprehensive Disease-Gene Finders 33
    2.4. Limited Power 44

3. DISEASE ALLELE DATABASES AND BIOMEDICAL ONTOLOGIES 47

    3.1. Disease Allele Databases 49
    3.2. Phenotype Descriptions 50
    3.3. Introduction to Ontologies 52

3.3.1. Ontology Structure 53
3.3.2. The Sequence Ontology 54
3.4. Biomedical Ontologies 57
3.5. Establishing Gene Function and Phenotype Using Bio-Ontologies 58
3.5.1. Gene Ontology 60
3.5.2. Pathway Ontology 60
3.5.3. Mammalian Phenotype Ontology 61
3.5.4. Rat Disease Ontology 61
3.5.5. Human Phenotype Ontology 61
3.5.6. Disease Ontology 62
3.5.7. Chemical Entities of Biological Interest 62
3.6. Using Bio-Ontologies 63
3.6.1. Overrepresentation Analysis 63
3.6.2. Semantic Similarity Analysis 64
3.7. Something Better Is Needed 65

4. PHEVOR: THE PHENOTYPE-DRIVEN VARIANT ONTOLOGICAL RE-RANKING TOOL 67

4.1. Why Did I Need Phevor? 68
4.2. Introducing Phevor 70
4.2.1. Just What Is a Phenotype? 72
4.3. The Phevor Algorithm 75
4.3.1. Connecting Biomedical Ontologies and Seeding Phenotype-Linked Genes 76
4.3.2. Ontology Propagation 79
4.3.3. Transferring Node Scores to Genes 83
4.3.4. Combining Propagations from Multiple Bio-Ontologies 84
4.3.5. Scoring and Ranking Expanded Candidate Genes 84
4.4. Combining Ontology and Variant Data 85

5. BENCHMARKING PHEVOR’S PERFORMANCE 87

5.1. Creating Benchmarking Datasets 87
5.2. Single Exome Benchmarking Results 89
5.2.1. Single Exome Benchmarking Results – Minor Allele Frequency 89
5.2.2. Single Exome Benchmarking Results – Phylogenetic Conservation 90
5.2.3. Single Exome Benchmarking Results – Amino Acid Conservation 90
5.2.4. Single Exome Benchmarking Results – Comprehensive Disease-Gene Finders 95
5.3. The Behavior of the Phevor Algorithm 100
5.3.1. Biomedical Ontology Seeding and Propagation 100
5.3.2. Best Nodes from Bio-Ontology Propagation Aid in Identifying the Disease-Gene 105
5.3.3. Distribution of Phevor Scores 107
5.4. PhevorAccuracy and Pheneotype Specificity 107

5.5. The Impact of Atypical Presentation and Misdiagnosis on Phevor's Accuracy 109
5.6. Novel Phenotype Association 112
5.7. Importance of Biomedical Ontology Structure on Phevor's Accuracy 116
5.8. Comparing Phevor to other Tools 118
5.9. Benchmarking Conclusions 119

6. PHEVOR AND THE UTAH GENOME PROJECT 122

6.1. The Utah Genome Project 122
6.2. Utah Genome Project – Two Diagnoses Missed During Clinical Screening 123
    6.2.1. Progressive Familial Intrahepatic Cholestasis 125
    6.2.2. Sick Sinus Syndrome 127
6.3. Utah Genome Project – Novel Genes and Atypical Phenotypes 129
    6.3.1. Common Variable Immunodeficiency (CVID) 129
    6.3.2. Immune Dysregulation, Polyendocrinopathy, Enteropathy 132
6.4. Utah Genome Project – Large Pedigrees 134
    6.4.1. Wolff-Parkinson-White Syndrome 135
    6.4.2. A Very Large Family Plagued with Early Onset Atrial Fibrillation ... 137
6.5. Conclusion 140

7. WHAT IS NEXT FOR PHEVOR? 141

7.1. Phevor and Cancer 141
7.2. Phevor and the Environment 143
7.3. Phevor and Pharmacology 144
7.4. Risk Assessment and Phevor 144
7.5. My Conclusions 145

REFERENCES 147

LIST OF FIGURES

2.1 Data Reduction of Exome Sequence Variants 18

2.2 Basic Methodology used by SIFT for Variant Interpretation 23

2.3 Performance Characteristics of Amino Acid Substitution Variant Interpretation 26

2.4 Performance Characteristics of Phylogenetic Conservation Variant Interpretation 30

2.5 Basic VAAST Methodology 38

2.6 Performance Characteristics of Comprehensive Disease-Gene Finders 45

3.1 OMIM Provided Clinical Descriptions and Phenotypes for Miller’s Syndrome 51

3.2 Generalized Ontology Structure and Example Node Attributes 55

3.3 Sequence Ontology Illustration of mRNA Processing and Components 56

4.1 General Phevor Methodology 71

4.2 Polycystic Kidney Disease Diagnostic Criteria and Phenotype 73

4.3 Base Node Seeding 78

4.4 Detailed Phevor Methodology 80

5.1 Recovery of Known Pathogenic Alleles Using Minor Allele Frequency (MAF) to Prioritize Variants under a Recessive Inheritance Model 91

5.2 Recovery of Known Pathogenic Alleles Using Minor Allele Frequency to Prioritize Variants under a Dominant Inheritance Model 92

5.3 Recovery of Known Pathogenic Alleles Using Phylogenetic Conservation to Prioritize Variants under a Recessive Inheritance Model ... 93

5.4 Recovery of Known Pathogenic Alleles Using Phylogenetic Conservation to Prioritize Variants under a Dominant Inheritance Model ... 94

5.5 Recovery of Known Pathogenic Alleles Using Amino Acid Conservation to Prioritize Variants under a Recessive Inheritance Model 96

5.6 Recovery of Known Pathogenic Alleles Using Amino Acid Conservation to Prioritize Variants under a Dominant Inheritance Model 97

5.7 Recovery of Known Pathogenic Alleles Using Comprehensive Disease-Gene Finders to Prioritize Genes under a Recessive Inheritance Model....... 99

5.8 Recovery of Known Disease Alleles Using Comprehensive Disease-Gene Finders to Prioritize Genes under a Dominant Inheritance Model....... 101

5.9 Distribution of Gene Annotated Nodes, Seeded Base Nodes, and Best Gene Annotated Nodes Postpropagation for the Human Phenotype Ontology 103

5.10 Example Phevor Report Detailing Bio-Ontology Evidence for Phevor Score and Ranking 106

5.11 Distribution of Phevor Scores... 108

5.12 Increaseing Phenotype Specificity Improves Phevor’s Identification of the Phenotype Responsible Allele 110

5.13 Increased Confidence in Variant Interpretation can overcome an Atypical or Inaccurate Phenotype Description... 111

5.14 Phevor Identifies Pathogenic Alleles with Limited Bio-Ontology Annotation 113

5.15 Accuracy of Ontology Design and Gene Annotations Impact Phevor’s Ability to Identify Pathogenic Genes 117

5.16 Comparison of Tools Using Phenotype to Find Candidate Genes 120

6.1 Schematic of the Utah Genome Project Analysis Pipeline........................... 124

6.2 VAAST and Phevor Results for a Single Proband with Intrahepatic Cholestasis 126

6.3 VAAST and Phevor Results for a Single Proband with Sick Sinus Syndrome 128

6.4 VAAST and Phevor Results for a Case-Control Cohort with Common Variable Immunodeficiency 131

6.5 pVAAST and Phevor Results for a Proband with Severe Enteropathy 133

6.6 pVAAST and Phevor Results for a Family with Wolff-Parkinson-White Syndrome 136

6.7 pVAAST and Phevor Results from a Pedigree with Atrial Fibrillation 138

LIST OF TABLES

1.1 Diagnostic Gene Panel Tests Available at Emory Genetics Laboratories..... 6

2.1 Alleles in the Cystic Fibrosis Mutation Database for *CFTR* Variant Interpretation.. 12

2.2 Classifications of Exome Sequencing Results by Variant Type 15

2.3 Classifications of Exome Sequencing Results by Variant Location 16

2.4 Variant Reduction and Filtering Steps by Annovar using a Recessive Inheritance Model... 36

3.1 Breakdowns of Alleles in Disease Databases... 48

3.2 Contents of Biomedical Ontologies with Gene Annotations........................ 59

ACKNOWLEDGEMENTS

It is unfortunate that genetics research is so focused on pathogenic genes and alleles. Too often, we in the medical genetics community fail to see the wonderful brilliance behind our genetic code. Fortunate for me, I am reminded every day. My two wonderful boys have inherited the best of me and all the love and compassion from my wife. They amaze me every day and I could never have made it this far without them to keep me laughing and on my toes. So too am I grateful for a wife that would allow her husband to be gone so long—working strange hours and always boring her with science talk.

I need to make a particular acknowledgement to my advisor and mentor, Mark Yandell. He has been the exact mentor that I needed to enter this new field of computational biology. He took a molecular biologist—and taught me how to hack some code together. I could not have asked for a supportive, patient yet demanding instructor. He has taken far more time to teach and help me grow as a scientist than would be expected. I am truly grateful for his tutelage and hope to continue working together for years to come.

CHAPTER 1

INTRODUCTION TO NEXT GENERATION SEQUENCING AND CLINICAL GENOMICS

Medical genetics is undergoing a revolution as Sanger sequencing is replaced by rapid, whole-genome sequencing technologies. This new breed of DNA sequencing technologies is termed Next Generation Sequencing (NGS). Next generation sequencing technologies provide fast, cost-effective approaches to sequence billions of short DNA fragments simultaneously. This revolution, however, has not been without its challenges. Sequencing errors, complicated bioinformatics analyses, and uncertain interpretation of the massive amount of data produced by NGS has left many in the community uneasy and longing for the simpler days of single gene sequencing. Like all revolutions, there is no turning back. Moreover, the improved biological understanding and improvements to personalized healthcare promised by whole genome sequencing makes it far too attractive even to want to turn back.

In 1965, Gordon E. Moore predicted that computing power would double every two years[1]. His prediction has now happened; yet, it is dwarfed by changes in DNA sequencing[2-4]. Completion of the Human Genome Project[5] took over 10 years and 3 billion dollars. Using today’s NGS technology, a human genome can be sequenced and assembled in only 48 hours for around 5,000 dollars[6]. The following chapter will provide an outline of NGS and how its adoption by the medical genetics community is forever changing diagnostics. I will also address some of the issues and complications inherent to NGS data that leave many clinicians uneasy[7]. I also will provide insight into what still needs to happen to realize the full potential of genomic sequencing.

1.1 Where to Start?

In 1977, Frederick Sanger and colleagues developed a method for sequencing DNA using chain-terminating nucleotides. These nucleotides terminate polymerase extension whenever the complementary base is encountered on the template molecule[8]. The resulting fragments can then be run through a size-selecting gel matrix, later supplanted by capillary arrays[9], to determine the exact sequence of the DNA fragment. With Sanger sequencing, it became possible to examine the sequence of a gene from a single individual on a nucleotide level. Sanger sequencing has its limitations. Even with improved polymerase design, chain-terminating nucleotides and fragment detection methods, Sanger sequencing can only accurately sequence 800 base-pair sized fragments of DNA. This limitation makes sequencing even moderately sized genes difficult, requiring many reactions that focus on the protein coding regions of the gene. Advances have been made to the Sanger method that improved speed and accuracy. Yet, Sanger sequencing was, and still is, a locus specific analytical method.

1.2 Locus Specific Testing

With the introduction of Sanger sequencing, the community began exploiting this technology for molecular diagnoses. Sanger sequencing and a host of other similar molecular tests that focus on a single gene or single allele is termed “locus specific tests.” Locus specific tests continue to be invaluable for establishing a molecular diagnosis in individuals with highly specific phenotypes; e.g., sequencing *PAH* (phenylalanine hydroxylase)[10] for diagnosis of *Phenylketonuria* (PKU)[11]. Locus specific tests fail in cases of heterogeneous disorders, or when a classic disease-allele presents a nonclassical phenotype. For example, individuals diagnosed with *Primary Ciliary Dyskinesia* (PCD)[12] can have mutations in more than 12 genes, including *CFTR*[13]—classically the cause for *cystic fibrosis*[14], a phenotypically similar, yet distinct disease. Using locus specific testing methods, all 12 genes need to be sequenced to establish a molecular diagnosis. Serialized testing of each PCD-associated gene will take months and costs thousands of dollars—without a guarantee of a conclusive result. Adding to these complications, the research community is continually making novel associations between phenotypes and genotypes. It is unwise to assume that the 12 genes we currently associate with PCD to represent a complete list. New genes will inevitably be discovered, and it will be challenging even for the most agile labs to integrate them into their locus specific testing panel. The capacity to sequence all genes simultaneously can alleviate much of the locus specific complications. This is the revolution promised by NGS.

1.3 Next Generation Sequencing

Next Generation Sequencing (NGS) is a scaled up version of the shotgun methods employed to complete the Human Genome Project[5]. Genomic DNA is first fragmented into smaller manageable pieces. The fragments are then sequenced and the sequence of each fragment is assembled back together, thus making a complete genome. Next generation sequencing has pushed this to massively parallel levels. As the Illumina paired-end sequencing technology currently dominates the market, I will focus on their NGS methodology[15]. Here, genomic DNA is broken into size-specific fragments, typically 500-1,000 bases in length. These fragments are then affixed to adaptor sequences and hybridized to a glass slide known as a flow cell. A single flow cell can contain billions of these short fragments. Starting at the five prime ends, all fragments are simultaneously sequenced one base at a time using fluorophore labeled nucleotides. Depending on the instrument and reagents used 100 – 300 bases are sequenced at a time. The fragments are then flipped over and sequenced for another 100 – 300 bases on the opposite end. Paired-end sequencing aids in the downstream alignment because we know the sequence at both ends of the fragment, as well as how big of a gap between the two to expect. Following sequencing, the raw sequence reads are aligned to the reference genome, and variants are called[16].

1.4 Next Generation Sequencing for Clinical Diagnosis

The key benefit of NGS methods is that there is far more data produced (i.e., genomic sequence) than with any other locus specific testing method. This advantage has led the medical genetics community to begin integrating NGS into routine clinical testing[17]. However, adoption of NGS technology has been slow. Many clinical laboratories are still under-utilizing the power of NGS. They are focusing on small subsets of phenotypically related genes (gene panels) as opposed to whole genome sequencing. Emory Genetics Laboratories currently offers 72 separate NGS gene panel tests. Having an average of 36 genes per panel, these gene panels are for diagnosis of highly specific phenotypes. Table 1.1 details the phenotype specific gene panel tests offered at Emory. Gene panels allow labs and clinicians to stay in comfortable territory; focusing on genes and alleles they are already familiar with. However, I believe this "gene panel" phase cannot last for long. It is expected that within the next decade a significant portion of individuals within the United States will have their genome sequenced. With the push for personalized medicine and the need to diagnose disorders that are genotypically heterogeneous, it seems inevitable that whole genome sequencing will rapidly become the norm rather than the exception. While the public waits for the medical genetics community to embrace whole genome sequencing, a compromise of sorts was reached, one I explain in more detail below.

1.5 Exome Sequencing

Nearly all known disease-alleles occur in protein coding regions[18]. Therefore, it is reasonable to suppose that if a molecular diagnosis is to be established, it will come from variants within the protein coding regions of the genome. Exome sequencing captures and then sequences[19] only the region of the genome where known protein coding genes reside (about 1% of the human genome[20]). As is so often the case, when you solve one problem you create two more. Exome sequencing is no exception[19]. The protein coding regions are preferentially enriched by probe hybridization or Polymerase Chain Reaction (PCR) before sequencing. Both these capture techniques have varying levels of effectiveness across the genome. Pseudogenes are mistakenly captured and then sequenced while regions of high GC content are often missed[21]. When aligned

Table 1.1 Diagnostic Gene Panel Tests Available at Emory Genetics Laboratories

Gene Panel Name	Number of Sequenced Genes
Autism Spectrum Disorders	60
Brain Malformations	50
Cardiomyopathy	103
- Arrhythmias	29
- Brugada Syndrome	8
- Dilated Cardiomyopathy	25
- Hypertropic Cardiomyopathy	14
- Long and Short QT Syndrome	12
- Pulmonary Arterial Hypertension	5
- Sudden Cardiac Arrest	10
Ciliopathies	112
Congenital Disorders of Glycosylation	66
Congenital Myasthenic Syndromes	11
Connective Tissue Disorders	29
Epilepsy and Seizure Disorders	108
Eye Disorders	204
- Albinism	5
- Bardet-Biedl Syndrome	18
- Congenital Stationary Night Blindness	15
- Flecked-retina Disorders	6
- Joubert Syndrome	18
- Leber Congenital Amaurosis	18
- Neuronal Ceroid-Lipofuscinoses	11
- Optic Atrophy	5
- Retinitis Pigmentosa	64
- Senior-Loken Syndrome	7
- Stickler Syndrome	5
- Usher Syndrome	13
- Vitreoretinopathy	9
Glycogen Storage Disorders: Comprehensive	20
- Glycogen Storage Disorders: Liver	11
- Glycogen Storage Disorders: Muscle	12
Hearing Loss	87
Hereditary Cancer Syndrome	46
Hereditary Neuropathies	70
Hereditary Periodic Fever Syndromes	7
Inflammatory Bowel Disease	48
Lysosomal Storage Disorders	55
Macrocephaly	11
Maturity Onset Diabetes of the Young	4
Multiple Epiphyseal Dysplasia	7
Neonatal and Adult Cholestasis	58

Table 1.1 Continued

Neurological Disorders	165
Neuromuscular Disorders	46
- Congenital Muscular Dystrophy	24
- Limb-Girdle Muscular Dystrophy	22
Neuromuscular Disorders- Expanded	79
Noonan Syndrome and Related Disorders	12
Pulmonary Disease	55
- Bronchiectasis	16
- Cystic Lung Disease	8
Short Stature Panel	45
Skeletal Dysplasia	163
- Disproportionate Short Stature	77
- Limb Malformation	46
Targeted Tumor Mutation	26
- Targeted Colorectal Tumor Mutation	13
- Targeted Lung Tumor Mutation	12
- Targeted Melanoma Mutation	8
- Targeted Gastric Tumor Mutation	7
- Targeted Ovarian Tumor Mutation	7
X-linked Intellectual Disability	92

back to the reference genome, exome sequencing has uneven depth of coverage that leaves gaps in the targeted sequence. Missing sequence and increased false variant calls add to the uncertainty many clinicians have about NGS diagnostic results[22].

1.6 Is Exome Sequencing Worth It?

Given the widely acknowledged pitfalls of exome sequencing, including the lack of confidence many clinicians share in making a molecular diagnosis on novel variant in known disease-genes, or still worse a novel variant in a novel disease-gene, is it worth it to sequence an individual's exome? Short answer: Yes. Even with its limited adoption, exome sequencing has shown an improved molecular diagnostic rate over traditional locus specific testing. The CLIA (Clinical Laboratory Improvement Amendments) and CAP (College of American Pathologist) certified clinical sequencing laboratory at Baylor College of Medicine released diagnostic rates from their first 250 sequenced exomes[7]. According to their report, 25% of individuals who had their exome sequenced were able to receive a molecular diagnosis. This is a marked improvement over the 3-15% diagnostic rate for locus specific Sanger sequencing. Baylor also reported that 80% of exomes sequenced were for diagnosis of intellectual disability, a highly heterogeneous disorder with minimal phenotypic differences[23,24]. Phenotypes like intellectual disability[25,26], nonsyndromic hearing loss[27], and autoimmune disorders[28] can only effectively be diagnosed molecularly through exome or genome sequencing. This is because there are far too many genes that might explain any of these phenotypes (e.g., over 200 genes are associated with hearing loss[27] and 400 phenotypically indistinguishable genes are associated with intellectual disability[23,24]).

Inspiring examples of the use of NGS can also be found at the clinical laboratories of Children's Mercy Hospital in Kansas City. At Children's Mercy, whole genome sequencing is carried out on newborns in the Neonatal Intensive Care Unit (NICU) to diagnose nearly 600 clinical phenotypes in as little as 50 hours[6]. With an estimated 20% of infant deaths in the United States being attributed to an inherited disorder[29-31], and over 3,500 monogenic disorders manifesting in the first 28 days of life[32], Children's Mercy is providing infants and their families with molecular diagnoses that can immediately impact their life.

1.7 Transforming the Future of Medical Genetics

Baylor College of Medicine and Children's Mercy are demonstrating the power and utility of next generation sequencing—yet they have only scratched the surface. Improved diagnostic rates reported by Baylor[7] are still limited to only known disease alleles. Under current American College of Medical Genetics (ACMG) guidelines[17], of the ~73,000 variants in any exome sequence, only ~1,600 are considered informative, and from those only variants previously associated with the described phenotype should be used to make a diagnosis. Clinical labs have become reliant on disease-allele databases[18,32,33] that provide variant-specific phenotypes and thus have no way to establish a novel relationship between phenotype and gene. Children's Mercy has expanded their known phenotype associations from single variants to entire genes but is still bound by the limited number of phenotypes and genes curated in their database[6]. It is of my opinion that for clinical genomics to continue along the path of success, changes are needed to the way a molecular diagnosis is established. Current ACMG guidelines are far too restrictive, and furthermore the medical genetics community will eventually have no choice but to abandon exomes in favor of whole genome sequencing. Research such as that conducted by the ENCODE project[34] is providing insight into the noncoding regions of our genome.

Improvements to sequencing, alignment, and variant calling technologies will continue to improve the quality of the data, helping increase certainty of diagnosis by eliminating false positives and negatives. However, without improvements to the methods used to relate a damaging variant's consequence to the patient's phenotype, clinical genomics will move at best, slowly forward—relying on curated disease-allele databases. My dissertation has focused on precisely this problem: To design better algorithmic means to prioritize genes and variants in light of phenotype. The Phenotype-Driven Variant Ontological Re-ranking tool (Phevor)[35] was developed to fill this need. Phevor integrates phenotype, gene function, and disease information with personal genomic data for improved power to identify pathogenic alleles. Phevor works by combining knowledge resident in multiple biomedical ontologies with the outputs of variant interpretation tools. It does so using an algorithm that propagates information across and between ontologies. This process enables Phevor to accurately prioritize potentially damaging alleles identified by variant interpretation tools in light of gene function, disease, and phenotype knowledge. As I will demonstrate, Phevor is especially useful for single exome and family trio-based diagnostic analyses, the most commonly occurring clinical scenarios, and ones for which existing personal-genome diagnostic tools are inaccurate and underpowered.

CHAPTER 2

GENOMIC VARIANT INTERPRETATION FOR CLINICAL DIAGNOSIS

Uncertainty surrounding variant interpretation is the single biggest reason clinical labs have been slow to embrace next generation sequencing tests. Locus-specific tests, like *cystic fibrosis* diagnosis through *CFTR*[36] sequencing rely on highly curated variant interpretation databases. The Cystic Fibrosis Mutation Database created by Sick Kids Hospital[37] for example, has nearly 2,000 characterized variants just for *CFTR* alone; summarized in Table 2.1. Variants not found in the database must be analyzed by the reporting medical director and fit into one of three classifications[17]: 1) Predicted Deleterious, 2) Predicted Benign, or 3) Variant of Unknown Significance (VUS). Initially, variants not found in allele interpretation databases are classified as a VUS and require secondary research to find evidence that suggests them as benign or deleterious. ACMG maintains strict criteria for classifying variants with the predicted moniker[17,38]. Variants that are silent (do not alter the amino acid sequence) and do not interfere with a potential splice site are classified as predicted benign, while nonsense, frameshift and splice site-interrupting variants are classified as predicted deleterious.

The *CFTR* gene has been studied for years. Thus, there is a wealth of information already available to assist with interpretation of variants in this gene. In

Table 2.1 Alleles in the Cystic Fibrosis Mutation Database for *CFTR* Variant Interpretation

Mutation Type	Number of Mutations	Frequency of Mutations
Missense	786	39.96
Frameshift	311	15.81
Splicing	228	11.59
Nonsense	162	8.24
In frame in/del	39	1.98
Large in/del	51	2.59
Promoter	15	0.76
Sequence variation	269	13.68
Unknown	106	5.39

contrast, for whole genomes and exomes, interpretation is much more difficult. One cannot simply query locus-specific databases for every gene—they do not exist. Thus, *ab initio* means of variant interpretation are needed to classify variants of unknown significance. This has been a major motivation behind the development of variant prioritization tools such as SIFT[39] and PolyPhen[40], and disease-gene search tools Annovar[41] and VAAST[42,43].

In this chapter, I will discuss the variant interpretation methodologies used by the few clinical laboratories currently performing diagnostic exome sequencing and describe the ways bioinformatics techniques have emerged as standards for interpreting variant impact. I will describe how these methods work and address their different strengths and weaknesses. I hope to make one point clear: accurate molecular diagnosis hinges on the ability to connect the damaging variant with the patient's phenotype. Means to accomplish this are currently very limited, *ad hoc*, and accuracy is poor. Better computational means to connect phenotype information to variants for will improve clinical diagnostics.

2.1 What Is in an Exome?

It is easy to understand why clinicians and sequencing labs are uneasy about exome sequencing once we consider the sheer number of variants they must interpret. To demonstrate the number of variants needing interpretation, sequence variants from 100 example exomes are detailed below. Exome capture was performed using the SureSelect hybridization[44] and sequenced using Illumina's 100 base paired-end[45] technology. Sequence reads were aligned, and variants called using the "best practice" guidelines laid out by the BROAD Institute[16]. From the aligned reads, three types of variants were identified: Single Nucleotide Variants (SNV), short (<25bp) insertions and deletions (indels) and no-call variants. Variants are classified as no-calls when a variant is called at a position in one of the 100 exomes, but there is insufficient coverage to determine the sequence at that same position in another. It is important that no-call variants be considered during exome analysis as they represent potential variants[46]. Exome variants were annotated using the latest RefSeq gene models[47].

On average, each exome contains over 73,000 variants as described in Tables 2.2 and 2.3. Single nucleotide variants comprise 59%, with indels at only 8%. No-call variants comprise 33% of the discovered variants, highlighting the inconsistencies in the exome capture technology. Exome capture targets the coding regions of the genome while intentionally reaching into the intronic, promoter, and untranslated regions of the genes. As indicated in Table 2.3 only 34% of the total variants are annotated to the protein coding sequence, demonstrating: 1) protein coding gene sequence is more conserved than noncoding sequence, and 2) uneven sequencing depth around the edges of “targeted” regions (just outside the protein coding sequence) result in many false variant calls. *Evidence that variant calling from whole genome sequencing is superior to exome sequencing.*

Following the ACMG guidelines[17], coding variants can be divided into two separate groups. About half are silent (synonymous), thus classified as predicted benign. The other half, classified as predicted deleterious, alter the amino acid sequence of the protein (missense, nonsense, frameshift, or inframe indels). Variants predicted to be deleterious leave nearly 13,000 variants that need their impact on gene function and

Table 2.2 Classifications of Exome Sequencing Results by Variant Type

Variant Type	Number of Variants	Percent Total Variants
Single Nucleotide Variants – SNV	42,846	59%
Insertions/Deletions – Indels	5,615	8%
No-Call Variants	24,557	33%
Total Variants	73,018	100%

Table 2.3 Classifications of Exome Sequencing Results by Variant Location

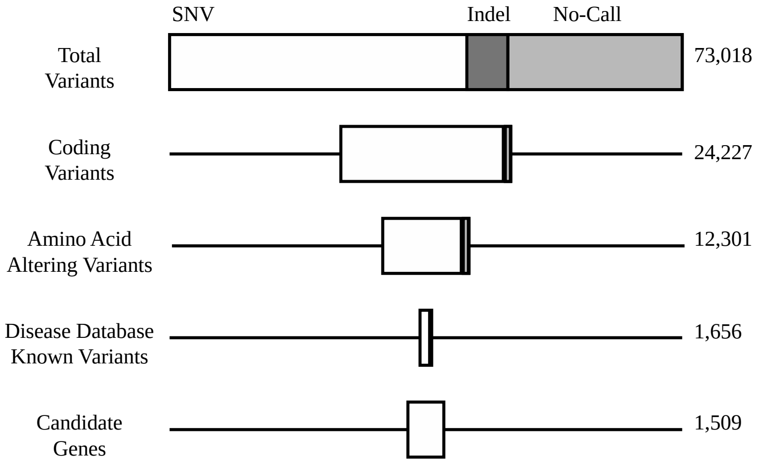
Variant Location	Number of Variants	Percent Total Variants
Coding Sequence	24,663	34%
Intronic Sequence	16,342	22%
Untranslated Sequence	27,100	37%
Splice Region Sequence	337	0.5%

contribution to phenotype carefully investigated. That is a lot of variants.

In contrast to locus-specific testing, where usually no more than 10 variants are identified, most of which are known polymorphisms[38], it is easy to see why variant interpretation for an exome's worth of variants is daunting. As a result, most laboratories offering clinical exome analysis have chosen to only focus on amino acid altering alleles found in disease databases. Using curated database like the Human Genetic Mutation Database (HGMD)[18] and ClinVar[33] these 13,000 variants can be narrowed down to only 1,600 variants with known disease consequence. This is a big data reduction, as detailed in Figure 2.1, but I would like to make two points.

The first point being that 1,600 variants are still far too many for manual analysis by a clinical geneticist. The reduction may seem huge in regards to the total magnitude, but it does not really solve the interpretation problem on its own. Variants remaining after the reduction lack any prioritization. Again, the key is to discover which of these 1,600 variants is responsible for the patient's phenotype. In some cases this is straightforward; *pancreatic insufficiency*[48] accompanied by a known pair of alleles in *CFTR*, both previously shown to cause disease, is sufficient for a confident diagnosis. However, problems arise for genetically heterogeneous diseases such as *primary ciliary dyskinesia*[49,50].

Secondly, many patients present with atypical phenotypes, or have combinations of phenotypes. This makes tying variants to disease far more complicated and diagnoses much less certain. Consider also, that patients with clear phenotypes are more likely to be tested using conventional diagnostic procedures; it is those patients with complicated, atypical phenotypes that are most likely to have their exome sequenced.



The result is that most patient’s exome sequences are uninformative for diagnosis using

Figure 2.1. Data Reduction of Exome Sequence Variants
Exome sequence variant analysis using the data reduction methods quickly narrows candidate alleles. Reducing the number of variants from 73,018 by: 1) coding variants only limits the data by 67%, 2) amino acid altering variants only by 83%, and 3) disease allele database variants only by 98%. By only considering the remaining 2% of alleles, only 5.5% of all protein coding genes potential candidates. Despite the massive data reduction, the remaining candidates are far too many for individual *ad hoc* interpretation.

today's methodologies.

2.2 Clinical Interpretation of Exome Sequence Variants

The majority of labs brave enough to venture into the exome sequencing waters have really only dipped their toe in, using exome sequencing only to help diagnose heterogeneous disorders having multiple candidate genes. For example, *primary ciliary dyskinesia* has 12 possible candidate genes[51]. The modest improvement in molecular diagnostic rates reported by Baylor College of Medicine is used to justify this (15% increased to 25%)[7]. However, focusing on only known disease alleles does not allow for novel associations between variants and their phenotypic consequence. In response, some labs have begun to employ "variant interpreters" that attempt to predict the impact of an uncharacterized variant's impact on gene function.

"Variant interpreters" are still bound to only known disease alleles, but instead of focusing on site-specific mutations (like those found in HGMD[18]), they expand phenotypic associations to other nearby deleterious variants found within the same gene. For instance, a single nucleotide change at coding position 905 causes an amino acid substitution from an arginine to a glutamine in *PRKAG2*, causing *familial hypertrophic cardiomyopathy*[52]. This mutation disrupts the cystathionine beta-synthase domain (CBS) of *PRKAG2*. When an individual with *cardiomyopathy* has a novel variant (not annotated in HGMD) in the CBS domain, and all evidence suggests this variant to be deleterious, the "variant interpreter" concludes this variant is pathogenic. Approaches like this can provide a novel molecular diagnosis where methods using databases of known disease alleles fail. Ambry Genetics was the first lab to successfully provide a molecular diagnosis using this technique[53]. Although this approach does expand the diagnostic range from known variants to known disease-genes, it is still bound to known disease databases, and is very time and personnel consuming. To fully realize the diagnostic power in exome sequencing, all variants not only need to be interpreted as deleterious or benign in an *ab initio* fashion, but then prioritized in the context of the patient's phenotype. In the paragraphs that follow, I first outline existing techniques for *ab initio* identification of deleterious variants. I then account for the shortcomings of these approaches and explain why integration of phenotype information into the process is so desirable.

2.3 Is This Variant Tolerated or Damaging?

Before damaging variants can be properly prioritized to identify those responsible for the patient's phenotype, they must first be interpreted as deleterious or benign. Clinical exome analyses use many bioinformatics techniques to interpret a variant's impact. These techniques can be combined to improve confidence.

2.3.1 Amino Acid Substitutions

The first algorithms used to interpret variants arose out of bioinformatics processes developed to accurately align homologous proteins. Long before next generation sequencing, researchers sought to determine the function of a protein by examining its structure. Structure often dictates function[54], therefore identifying structural similarities that proteins share aids in understanding function. The problem then became: how do we accurately align two homologous but not identical proteins?

Amino acid substitution matrices were designed to penalize mismatches during the alignment process[55,56]. Using the frequency an Amino Acid Substation (AAS) was found in proteins with 85% similarity, early substitution matrices like the Point Accepted Mutation (PAM) matrices[57] provided a way to align proteins with similar structure. The work of Henikoff and Henikoff in 1992[58] greatly improved upon these substitution matrices by analyzing similar blocks opposed to whole proteins. Breaking proteins into ungapped blocks identified several-hundred conserved protein blocks. These blocks were then used to calculate the observed and expected AAS rates. Represented in a substitution matrix as the Logarithm of Odds (LOD), each possible AAS has a calculated score detailing the frequency at which the observed differs from what is expected. By altering the required level of similarity between the protein blocks prior to calculating the LOD, separate matrices can be made, each having its own level of specificity. The BLOSUM matrices (BLOSUM62, BLOSUM85)[59] serve as guides for sequence alignment, making it possible to align distant protein homologues with improved accuracy. Two widely used algorithms (SIFT[39] and PolyPhen[40]) have exploited the methodology behind BLOSUM for predicting how well variants are tolerated in humans.

2.3.1.1 SIFT

Originally designed to align homologous proteins, BLOSUM matrices measure how well Amino Acid Substitutions (AAS) are tolerated. Tolerance scores were optimized to predict how well amino acid altering variants are tolerated in human genes. The SIFT algorithm (Sorting Intolerant from Tolerant)[39] was the first developed to interpret impact (tolerated/benign or intolerant/deleterious) of amino acid altering human genetic variants. When a variant causes an AAS, SIFT builds a dataset of structurally similar proteins. At the position of the amino acid altering variant, SIFT calculates the probability of observing each of the possible 20 amino acids based upon the gapped multiple alignment. Probabilities are normalized to the most frequent change and returned as a SIFT score. If the returned SIFT score is above or below established thresholds, the variant is predicted as benign or damaging, respectively. The general methodology used by SIFT is indicated in Figure 2.2. SIFT scores are bound between 0 and 1, 0 being the highest confidence for damaging and 1 being benign. Unfortunately, predictions are only made to the top and bottom 5%, leaving most SIFT scores classified as uncertain (Figure 2.2). SIFT does provide a binary classification for each variant it can score, however, anything outside the top and bottom 5% do not receive a prediction.

SIFT is limited by the type of variant it can score. SIFT calculates the probability a given AAS is observed at a given position. Therefore, SIFT cannot score variants that do not alter amino acids. Silent variants, those that alter splice sites, insertions and deletions and variants outside the coding region are simply ignored by SIFT. This immediately excludes 83% of variants in the example 100 exomes dataset I described above, and risks missing biologically significant variants. Of the variants that can be scored (single nucleotide coding variants only) SIFT is unable to score an additional 2.5% because they reside in proteins without sufficient conservation to establish a score. Of the score-able variants, SIFT predicts 28% as damaging, 8% as

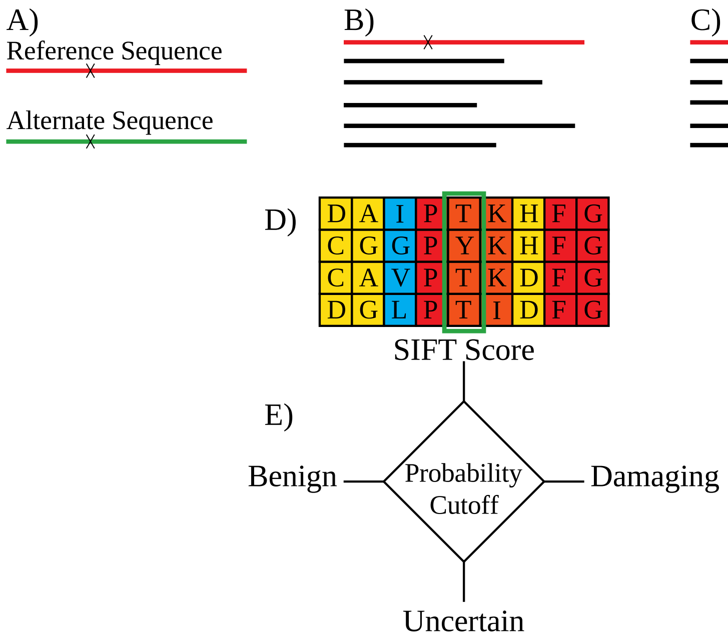

A)
Reference Sequence
Alternate Sequence
B)
C)
D)
D A I P T K H F G
C G G P Y K H F G
C A V P T K D F G
D G L P T I D F G
SIFT Score
E)
Benign
Probability Cutoff
Damaging
Uncertain

Figure 2.2. Basic Methodology used by SIFT for Variant Interpretation
SIFT predicts a variant's tolerance by: A) Identify the amino acid sequence caused by the reference and variant alleles. B) Retrieve proteins with similar structure from a large protein database. C) Generate gapped alignments to the ungapped reference sequence. D) Calculate amino acid substitution rates for all 20 amino acids at the position of the variant. E) Predictions are returned if the SIFT score meets established cutoffs.

benign. The remaining variants are classified as uncertain, having a SIFT score between 0.05 and 0.95. In a review of their own tool, Ng and Henikoff report SIFT being able to score only 60% of single nucleotide coding positions[60]. Along with a self-reported 31% false negative and 20% false positive rate[60], SIFT alone is clearly an inadequate tool for clinical interpretation of exome variants. Even when SIFT is able to score the allele responsible for the phenotype, several thousand variants are likely to have comparable or better SIFT scores, increasing the difficulties identifying those responsible for the phenotype.

2.3.1.2 PolyPhen

Incremental improvements were made to interpretation of amino acid altering variants with the introduction of PolyPhen[40]. Like SIFT, PolyPhen builds upon the BLOSUM methodology and calculates AAS rates using similar proteins. However, PolyPhen only uses short fragments of these proteins to calculate the AAS probabilities. PolyPhen does not require similarity protein wide, only similar protein parts, i.e., protein domains. Aligning short protein fragments helps PolyPhen calculate scores on more of the coding sequence, reporting an improved number of possible single nucleotide coding variants that could be scored over SIFT from 60% to 80%[60]. Although this did not appear to be true when variants from the example exomes were analyzed, SIFT was able to provide a score for 15.5% of all variants, where PolyPhen only scored 15% of the same set.

PolyPhen takes variant interpretation one-step further by classifying the variant into the protein domain where it is found, e.g., disulfide, nucleotide binding, transmembrane, or signaling. Variants found in these domains receive further analysis. PolyPhen compares the three-dimensional structure of the domain, and how the introduced AAS alters the structure. Profiling structural changes in the domain, PolyPhen reduces false positives that do not alter the structure. As with SIFT, PolyPhen scores are bound between 0 and 1, but PolyPhen returns three different prediction categories: probably damaging (1-0.957), possibly damaging (0.956-0.453) and benign (0.452-0). PolyPhen avoids the uncertain classification, lending more confidence in its predictions. However, suffering from the same false negatives as SIFT (i.e., unable to score silent, noncoding, insertions, deletions and splicing variants) leaves only predictions for amino acid altering single nucleotide variants. Comparing PolyPhen's performance with the same criteria as SIFT, Ng and Henikoff report an improved false positive rate of 9%, but a similarly high false negative rate of 31%[60].

Using similar methodologies for variant interpretation, the performance characteristics for SIFT and PolyPhen have similar weaknesses. The limitations of these two tools can clearly be seen when they are used to analyze known damaging alleles found in HGMD, and those in the 100 example exomes dataset described above. Out of all exome variants in this dataset, SIFT and PolyPhen were only able to interpret 15.5 and 15%, respectively (Figure 2.3A). Using the example exome variants as true negatives, and known single nucleotide coding alleles from HGMD as true positives PolyPhen shows increased accuracy to SIFT. However, in order to return 80% of the true positives, SIFT suffers from a 67% false positive rate, while PolyPhen returns 37% false positives (Figure 2.3B). Coupling the variant type limitations and the high rate of false positive candidates leaves the results from these tools suspect at best. Utilizing

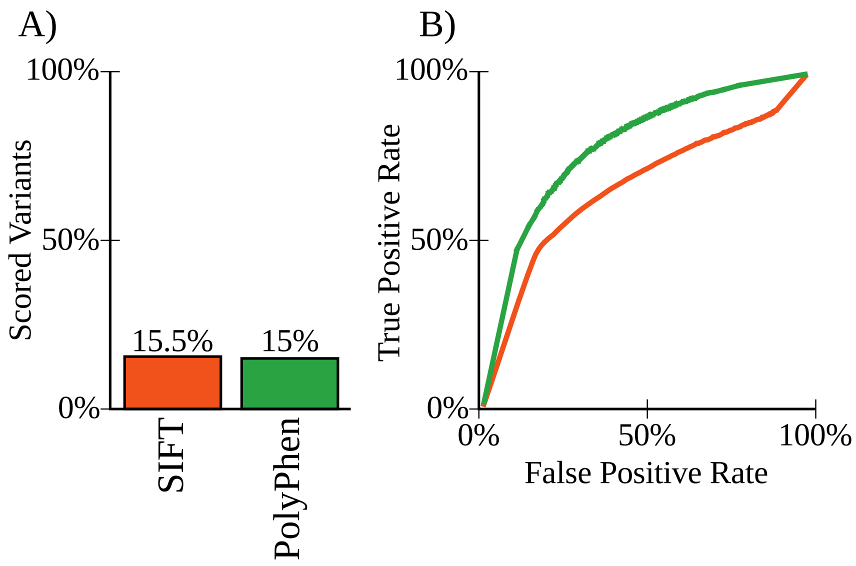


Figure 2.3. Performance Characteristics of Amino Acid Substitution Variant Interpretation

Variants from 100 healthy exomes were analyzed using SIFT and PolyPhen. A) Both algorithms are limited to scoring only amino acid altering, single nucleotide variants—failing to provide an interpretation to 83% of all exome sequence variants. Additionally, the lack of sequence conservation for many protein-coding genes further limits the number of variants that could be score. B) Using coding single nucleotide variants from HGMD (true positive) and the same from 100 healthy exomes (true negative), PolyPhen outperforms SIFT, yet both have considerable false positive rates for the limited variants they are able to score. To recover 80% of the true positives, the SIFT results are contaminated with 67% false positives and PolyPhen with 37%.

either tool unilaterally is impractical for establishing a molecular diagnosis.

2.3.2 Phylogenetic Conservation

Using variant interpretation algorithms based on amino acid substitution rates has limited utility. Despite most known disease causing alleles being found in the coding sequence, and with most of them creating an amino acid change, neglecting 83% of exome variants is simply a nonstarter and likely to lead to misdiagnoses. Studies like the ENCODE project[34] have hoped to shed light on functional characteristics of noncoding regions of the genome. Their results have been less than conclusive, and today, most heritable disorders can be traced to protein coding genes[18]. As exome and whole genome sequencing continue to expand, our emphasis on amino acid altering variants must be expanded to all variant types. Several methods using phylogenetic conservation have been developed using evolutionary signatures to predict the impact of a variant. Observing how a region of the genome has evolved allows for estimates of positive and negative selection that can predict how variation in these regions is tolerated. There are numerous tools that calculate phylogenetic conservation across the human genome, two of them being GERP+[61] and phyloP[62]. These tools are integrated into the UCSC Genome Browser[63] for the analysis of any gene of interest.

2.3.2.1 GERP+

The logic behind using phylogenetic conservation to predict the tolerance of sequencing variants is relatively simple. Compare the genomes of multiple related species and identify genomic regions where variation is lower than normal—i.e., under purifying selection[64]. Variants found in these purifying selection regions will not be tolerated and thus are likely damaging. The GERP+ (Genomic Evolutionary Rate Profiling)[61] tool generates a human centric calculation of phylogenetic conservation. Attempting to improve statistical robustness, resolution, and convey the intensity of conservation, GERP+ returns a score called "rejected substitutions." Sequences from a diverse selection of organisms are aligned to an ungapped human reference. From the multiple-sequence alignment, neutral phylogeny is assumed for those with ungapped alignments. The Rejected Substitution (RS) scores are calculating the rate of change in the neutral phylogeny compared to the rate of change in all aligned sequences. Simply put, RS scores calculate how tolerant substitutions are in a particular genomic region using closely related species, opposed to all species. GERP+ is unable to calculate a RS score on every position in the human genome (e.g., repetitive regions, centromeres, telomeres), but is better suited than SIFT or PolyPhen for scoring variants discovered though exome and genome sequencing. All variants in the 100 example exomes are scored by GERP+ (Figure 2.4A). Unfortunately, the GERP+ calculation is made on a broad evolutionary scope and lacks sensitivity within closely related species[62]. Additionally, because GERP+ uses neutral selection to define regions of purifying selection, it misses evolutionary active regions. Meaning, RS scores at either extreme are predictive and accurate, but middle ranged scores have little to no predictive power.

2.3.2.2 phyloP

GERP+ uses genomic regions evident of purifying selection to predict the tolerance of any given variant, but negates to account for genomic regions displaying faster than normal evolution, e.g., positive selection[65]. Phylogenetic conservation scores generated by the phyloP tool[62] attempt to account for actively evolving regions of the genome by scanning aligned genomic sequencing for fast-evolving regions. phyloP combines four distinct tests into a single score of nonneutral selection rates. The four tests are: 1) a likelihood ratio[66], 2) score[67], 3) SPH[68] and 4) a GERP-like tests[61]. Despite some overlapping features, each one of these tests measures nonneutral selection differently. The likelihood ratio test returns a p-value that the data fit the alternate model better than the null model; the null model being a neutral rate of selection and the alternative is an accelerated or reduced rate. Neutral selection is measured using the score test (Rao's score test). The score test is similar to the likelihood test, but only tests that the data fit the null model, thus measuring neutral selection. The SPH test models the rate of substitutions occurring along all branches of the phylogeny, including distally related organisms and those within the same subtree. Incorporating the SPH test permits phyloP to improve on the inadequacies of GERP+[61]. The final scoring method is a GERP-like test where "rejected substitution" (RS) scores are calculated. All four calculations are coupled into a single phyloP score. Detailing the divergence from neutral selection, the phyloP score records purifying and accelerated evolution. phyloP can reliably interpret all variants (noncoding and coding) in the example exomes (Figure 2.4A).

Without limitations on variant type, performance characteristics were detailed using all variants in the example exomes. Both tools were able to generate a score on every variant, coding or noncoding (Figure 2.4A). I used GERP+ and phyloP to score 97,000 known disease variants from the HGMD database as true positives, and non-

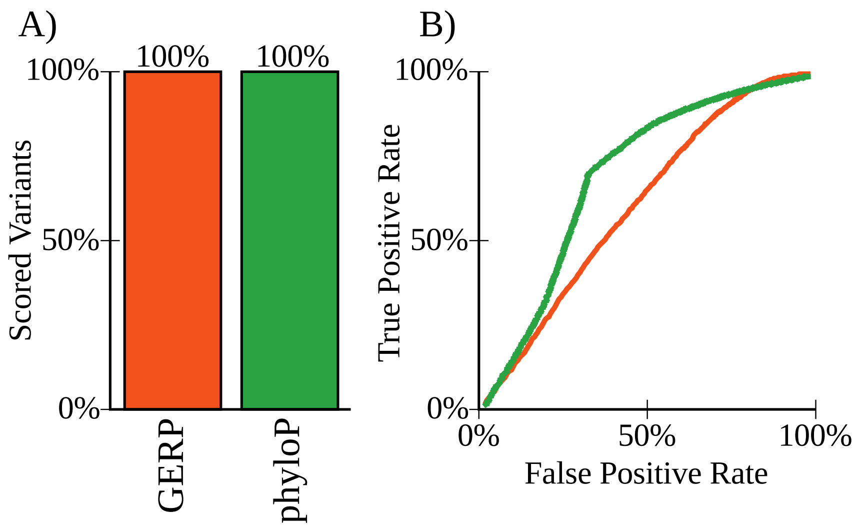


Figure 2.4. Performance Characteristics of Phylogenetic Conservation Variant Interpretation

Variants from 100 healthy exomes were analyzed using GERP+ and phyloP. A) Without variant type or location restrictions, both tools were able to score all variants found in the exomes. B) Using disease alleles from HGMD (true positive) and the 100 healthy exomes (true negative), phyloP outperforms GERP+, yet both have considerable false positive rates. Using phylogenetic conservation to interpret variants, recovering 80% of the true positive results in 64% and 43% false positives using GERP+ or phyloP are used, respectively.

HGMD alleles from the example 100 exomes as true negatives. Suffering from excessive false positives, using these tools alone is ineffective at identifying pathogenic alleles. Interpreting variants phylogenetically, to return 80% of the true positives, GERP+ is contaminated with 64% of the benign alleles predicted as damaging. phyloP does considerably better but still returns 43% of the true negatives as damaging (Figure 2.4B).

2.3.3 Limiting the Variant Search Space

As I have shown, using the interpretation tools described above results in too many variants predicted as damaging to effectively identify the alleles responsible for the phenotype. Databases cataloguing allele frequencies across the population can be used to help reduce the number of damaging alleles. Typically, clinical exome sequencing is performed to diagnose rare or uncommon disorders. As the disorder is infrequent, it is unlikely to be caused by a common variant[69]. Therefore, the frequency at which variants are found in the population can predict how likely it is the variant that causes disease. Unlike described conservation based tools that interpret variants using comparisons to nonhuman sequences, population databases are human specific. Large population-based sequencing projects, e.g., 1,000 Genomes Project (TGP)[70], and the NHLBI GO Exome Sequencing Projects (NHLBI)[71], provide Minor Allele Frequencies (MAF) that can be used to filter exome variants that are common in the population.

2.3.3.1 1000 Genomes Project

The 1000 Genomes Project[70,72] was charged with cataloguing common genomic variants across ethnically diverse human populations. Identifying variants in the population with a frequency at least 1%, the 1000 Genomes Project provides a much-needed atlas of tolerated genomic variation. The initial phase of the 1000 Genomes Project provided low coverage genomic sequencing for variant calls. Unfortunately, low coverage sequencing resulting in significant false negative/positive rates, but was quickly corrected with deeper exome sequencing. Currently there are almost 40 million unique variants detailed in the 1000 Genomes dataset, representing several different ethnic backgrounds. For example, the 1000 Genomes Project has reported observing 79% of the variants found in my 100 example exomes. A clinical exome sequenced for diagnosis of a rare disorder could be filtered against the 1000 Genomes to eliminate common variants unlikely responsible for the phenotype. The 1000 Genomes Consortium periodically releases all sequencing data relating to the project. This provides for the construction of individual specific genotype across genes and haplotypes. While the medical genetics community rarely uses these data, it is immensely valuable for prioritizing genes by their variant burden[43].

2.3.3.2 NHLBI GO Exome Sequencing Project (ESP)

Similar to the 1000 Genomes Project, but more narrowly focused, the NHLBI GO Exome Sequencing Project[73] seeks to establish a variant frequency dataset for individuals with heart, lung and blood disorders[71]. The primary goal of the ESP is to sequence well phenotyped individuals in order to find genes and variants responsible for their disorder. To date 6,400 exomes have been sequenced and the resulting allele frequencies made available. Making the wealth of variant information available provides an informative background of common variants. Unfortunately, the NHLBI group has not released the data for constructing individual haplotypes. With nearly 2 million exome-specific variants, the ESP is rapidly becoming the favored allele database for clinical laboratories performing diagnostic sequencing. Despite only 42% of the example exome variants being shared by the ESP database, nearly all are within the coding regions. Well-established variant frequencies have created an invaluable description of tolerated human variation.

2.3.4 Comprehensive Disease-Gene Finders

Approaches that employ allelic population frequency coupled with the variant interpretation tools (SIFT, PolyPhen, GERP+ and phyloP) to filter out common and benign alleles provide effective methods for finding rare disease alleles[74]. Many recently discovered disease alleles were identified using such methods. It is unclear, however, how often these filtering methods fail. Filtering variants based on their population frequency (MAF) risks losing true pathogenic alleles. Once the disease allele is filtered out, all hope is lost for accurate diagnosis, resulting in one of two scenarios: inaccurate diagnosis or an inconclusive result. Both will likely have dramatic impact on an individuals continued care. In spite of risks associated with filtering variants, most exome analysis (research and clinical) is done using filtering. The popular tool Annovar[41] employs a simple automated filtering process.

2.3.4.1 Annovar

Quick and automated variant annotation and filtering is highly appealing to most clinical genomic labs, as annotating exome variants continues to remain a challenging endeavor[75]. Annovar[41] performs two bioinformatics functions: 1) functional annotation of variants using current gene models, and 2) variant reduction through serialized filtering steps to select candidate disease alleles. Annovar's popularity stems from its rapid functional annotation methods (e.g., assigning genes, coding positions, and amino acid changes), methods using gene models downloaded directly from the UCSC Genome Browser[63]. Annovar is able to functionally annotate known and novel exome variants. Clinical laboratories that do not have sufficient informatics capability to annotate exome variants themselves utilize Annovar's fast analysis methods. Unfortunately, Annovar does not incorporate no-call variants, thus missing many potentially informative variants.

Mirroring the data reduction methods used by many clinical laboratories, Annovar has been cited in many disease gene discoveries[76,77]. Reducing the number of variants, Annovar attempts to identify rare damaging alleles potentially responsible for an individual's disease. To do so, it filters exome variants against population databases, disease allele databases, phylogenetic conservation scores, amino acid substitution scores and their function annotation. Filtering is done in a series of steps, returning ever fewer number of potential candidate genes at each step. Users specify the order and what filtering criteria to use as well as the inheritance model that the remaining variants should fit.

To demonstrate Annovar's variant reduction functionality, I used the 100 exomes example dataset assuming recessive inheritance and the default filtering criteria as suggested by developers of Annovar. The massive data reduction is detailed in Table 2.4. Beginning with functionally annotated variants, Annovar filters out any that are not in the coding or splicing regions, reducing the variant pool by 76%. Hoping to account for sequencing and alignment errors, Annovar filters variants found in segmental duplication regions, dropping the variant pool by another 2%. Using phyloP[62] scores Annovar filters out variants unless they are in “rejecting variants” regions that are intolerant to variation. Phylogenetic conservation filtering reduces the pool of variants to only 9% of the original. Annovar is able to filter on the minor allele frequency for many population databases. Filtering variants found with a minor allele frequency greater than 1%, Annovar filters out variants found in the 1000 Genomes[72], followed by dbSNP[78]. It is important to note that dbSNP filtering excludes previously associated or flagged disease alleles. Post minor allele frequency filtering, only 1% of the original variant pool remains. The final steps remove variants that do not alter the amino acid sequence or reside in the splicing regions, then assign the variants to their annotated genes and check that the remaining fit the inheritance model. From the original 73,000 variants, Annovar filters out all but 629 variants annotated in 60 candidate genes. Variant reduction by filtering returns discrete results. Each variant escaping filtering is a potential candidate, and all filtered variants are not. There is no additional ranking or scoring of the candidates within the final candidate gene list.

Annovar’s process is simple, straightforward and easy to perform, however, the drastic filtering has many risks that could alter the patient’s diagnosis and their treatment, even possibly cause more harm. Unfortunately, exome analysis by filtering

Table 2.4 Variant Reduction and Filtering Steps by Annovar using a Recessive Inheritance Model

Annovar Step	Filtering Criteria	Remaining Variants	% Filtered
0 – Initially Discovered Variants	None	73,018	0%
1 – Annovar annotation	Remove no-call variants	50,048	31%
2 – Intronic/Exonic	Must be in coding or splice region	12,248	83%
3 – Segmental Duplications	Must not be in segmental duplication region	10,865	85%
4 – Phylogenetic Conservation	Must have positive phyloP score	4,435	94%
5 – 1000 Genomes	Must not have a MAF > 1%	784	98.9%
6 – dbSNP	Must not have a MAF > 1%	647	99.0%
7 – Variant Type	Must alter amino acid sequence or splicing sites	629	99.1%
8 – Final Candidate Gene List	Must match recessive inheritance model	60 – Genes	99.7%

has become the standard for most labs[7]. Only when every variant is accurately interpreted, will the diagnostic potential of exome sequencing be achieved. Genome wide variant interpretation requires a subsequent prioritization of the variant, followed by logical connections between the gene and variant(s) responsible for the phenotype. Filtering alone will never be able to establish the variant-to-phenotype relationship.

2.3.4.2 VAAST

The Variant Annotation, Analysis and Search Tool (VAAST)[42,43] provides a probabilistic framework to analyze every variant in a personal genome sequence. Furthermore, VAAST returns a prioritized list of all genes ranked according to their likelihood of being damaged. VAAST represents a new breed of disease-gene finder, one that incorporates all the power and sensitivity from phylogenetic and amino acid conservation, combined with target specific analysis using population allele frequencies. Similar to Annovar, the VAAST suite of tools includes a functional annotator. Known as VAT, the VAAST annotator is able to functionally annotate genomic variants using the chosen gene model. Unlike Annovar, VAAST is able to combine sequencing results from multiple individuals into a single target, e.g., variants from a cohort with the same phenotype, or parent-child cohorts. Using the VAAST tool VST, variants from a cohort can be combined, intersected, or compared appropriately for analysis. VST allows for specific target (affected) and background (unaffected) populations to be established (Figure 2.5). Creating separate target and background files eliminates dependence on third-party databases (e.g., dbSNP) and allows for

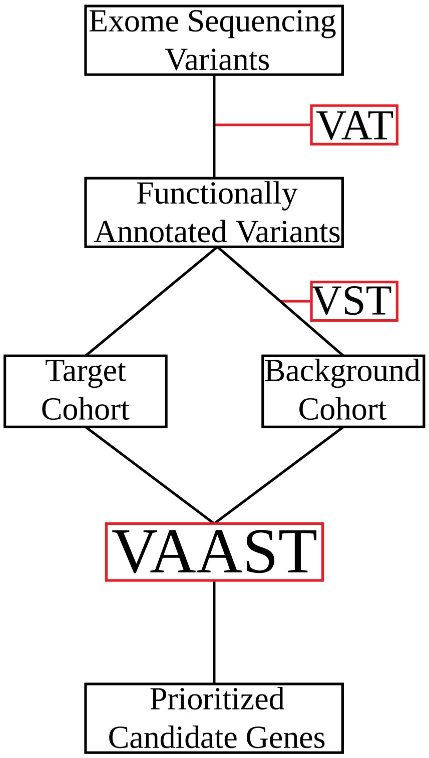

Exome Sequencing Variants
VAT
Functionally Annotated Variants
VST
Target Cohort
Background Cohort
VAAST
Prioritized Candidate Genes

Figure 2.5. Basic VAAST Methodology Sequence variants are functionally annotated using VAT. Using VST the union, intersection or complement of variants from multiple individuals creates target and background cohorts. These cohorts are given to the VAAST probabilistic disease-gene finder.

proper matching between cases and controls. Target and background cohorts are used by VAAST to establish minor allele frequencies and to develop *empirical* amino acid substitution rates to identify the pathogenic gene.

VAAST is a probabilistic disease-gene finder designed for the analysis and prioritization of deleterious variants in a personal genome or exome sequence. Using gene features (e.g., transcript, exon, coding sequence) VAAST computes a composite likelihood ratio for each gene and returns a prioritized list of genes ranked by the likelihood of being deleterious. Computing on genes and their features, as opposed to individual variants, VAAST is able to sidestep massive Bonferroni correction that so plagues Genome-Wide Association Study (GWAS) approaches[79]. VAAST's Composite Likelihood Ratio Test (CLRT) measures the likelihood that the impact of variants in a gene in the target differs from those in the background. All three of the previously described techniques for variant interpretation are incorporated into the VAAST CLRT: population frequency, amino acid substitution rates, and phylogenetic conservation. Significance of the CLRT (VAAST score) is determined through permutation, which also controls for linkage disequilibrium[80].

2.3.4.2.1 Composite likelihood ratio test – population frequency. The CLRT is used to determine if the allele frequency difference between the background and target agrees with the null model (there is no difference) or the alternative model (there is a difference). For each variant the log likelihood ratio (λ) is the natural log of the ratio of the null model likelihood for the variant (Li) over the alternative likelihood.

$$\lambda_i = \ln\left(\frac{L_i^{NULL}}{L_i^{ALT}}\right) \tag{2.1}$$

The null model is the likelihood that variant frequency in the target (*T*) and background (*B*) represent the same population. Null model $\left(L_i^{NULL}\right)$ is calculated by combining the likelihood of observing the target allele at position *i* in the target and background with the likelihood of observing the background allele at position i in the target and background.

$$L_i^{NULL} = (L_i^T | B \cap T) \times (L_i^B | B \cap T) \qquad (2.2)$$

To calculate the likelihood of observing the target allele at position i in the background and target, 1) the number of different ways a target allele (n_T) can be chosen out of all target alleles (k_T) is combined with 2) minor allele frequency in the target and background $\left(\hat{P}_i\right)$ raised to the number of minor alleles in the target (k_T) and the 3) major allele in the target and background $\left(1 - \hat{P}_i\right)$ is raised to number of major alleles in the target $(n_T - k_T)$.

$$(L_i^T | B \cap T) = \binom{n_T}{k_T} \times \hat{p}_i^{k_T} \times (1 - \hat{p}_i)^{n_T - k_T} \qquad (2.3)$$

This is also done to calculate the likelihood of observing the background allele at position i in the background and target $(L_i^B | B \cap T)$.

$$(L_i^B | B \cap T) = \binom{n_B}{k_B} \times \hat{p}_i^{k_B} \times (1 - \hat{p}_i)^{n_B - k_B} \qquad (2.4)$$

The alternative model is the likelihood that the variant frequency in the target (*T*) is from a different population than the background (*B*). To calculate the alternative model, the likelihood that the target allele is observed at position *i* is combined with the likelihood that the background allele is observed at position i in the background.

$$L_i^{ALT} = (L_i^T|T) \times (L_i^B|B) \qquad (2.5)$$

Calculating the likelihood of observing the target allele at position i in the target combines 1) the number of alleles in the target (n_T) given the total possibilities (k_T) with 2) minor allele frequency from the target (p_i) raised to the number of target minor alleles (k_T) and the 3) target major allele frequency $(1 - p_i)$ raised to the number of major alleles in the target $(n_T - k_T)$.

$$(L_i^T|T) = \binom{n_T}{k_T} \times p_i^{k_T} \times (1 - p_i)^{n_T - k_T} \qquad (2.6)$$

Again, the same equation can be applied to the background to calculate the likelihood of observing the allele at position i in the background.

$$(L_i^B|B) = \binom{n_B}{k_B} \times p_i^{k_B} \times (1 - p_i)^{n_B - k_B} \qquad (2.7)$$

2.3.4.2.2 Composite likelihood ratio test – amino acid substitution rates. Adding *ad hoc* generated amino acid substitution rates as extensions to the CLRT allows VAAST to improve its ability to identify damaging alleles. The previously describe variant interpreters that use amino acid substitution rates (SIFT and PolyPhen) are generated from multispecies alignments. These methods fail to detail the level of amino acid substitution tolerance within humans; for example, stop codons are never seen in phylogenetic multiple alignments. VAAST takes an *ad hoc* approach to amino acid substitution rates using rates observed in the background population compared to alleles known to cause disease. The substitution rates of amino acid changes observed in the target and background cohort (a_i) are multiplied to the null model, and the substitution rates, from known disease alleles in HGMD (n_i) are used in the alternative model.

$$\lambda_i = \ln\left(\frac{\boldsymbol{a_i} L_i^{NULL}}{\boldsymbol{n_i} L_i^{ALT}}\right) \tag{2.8}$$

For example the amino acid change glycine to tyrosine is observed in the 1000 Genomes72 variants (background) at a frequency of 0.001817 but in the HGMD[18] dataset (disease) it is observed with a frequency of 0.02166. Multiplying these *empirical* substitution rates to the CLRT improves VAAST's ability to distinguish between deleterious and benign alleles.

2.3.4.2.3 Composite likelihood ratio test – phylogenetic conservation. Additional specificity can be had using phylogenetic conservation. PhastCons scores[81] measures sequence conservation shared across vertebrates. Primarily, PhastCons scores highlight regions of purifying selection, i.e., regions intolerant to variation. As PhastCons scores are available for all informative positions in the genome (including noncoding regions), they can be used to extend the CLRT when scoring each variant. Extending the CLRT with PhastCons increases the accuracy for scoring both noncoding and coding variants.

2.3.4.2.4 Composite likelihood ratio test – significance. To determine significance of the composite likelihood ratio, permutation is performed, randomizing the affected status (target or background) and recalculating the CLRT. Combinations of target and background sets that score higher than the original CLRT reduce the score's significance. By permuting on the randomized affected status, VAAST is able to control for linkage disequilibrium between variants. Controlling for linkage disequilibrium is only possible when the entire genotype of a gene is known, the 1000

Genomes Project makes these data available while NHLBI does not.

VAAST ranks candidate genes by their significance followed by their score. Ranking this way provides effective and accurate prioritization of genes according to how likely the variants within them are deleterious. Because the significance of the CLRT is dependent on permutation testing, the number of individuals in the target and background limits VAAST's power. Thus, VAAST's accuracy suffers when limited data are available.

Performance characteristics of Annovar[41] and VAAST[42,43] greatly exceed those of any single variant interpretation algorithm, e.g., SIFT, PolyPhen, etc. VAAST and Annovar are comprehensive disease gene finders that incorporate the features of the bioinformatics tools describe above. Using known disease alleles from HGMD inserted into the 100-example exomes dataset, VAAST and Annovar were used to recover the disease causing alleles. Because Annovar does not incorporate no-call variants into its disease gene search, it is only able to interrogate 69% of the variants from the 100 example exomes, as seen in Figure 2.6A. Annovar's performance is impressive compared to any single variant interpreter. However, the returned results from Annovar are discrete; meaning a gene is considered a candidate, or it is not. VAAST demonstrates the best performance of any tool described, resulting in only 7% false positives to return 80% of the true positives. Lower than any other described tool, the 7% returned false positives by VAAST still accounts for over 1,500 genes (Figure 2.6B). Simply put, with only a single exome, no existing tool is able to accurately interpret and prioritize variants effectively to establish an accurate molecular diagnosis. Described as comprehensive disease-gene finders, VAAST and Annovar neglect to incorporate the single most valuable part of any disease gene search: the patient's phenotype.

2.4 Limited Power

The power to find the allele responsible for an individual's phenotype is unfortunately limited by the size of the case cohort. As I will explain in the following chapters, with a cohort of only a single individual, every diagnostic tool is underpowered for *ab initio* association of a variant with disease. Because of this, most successful exome based disease-gene searches have included sequences either from multiple unrelated individuals or from family members in addition to the proband. Unfortunately, clinical genomic testing is all about single affected individual, or best case scenario, an affected individual and their unaffected parents[7]. In such cases, neither VAAST nor Annovar offers the clinical laboratory much power to find the variant responsible for the phenotype, particularly if the allele they are searching for is novel. The lack of power helps explain why most clinical labs focus only on known disease alleles. However, as discussed, focusing only on known alleles leaves many patients undiagnosed, and eliminates the possibility of ever making a novel diagnosis. An estimated 8% of the population has a diagnosable inherited disorder[82], but clinical genomics is only reporting a 25% molecular diagnostic rate[7] of rare phenotypically severe disorders. This rate obviously needs to improve. This improvement will not come from better variant calling or variant prioritization—as I have explained, the information is simply not there for diagnosis using only a single proband sequence. To improve diagnostic rates, computational means are needed that can make the connection

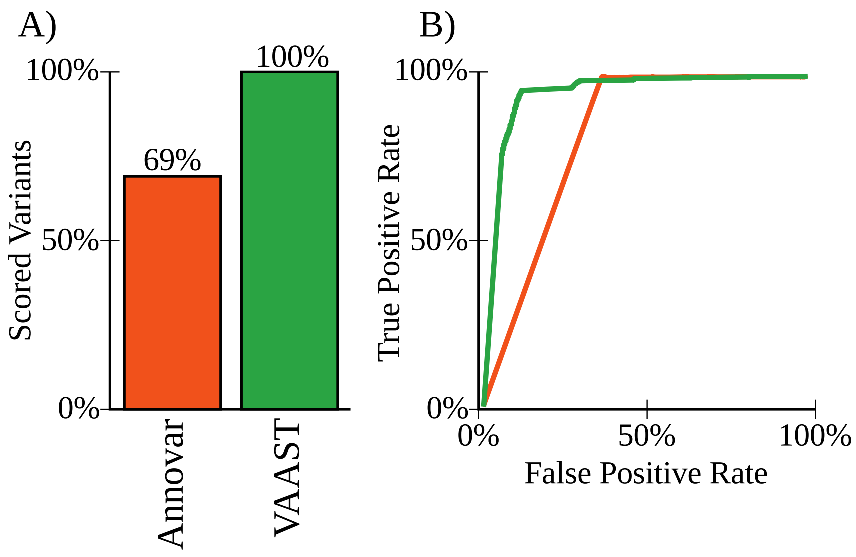


Figure 2.6 Performance Characteristics of Comprehensive Disease-Gene Finders

A) Variants from the 100 healthy exomes were analyzed by Annovar and VAAST. Unlike VAAST, Annovar is unable to annotate or interpret no-call variants, missing 31% of potential damaging alleles. VAAST can provide interpretation results on all variants. B) Known disease alleles from HGMD inserted into one of 100 exomes used to characterized Annovar and VAAST's performance. Annovar's filtering and unordered candidate gene list returns many false negatives, while VAAST's superior probabilistic scoring method accurately returns all known alleles at the cost of many false positives.

between genes having damaging variants to the patient’s observed phenotype. My dissertation work has attempted to address this problem. Phevor, the Phenotype-Driven Variant Ontological Re-ranking tool[35], is the result. As I will demonstrate in the following chapters, my results to date indicate that given the right approach, combining variant interpretation and prioritization with phenotype information can dramatically improve the accuracy of molecular diagnoses for rare diseases. I will also present unpublished results that strongly suggest Phevor also has a significant role to play in diagnosis of common disorders like diabetes and autism.

CHAPTER 3

DISEASE ALLELE DATABASES AND BIOMEDICAL ONTOLOGIES

The objective of clinical genomics is to identify variant(s) that can explain an individual's phenotype. As discussed, there are many ways to identify likely damaging variants, but as explained in the previous chapter every individual contains thousands of such variants in their exome. Thus, the challenge is determining which of the possibly damaging variants is actually responsible for the phenotype.

An accurate diagnosis hinges on the ability to correlate the damaged gene with the individual's phenotype. Currently, clinically sequenced exome analysis relies heavily on databases of known disease alleles to make this correlation. Known alleles, however, represent only a fraction of inheritable disease (Table 3.1), leaving many patients undiagnosed. Phevor[35] is not bound by the limitations of existing associations between known variants and their distinct phenotypes. Rather, it extrapolates biological properties and functions from biomedical ontologies to connect variants with phenotypes. Phevor does so using information resident in biomedical ontologies. To understand how Phevor relates a gene's properties to its phenotype some general knowledge is needed about ontologies and how ontology format improve upon existing

Table 3.1 Breakdowns of Alleles in Disease Databases

Gene Specific Disease Databases		
OMIM[32]	Genes	14,537
	Phenotype with molecular basis known	4,076
	Phenotype with molecular basis unknown	1,708
Orphanet[83]	Genes	3,130
	Phenotype with molecular basis known	6,207
Variant Specific Disease Databases		
HGMD[18]	Genes	6,137
	Variants	148,413
ClinVar[33]	Genes	18,714
	Variants	99,880

disease databases.

3.1 Disease Allele Databases

To employ a database of disease alleles into your exome analysis, you need to know exactly what you are trying to find. With 73,000 variants in a typical exome sequence, sometimes you are fortunate and find your patient has a known disease allele—and that the disease allele causes the phenotype you are looking for. Unfortunately, the situation is usually much less clear. Consider, for example, an individual with the phenotypes of *long QT syndrome* and *bilateral hearing loss*, also known as *Jervell and Lange-Nielson syndrome*. Querying this patient's variants against a disease database finds one variant annotated with the disease *cardiomyopathy*. While it is true that *long QT syndrome* can be a characteristic of *cardiomyopathy*, from this annotation there is no way to know if the variant is responsible for both phenotypes, or an entirely different *cardiovascular abnormality*. Certainly the disease database annotation is not enough to make a molecular diagnosis of *Jervell and Lange-Nielson syndrome*[84], and possibly too vague to make any diagnosis. Confusion in allele annotations like this highlight one of the biggest problems with using today's disease allele databases to identify variants responsible for the patient's disease—another is the laxity of phenotypic descriptions. Often these are so vague as to be downright misleading.

3.2 Phenotype Descriptions

The phenotypes found in disease allele databases are intended to aid in the diagnosis. Unfortunately, these descriptions are usually less than helpful. Take for instance the Online Mendelian Inheritance of Man (OMIM)[32] entry #263750 – POSTAXIAL ACROFACIAL DYSOSTOSIS; PODS, better known as *Miller's syndrome*[85]. Despite *Miller's syndrome* being extremely rare disease, seven separate clinical descriptions exist in OMIM. Looking closer at three of these descriptions it is difficult to know for sure what the phenotype for *Miller's syndrome* truly is, or if all three entries are even describing the same disorder (Figure 3.1). Still worse, many older OMIM entries are idiosyncratic, having extreme deviations in style and vocabulary between clinical descriptions.

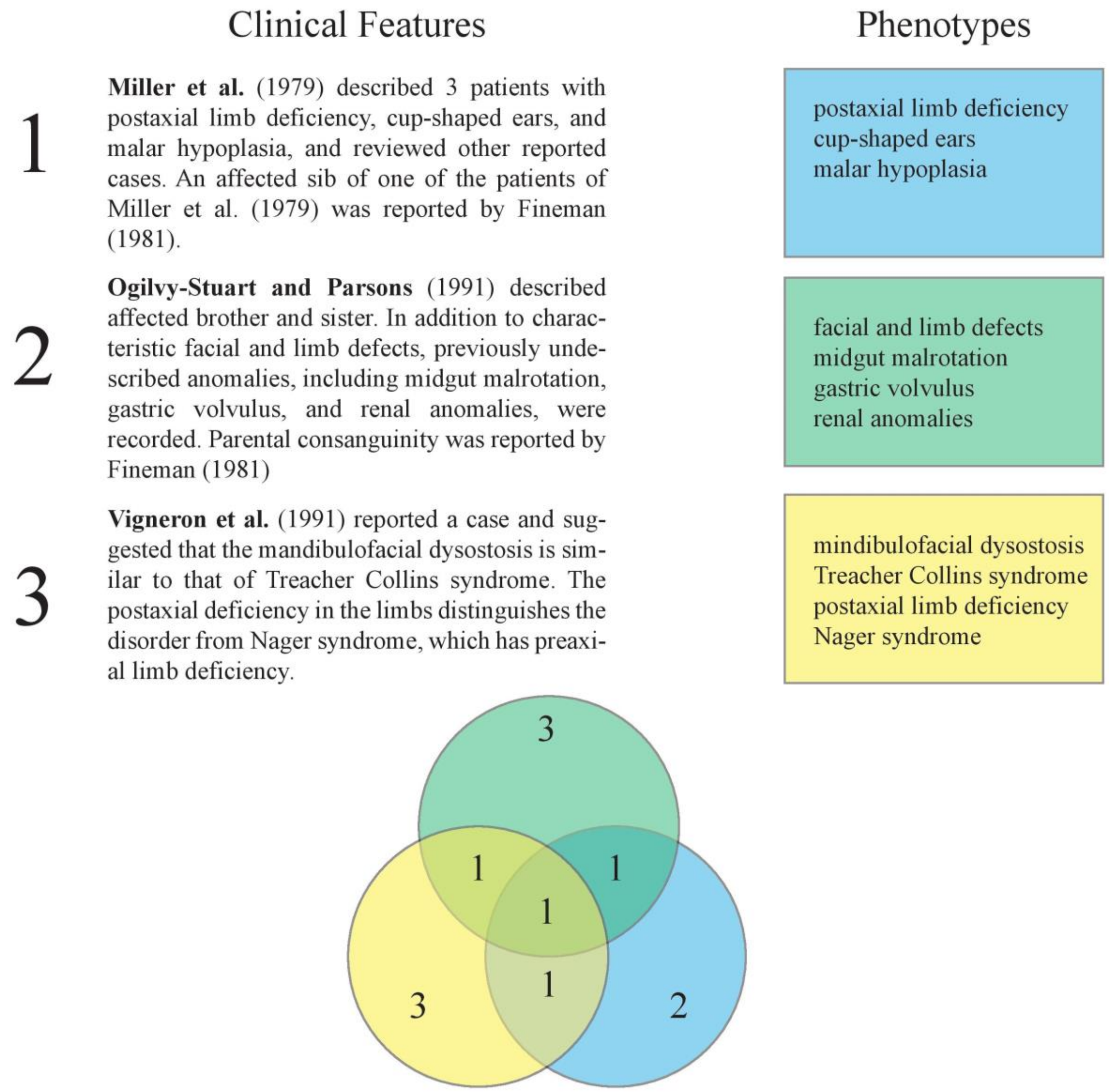
In response to criticism, OMIM has recently tried to improve its clinical descriptions by aggregating every phenotype associated with a given disease into a single clinical synopsis. These clinical synopses use controlled terminologies to describe phenotypes. A marked improvement certainly, but every clinical description is aggregated neglecting to properly account for the accuracy of the descriptions. This work, still underway, will result in a huge improvement to OMIM, but it still falls short.

Employing controlled terminologies is a step forward, but still missing from this approach are means to model relationships between the terms. As I explain below, ontologies provide a solution to this problem.

3.3 Introduction to Ontologies

Ontologies are relatively new to the medical genetics and bioinformatics communities, yet their origins are literally ancient. Originally defined by Aristotle and Plato, "*as the metaphysical study of being*" (384 BC – 322 BC), i.e., ontology is the philosophical study of being, and the relationships therein. Distilled down to one sentence, an ontology is a categorized set of ideas, terms, or themes that have hierarchical relationships to one another. Aristotle proposed that there are ten high-level categories that describe all being, thus creating the first ontologically organized database. Aristotle's approach was reborn in the late 1990s with the growth of the world wide web.

As the Internet grew in size, it became difficult to search for web content using exact or simple string matching, similar to the difficulties encountered using disease allele databases. The search engine Yahoo![86] was among the first to begin categorizing documents and web content for fast searching. By categorizing each entry, Yahoo! could return an exact match along with related content. Returning related content became especially important as e-commerce developed. Today, most e-commerce and Internet search engines utilize an ontological structure in some form[87,88]. The use of ontologies to describe the contents of databases has become so widespread that a particular term, "Knowledgebase" has come into use to describe such databases. Knowledgebases[89,90] lie at the heart of just about every e-commerce business. For example, using eBay to get an original print edition of "Origin of the Species" by Charles Darwin, eBay will return all results annotated in the category "Antiquarian & Collectible" under the category of "Biology," then, "Books." Categorizing items on e-



1 **Miller et al.** (1979) described 3 patients with postaxial limb deficiency, cup-shaped ears, and malar hypoplasia, and reviewed other reported cases. An affected sib of one of the patients of Miller et al. (1979) was reported by Fineman (1981).

2 **Ogilvy-Stuart and Parsons** (1991) described affected brother and sister. In addition to characteristic facial and limb defects, previously undescribed anomalies, including midgut malrotation, gastric volvulus, and renal anomalies, were recorded. Parental consanguinity was reported by Fineman (1981)

3 **Vigneron et al.** (1991) reported a case and suggested that the mandibulofacial dysostosis is similar to that of Treacher Collins syndrome. The postaxial deficiency in the limbs distinguishes the disorder from Nager syndrome, which has preaxial limb deficiency.

postaxial limb deficiency
cup-shaped ears
malar hypoplasia

facial and limb defects
midgut malrotation
gastric volvulus
renal anomalies

mindibulofacial dysostosis
Treacher Collins syndrome
postaxial limb deficiency
Nager syndrome

Figure 3.1. OMIM Provided Clinical Descriptions and Phenotypes for Miller's Syndrome

Clinical descriptions curated by OMIM contain idiosyncratic language open to interpretation of the reporting author. Three separate entries in OMIM describing Miller's syndrome demonstrate the subjective descriptions of the same disorder. From these three entries, only one phenotype overlaps, exemplifying the difficulties using OMIM as a diagnostic guide.

commerce sites allows for fast accurate searching, and returns related entries when the exact search criteria cannot be met. Returning related results give the shopper a choice when their specific request is unavailable. Utilizing today's disease allele databases, when nothing exact is found, nothing is returned. If the disease database was organized as an ontology, related diseases and disease alleles could be returned even when an exact match is not available.

3.3.1 Ontology Structure

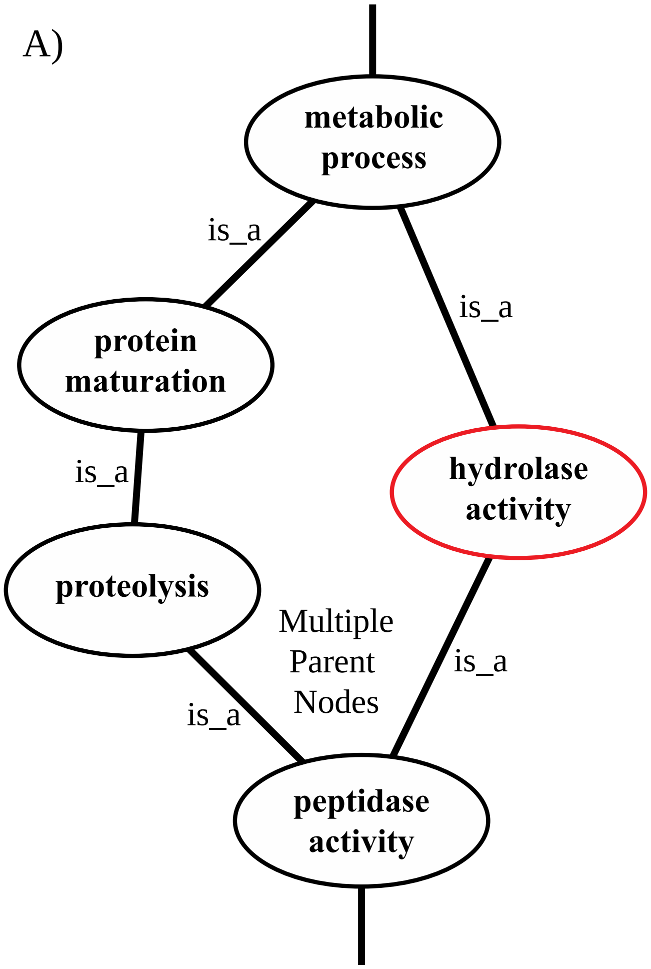
Ontologies also have many other features, besides those described above making knowledgebases superior to flat databases such as OMIM[32] or HGMD[18]. One notable feature of ontologies is that they provide a controlled vocabulary in which to describe the contents of a database. Another key feature is that ontologies relate the terms comprising their controlled vocabularies to one another using a second controlled vocabulary of relationship terms.

As mentioned previously, OMIM is improving its utility by translating the author clinical entries into descriptive controlled vocabularies, thus reducing the confusion caused by subjective or idiosyncratic descriptions. Doing so permits machine-readable parsing and data mining of these OMIM entries. However, controlled vocabularies alone are not what make knowledgebases superior to flat databases. Ontologies also relate each entry to all other entries via their relationship terms. Ontologies organize their data using each entry's relationship to all other entries[91]. That is to say, using terms such as, ***is a, part of, starts during, positively regulates, etc***., each entry or term in the ontology can be related to every other term. From these relationships, organizational trees can be constructed, relating each term to a common parent that describes all terms within the ontology. Ontology trees are organized as Directed Acyclic Graphs (DAG)[92] where each entry (term) can have multiple parents, each parent thus being less specific in nature and containing all the properties and attributes of each child (Figure 3.2A). At each term or node of the ontology, multiple attributes can describe the term, including definitions, cross-references, primary literature, and of course gene annotations (Figure 3.2B).

3.3.2 The Sequence Ontology

Exemplifying the directed acyclic graph structure of an ontology, the Sequence Ontology[93] provides biologically relatable terms and relations used to describe genomic sequence annotations. Originally designed by the Gene Ontology Consortium[94], the Sequence Ontology provides a means to unify annotation descriptions across the model organism communities. It became clear that each group was using different terminology to describe the same sequence attribute; e.g., translation start, coding start and translation initiation site all describe the Sequence Ontology term coding_start. The term coding_start only has three synonyms, trying to account for all the differing terminologies across all model organisms would elevate confusion beyond reason. Standardizing terminology is the first step in diminishing complexity, but where the Sequence Ontology excels is in its organization of the relationships between annotation terms. Clinical genomics focuses its efforts on protein coding genes; therefore, to illustrate the logic built into ontologies the example of *mRNA* is detailed in Figure 3.3.

The Sequencing Ontology defines *mRNA* as: "*an intermediate molecule between*


A)
metabolic process
is_a
is_a
protein maturation
is_a
hydrolase activity
proteolysis
Multiple Parent Nodes
is_a
is_a
peptidase activity

B) id: GO:0016787
name: hydrolase activity
namespace: molecular_function
def: "Catalysis of the hydrolysis of various bonds, e.g. C-O, C-N, C-C, phosphoric anhydride bonds, etc. Hydrolase is the systematic name for any enzyme of EC class 3." [ISBN:0198506732]
xref: Reactome:REACT_15331 "Hydrolysis of phosphatidylcholine, Homo sapiens"
is_a: GO:0008152 ! metabolic process

Figure 3.2 Generalized Ontology Structure and Example Node Attributes
A) Adhering to the rules governing directed acyclic graph structure, parental nodes describe less specific terminology then their child nodes; e.g., *metabolic process* is less descriptive than *protein maturation* or *hydrolase activity*. Ontology nodes can have multiple parents, e.g., *peptidase activity* is a child node of both *proteolysis* and *hydrolase activity*. B) The gene ontology node *hydrolase activity* contains the following attributes. namespace: the root node that own this term, def: description of node's meaning, xref: cross-reference supplying reasoning behind node creation, is_a: parental node ID number(s) and name(s).

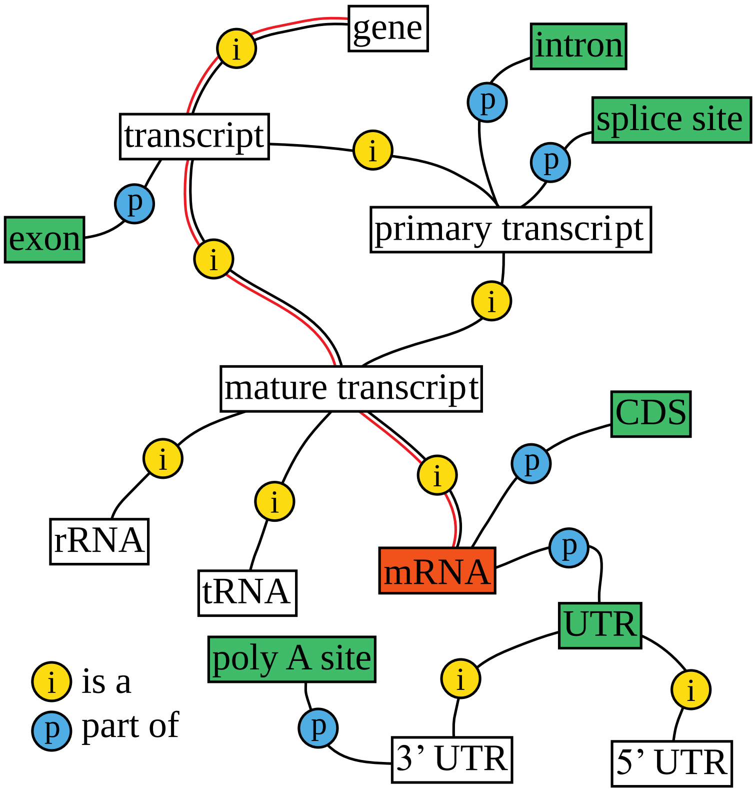

gene
intron
transcript
splice site
exon
primary transcript
mature transcript
CDS
rRNA
tRNA
mRNA
UTR
poly A site
i is a
p part of
3' UTR
5' UTR

Figure 3.3. Sequence Ontology Illustration of mRNA Processing and Components

Using the Sequence Ontology's representation of terms connected to *mRNA*, the different components and processing steps are easy to follow. *mRNA* is generated from a gene, first as a *transcript* or *primary transcript*, then following processing—*a mature transcript* (red line). The final *mRNA* has various parts, including the Untranslated Regions (*UTR*) and Coding Sequence (*CDS*). Parental terms contain the attributes of their child terms; the *mature transcript* contains all of *mRNA's* elements, but could also contain all the attributes of *rRNA* and *tRNA*. Child nodes do not inherit the attributes of their parents, for example, *primary transcript*; the parent node of *mature transcript* has *introns* and *splice sites* as parts. The *mature transcript* does not contain either of these components, but does contain the *CDSs*, and *UTRs*, which are parts of the child nodes of both *mature transcript* and *primary transcript*.

DNA and protein. It includes UTR and coding sequences. It does not contain introns." We can conclude that *mRNA* is the postprocessed result of a *transcript*, transcribed from a *gene*. In ontological terms, *mRNA* **is a** *mature transcript*, which **is a** *transcript* or *primary transcript*, which **is a** *gene*. Of course this is a simple of components of mRNA and transcripts. We know that the coding sequence (*CDS*) is part of *mRNA*, as well as the untranslated regions (*UTR*) made up of five and three prime UTRs. The DAG structure of the Sequence Ontology dictates each parent term have all the components and attributes of the child terms, but the child does not inherit all the parent's attributes. Therefore, *CDS* and *UTR* are **part of** *mRNA*; they are likewise **part of** the *mature transcript* and the *primary transcript*. However, *introns* and *splice sites* are **part of** the *primary transcript* but are removed from the *mature transcript*, and not in the *mRNA*. Sequence Ontology descriptions make it easy to represent complex interactions and relationships like this in a meaningful way. Trying to represent the interconnected components and actions that go into mRNA processing in a predetermined database would require complicated connections difficult to interpret.

3.4 Biomedical Ontologies

The interconnected nature of any biological system makes them difficult to represent accurately in a flat database structure. On the other hand, the relationship structure of an ontology is very suitable for describing biological ideas. The Gene Ontology[94], for example, was created to provide a unified resource for managing the highly interconnected biological properties of genes in a machine-readable fashion. This created a powerful tool for exploring and interrelating genes' biological processes, molecular functions and cellular components. Properties of genes studied in *Drosophila*, for example, can be related to their human homologues, thus accelerating the understanding of the human genome. Importantly this need not be done manually—the Gene Ontology made it possible for software to automate what was previously performable only by knowledgeable scholars.

The success of the Gene Ontology[94] demonstrated the utility of organizing biological data into an ontological format and spurred the development of other biologically related ontologies. In 2007, the Open Biological Ontologies (OBO) Foundry[95] was established to integrate and organize bio-ontologies into a single repository. Each ontology housed under the OBO Foundry banner has differing domains of interest and specificity—e.g., the Drosophila Development Ontology[96], Human Phenotype Ontology[97] and Neuro Behavior Ontology[98]. Ontologies found on the OBO Foundry can be considered works-in-progress that are periodically updated.

3.5 Establishing Gene Function and Phenotype Using Bio-Ontologies

Bio-ontologies span many biological subspecialties, but only a subset has gene annotations included as one of their attributes, a requirement for connecting damaged genes to phenotype. Phevor[35] (Chapter 4) incorporates several gene-annotated bio-ontologies: Gene Ontology[94], Pathway Ontology[99], Mammalian Phenotype Ontology[100], Rat Disease Ontology[101], Human Phenotype Ontology[97], Disease Ontology[102], and Chemical Entity of Biological Interest[103]. Phevor combines the information resident in each bio-ontology to accurately identify genes that will result in any given phenotype. Below I describe each of the ontologies employed by Phevor. Detailed in Table 3.2 are

Table 3.2 Contents of Biomedical Ontologies with Gene Annotations

Bio-Ontology	Nodes	Synonyms	Depth	Genes
Gene Ontology[94]	40,349	93,644	16	18,927
Human Phenotype Ontology[104]	10,324	6,480	15	2,872
Disease Ontology[102]	8,713	17,678	12	4,340
Mammalian Phenotype[100]	10,096	16,685	15	7,452
Rat Disease Ontology[101]	12,092	65	11	4,399
Pathway Ontology[105]	1,507	731	9	4,733
Chemical Entity of Biological Interest[106]	44,309	266,956	28	19,253

various metrics defining each of the described bio-ontologies. In Chapter 4, I explain how Phevor employs these ontologies for disease-diagnosis.

3.5.1 Gene Ontology

The Gene Ontology[94] is the most commonly used and mature bio-ontology. It was established to provide a structured unification of gene annotations and their products. Within the Gene Ontology, genes are cataloged by their functional properties, including their biological process, molecular function and cellular components. Prior to the Gene Ontology, there were several hierarchical datasets describing functional properties of genes and their products, but the Gene Ontology was the first to do so using a directed acyclic graph[92]. Representing functional properties of genes in an ontological format allows for biologically relevant analyses of genes. The Gene Ontology continues to be maintained by the GO Consortium—a community of researchers actively involved in biological research across many model organisms.

3.5.2 Pathway Ontology

The Pathway Ontology[99] was developed and is currently maintained by the Rat Genome Database. Analogous to the Gene Ontology, terms within the Pathway Ontology describe biological properties but do so as interconnected reactions and interactions describing the working relationships between biomolecules. High-level terminology utilized by the Pathway Ontology includes classical metabolic pathway, disease pathway, drug pathway, regulation pathway and signaling pathway. From these high level terms, biologically active genes and their products are further classified into increasingly more precise pathway components.

3.5.3 Mammalian Phenotype Ontology

Developed and maintained by the International Mouse Phenotype Consortium[107] (IMPC), the Mammalian Phenotype Ontology[100] organizes the finding of the IMPC's attempt to identify the function of every gene in the mouse genome. Mouse is possibly the best model organism for modeling human disease, and phenotypic insight gained from its study is invaluable to predict human disease.

3.5.4 Rat Disease Ontology

Comparable to the Mammalian Phenotype Ontology, the Rat Disease Ontology[101] documents phenotypes observed in rats. Maintained and promoted by the Rat Genome Database, this ontology contains less human gene annotations than the Mammalian Phenotype Ontology (genes homologous to rat genes), but has a higher level of specificity in its terminology. There are overlaps between the Mammalian Phenotype Ontology and the Rat Disease Ontology, however, as I explain in subsequent chapters, Phevor's use of these two ontologies improves its ability to discover factual annotations and relationships while weakening inaccuracies.

3.5.5 Human Phenotype Ontology

Inconsistent phenotype description create problems using disease databases like OMIM[32]. Using a controlled vocabulary, the Human Phenotype Ontology[104] tries to correct inconsistencies in terminology and relate phenotypes to one another. When human phenotypes are organized in a directed acyclic graph format, multiple signs and symptoms can be deemed related or unrelated to each other. An example is as follows: The symptoms of *fever, cough* and *runny nose* can be aggregated into a single phenotype of *cold*, whereas *cough* and *abnormally short femur* should be considered separate phenotypes. Gene annotations to Human Phenotype Ontology terms have proven useful to identify candidate genes. There are more than 10,000 descriptive phenotype terms in the ontology, describing all monogenic disorders found in OMIM.

3.5.6 Disease Ontology

Sometimes referenced as the Human Disease Ontology, the Disease Ontology[102] was developed by NUgene project. The goal of the Disease Ontology was to provide a hierarchical controlled vocabulary for human disease. Syndromes and disorders are often complex groupings of several phenotypes (i.e., *hypertension*, *hematuria* and *recurrent urinary tract infections* are phenotypes for *polycystic kidney disease*). Annotating genes to diseases, opposed to phenotype, adds another level of detail—exploited by Phevor when making correlations between phenotypes and genes.

3.5.7 Chemical Entities of Biological Interest

The Chemical Entities of Biological Interest[103] ontology targets nongene chemicals, and their interactions with genes and gene products. It may not be clear how using interactions between gene products and nongene chemicals can assist in identifying damaged genes. However, often, specific interactions reveal relationships and the underlying biological properties overlooked or hidden from model organism based ontologies like the Gene Ontology or Mammalian Phenotype Ontology. The

Chemical Entities of Biological Interest (CHEBI) ontology creates a controlled vocabulary for biologically relevant chemicals, e.g., medications, toxins, environmental agents and non-gene cellular components. On average, each term in the CHEBI ontology has six synonyms, too many iterations of the term to effectively interrogate without the controlled vocabulary provided.

3.6 Using Bio-Ontologies

The wealth of biological properties and their relationships to one another makes ontologies a valuable resource for genomic research. Unfortunately, ontologies are often thought as tools tailored to gene expression analyses and are generally uninformative for inherited disease-gene searches. Before the release of Phevor, the use of bio-ontologies has been confined to overrepresentation analyses and discovering semantic similarities between phenotype and disease. Because bio-ontologies have been considered tools for gene expression analysis, most of the bioinformatics tools do this, and little else.

3.6.1 Overrepresentation Analysis

Discovering differentially expressed genes, and the biological pathways they influence has become the primary use of bio-ontologies to date. Next generation sequencing technology has accelerated these efforts with tools like RNA-Seq. A single RNA-Seq experiment results in hundreds or thousands of differentially expressed genes, which can be categorized into bio-ontology terms. Once categorized, overrepresented bio-ontology terms can be used to find properties common to the differentially expressed genes. Overrepresentation analyses have also been used to look for overrepresentation of deleterious alleles in genes sharing properties in exome sequencing results, but these analyses require large cohorts to produce significant signals, and, therefore, are not useful for clinical diagnostics.

3.6.2 Semantic Similarity Analysis

More in line with medical genetics is the use of semantic similarity to identify candidate disease genes. Converting the observed phenotype into the controlled vocabulary of the Human Phenotype Ontology, similarities can be calculated across characterized genetic diseases in disease allele databases. Calculated similarities can be used to identify genes "likely" to cause the phenotype, because they have already been associated with neighboring phenotypes in a phenotype ontology, i.e., the two genes are "semantically similar" to one another. Outside of Phevor, no tool utilizing bio-ontologies to date has incorporated the genotype results from personal sequencing into a similar search. Thus, calculating semantic similarities between a phenotype and set of known diseases can only return a list of prospective candidate genes, blinded to their genotype. Additionally, because semantic similarities utilize known disease databases to establish the gene to phenotype relationships, they are restricted to only known disorders.

3.6.2.1 Phenomizer

Phenomizer[104] is a web based search tool optimized to perform semantic similarity searches between the Human Phenotype Ontology[97] terms and disease databases (OMIM[32], Orphanet[83], and DECIPHER[108]). Users input the observed phenotype information into Phenomizer interface. A mode of inheritance can be assigned and then the user is returned a list of probable diseases. Because Phenomizer is designed to look for semantic similarities between Human Phenotype Ontology terms and known diseases, and many unknown diseases do not have responsible genes assigned to them, a Phenomizer search is not guaranteed to return candidate genes. When candidate genes are returned, three crucial questions remain: 1) Are there damaging variants in any of these candidate genes? 2) Are the candidate genes responsible for all, some, or just associated with the entered phenotypes? 3) Are additional genes missed by Phenomizer that have yet to be linked to the phenotype in the disease databases? As I explain in Chapter 4, Phevor directly addresses all of these concerns.

3.7 Something Better Is Needed

Exome sequencing is rapidly becoming the norm for molecular diagnostics. It is this imperative to have a fast and accurate method to identify genes responsible for the individual’s phenotype. For the reasons I have enumerated in this chapter, clinical laboratories reliant upon disease databases, regardless of how up to date, will not be able to improve diagnostic rates sufficiently to justify exome sequencing. As I will show in the coming chapters, using the relationships and vast biological information resident to bio-ontologies, genes responsible for the phenotype can quickly be identified, and phenotype associations for genes that have not been established can be done with confidence. My dissertation work, e.g., Phevor, has provided a means to accomplish these goals. In the succeeding chapters, the Phevor algorithm will be described in detail, and along the way, I will explain how it uses bio-ontologies differently than existing tools. The utility of Phevor will be demonstrated using benchmark datasets, and real analyses using individuals from the Utah Genome Project.

CHAPTER 4

PHEVOR: THE PHENOTYPE-DRIVEN VARIANT ONTOLOGICAL RE-RANKING TOOL

Next Generation Sequencing (NGS) has paved the way for personalized health care; unfortunately, with today's methods, most clinical labs are ill equipped to connect damaging variants to the patient's disease. Accurate molecular diagnostics using whole genome and exome sequencing will improve treatment and provide answers for many with inherited disorders. Despite ever-improving variant interpretation tools, and disease-gene finders, connecting the damaged gene responsible with the patient's phenotype remains the biggest challenge, especially when trying to diagnose a single patient. One approach is to rely on disease-allele databases, but as I have explained previously this eliminates any chance for new diagnostic discoveries.

To fully realize the power of Next generation sequencing, medical genetics needs reliable *ab initio* methods to relate genes and alleles to phenotypes. This has become the goal of my dissertation work. Phevor, the Phenotype Driven Variant Ontological Re-ranking tool[35] integrates phenotype, gene properties and disease knowledge with personal genomic sequencing data for accurate diagnosis of pathogenic alleles. Using knowledge resident in bio-ontologies, Phevor makes connections between various biological properties (e.g., pathway, chemical interactions or molecular function) and phenotypes to associate genes and pathogenic alleles with phenotypes. This allows Phevor to make accurate diagnoses without reliance on disease-allele databases. This chapter will describe the Phevor algorithm, detailing each step of the process. The algorithm behind Phevor is the first of its kind—making logical connections between gene functions and phenotypic consequence for accurate molecular diagnosis.

4.1 Why Did I Need Phevor?

As the saying goes, "*necessity is the mother of invention*." Phevor was certainly born out of necessity. For nearly a year, I worked with collaborators at ARUP to identify novel genes responsible for the autosomal dominant disorder *Hereditary Hemorrhagic Telangiectasia* (HHT)[109,110]. Before this project, three known HHT genes had been identified, but alleles in these genes only explain 85% of HHT diagnoses. Although the objective of our study were clear—identify novel genes that result in the HHT phenotype—as it turned out, the phenotype of HHT was not as straightforward as the diagnostic criteria might leave one to believe.

A clinical diagnosis of HHT requires three out of four diagnostic criteria[111]: 1) spontaneous nosebleeds (epistaxis), 2) multiple telangiectasias—small dilated blood vessels near the surface of skin on the lip, face or hand, 3) arteriovenous malformations in the internal organs, or 4) family history for HHT. Routinely, locus specific testing is conducted on the three genes known to cause HHT: *ENG*[112], *ACVRL1*[113] and *SMAD4*[114]. Yet, using Sanger sequencing, nearly 15% of those clinically diagnosed with HHT had no explanatory mutations. From the pool of patients with an HHT diagnosis, but without a causal mutation, probands were selected for follow-up with exome sequencing. Recruiting probands with affected family members, sequencing was performed on multiple affected individuals (2-3), and in several different families. In a misguided attempt to recruit more probands, diagnostic criteria were relaxed. Individual probands were included if they had two of the four diagnostic criteria. Moreover, the two diagnostic criteria did not necessarily represent the same for all those that were included.

Variant interpretation and candidate gene prioritization were performed two ways: 1) variants were queried against disease databases[18,32,33] (similar to today's clinical exome analysis), 2) variants were interpreted and genes prioritized using the comprehensive disease-gene finders Annovar[41] and VAAST[43]. Disease databases immediately identified variants in several probands known to cause disease. However, none associated with HHT. As it turned out, several HHT probands had causative mutations for entirely different, yet phenotypically similar disorders. This became our first indication that this search would not be as straightforward as we initially hoped. VAAST and Annovar returned many "interesting" candidate genes, but I was unable to establish a clear link between any of these candidate genes and the HHT phenotype.

These results illustrate two major complications with exome analyses and demonstrated the need for a better methodology: 1) The clinical diagnosis is often less specific than might be assumed, and 2) making the connection between damaged genes and phenotype is very difficult. Because probands included in the study only needed two of the four diagnostic criteria, the affected population did not represent a single disease, or a single set of phenotypes. True they were all were diagnosed with HHT, but their disease symptoms varied from occasional nosebleeds to life threatening arteriovenous malformations. Often the diagnosing clinician got the diagnosis wrong, e.g., the individuals had variants that could explain phenotype, but not an HHT diagnosis. This illustrates two basic truths of diagnosis: Diseases are collections of phenotypes and different diseases are phenotypic pleomorphs. Grouping individuals having what “looked like” HHT as a single population ultimately diluted our chances of finding the responsible genes. From the failed HHT search, Phevor was born.

4.2 Introducing Phevor

The Phenotype-Driven Variant Ontological Re-ranking tool, affectionately called Phevor[35], integrates phenotype data with genotype data to find pathogenic alleles. Phevor’s novel algorithm combines human phenotype knowledge with information about gene function, human disease, chemical interactions, and phenotype data from other mammals[18,32,33,83]. As illustrated in Figure 4.1, inputs to Phevor are a patient’s phenotype and their variant interpretation/prioritization results. Clinical observations are converted into phenotypes using controlled vocabulary terms provided by the Human Phenotype Ontology. Briefly, the terms are used to “seed” multiple bio-ontologies. Phevor then propagates these seeds to expand the candidate gene list. As set out in the details below, Phevor can make novel connections between gene and phenotype. Breaking away from the limitations inherent to candidate-gene or disease lists, Phevor promises to expand molecular diagnostics to the entire exome.

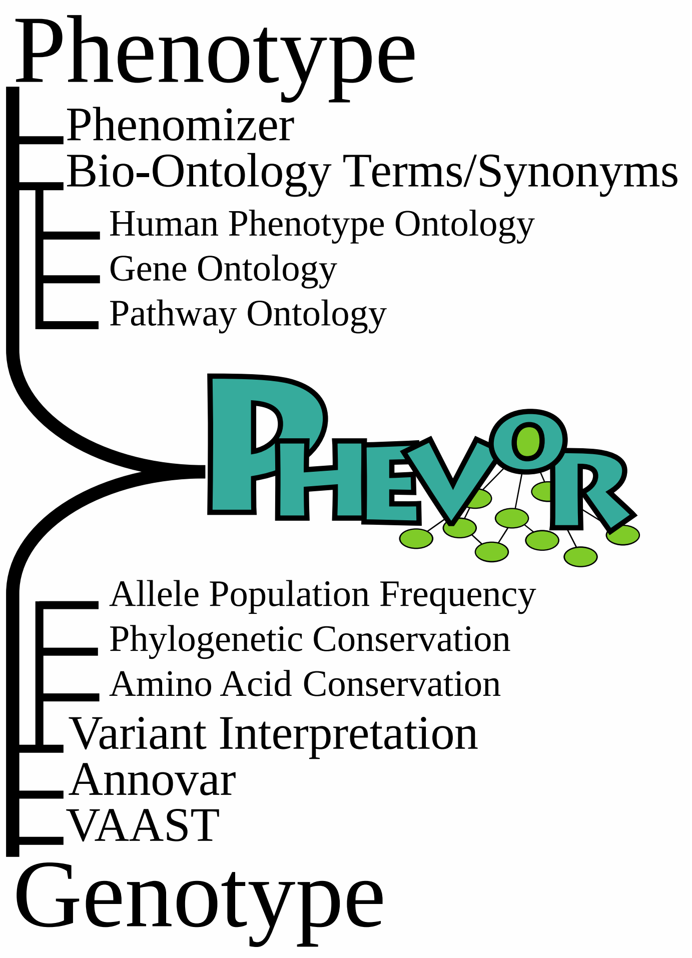

Phenotype
Phenomizer
Bio-Ontology Terms/Synonyms
Human Phenotype Ontology
Gene Ontology
Pathway Ontology
PHEVOR
Allele Population Frequency
Phylogenetic Conservation
Amino Acid Conservation
Variant Interpretation
Annovar
VAAST
Genotype

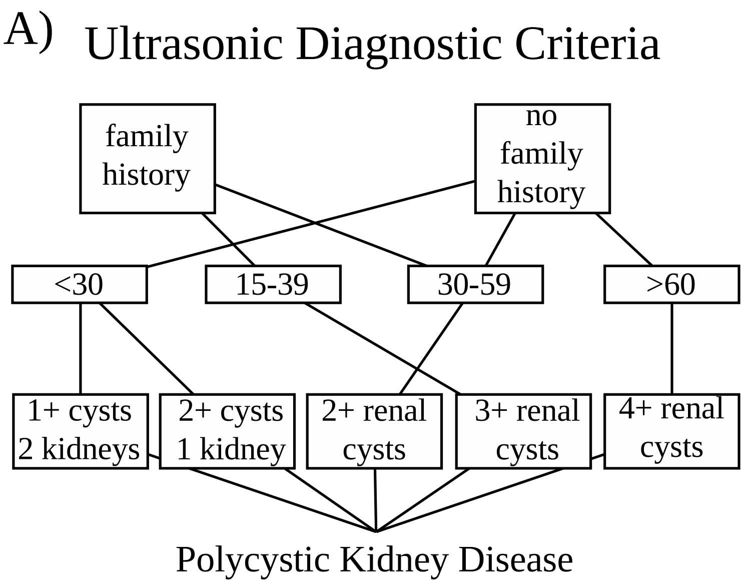
Figure 4.1 General Phevor Methodology
Phevor combines phenotype, and genotype to identify pathogenic alleles in personal genomic sequencing. The phenotype is used to create a list of phenotype-linked genes that are propagated across the bio-ontologies, expanding the candidate gene list. Phenotype-linked gene lists can be generated externally (Phenomizer[104]) or pulled directly from bio-ontologies by Phevor. Variants from personal genomic sequencing are interpreted and prioritized by one of many variant interpretation or gene prioritizing tools. Phevor combines the expanded candidate genes with their prioritized genotype. Phevor can accurately identify the gene responsible for the described phenotype(s).

4.2.1 Just What Is a Phenotype?

Describing phenotypes in a machine-readable fashion is a major research problem in and of itself. Although geneticists and physicians generally have a strong idea of what the word phenotype means, and what a patient phenotype is, translating these opinions into machine-readable form is a process fraught with difficulty. Describing the patient's phenotype accurately is essential for an accurate diagnosis, clinically and molecularly. Often it is difficult to say which clinical traits are connected and which are confounding.

Clinicians use diagnostic criteria, analogous to those described for HHT to aid in making a diagnosis, but often these criteria do not fully explain the phenotype. Take for example the diagnostic criteria and phenotype for *Polycystic Kidney Disease* (PKD)[115]. Figure 4.2 shows the ultrasonic diagnostic criteria needed to make a PKD diagnosis. Depending on family history and age, a specific number of renal cysts need to be observed to make a diagnosis (Figure 4.2A). However, PKD has other phenotypes that cannot be directly evaluated using these criteria. Part B of Figure 4.2 details the phenotype as reported by OMIM. Using only the ultrasonic diagnostic criteria, an accurate diagnosis could be missed. Additionally, heart, liver and colon phenotypes could confound a diagnosis of PKD; likewise, the insufficient number of renal cysts could lead to misdiagnosis.

Named diseases are collections of specific phenotypes that aggregate to a generalized medical term summarizing the patient's problem. Because many diseases have overlapping phenotypes, finding a molecular diagnosis from a named disease can be difficult. For example, a patient with *abnormal muscle coordination*, *paralysis of the*


A)
Ultrasonic Diagnostic Criteria
family history
no family history
<30
15-39
30-59
>60
1+ cysts 2 kidneys
2+ cysts 1 kidney
2+ renal cysts
3+ renal cysts
4+ renal cysts
Polycystic Kidney Disease

B) OMIM Phenotype Description

System	**Organ**	**Features**
Cardiovascular	Heart	Increased prevalence of valvular disease
	Vascular	Intracranial aneurysm
Abdomen	Liver	Hepatic cysts
	Gastrointestinal	Colon diverticula
Genitourinary	Kidneys	Polycystic kidney
		Renal failure

Figure 4.2. Polycystic Kidney Disease Diagnostic Criteria and Phenotype
A) Ultrasonic diagnostic criteria for polycystic kidney disease (PKD) depend on family history, age, and the number of renal cysts observed. B) Phenotype description of PKD—as found in OMIM—indicates several phenotypes not accounted for by the diagnostic criteria. Diagnosing a patient with PKD using the diagnostic criteria alone fails to consider any of the confounding phenotypes, potentially resulting in misdiagnosis.

eye muscles, and the *absence of the tendon reflexes* could be diagnosed with Guillain-Barre syndrome[116] or the phenotypically similar Miller Fisher syndrome[117]. Both diseases have similar phenotypes, but different etiologies. Using either of these named diseases as the phenotype for a patient risk restricting the chances of making a molecular diagnosis.

Using the phenotype of the patient, as opposed to a named clinical disease, helps to reduce the confusion. Moreover, phenotype terms can be combined to terminology that is more inclusive if they share properties. The phenotype of a patient presenting with *uterine leiomyoma* (a benign neoplasm in the uterine wall) and *hypoplasia of the uterus* (abnormally small uterus), for example, can be aggregated to *abnormality of the uterus*, a phenotypic term that includes both characteristics. As one would expect, the specificity of a phenotype description will have an effect on Phevor's ability to identify the gene responsible. In the next chapter, this effect will be further documented, however, providing Phevor generalized phenotypic descriptions have minimal impact on performance.

Phenotype descriptions are not limited solely to medical terminology or morphological descriptions. Phevor utilizes the information resident to multiple bio-ontologies, and, therefore, can use terminology from each as a phenotypic description. For example, an *allergy to penicillin* can be invoked as a phenotype. Using the Chemical Entity of Biological Interest[103] ontology penicillin is known to interact with several gene products, thus aiding in the identification of genes responsible for an *allergy to penicillin*. Likewise, the Gene Ontology[94] term *baroreceptor detection of decreased arterial stretch* can be used to describe abnormal blood pressure measurements. Using any number of combinations of terms, phenotype, gene function or chemical interactions, Phevor can return the appropriate candidate genes in a probabilistic fashion, even if they have not been previously explicitly associated with a particular disease or phenotype.

4.3 The Phevor Algorithm

Phevor works by combining the prioritized results of widely-used variant interpretation tools with knowledge resident in diverse biomedical ontologies, such as the Human Phenotype[118], the Mammalian Phenotype[119], the Disease[120], the Gene [121], the Rat Disease[101], and the Chemical Entity of Biological Interest[103] ontologies. Described in more detail in Chapter 3, ontologies are graphical representations of the knowledge in a given domain, such as gene function or human phenotype. They organize this knowledge using directed acyclic graphs wherein concepts/terms are nodes in the graph connected by the logical relationships between them.

Phevor propagates an individual's phenotype information across and between biomedical ontologies. This process enables Phevor to accurately reprioritize candidates identified by variant interpretation tools in light of knowledge contained in the ontologies. This process permits Phevor to discover emergent gene properties and latent phenotypic information by combining ontologies, further improving its accuracy.

Phevor does not replace existing interpretation tools; rather, it provides the general means to improve every tool's performance. Phevor also differs from tools such as Phenomizer[122] in that it does not postulate a set of fixed associations between genes, phenotypes or diseases. Rather, Phevor dynamically integrates knowledge resident in the biomedical ontologies into the variant interpretation process. Phevor provides means to integrate ontologies into the disease-gene search process, such as the Gene Ontology, which contains knowledge not explicitly linked to the phenotype. This enables Phevor to improve diagnostic accuracy not only for established disease phenotypes, but also for previously undescribed and atypical disease presentations.

4.3.1 Connecting Biomedical Ontologies and Seeding Phenotype-Linked Genes

Biomedical ontology annotations are now readily available for many human and model organism genes. Phevor employs these annotations to associate phenotype descriptions (phenotype translated to ontology nodes) to genes, and vice versa. Consider the following example of a phenotype description consisting of two Human Phenotype Ontology terms: *Hypothyroidism* and *Abnormality of the intestine*. If genes were previously annotated to these two nodes in the ontology, Phevor saves those genes in an internal list. In cases where no genes are annotated to a user-provided ontology term, Phevor traverses that ontology beginning at the provided term and proceeds toward the ontology's common parent until it encounters a node with annotated genes, adding those genes to the list. At the end of this process, the resulting gene list is then used to seed nodes in the other ontologies, the Gene Ontology, the Mammalian Phenotype Ontology and the Disease Ontology, for example.

Phevor can connect superficially unrelated bio-ontologies by shared gene annotations. For example, *ACTN2* is linked to *lipoatrophy* (localized loss of fat) in the Human Phenotype Ontology, and *platelet degranulation* in the Gene Ontology.

Connecting bio-ontologies allows Phevor to explicitly relate the phenotype and molecular function. Phevor relates different ontologies via their common gene annotations, and uses these common gene annotations to selectively seed phenotype-linked genes.

Selectively seeding only nodes with these genes as attributes make immediate connections between the phenotype and the biological properties inherent each ontology. Using phenotype-linked genes for the phenotype *atrial fibrillation* to seed the Gene Ontology's[94] biological process vein, enriches the nodes *regulation of heart rate by cardiac conduction* and *cardiac muscle contraction*. Seeded into the Pathway Ontology[99] *hypertrophic cardiomyopathy pathway* and *dilated cardiomyopathy pathway* nodes are enriched. This may seem like obvious connections between the phenotype *atrial fibrillation* and the enriched ontology terms, but note: at each of these nodes, there are additional genes, likely to give rise to the *atrial fibrillation* phenotype never before attributed to the human phenotype.

To sustain a high level of specificity, Phevor only seeds "base nodes." Base nodes are nodes annotated with one or more of the candidate genes—distal to the common parent, where no child node is attributed with the equivalent gene annotation. To illustrate, *ENPP1* is annotated to various nodes in the Gene Ontology including *protein binding*, *protein dimerization activity*, *identical protein binding*, and *protein homodimerization activity* (Figure 4.3). The child node to *protein homodimerization* is *actin homodimerization activity* but does not possess the gene annotation of *ENPP1* as one of its attributes. Therefore, *protein homodimerization activity* is the base node for *ENPP1*.

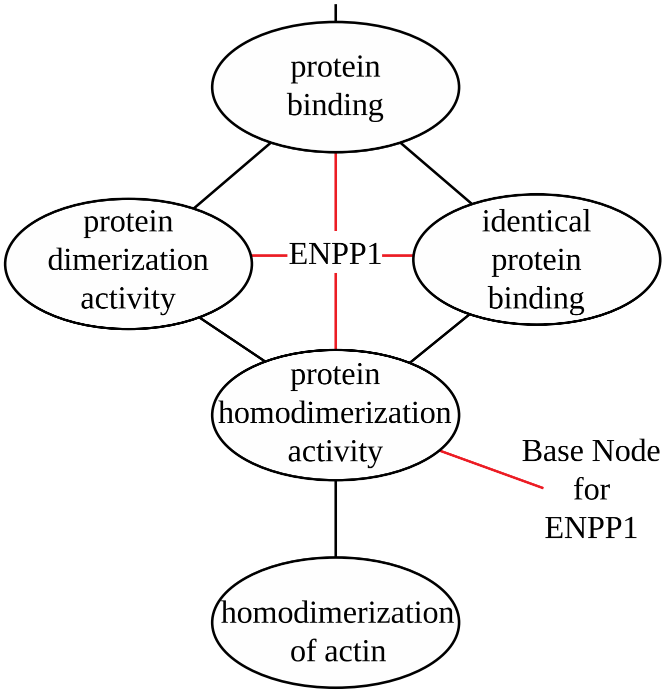

protein binding
protein dimerization activity
ENPP1
identical protein binding
protein homodimerization activity
Base Node for ENPP1
homodimerization of actin

Figure 4.3. Base Node Seeding
Attributes of ontology nodes are transitive with parental nodes. The gene annotation *ENPP1* is an attribute for *protein homodimerization activity*, and therefore is an attribute for all parental nodes. *ENPP1* annotation is not transferred to the child node *homodimerization of actin*, leaving protein *homodimerization activity* the base node for nodes having the attribute of *ENPP1* gene annotation.

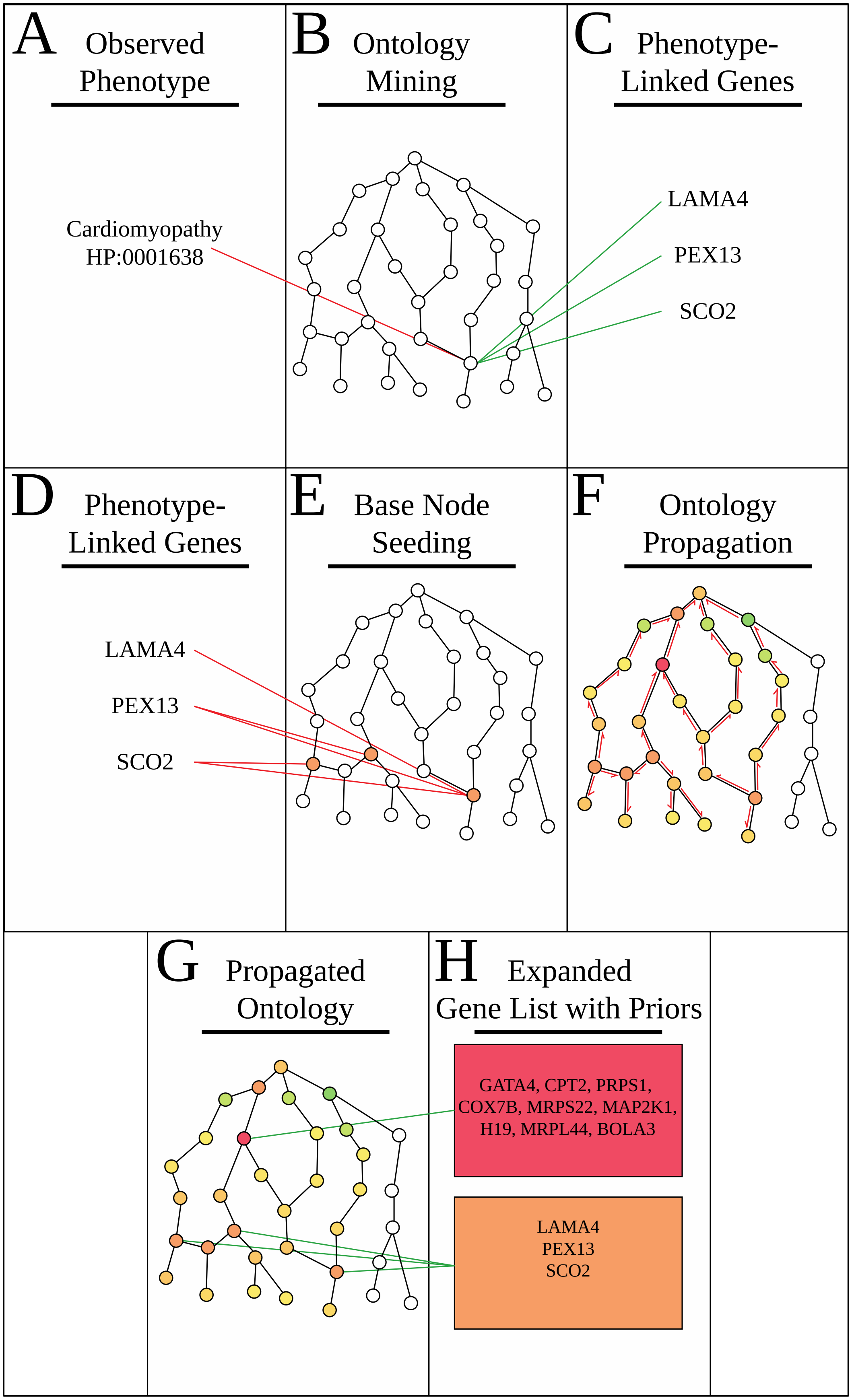
Seeded base nodes detail various biological properties contributing to the phenotype. Higher scores indicate a tighter link between the described biological property and the phenotype. Seeding only base nodes allows Phevor’s entry into each ontology to be as deep (specific) as possible before the ontology propagation. Illustrated in Figure 4.4D, E, the phenotype-linked genes are seeded to this ontology at only their base nodes. The ontology propagation expands candidate genes to those with similar biological properties and makes it possible to establish novel phenotype relationships.

4.3.2 Ontology Propagation

Phevor extends phenotypes to genes previously not associated by propagating the base node scores across the ontology. Phevor’s ontology propagation sets it apart from other tools attempting to find candidate genes using phenotype. Phevor is not restricted to what is known in disease or phenotype databases. Most protein coding genes or their homologues in model organisms have at least minimal knowledge about their structure, location and function. Phevor uses these known biological properties to connect their phenotype implications. For instance, 20 genes are phenotypically linked to *pulmonary fibrosis*. Looking at the molecular function of these genes, 18 of the 20 are annotated to the Gene Ontology[94] node: *ion channel activity*. Of course, this does not mean all genes annotated to *ion channel activity* cause *pulmonary fibrosis*, only that there is a prior expectation that it is reasonable to expect novel *pulmonary fibrosis* causing genes are likely to have *ion channel activity*.

Once a set of starting base nodes is established for each bio-ontology, i.e., those provided by the user in their phenotype list, or derived from it by the cross-ontology

Figure 4.4. Detailed Phevor Methodology
A) Clinical descriptions of the phenotype are translated into bio-ontology nodes using the controlled vocabulary of the ontologies. B) From bio-ontology nodes, gene annotations are extracted from the ontology. C) Extracted gene annotations are phenotype-linked genes. D) Phenotype-linked genes are used to link multiple bio-ontologies at nodes sharing gene annotations. E) This ontology is seeded using the phenotype-linked genes at base nodes for each gene. F) Seeded base nodes values are propagated across the ontology i.e., towards child and parental nodes, where they are penalized the further they get from the base node. Intersecting propagation paths are combined, intensifying similarities between seeded nodes. G) Intersecting propagation paths identify latent similarities between phenotype-linked genes and expand the potential candidates—thus relating biological properties to the phenotype.


A Observed Phenotype
Cardiomyopathy
HP:0001638
B Ontology Mining
C Phenotype-Linked Genes
LAMA4
PEX13
SCO2
D Phenotype-Linked Genes
LAMA4
PEX13
SCO2
E Base Node Seeding
F Ontology Propagation
G Propagated Ontology
H Expanded Gene List with Priors
GATA4, CPT2, PRPS1, COX7B, MRPS22, MAP2K1, H19, MRPL44, BOLA3
LAMA4
PEX13
SCO2

linking procedure described in the preceding section, Phevor next propagates this information across each ontology by means of a process termed *ontological propagation.* Consider the example illustrated in Figure 4.4. Here, three base nodes in some ontology have been identified, and in two of the nodes multiple genes have been previously annotated. Each base node is seeded and this information is then propagated across the ontology as follows. Proceeding from each base node toward its parents and children, each time an edge is crossed to a neighboring node, the current value of the previous node is divided by two. This process continues until a terminal leaf or common parent node is encountered (Figure 4.4F). In practice, there can be many seed nodes. In such cases, intersecting threads of propagation are first combined, and the process of propagation proceeds as described. One interesting consequence of this process is that nodes far from the original seeds can attain high values, greater even than any of the starting seed nodes. The darker red node in Figure 4.4G and H, in which propagation has identified additional gene-candidates, not associated with the original base nodes, illustrates the phenomenon.

Propagating information across the ontologies can be likened to dropping rocks into a puddle. Where the rock falls in a puddle, waves of water propagate outward, decreasing in intensity as they spread. When multiple rocks are dropped into a puddle, where multiple waves intersect the wave intensity is appropriately increased. Likewise, as the score is propagated outwards, away from the base nodes the scores are penalized (Figure 4.4F). However, when propagation chains originating from different base nodes intersect, the score is intensified. This amplifies commonalities between nodes seeded with the phenotype-linked genes. Unlike the puddle analogy, ontology propagation stops when a terminal node is reached, i.e., common root of all nodes, or nodes without a child node (Figure 4.4F). Propagating away from the base nodes does not guarantee that every node of the ontology receives a score. This is done by design (Figure 4.4G). If all phenotype-linked genes are known to methylate DNA, it is pointless to move back downward in the ontology to interrogate genes involved with lipid metabolism.

4.3.3 Transferring Node Scores to Genes

Every gene annotated within the ontology will have at least one node receiving a value from the propagation. At the very least, the root node of the entire ontology will get a score. The root node contains all the attributes of the entire ontology, thus all contained gene annotations. Depending on the path the propagation took, many genes will be annotated to multiple scored nodes. Using the best scoring node as the representative for the gene, the phenotype linkage score is transferred from the node to the gene. Propagating phenotype-linked genes for *intrahepatic cholestasis* through the Human Phenotype Ontology[97] scores *ABCB11* annotated nodes: *intrahepatic cholestasis* with 2.90, *conjugated hyperbilirubinemia* with 2.95 and *cholestasis* of 3.03. Therefore, the ontology node *cholestasis* will be transferred to *ABCB11* as the best node representation from this ontology. Gene specific propagation scores are computed for each ontology used by Phevor. These scores describe how connected the biological properties are to the original phenotype. Genes absent from the ontology are assigned default phenotype linkage scores—as indicated in Figure 4.4.

4.3.4. Combining Propagations from Multiple Bio-Ontologies

Other tools using bio-ontologies consider them separate entities and use them independent to each other. Phevor uses the differences and similarities between bio-ontologies to its advantage. Differences in design and scope allow Phevor to encourage accurate annotations and relationships, whilst limiting the impact of errors in ontology design and annotations. The final propagation score for each gene is the sum of all its ontology specific propagation scores normalized to the sum of all propagation scores.

Finally, to account for genes without an annotation in one or multiple ontologies, default propagation values are assigned to unannotated genes. Default values are calculated for each bio-ontology by weighting the median propagation score against the probability any given gene is annotated to that ontology. For example, the Gene Ontology[94] accounts for ~18,000 human genes, therefore, its default is well below the median. The Human Phenotype Ontology[97], on the other hand, only has ~2,800 gene annotations, resulting in a default above the median. Assigning a default value for unannotated genes allows Phevor to make novel phenotype associations even for genes with limited bio-ontology annotations.

4.3.5 Scoring and Ranking Expanded Candidate Genes

Post-ontology propagation, the gene's final propagation score is combined with its genotype, as scored using any of several variant interpretation tools, e.g., VAAST[42,43], Annovar[41], SIFT[39], PolyPhen[40], etc. (Chapter 2). Phevor first calculates a disease association score for each gene,

$$D_g = \left(1 - V_g\right) \times N_g \tag{4.1}$$

where Ng is the percentile rank of the renormalized gene sum score derived from the ontological propagations procedures described in Figure 4.4, and V_g is the percentile rank of the gene provided by the external variant interpretation or search tool, e.g., Annovar, SIFT and phyloP. Phevor then calculates a second score summarizing the weight of evidence that the gene is not involved with an individual's illness, H_g, i.e., the variant(s) nor the gene is involved in the individual's disease.

$$H_g = V_g \times (1 - N_g) \qquad (4.2)$$

The Phevor score is the log10 ratio of disease association score (D_g), and the healthy association score (H_g),

$$S_g = \log_{10} D_g / H_g \qquad (4.3)$$

These scores are distributed normally (Chapter 5). The performance benchmarks presented in Chapter 5 provide an objective basis for evaluating the utility of S_g.

4.4 Combining Ontologies and Variant Data

Upon completion of all ontology propagations, and their subsequent combination and gene scoring steps described in the preceding paragraphs, genes are ranked using their Phevor scores. Each gene's percentile rank from the ontology propagation is combined with variant or gene prioritization score using a naïve Bayes approach, whereby the ontology derived gene percentile ranks are used as priors. Just as Phevor can employ multiple ontologies, it can also employ multiple variant prioritization and search tools, including, SIFT[39], phyloP[62], Annovar[41]and VAAST[42,43]. Variant prioritization "scores" are used as the forward probabilities. The percentile score ranks are used for VAAST, Annovar, SIFT and phyloP. Phevor uses these data to calculate the posterior probabilities of two models for each gene, P^D, that the gene is responsible for the disease, and P^H, the null model, i.e., neither the variants nor the gene are involved in the disease. The Phevor score reports the $\log_{10}$ Bayes Ratio[123] of these two models. The performance benchmarks presented in the following chapter provides an objective basis for evaluating the utility this approach.

In summary, free from the limitations imposed by restricting the search space to the list of candidate genes drawn from disease allele databases, Phevor is able to make logical connections between a damaged gene's biological properties and the resulting phenotype, even when no previous association has been made. The following chapters benchmark Phevor's abilities using simulated data, and real clinical examples from the Utah Genomes Project.

CHAPTER 5

BENCHMARKING PHEVOR’S PERFORMANCE

Phevor’s ability to connect damaged genes with phenotypes will dramatically improve diagnostic exome sequencing. As I demonstrate below, Phevor is unique in its ability to make novel phenotype associations. The benchmarks presented in this chapter push Phevor’s abilities, expose its limitations and demonstrate that Phevor works as described. Simulated data were required to test Phevor’s functionality fully. In this chapter, I will describe how the benchmark datasets were created and how they were used. The preceding chapter will present several actual patient diagnoses made through the Utah Genome Project, which employs Phevor as part of its analysis pipeline.

5.1 Creating Benchmarking Datasets

Exome sequencing was performed on 100 unrelated individuals at the Clinical Laboratory Improvement Amendments (CLIA) and College of American Pathologist (CAP) certified laboratory at ARUP. Using Agilent’s “SureSelect (XT) Human All Exon v5 plus UTRs capture”[20], the exome was enriched and sequenced on an Illumina HiSeq instrument programmed to perform 100 cycles paired-end sequencing. Sequence reads were aligned and variants called following the best practice guidelines provided by the BROAD Institute[16].

Known pathogenic alleles form the Human Genetic Mutation Database (HGMD)[18] were then inserted into these 100 exomes. HGMD variants were then recovered using variant interpretation or disease-gene finding tools with and without the help of Phevor. Because a single affected exome is the most commonly observed scenario in the clinical genomics laboratory, and a single affected exome is the most difficult diagnostic scenario for today's variant interpretation methods, single affected exomes were used for most of the benchmarks described below.

As discussed in Chapter 2, many of today's variant interpretation tools[39,40,61,62] are limited by what kind of variant they can score; thus to provide every tool with a level playing field, only coding single nucleotide variants were spiked into the benchmarking exomes. No additional restrictions were placed on the genomic background in which the known alleles were inserted. All discovered variant types—single nucleotide variants, insertions, deletions, no-calls—were retained to simulate the appropriate genetic background encountered during exome analysis. Benchmarking was performed using one, or more randomly selected exomes combined with one or two randomly selected known pathogenic alleles from HGMD—mimicking dominant and recessive disease, respectively. Each exome contained pathogenic allele(s) in a single gene. In total 6,000 different genes—each with a specifically known phenotype—were used in these benchmarks.

Phenotypes for each of the 6,000 known disease alleles were extracted from OMIM[32] entries. For example, alleles in *PIGT*[124,125] annotated to the disease p*aroxysmal nocturnal hemoglobinuria*[126]m characterized by the phenotypes of *dyspnea, abdominal pain, diarrhea, arthralgia, urticarial, headaches* and *fatigues*. Note the complexity of this phenotype. Computing on complex phenotypes such as this is one of the central challenges encountered in trying to build a tool such as Phevor. As I show below, it is a challenge Phevor has met.

5.2 Single Exome Benchmarking Results

How often the pathogenic gene was recovered as a top candidate was used to measure the success rates of each tool—with and without the aid of Phevor. Each benchmarking exome was analyzed and candidate genes prioritized using each of the variant interpretation tools are described below. The outputs of each tool were then given to Phevor along with the OMIM[32] derived phenotype. Success was measured according to how often the pathogenic alleles were returned as the top candidate, within the top ten candidates, within the top 100 candidates, or able to be scored at all.

5.2.1 Single Exome Benchmarking Results – Minor Allele Frequency

Because of selective pressure tolerated variants are expected to be found in the population at greater frequency than those that damage gene function, thus minor allele frequency (MAF) can be used to prioritize candidate alleles. Incidentally, this assumption was one of the main motivations during the design of VAAST[42,43].

The frequency of each variant in the target exome was determined using two population databases: the 1000 Genomes Project[72] and the NHLBI GO Exome project[69] (see Chapter 2 for additional details). Prioritizing variants in the benchmarking exome only by their MAF as reported in either databases is not effective. Between 21% and 58% of a given exome’s variants are novel—i.e., not found in 1000 Genomes or NHLBI

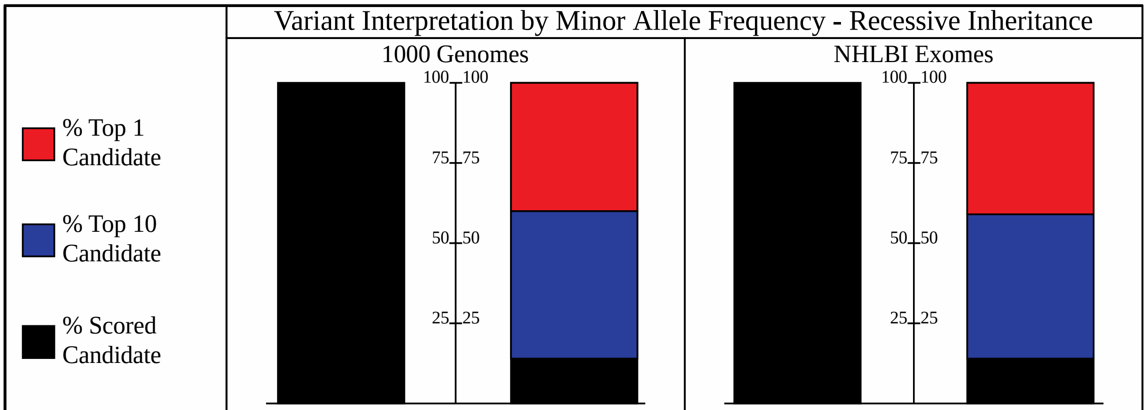
GO Exomes, respectively. Thus, genes containing novel variants will be tied as top ranked candidates. Hence, the poor performance of this simple approach as illustrated in Figures 5.1 and 5.2. In contrast, under the recessive inheritance model—i.e., candidate genes are required to have two variants—Phevor recovered most of the known alleles (86%) in the top 10 candidates when alleles were prioritized by MAF alone. These results demonstrate that using the phenotype as a prior, Phevor is able to dramatically improve the average ranking of the known alleles. The same is true for dominantly inherited disorders.

5.2.2 Single Exome Benchmarking Results – Phylogenetic Conservation

Like MAF, phylogenetic conservation provides another simple means to identify genes harboring damaging alleles. As I show in Figures 5.3 and 5.4, phylogenetic conservation, like MAF, provides little power for identifying damaged genes with pathogenic alleles. Figure 5.3 and Figure 5.4 show the results of using two phylogenetic variant interpretation tools, GERP+[61] and phyloP[62], on recessive and dominantly inherited disorders. Even when the phenotype is used as a prior, Phevor is only able to rank 2% of the known alleles in the top 10. This is true regardless of inheritance pattern or interpretation algorithm used. These results however, suggest Phevor will be able to aid in the search for noncoding pathogenic variants.

5.2.3 Single Exome Benchmarking Results – Amino Acid Conservation

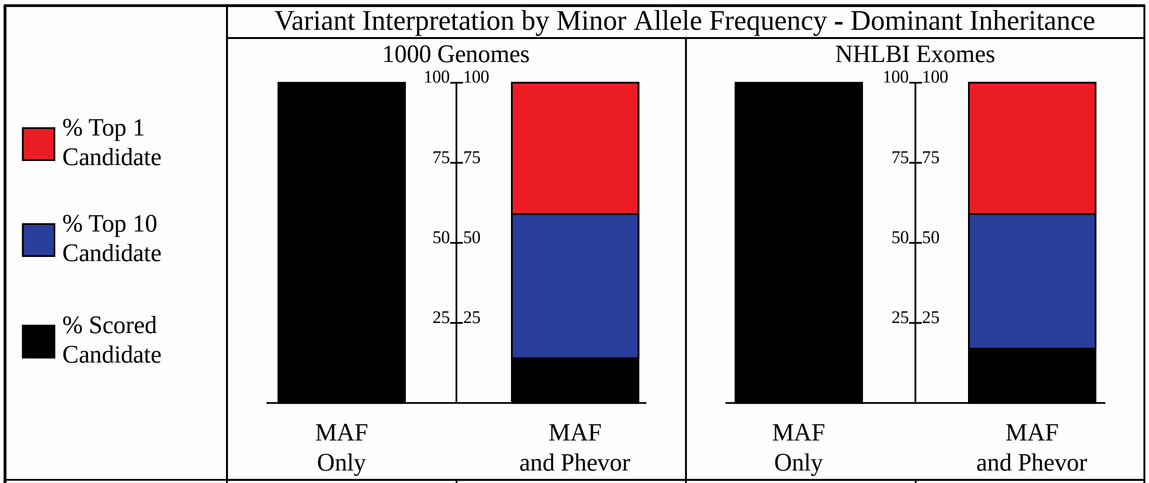
In principle, amino acid conservation provides powerful means to prioritize variants. The drawback is that unlike conservation and MAF, this approach is only



	1000 Genomes		NHLBI Exomes	
	MAF Only	MAF and Phevor	MAF Only	MAF and Phevor
Top 1 Candidate	0%	40%	0%	41%
Top 10 Candidates	0%	86%	0%	86%
Top 100 Candidates	0%	97%	0%	96%
Scored Candidates	100%	100%	100%	100%
Average Rank	10,637	206	11,467	177

Figure 5.1. Recovery of Known Pathogenic Alleles Using Minor Allele Frequency (MAF) to Prioritize Variants under a Recessive Inheritance Model

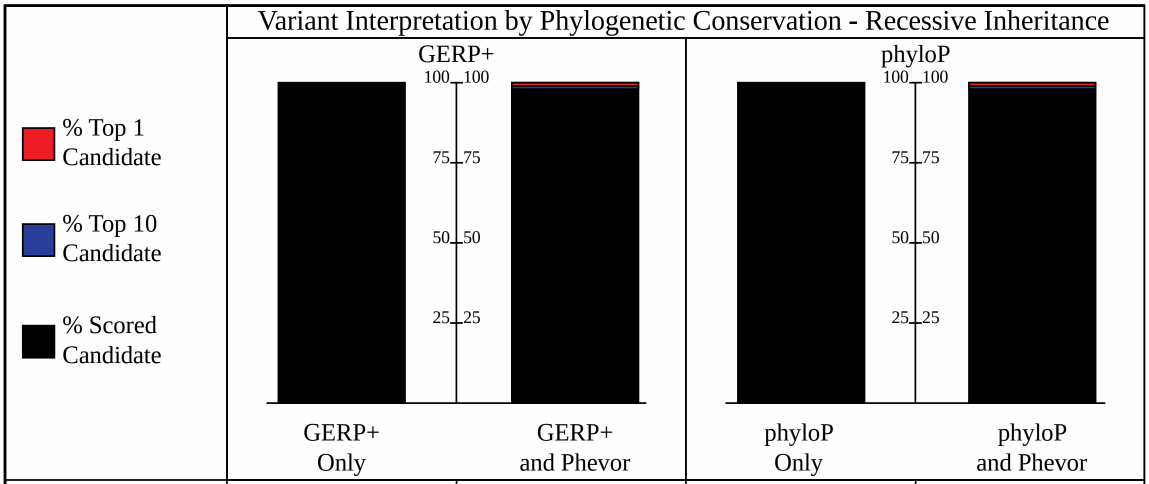
Benchmarking exome variants were prioritized using the MAF reported in the 1000 Genomes or NHLBI Exomes databases. For recessive inheritance, each ranked candidate gene was required to have two variants, i.e., single homozygote or two heterozygotes. Prioritizing candidates by minor allele frequency alone fails to rank any of the genes containing the pathogenic alleles among the top 100 candidates, however, with the aid of Phevor 86% of known alleles are recovered in the top 10 candidates.



	1000 Genomes		NHLBI Exomes	
	MAF Only	MAF and Phevor	MAF Only	MAF and Phevor
Top 1 Candidate	0%	41%	0%	41%
Top 10 Candidates	0%	86%	0%	83%
Top 100 Candidates	0%	96%	0%	95%
Scored Candidates	100%	100%	100%	100%
Average Rank	10,738	242	11,614	251

Figure 5.2. Recovery of Known Pathogenic Alleles Using Minor Allele Frequency to Prioritize Variants under a Dominant Inheritance Model

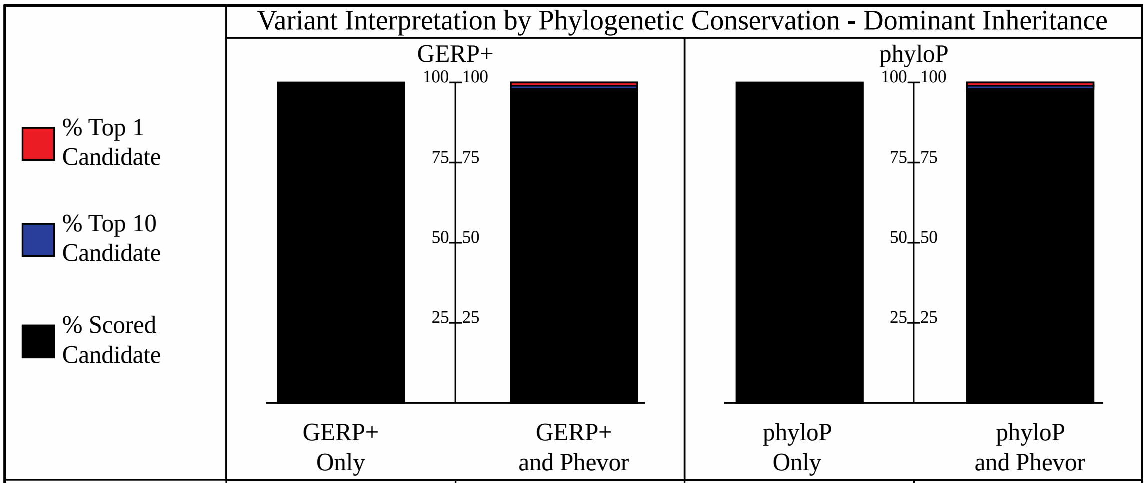
Variants from benchmarking exomes were prioritized according to their MAF reported in the 1000 Genomes or NHLBI Exomes databases. For dominant inheritance, each ranked candidate gene required a single variant, i.e., single heterozygote. Minor allele prioritization of variants fails to rank any known alleles in the top 100. Aided by the phenotype, Phevor can rank 86 and 83% of known alleles in the top 10 candidates when 1000 Genomes and NHLBI Exomes are used to prioritize, respectively.



	GERP+ Only	GERP+ and Phevor	phyloP Only	phyloP and Phevor
Top 1 Candidate	0%	1%	0%	1%
Top 10 Candidates	0%	2%	0%	2%
Top 100 Candidates	0%	6%	0%	20%
Scored Candidates	100%	100%	100%	100%
Average Rank	7,758	418	7,262	291

Figure 5.3. Recovery of Known Pathogenic Alleles Using Phylogenetic Conservation to Prioritize Variants under a Recessive Inheritance Model

Variants from benchmarking exomes were prioritized using the phylogenetic conservation tools GERP+ and phyloP. For the recessive inheritance model, both tools fail to rank any of the target genes in the top 100 candidates. Aided by Phevor, the average rank of known alleles dramatically improves; yet, only 1% of the time is the target gene is ranked as the top candidate. These results demonstrate Phevor’s ability to assist in identifying noncoding functional variants.



	GERP+ Only	GERP+ and Phevor	phyloP Only	phyloP and Phevor
Top 1 Candidate	0%	1%	0%	1%
Top 10 Candidates	0%	2%	0%	2%
Top 100 Candidates	0%	4%	0%	14%
Scored Candidates	100%	100%	100%	100%
Average Rank	8,473	432	8,748	326

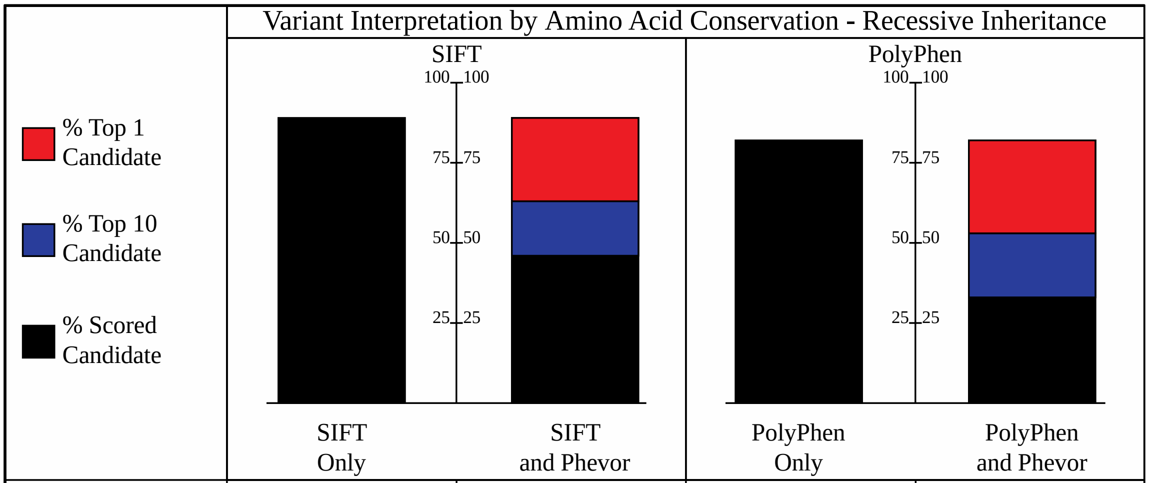
Figure 5.4. Recovery of Known Pathogenic Alleles Using Phylogenetic Conservation to Prioritize Variants under a Dominant Inheritance Model
Benchmark exome variants were interpreted and prioritized using the phylogenetic conservation tools GERP+ and phyloP under a dominant inheritance model, each tool fails to rank any known alleles in the top 100. Adding phenotype as a prior, Phevor improves the average rank of known alleles but is still underpowered to identify the responsible allele in the top 10 candidates.

applicable to coding variants. Tools such as SIFT[39] and PolyPhen[40], for example, are severely limited because they score only amino acid altering variants. Yet, despite these limitations, both these tools are routinely used to interrogate exome variants. To give SIFT and PolyPhen the best chance of recovering the known pathogenic allele, only coding variants were used to produce Figures 5.5 and 5.6. Under the recessive model, SIFT is still limited, only scoring 89% of genes with a known allele—this value is not 100% because coding variants falling in unconserved positions in human proteins cannot be scored by SIFT. Likewise, PolyPhen is able to score only 83%. High false negative rates force the average rank of known alleles out of an acceptable range. Reprioritization by Phevor recovers most of the "scored" known alleles in the top 10 candidates for both inheritance models. However, Phevor cannot overcome these tools inherent false negative rates. This is why the bar plots in Figures 5.5 and 5.6 do not reach 100%.

5.2.4 Single Exome Benchmarking Results – Comprehensive Disease-Gene Finders

Given the results presented in Figures 5.1-5.6, it is clear that none of these variant interpretation techniques is sufficiently powerful for molecular diagnosis. Combining methods is one way improve exome analyses. Currently there are two main tools for doing so: Annovar and VAAST.

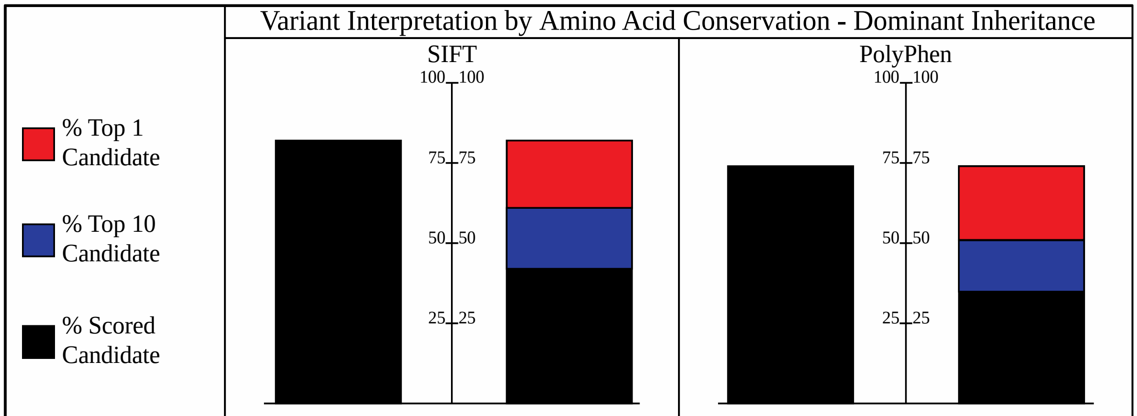
Annovar[41] is a filtering tool while VAAST[42,43] is a probabilistic disease-gene search tool. As I show below, both tools dramatically improve upon the variant interpretation methods described above. However, neither of these tools have sufficient



	SIFT Only	SIFT and Phevor	PolyPhen Only	PolyPhen and Phevor
Top 1 Candidate	0%	26%	0%	29%
Top 10 Candidates	0%	43%	0%	49%
Top 100 Candidates	0%	46%	0%	53%
Scored Candidates	89%	89%	83%	82%
Average Rank	9,974	4,539	10,038	5,441

Figure 5.5. Recovery of Known Pathogenic Alleles Using Amino Acid Conservation to Prioritize Variants under a Recessive Inheritance Model

Benchmarking exome variants were interpreted and prioritized using the amino acid conservation tools SIFT and PolyPhen. Genes were prioritized using the percent rank of the two best scoring alleles as interpreted by SIFT and PolyPhen. Under recessive inheritance, both tools fail to rank any genes with known alleles in the top 100. Aided by Phevor, the majority of known alleles that could be scored are ranked in the top 10.



	SIFT Only	SIFT and Phevor	PolyPhen Only	PolyPhen and Phevor
Top 1 Candidate	0%	21%	0%	23%
Top 10 Candidates	0%	39%	0%	39%
Top 100 Candidates	0%	40%	0%	45%
Scored Candidates	82%	82%	75%	74%
Average Rank	10,840	6,033	11,406	7,261

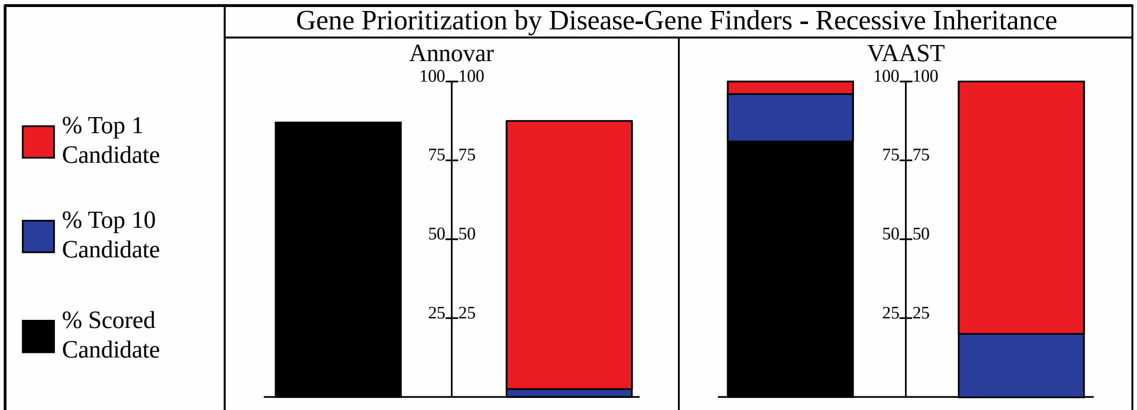
Figure 5.6. Recovery of Known Pathogenic Alleles Using Amino Acid Conservation to Prioritize Variants under a Dominant Inheritance Model

Variants from the benchmarking exomes interpreted and prioritized with the amino acid conservation tools SIFT and PolyPhen failed to rank any known alleles in the top 100. Fewer dominant alleles could be scored using SIFT and PolyPhen because candidate genes require a single scored heterozygous variant. Using the phenotype as a prior Phevor ranks the majority of known alleles in the top 10.

power with only a single exome to confidently identify a pathogenic allele. Still, the power of these two approaches is considerable. For recessive diseases, Annovar and VAAST rank 87% and 67% of the known alleles in the top 100 candidates, respectively. VAAST has the advantage over Annovar as it prioritizes candidate genes, whereas Annovar returns simple unordered list. At first glance, in Figure 5.7 and Figure 5.8 Annovar appears to outperform VAAST. However, Annovar's serialized filtering inadvertently removes many true positives—resulting in only 87% of recessive known alleles and 82% of the dominant alleles being recovered. VAAST suffers no such shortcoming and is able to score every gene. False negatives penalize Phevor's ability to reprioritize the Annovar results in light of the phenotype. Assigning false negatives with the median rank of all genes, Annovar alone ranks recessive known alleles at 2,842 and only 2,788 with Phevor's assistance.

In contrast, VAAST's probabilistic approach does not suffer from the same false negative problems that plague Annovar. Additionally, VAAST's output is prioritized—i.e., it orders genes by how damaging VAAST believes their variants to be. Thus, VAAST is able to identify 20% of known recessive alleles in the top 10 candidates (Figure 5.7), whereas Annovar cannot identify any. For dominant disorders (Figure 5.8), VAAST fails to rank any known alleles in the top 10 candidates, but it does produce an average rank better than any other variant interpretation tool.

Using Phevor in tandem with either of these tools greatly increases the accuracy of diagnosis. For recessive diseases, VAAST plus Phevor ranks 100% of target genes among the top 10 candidates in every run. The average rank for known recessive alleles when variants are interpreted by VAAST and Phevor reprioritizes candidate genes is 1.4



	Annovar Only	Annovar and Phevor	VAAST Only	VAAST and Phevor
Top 1 Candidate	0%	85%	4%	80%
Top 10 Candidates	0%	87%	19%	100%
Top 100 Candidates	87%	87%	67%	100%
Scored Candidates	87%	87%	100%	100%
Average Rank	2,842	2,788	172	1.4

Figure 5.7. Recovery of Known Pathogenic Alleles Using Comprehensive Disease-Gene Finders to Prioritize Genes under a Recessive Inheritance Model

Variants from benchmarking exomes were interpreted and genes prioritized using the comprehensive disease-gene finders Annovar and VAAST. Neither tool has sufficient power with only a single affected exome to accurately identify the disease-allele. Annovar has the added handicap of mistakenly filtering out true positives. With phenotype as a prior, Phevor returns the known pathogenic alleles in the top 10 candidates except those filtered by Annovar. Using VAAST to score variants and prioritize genes, when coupled with Phevor presents an efficient and reliable method to make a molecular diagnosis with only a single exome.

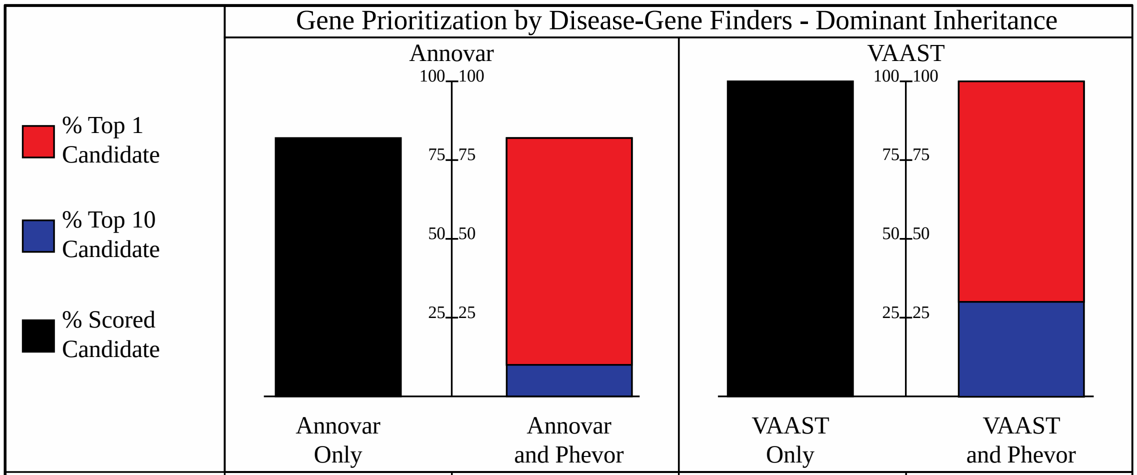
(Figure 5.7). Similarly, impressive results are observed for benchmarking analyses using dominant inheritance where the average ranking is 1.7 (Figure 5.8). Annovar’s performance is also markedly improved when used in conjunction with Phevor. In fact, as regards top candidates, it slightly outperforms the VAAST/Phevor combination, the target gene is ranked first by Phevor’s output 85% of the time for recessive diseases, and 72% of the time for dominants. By comparison, VAAST/Phevor only achieves this 80% and 70% of the time for recessive and dominant cases, respectively. However, Annovar’s false negative rate limits the power of the Annovar/Phevor duo for recessive diseases. Only 87% of the time the target gene is ranked in the top 10 candidates, whereas the success rate is 100% when using VAAST/Phevor combination.

5.3 The Behavior of the Phevor Algorithm

The benchmarking data presented above make it clear that Phevor works, and works quite well. However, how it accomplishes this is not always clear. Although I have done my best to describe the essentials of the Phevor algorithm in Chapter 4, what happens in practice is not immediately clear from the algorithm explanation alone. In this sense, Phevor resembles many other complex algorithms—dynamic programming and hidden Markov models, for example. The following sections try and demystify Phevor by illustrating the algorithm in action.

5.3.1 Biomedical Ontology Seeding and Propagation

As explained in Chapter 4, Phevor expands phenotype associations to genes with similar biological properties as seeded node scores are propagated across the bio-ontologies. Before propagation, Phevor identifies starting base nodes as the nodes most



	Annovar Only	Annovar and Phevor	VAAST Only	VAAST and Phevor
Top 1 Candidate	0%	72%	0%	70%
Top 10 Candidates	0%	82%	0%	100%
Top 100 Candidates	0%	82%	20%	100%
Scored Candidates	82%	82%	100%	100%
Average Rank	4,370	3,985	565	1.7

Figure 5.8. Recovery of Known Disease Alleles Using Comprehensive Disease-Gene Finders to Prioritize Genes under a Dominant Inheritance Model

Variants from benchmarking exomes were prioritized using the comprehensive disease-gene finders Annovar and VAAST. Using VAAST to score variants and prioritize genes when coupled with Phevor presents an efficient and reliable method to make a molecular diagnosis with a single exome. Serialized filtering by Annovar results in may false negatives. Therefore, Annovar is only able to return 82% of the known alleles.

distal to the ontology's root having a phenotype-linked gene association. In other words, the node has no child node with the same gene annotation. This behavior allows Phevor to begin propagation from the most specific nodes in the ontology. For example, searching distally in the Gene Ontology from its root, the gene *CSF1R* is first annotated to the node *tyrosine protein kinase activity* along with 140 other genes, descending further, the node *macrophage colony-stimulating factor receptor activity* is encountered, and *CSFIR* is also annotated to this node; thus propagation would begin from this node. Figure 5.9 illustrates these points using the Human Phenotype Ontology.

The black line in Figure 5.9 plots the percentage of genes annotated to the Human Phenotype Ontology as a function of the gene annotated node's depth in the Human Phenotype Ontology. As can be seen, the majority of genes are annotated to nodes located around five nodes distal to the Human Phenotype Ontology's root node. The red line in Figure 5.9 shows the depth distribution for the seeded base nodes selected by Phevor—note that these are generally deeper within the Human Phenotype Ontology. Postpropagation, each gene in the ontology is assigned its best scoring node to act as that gene's ontology representative. Looking at the distribution of these best nodes relative to their depth within the ontology, a drastic shift is observed—away from the seeded nodes towards the common parent (green line). This illustrates how propagation paths intersect to identify new nodes located at phenotypic "crossroads" in the ontology. However, for the pathogenic genes inserted into the exome backgrounds i.e., the target of the Phevor search (yellow line), the best ontology nodes tend to be found deeper into the ontology.

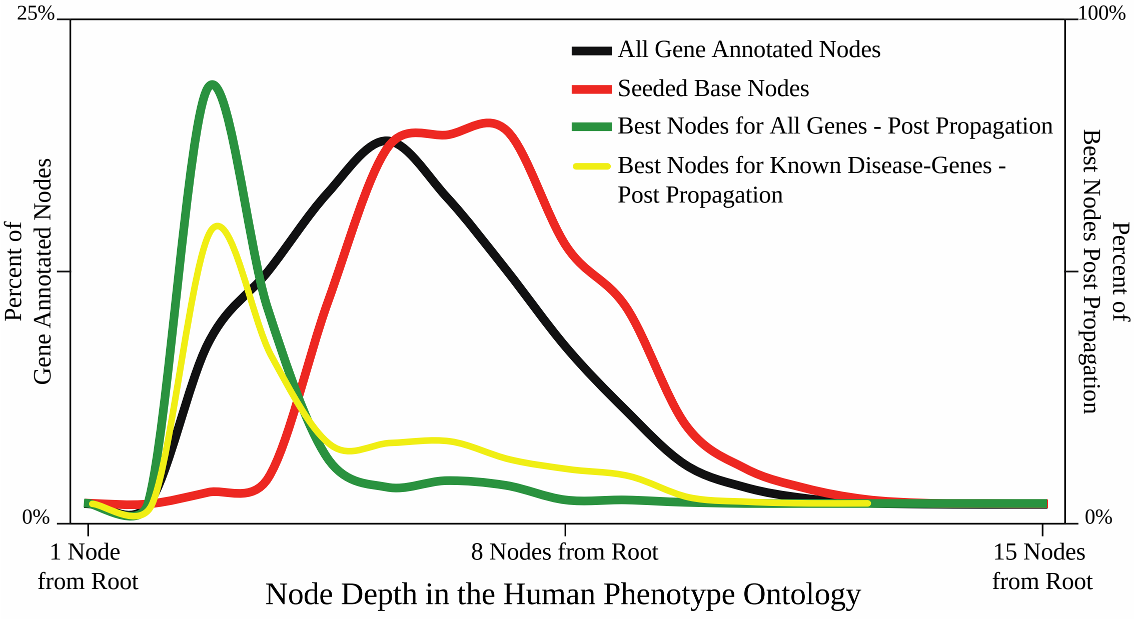


Figure 5.9. Distribution of Gene Annotated Nodes, Seeded Base Nodes, and Best Gene Annotated Nodes Postpropagation for the Human Phenotype Ontology

Plotting the distribution of gene annotated nodes, relative to the depth within the Human Phenotype Ontology, for all nodes with a gene annotation (black). The majority of nodes are found in the middle of the ontology, having only average specificity. Phevor seeds base nodes before ontology propagation (red). Seeded nodes have a deeper ontology depth than the average node—i.e., have greater specificity than the average ontology node. Postpropagation, the best nodes for all genes show a dramatic shift (green) towards the common parent—i.e., less specific nodes at the phenotypic "crossroads" of the seeded nodes. These results demonstrate Phevor's ability to find latent commonalities between phenotype-linked genes and additional candidates. The best nodes for the known pathogenic alleles (targeted phenotypes) have a distribution shifted away from the common parent (yellow). These nodes are more specific and show Phevor is finding latent connections shared by phenotype-linked genes with a high level of specificity.

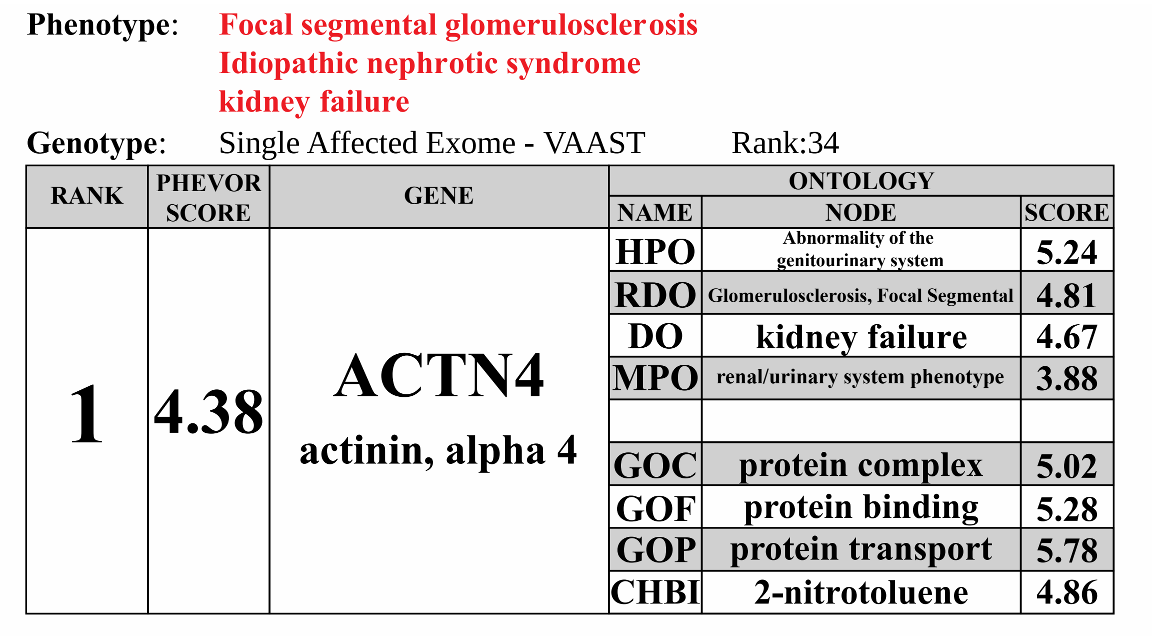
Crossroad nodes are especially important for Phevor's power. As discussed previously, phenotypes are often collections of apparently unrelated terms, e.g., *high blood pressure*, *deafness* and *mental retardation*. These crossroad nodes, identified during propagation, represent those nodes in the ontology that synthesize unifying phenotype terms. Genes attached to these nodes may never previously have been associated with any of the original phenotype terms. This is one of the ways Phevor uncovers latent knowledge within ontology to improve its diagnostic power.

Phevor also has the ability to combine ontologies that model knowledge in different domains of biology for still greater power, e.g., Human Phenotype and Gene Ontologies. Combining ontologies is accomplished by extending the propagation process across ontologies via their shared gene-annotations. This is a complex point, and an illustration is helpful. Consider the case for two *potassium transporters*, A and B. Deleterious alleles in one (A) are known to cause *cardiomyopathy*, whereas B, as of yet, has no disease associations. If A and B are both annotated in the Gene Ontology as *potassium transporters*, when Phevor propagates the Human Phenotype Ontology, associations of A to Gene Ontology, the Gene Ontology node *potassium transporter* will receive some score, which in turn will be propagated to B. Thus, even though B was absent from the Human Phenotype Ontology, its Phevor disease association score will increase because of its Gene Ontology annotation. This illustrates the simplest of cases. Many, more complex scenarios are possible. For example, A and B might be annotated to different nodes in Gene Ontology, with B's disease association score being increased proportionally following propagation across Gene Ontology. Importantly, neither of these scenarios is mutually exclusive.

5.3.2 Best Nodes from Bio-Ontology Propagation Aid in Identifying the Disease-Gene

Following propagation, the best scoring nodes in each ontology (corresponding to the green and yellow lines in Figure 5.9) are returned in the Phevor report. Consider Figure 5.10, which shows the top-scoring gene from a Phevor report. This gene (*ACTN4*) was ranked 34th on the original VAAST search (data not shown). Passing the VAAST report along with the patient phenotype (in this case, focal *segmental glomerulosclerosis*, *idiopathic nephrotic syndrome* and *kidney failure), ACTN4* is the top-scoring Phevor hit.

The Phevor score for *ACTN4* is 4.38—keep in mind that Phevor Scores are $\log_{10}$ Bayes ratios of the posterior probability of the disease model to that of the healthy model—see Chapter 4 and discussion of Equation 4.1-4.3 for more on this point. Thus, the Phevor score of 4.38 means that the disease model was 24,000 times more likely than the healthy model. On the right of Figure 5.10 are shown the best node from each ontology, and its Phevor "node" score. Topping the list are bio-ontologies containing human and model-organism phenotype information, below are gene properties and interactions. It is clear from the human and model organism phenotypes that *ACTN4* is very likely to cause the observed phenotypes. These nodes are the "crossroads" nodes discussed above. The biological properties found in the Gene Ontology and Chemical Entities of Biological Interest help support this conclusion as *focal segmental glomerulosclerosis* are proteins found in the urine, i.e., protein binding, transport and complexes. Additionally, 2-nitrotolune localizes to the kidney and causes a down-regulation of *ACTN4*. Collectively the results shown in Figure 5.10 illustrate how



Phenotype: **Focal segmental glomerulosclerosis**
Idiopathic nephrotic syndrome
kidney failure

Genotype: Single Affected Exome - VAAST Rank:34

RANK	PHEVOR SCORE	GENE	ONTOLOGY		
			NAME	NODE	SCORE
1	**4.38**	**ACTN4** **actinin, alpha 4**	**HPO**	**Abnormality of the genitourinary system**	**5.24**
			RDO	**Glomerulosclerosis, Focal Segmental**	**4.81**
			DO	**kidney failure**	**4.67**
			MPO	**renal/urinary system phenotype**	**3.88**
			GOC	**protein complex**	**5.02**
			GOF	**protein binding**	**5.28**
			GOP	**protein transport**	**5.78**
			CHBI	**2-nitrotoluene**	**4.86**

Figure 5.10. Example Phevor Report Detailing Bio-Ontology Evidence for Phevor Score and Ranking

The Phevor report ranks candidate genes by their Phevor score. The Phevor score is the $\log_{10}$ Bayes ratio of the disease model to the healthy model. Also reported are the best nodes from each bio-ontology and their ontology specific Phevor scores. Returning the best bio-ontology node annotated to each gene postpropagation helps provide evidence that the phenotype causing allele has been identified.

Phevor is able to rationally and reliably extract both explicit and latent knowledge modeled by bio-medical ontologies and use this information to reprioritize the outputs of a variant prioritization tool such as VAAST.

5.3.3 Distribution of Phevor Scores

As explained in the previous section, the Phevor score is the $\log_{10}$ ratio of the disease model to the healthy model. These two models are inversely correlated with each other, as easily seen in Figure 5.11A. Plotting the posterior probabilities of the disease and healthy models as a function gene rank in a Phevor report, and looking only at the disease model might give the erroneous impression that many genes are implicated by Phevor in the phenotype. However, once the ratio of the two models is taken, it is clear (Figure 5.11B) that only a few select candidate genes truly have a significantly higher disease model posterior probability compared to the healthy model. This quick drop off of Phevor scores leaves only a few gene candidates—clearly a desirable feature of the tool.

5.4 Phevor Accuracy and Pheneotype Specificity

Human Phenotype Ontology terms deeper in the ontology—distal to the common root node—describe more specific phenotypes. For example, *abnormality of cardiovascular morphology* is less specific than *ventricular septal defects*. Accurately describing the phenotype improves Phevor’s chance at identifying the damaged pathogenic allele. Figure 5.12 demonstrates the impact of phenotype specificity on Phevor’s accuracy. In this described experiment; as phenotype specificity increases,

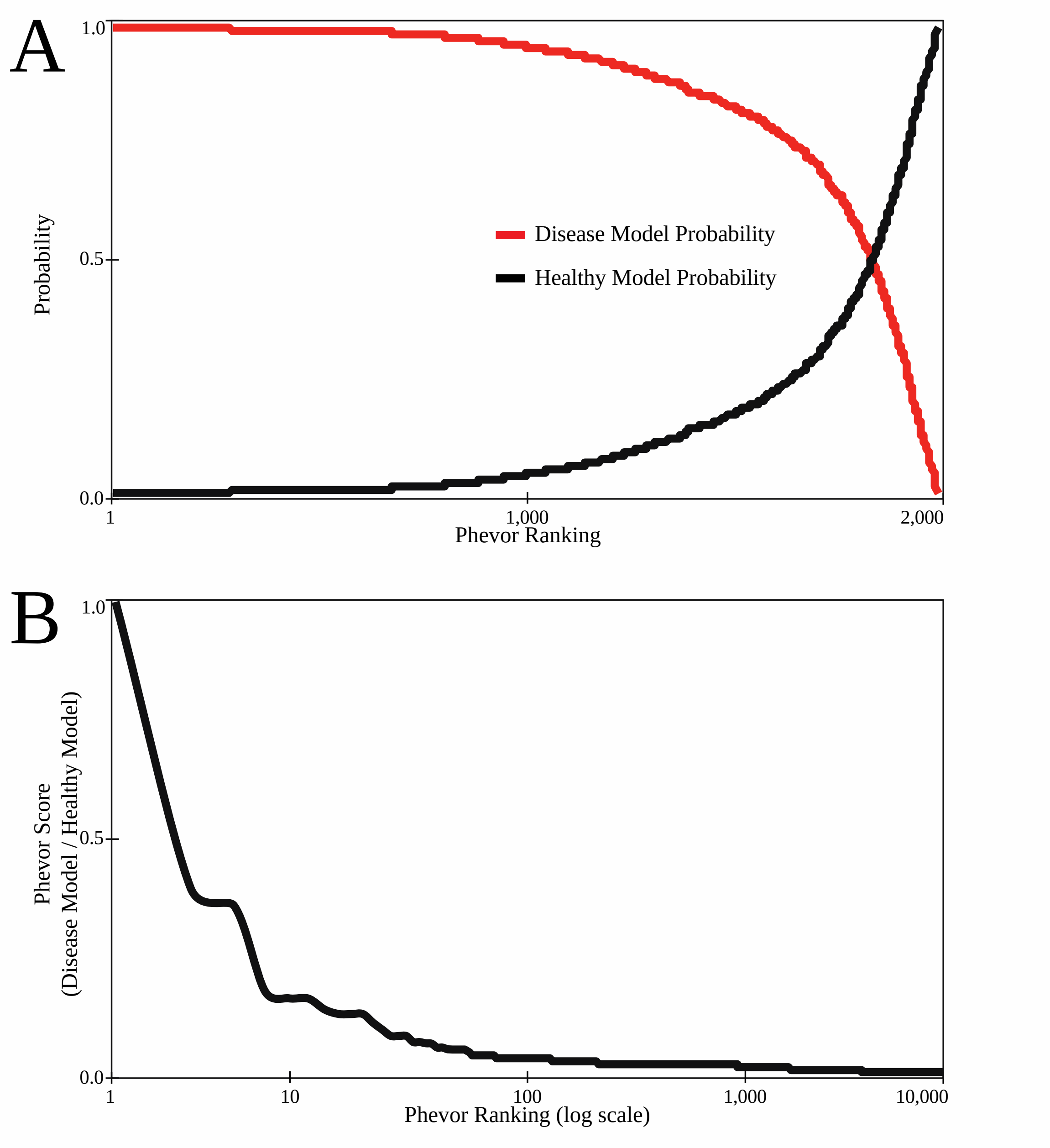


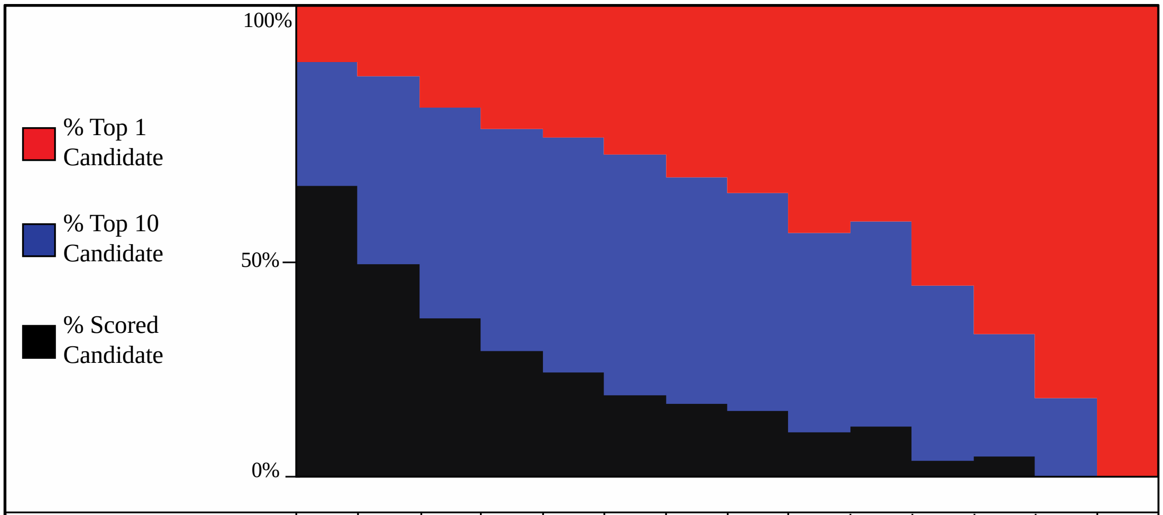
Figure 5.11 Distribution of Phevor Scores

A) Plotting the posterior probabilities of the disease and healthy models against their ranks in the Phevor report show the two models to be inversely correlated. Taking the log10 ratio of the two models B) reveals only a few candidate genes with significant Phevor scores. The illustrated Phevor scores represent the averages across all recessive benchmark exomes presented in Figure 5.7—variant interpretation and gene prioritization performed by VAAST and then analyzed by Phevor.

Phevor seeds fewer base nodes, thereby condensing the seeded nodes to only those with specific application to the biological properties driving the phenotype. At its deepest point, the Human Phenotype Ontology is 15 nodes deep. As can be seen, simply narrowing the phenotype to a first degree child node (see Figure 5.12) results in an average ranking of 27—giving marked improvement over VAAST alone. On average, the phenotype terms described in OMIM are at least six levels deep in the Human Phenotype Ontology. Thus, even though, as previously discussed in Chapter 3, OMIM phenotype descriptions are often vague they clearly provide Phevor with useful starting points. In summary, Figure 5.12 demonstrates that Phevor works well when provided with even relatively vague phenotype descriptions, and that it works progressively better as the specificity of these descriptions improves.

5.5 The Impact of Atypical Presentation and Misdiagnosis on Phevor's Accuracy

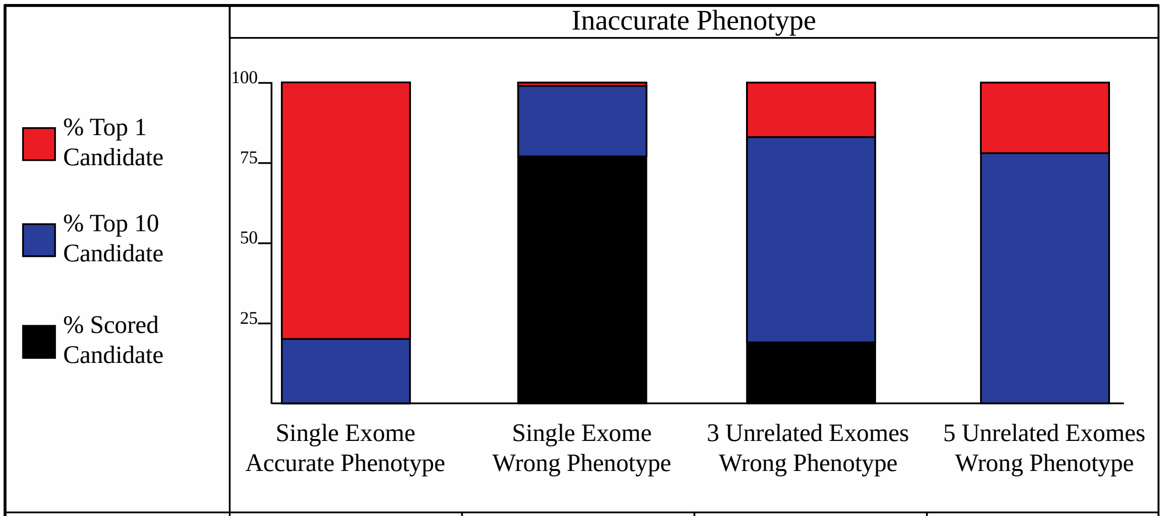
Figure 5.13 assays the impact of atypical presentation and misdiagnosis on Phevor's accuracy. The term atypical presentation refers to cases in which an individual has a known genetic disease but does not present with the typical disease phenotype. Reasons include novel alleles in known disease genes, novel combinations of alleles, ethnicity (genetic background effects), environmental influences, and in some cases, multiple genetic diseases presenting in the same individual(s), to produce a compound phenotype[127]. Atypical presentation resulting from novel alleles in known disease genes and compound phenotypes due to pathogenic alleles are emerging as a common occurrence in personal genome diagnosis[127-129]; thus, Phevor's performance in such



Nodes Away From Root	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# Seed Nodes	4,377	3,410	2,717	2,286	2,023	1,752	1,542	1,441	1,134	1,230	878	690	801	272
# Scored Nodes	3,842	3,739	3,531	3,346	3,220	3,120	3,012	2,962	2,715	2,749	2,458	2,236	2,345	1,105
Average Rank	**27**	**17**	**13**	**10**	**8**	**7**	**6**	**6**	**5**	**5**	**3**	**2**	**1**	**1**

Figure 5.12. Increasing Phenotype Specificity Improves Phevor's Identification of the Phenotype Responsible Allele

Using increasingly specific Human Phenotype Ontology terms to describe the phenotype of 1000 benchmarking exomes, Phevor's accuracy improves. More specific phenotype terms result in fewer seeded bio-ontology base nodes. A lower number of seeded base node results in fewer ontology nodes scored during the propagation, amplifying the biological properties responsible for the gene's phenotype.



Top 1 Candidates	80%	1%	17%	22%
Top 10 Candidates	100%	23%	81%	100%
Scored Candidates	100%	100%	100%	100%
Average Rank	1.4	57	7	4

Figure 5.13. Increased Confidence in Variant Interpretation can overcome an Atypical or Inaccurate Phenotype Description
Phevor’s accuracy suffers when intentionally inaccurate phenotypes are used to diagnose 1000 benchmark exomes. Accuracy can be recovered by increasing the confidence in the variant interpretation by VAAST. By sequencing and analyzing multiple unrelated cohorts, the variant interpretation scores (forward probability) can outweigh an inaccurate phenotype (prior probability).

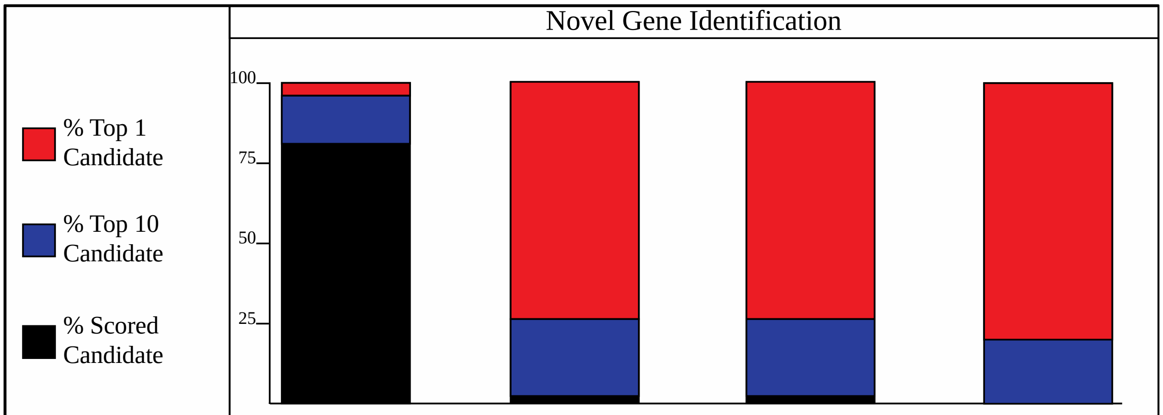
situations is of interest.

Figure 5.13 addresses the impact of atypical presentation on Phevor for case cohorts of one, three and five unrelated individuals, using the same benchmarking methodology as in Figures 5.1-5.8. For this experiment, however, I randomly replaced each disease-gene's OMIM derived phenotype description with another's, thereby mimicking an extreme scenario of atypical presentation/misdiagnosis, whereby each individual presents with not only an atypical phenotype, but still worse, one normally associated with some other known genetic disease. Unsurprisingly, this significantly influences Phevor's' diagnostic accuracy. Using VAAST outputs, for a single affected individual, accuracy declines from the damaged gene being ranked in the top 10 candidates genome-wide for 100% of the cases to 26%. More surprising is that Phevor is still able to improve on VAAST's performance alone, a phenomenon resulting again from Phevor's use of Gene Ontology, a point that I addressed above in section 5.3.1.

The remaining columns in Figure 5.13 measure the impact of increasing case cohort size. As can be seen, with three or more unrelated individuals all with the same (shuffled) atypical phenotypic presentation, Phevor performs very well, even when the phenotype information is misleading. Thus, these results demonstrate how Phevor's ontology-derived scores, are gradually overridden in the face of increasing sequence-based experimental data to the contrary—a clearly desirable behavior.

5.6 Novel Phenotype Association

The question naturally arises as to how dependent is Phevor upon the disease gene having been previously annotated to an ontology. Figure 5.14 addresses this issue.



	VAAST Only	Gene Ontology Only	Phenotype and Disease Ontologies	All Avalible Bio-Ontologies
Top 1 Candidate	4%	74%	74%	80%
Top 10 Candidates	19%	98%	98%	100%
Top 100 Candidates	67%	100%	100%	100%
Scored Candidates	100%	100%	100%	100%
Average Rank	172	38	2	1.4

Figure 5.14. Phevor Identifies Pathogenic Alleles with Limited Bio-Ontology Annotation

Masking the targeted alleles from the various bio-ontologies demonstrates Phevor's ability to identify novel pathogenic alleles that do not have a known phenotype or biological properties. Having only basic biological properties, resident to the Gene Ontology, Phevor can rank 98% of known alleles in the top 10. Having only phenotype and disease associations Phevor again ranked 98% in the top 10 candidates. These results exemplify Phevor's ability to make novel associations between damaged genes and their phenotypic consequence using latent relationships discovered in the bio-ontology propagation.

Figure 5.14 employs the same procedure used to produce Figures 5.1-5.8, but with the disease-gene removed from one or more of the ontologies prior to running Phevor. This makes it possible to evaluate the ability of Phevor to improve the ranks of a disease gene in the absence of ontological assignments (i.e., as if it were a novel disease gene, never before associated with a disease or phenotype). For these benchmarks, I investigated not only the impact of simultaneously masking the gene's Human Phenotype, Mammalian Phenotype and Disease Ontologies phenotype annotations, but its Gene Ontology annotations. Because VAAST has proven to be the superior variant interpretation and gene-prioritizing tool, Figure 5.14 presents the results of these experiments using only VAAST outputs.

As can be seen, removing the gene from one or more ontologies does decrease Phevor's power to identify the gene, but does not eliminate it; this demonstrates that Phevor is gaining power by combining multiple ontologies. Removing the target gene from Gene Ontology, and using only the phenotype ontologies (Human Phenotype, Mammalian Phenotype, Disease Ontologies etc.) the target disease gene is still ranked in the top 10 candidates 98% of the time, and among the top 100 candidates 100% of the time. By comparison, using VAAST alone the target gene is ranked among the top 10 and 100 candidates 19% and 67% of the time, respectively. The false negative rate is an artifact of the benchmark procedure and results from removing the gene from the Gene Ontology. Briefly, because the majority of human genes (18,824) are already annotated to the Gene Ontology, the prior expectation is that a novel disease gene is also more likely to be annotated to Gene Ontology than not, causing Phevor to prefer candidates already annotated to the Gene Ontology in this benchmarking scenario.

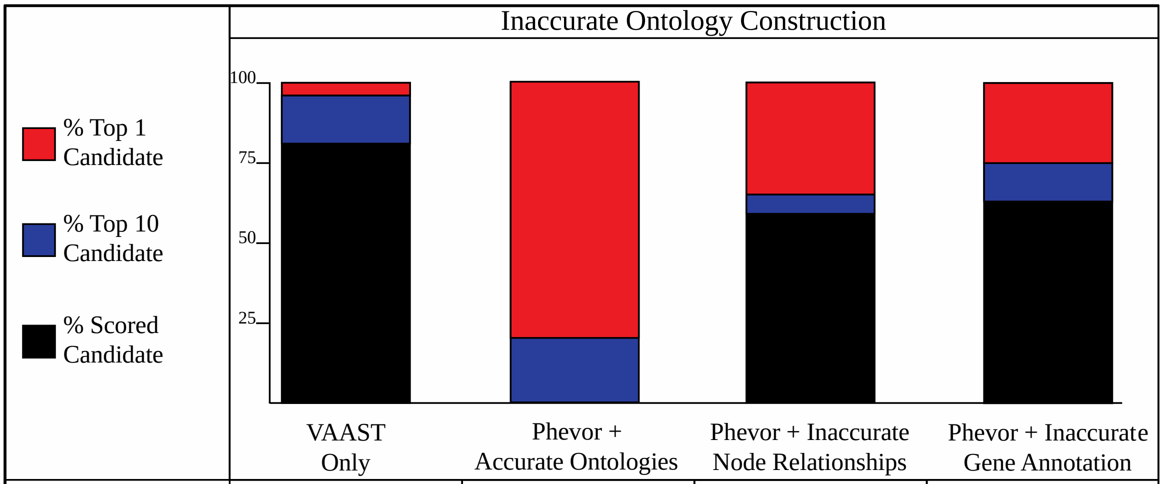
Similar trends are seen using the Gene Ontology[121] alone. This time removing the gene from the Mammalian Phenotype, Human Phenotype and Disease Ontologies, Phevor places the disease gene among the top ten candidates 98% of the time and among the top 100 candidates 100% of the time. Recall that for this analysis, Phevor was provided with only a phenotype description—not Gene Ontology terms—and that the disease gene was removed from every ontology containing any phenotype data, e.g., the Human Phenotype Ontology, the Disease Ontology and the Mammalian Phenotype Ontology. Thus, this increase in ranks (e.g., 19% vs. 98% in the top 10) is solely the result of Phevor's ability to integrate the Gene Ontology into a phenotype driven prioritization process, demonstrating that Phevor can use the Gene Ontology to aid in the discovery of new disease-genes and pathogenic alleles. Collectively, these results demonstrate that a significant portion of Phevor's power is derived from its ability to relate phenotype concepts in the Human Phenotype Ontology to gene function, process and location concepts modeled by the Gene Ontology.

Figure 5.14 demonstrates that Phevor improves the performance of the variant prioritization tool for novel disease-genes. This is possible because, even when a (novel) disease gene is absent in the Human Phenotype Ontology, Phevor can nonetheless assign it a high score for disease association after information associated with its paralogs are propagated by Phevor from the Human Phenotype Ontology to the Gene Ontology.

5.7 Importance of Biomedical Ontology Structure on Phevor's Accuracy

Accurate gene annotation to bio-ontology nodes is just as important as the relationship structure of the ontology. Phevor relies on proper gene annotations to seed the proper nodes before propagation and connect each of the bio-ontologies. Correct annotation is again critical when assigning the best ontological node and calculating the Phevor score. When propagating the seed scores across the ontologies, the relationships between nodes are just as important as what nodes were seeded. Using multiple bio-ontologies helps to limit the impact of annotation and relationship errors.

To assess the impact of poor ontology construction on Phevor's accuracy, two experiments were performed (Figure 5.15) in which the bio-ontology structure was intentionally broken. In the first experiment, the gene annotations attached to each ontological node were randomized; thus destroying accurate representation of a gene's true biological properties (Gene Ontology) or phenotype associations (Human Phenotype Ontology). In the second experiment, the correct gene annotations were retained, but the relationships within the ontology were randomized. Again, this affects Phevor's power to identify the pathogenic alleles. Interestingly, as seen in Figure 5.15, both experiments have about equal impact on Phevor's performance, demonstrating equal importance of ontology design and gene annotation. Incidentally, to my knowledge, this is the first ever measurement of the proportion of knowledge modeled within a biomedical ontology by its node-gene associations, compared to the quantity of knowledge contained in the ontology's relationships. Much blood has been spilled over the years as to the true value of ontologies versus controlled vocabularies (that lack



	VAAST Only	Phevor + Accurate Ontologies	Phevor + Inaccurate Node Relationships	Phevor + Inaccurate Gene Annotation
Top 1 Candidate	4%	80%	35%	25%
Top 10 Candidates	19%	100%	41%	37%
Top 100 Candidates	67%	100%	49%	48%
Scored Candidates	100%	100%	100%	100%
Average Rank	172	1.4	618	433

Figure 5.15. Accuracy of Ontology Design and Gene Annotations Impact Phevor's Ability to Identify Pathogenic Genes

Phevor's accuracy is impacted when ontology relationship structure or gene annotations are inaccurate. VAAST performed variant interpretation and using the OMIM derived phenotype of the pathogenic allele Phevor reprioritized genes. Randomizing the gene annotations across the bio-ontologies decreases but does not destroy Phevor's ability to recover the known alleles. Likewise, randomizing the bio-ontologies node relationships has a similar effect on Phevor's ability to make connections between phenotype-linked genes and their biological properties. These results suggest half the information contained in bio-ontologies reside in the controlled vocabulary and gene annotations, the other half in the relationships between terms.

relationships between terms) to annotate biomedical data. The results shown in Figure 5.15 make it clear that (at least for these analyses) the node relationships approximately double total knowledge content.

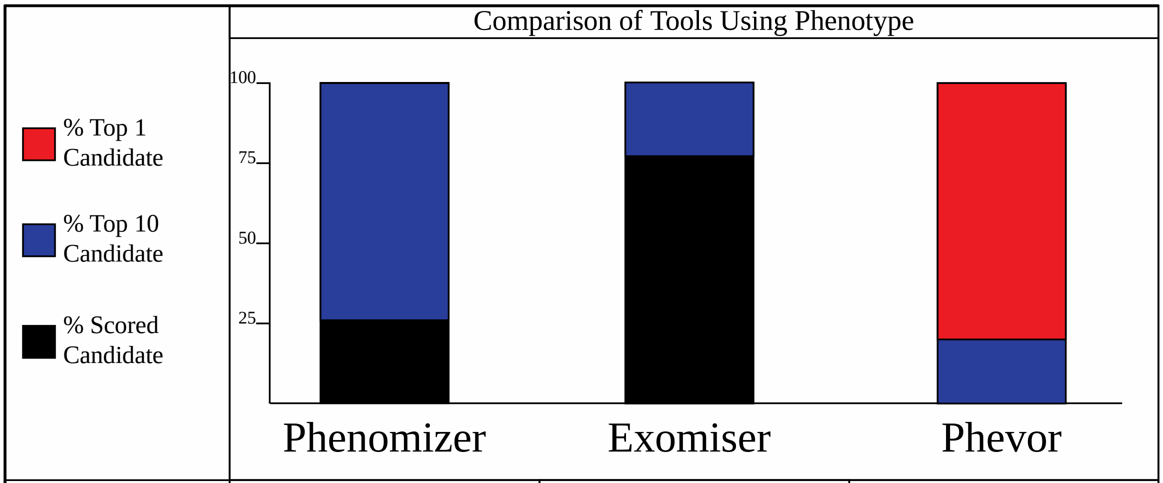
5.8 Comparing Phevor to other Tools

Phevor has a unique approach to its methodology; however, other tools that attempt to employ phenotype information for improved variant prioritization do exist. Two such tools are Phenomizer[104] and Exomiser[130]. Phenomizer was briefly discussed in Chapter 3—it is a web-based phenotype search tool that queries phenotype terms against disease databases to return candidate genes. Candidate gene ranks and their significance is calculated using a semantic similarity metric[104]. The results returned by Phenomizer consist of named diseases with similar phenotypes to the query phenotype along with genes previously shown to play a role in producing those phenotypes/diseases. Phenomizer does not offer any means to compute upon sequence data such as exomes or whole genomes. Exomiser[130] is another web-based tool and does analyze an individual's genotype. Exomiser uses the comprehensive disease gene finder Annovar[41] to filter variants that meet user-defined criteria such as MAF. Before filtering, Exomiser uses the phenotypes of mouse models to pre-identify likely candidates. This is done using the semantic similarities between the patient's phenotype and the mouse models and hence sheds light on the underlying cause of the patient's phenotype. Both Phenomizer and Exomiser, unlike Phevor, are bound to known diseases and phenotypes. There is no possibility of making a novel phenotype association using either of these tools.

Additionally, as detailed in Figure 5.16, neither tool is able to match the performance and accuracy of using VAAST to interpret variants and Phevor to reprioritize genes. First, I once again inserted two copies of known pathogenic allele randomly selected from HGMD [131] (see methods for details) into a target exome, repeating the process 100 different disease genes. To produce this figure, I then passed Phenomizer the OMIM phenotype for each of these 100 diseases. Those results are shown on the left. Note that although Phenomizer does not use genotype data, using phenotype information alone it was able to rank the target gene among the top 10 candidates 74% of the time, but never succeeded in placing the candidate gene first in the report. Surprisingly Exomiser, even though it also employs the variant data, performed much more poorly, placing the target gene among its top 10 candidates only 23% of the time. In contrast, the VAAST + Phevor duo placed the target gene among the top 10 candidates 100% of the time and first in its list 80% of the time.

5.9 Benchmarking Conclusions

This chapter presented a series of benchmarks and demonstrates that Phevor provides an effective means to improve the diagnostic power of widely used variant interpretation tools. As I have shown, Phevor's ability to improve the accuracy of variant interpretation tools is the result of its ability to relate phenotype and disease concepts in ontologies such as Human Phenotype Ontology to gene function, process and location concepts modeled by the Gene Ontology. This allows Phevor to model key features of genetic disease that are not taken into account by existing methods[104,132,133] that employ phenotype information for variant prioritization. For example, paralogous



Top 1 Candidate	0%	0%	80%
Top 10 Candidates	74%	23%	100%
Top 100 Candidates	94%	47%	100%
Scored Candidates	100%	100%	100%
Average Rank	112	5,719	1.4

Figure 5.16. Comparison of Tools Using Phenotype to Find Candidate Genes

Phevor was evaluated against two other tools that use phenotype to identify candidate genes. Phenomizer accepts phenotype terms, searches for similarities in known diseases, and returns candidate diseases and genes. Phenomizer does not incorporate genotype data. Exomiser does incorporate genotype using Annovar’s filtering methodology after identifying likely candidate genes whose mouse model homologues have a phenotype with semantic similarity to the patient’s phenotype. Phevor’s ability to identify the pathogenic allele is superior to Phenomizer or Exomiser when given same phenotypes and benchmarking exomes.

genes often produce similar diseases[134] because they have similar functions, operate in similar biological processes and are located in the same cellular compartments.

Phevor scores take into account not only weight of evidence that a gene is associated with the patient's illness, but that it is not. In typical whole exome, searching every variant interpretation tool identifies many genes harboring what it considers deleterious mutations. Often the most damaging of them are found in genes without any known phenotype associating them with the disease of interest; moreover, in practice, highly deleterious alleles are also often false positive variant calls. Phevor successfully down weights these genes and alleles, with the target disease gene's rank climbing as an indirect result. This phenomenon is well illustrated by the fact that Phevor improves the accuracy of variant interpretation even when provided with an incorrect phenotype description, e.g., Figure 5.13. This result underscores the consistency of Phevor's approach; it also has some important implications. Namely, that lack of previous disease association, weak phylogenetic conservation, and lack of Gene Ontology annotations for a gene are (weak) *prima facie* evidence against disease association.

CHAPTER 6

PHEVOR AND THE UTAH GENOME PROJECT

Phevor's inclusion in the Utah Genome Project's analysis pipeline has demonstrated its diagnostic utility using real cases. Here, I describe the Utah Genome and present six clinical cases where Phevor aided in the diagnosis.

6.1 The Utah Genome Project

The Utah Genome Project (UGP) is a large-scale, intramural genome-sequencing project, the aim of which is to discover new pathogenic genes, and diagnose inherited diseases. The State of Utah is especially well suited for the study of inherited diseases. Originally settled by the Mormon population, much of the populace of Utah belongs to the Church of Jesus Christ of Latter Day Saints (LDS); a culture very interested in genealogy. Additionally, members of the LDS church are often part of very large families. Large families and extensive genealogical knowledge combine to make Utah the perfect place for studies of genetic disease; indeed, many famous discoveries involving genetic disease have been made at the University of Utah over the years[85,135-138]. Utah has also produced world-renowned genetic testing companies, including Myriad Genetics, ARUP Laboratories and Ansestory.com.

In lock-and-step with the rise in next generation sequencing and the genomics era, Utah has worked to once again, position itself as a center for genetic discovery. The UGP seeks to find genetic signatures of disease, large families and deep genealogical knowledge to amplify signals that are absent in case-control cohorts. The UGP has developed an alignment and variant calling pipeline that mirrors many of the techniques proposed by the BROAD Institute[16]. What sets the Utah Genome Project apart from similar projects is its variant interpretation and diagnostic pipeline, that includes cutting edge bioinformatics tools—including VAAST[42,43] and Phevor[35], and an HIPPA compliant diagnostic reporting interface, Opal[139].

Figure 6.1 presents a schematic of the UGP's analysis pipeline. Briefly, single probands and small-unrelated case-control cohorts are analyzed using VAAST. Nuclear families and large pedigrees are analyzed using the newly developed tool pVAAST[139,140]. pVAAST is an extension of the VAAST algorithm that incorporates linkage information for greater power. VAAST and pVAAST results are then passed to Phevor along with the patient's phenotype for diagnosis. Below I describe six of such cases that demonstrate the role and utility of Phevor for the UGP pipeline.

6.2 Utah Genome Project – Two Diagnoses Missed During Clinical Screening

Below I describe cases presented to the Utah Genomes Project (UGP) where diagnoses were made using exome sequencing after locus-specific testing proved inconclusive. These cases highlight the limitations of locus-specific testing and the advantages of diagnostic exome sequencing.

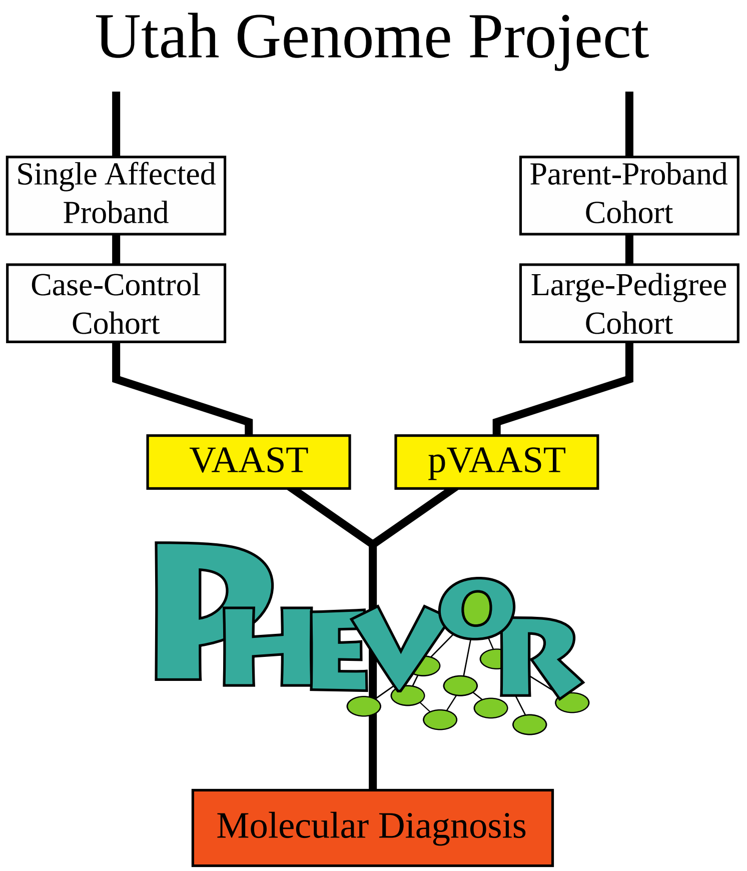

Utah Genome Project
Single Affected Proband
Parent-Proband Cohort
Case-Control Cohort
Large-Pedigree Cohort
VAAST
pVAAST
PHEVOR
Molecular Diagnosis

Figure 6.1. Schematic of the Utah Genome Project Analysis Pipeline

Single affected probands and case-control cohorts are analyzed used VAAST. Families and extended pedigrees are analyzed using pVAAST. Results from VAAST and pVAAST are analyzed together with phenotype descriptions using Phevor for an accurate molecular diagnosis.

6.2.1 Progressive Familial Intrahepatic Cholestasis

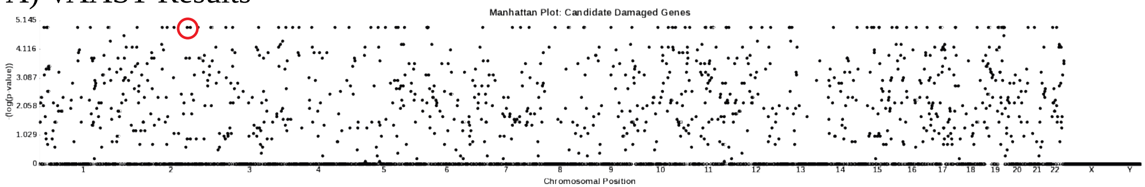
A six-month old infant with a liver disease was presented to the UGP by Stephen Guthery and Karl Volkerding and was suspected of having Familial Intrahepatic Cholestasis (PFIC)[141,142]. Clinical presentation of PFIC includes *failure to thrive*, *growth retardation*, *intrahepatic cholestasis*, *jaundice* and *hepatomegaly*. The treating physician recognized the symptoms and made a clinical diagnosis of PFIC but was not in a position to make a molecular diagnosis.

Traditional locus specific testing for alleles known to cause this phenotype was inconclusive. The proband's exome was then sequenced. Using only a single affected exome VAAST was utilized to interpret variants and prioritize damaged genes. As already discussed, VAAST is underpowered using only a single affected exome. VAAST returned 105 candidate genes tied for first place (Figure 6.2A).

The Phevor analysis used only a single term to describe the patient's phenotype: *Intrahepatic cholestasis*. Phevor ranked an ATP-binding cassette gene, *ABCB11* as the top ranked candidate (Figure 6.2B). *ABCB11* is known to cause intrahepatic cholestasis[143] and was included, but missed, in the locus-specific testing performed on the patient prior to exome sequencing.

Sanger conformation of the variant and follow-up sequencing of the child's parents confirmed the Phevor diagnosis. Using these data, VAAST identified two variants in *ABCB11* as damaging. Each inherited from a different parent forming a compound heterozygote in the proband. The paternal variant (NM_003742.2:c.3332T>C; p.Phe1111Ser) and the maternal variant (c.890A>G; p.Glu297Gly) are both considered highly damaging by VAAST. The maternal variant is

A) VAAST Results



B) Phevor Results

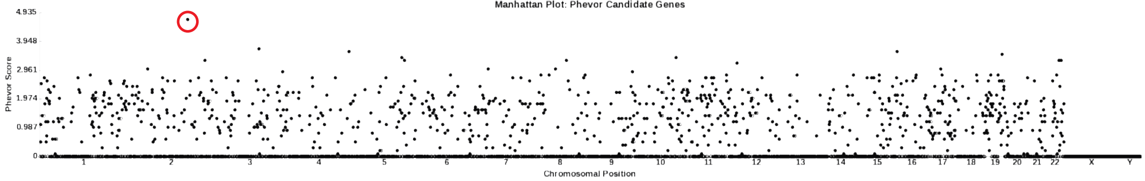


Figure 6.2. VAAST and Phevor Results for a Single Proband with Intrahepatic Cholestasis

Represented as a Manhattan plot, the VAAST p-value or Phevor score is plotted against its genomic position. A) VAAST returns *ABCB11* tied with 104 other candidate genes. B) VAAST output and the phenotype of *intrahepatic cholestasis* passed to Phevor. *ABCB11* (circled in red) is the top candidate.

a known disease-allele, shown to cause intrahepatic cholestasis while the paternal mutation has not been previously reported.

These results demonstrate the utility of Phevor to identify novel mutation in a gene known to be responsible for the disease. Additionally, VAAST and Phevor together identified a novel disease-allele in trans to a known pathogenic allele using only a single affected exome.

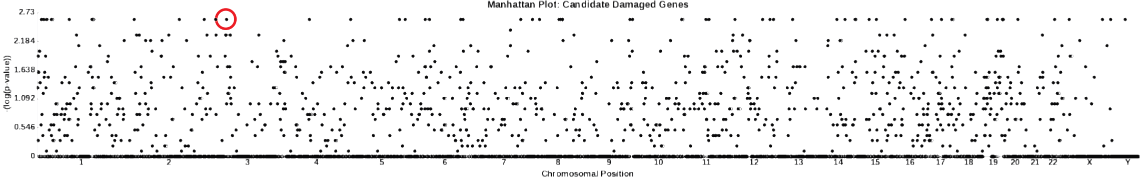
6.2.2 Sick Sinus Syndrome

Martin Tristani-Firouzi and colleagues presented an individual to the UGP as a pilot for a potentially larger study whose aim was to discover novel causes of sick sinus syndrome and other associated cardiac defects. Clinical presentation of the individual included: *Slower than normal pulse*, *dizziness or fainting*, *shortened breath* and *heart palpitations*.

Locus-specific testing was performed as part of clinical screening. No known mutations associated with sick sinus syndrome were identified. Exome sequencing was performed on the proband as a preparation for a larger UGP family-based sequencing analysis. Variant interpretation and gene prioritization by VAAST on the proband exome identified 66 genes tied as potential candidates (Figure 6.3A).

Phevor analysis was then performed using the same VAAST output together with the phenotypes *prolonged QT intervals*, *sick sinus syndrome*, *ventricular escape rhythms*, *atrioventricular block*, and *sinus bradycardia*. Phevor returned a gene known to cause sick sinus syndrome ranked as second (Figure 6.3B). Variants in this gene, *SCN5A*, are known to cause sick sinus syndrome[144-146] and were part of the mutation

A) VAAST Results



B) Phevor Results

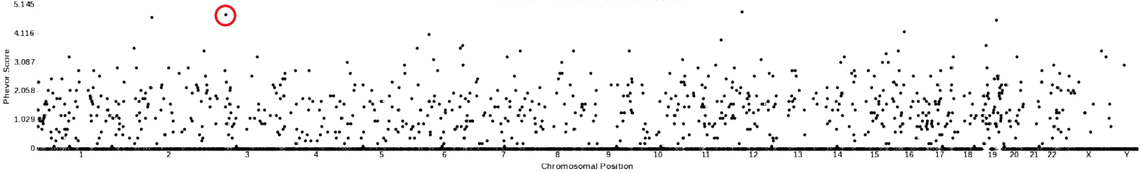


Figure 6.3. VAAST and Phevor Results for a Single Proband with Sick Sinus Syndrome Represented as a Manhattan plot, the VAAST p-value or Phevor score is plotted against its genomic position. A) VAAST returns *SCN5A* (circled in red) tied with 65 other candidate genes. B) The same VAAST output file and the phenotype of *prolonged QT intervals*, *sick sinus syndrome*, *ventricular escape rhythms*, *atrioventricular block*, and *sinus bradycardia* were used by Phevor to identify *SCN5A* as the second ranked candidate in panel B. *SCN5A* has previously been associated with sick sinus syndrome.

panel tested, and missed during the clinical screening process. It is unclear why the locus specific testing failed to identify these alleles. The allele was confirmed by Sanger sequencing and is a known pathogenic variant (NM_000335.4:c.4255G>C; p.Gly1419Arg). Follow-up studies confirmed the association of this allele with affected members of the family.

These results highlight once again the limitations of locus-specific testing and in particular, mutation panels for a diagnosis. They also illustrate the power of VAAST and Phevor for accurate diagnosis using only a single proband's exome.

6.3 Utah Genome Project – Novel Genes and Atypical Phenotypes

A major research goal of the Utah Genome Project is to identify new disease-genes. Phevor's ability to use phenotypes and biological information for identification of pathogenic alleles provides the UGP a unique tool. Presented here are two such cases illustrating the power of Phevor to enable novel discovery.

6.3.1 Common Variable Immunodeficiency (CVID)

Karin Chen presented two separate families to the UGP and colleagues with members having a possible autosomal dominant disorder characterized by early-onset *hypogammaglobulinemia* with *variable autoimmune features* and *adrenal insufficiency*. The first family had an affected mother and two affected children, while the father was unaffected. The second family had a single affected individual with the same phenotype.

All four affected individuals were used as cases and the unaffected father as the control. As pVAAST was not available at the time the original analyses were carried out, VAAST was used to interpret variants and prioritize genes[147]. Recall that unlike pVAAST, VAAST is designed for cohorts of unrelated individuals, and its statistical approach does not take into account relatedness among cases or controls. As a result, VAAST identified an excess of candidates—genome wide significance was reached for 168 candidate genes, with 17 sharing the top significance level (Figure 6.4A).

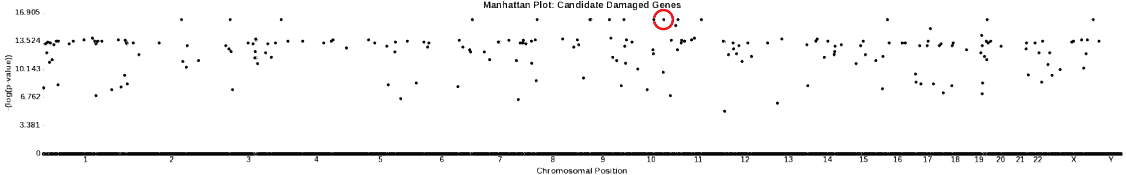
Later analyzed using the same VAAST results in conjunction with the phenotype terms *abnormality of humoral immunity* and *combined B and T cell immunodeficiency,* Phevor identified *NFKB2* as the most likely candidate gene (Figure 6.4B).

VAAST identified two damaging variants (one from each family) in *NFKB2*. The first family had a single base deletion resulting in a frameshift mutation (NM_002502.4:c.2564delA; p.Lys855Serfs*7) and the second family had a nonsense mutation just two amino acids upstream from the frameshift (NM_002502.4:c.2557C>T; p.Arg853*).

Subsequent immunoblot analysis and immunofluorescence microscopy of transformed B cells from affected individuals showed that the *NFKB2* mutations affect phosphorylation and proteasomal processing of the p100 NFKB2 to its p52 derivative and, ultimately, p52 nuclear translocation established *NFKB2*. Hence the noncanonical NF-κB signaling pathway, as a genetic etiology for this primary immunodeficiency syndrome[147].

These mutations in were the first ever association of *NFKB2* with CVID. Indeed, it is wholly absent from the Human Phenotype, Disease and Mammalian

A) VAAST Results



B) Phevor Results

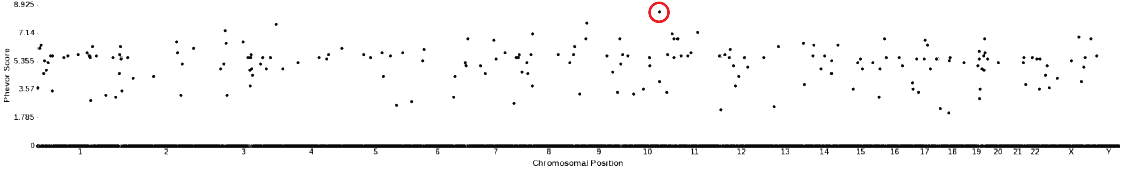


Figure 6.4. VAAST and Phevor Results for a Case-Control Cohort with Common Variable Immunodeficiency

Manhattan plot of the VAAST p-value or Phevor score is plotted against each gene’s genomic location. A) VAAST returns *NFKB2* tied with 16 additional candidate genes, and 168 candidate genes attain genome-wide significance. B) VAAST prioritized genes and the phenotype of *abnormality of humoral immunity* and *combined B and T cell immunodeficiency* were used by Phevor to identify *NFKB2* as the top candidate and provide a molecular diagnosis.

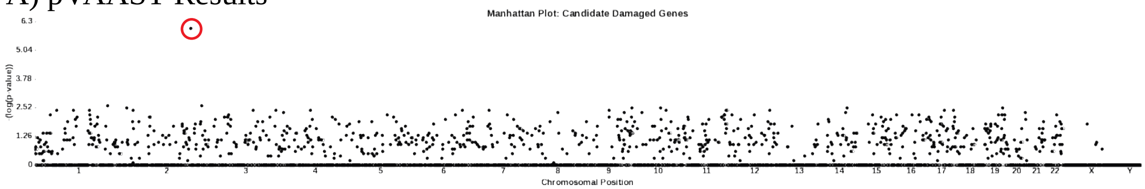
Phenotype Ontologies. The role played by *NFKB2* in the noncanonical nf-kappa-beta signaling pathway, however, is well studied and the gene is well annotated in the Gene Ontology. These results perfectly demonstrate Phevor's ability to discover latent information scattered across multiple ontologies belonging to different domains of knowledge—in this case gene function information never previously explicitly linked to a phenotype—to discover novel phenotype associations.

6.3.2 Immune Dysregulation, Polyendocrinopathy, Enteropathy

A severely sick 12 year old male was presented to the UGP by Stephen Guthery and colleagues having *severe diarrhea*, *intestinal inflammation*, *total villous atrophy* and *hypothyroidism*. Initially diagnosed with immune deregulation, polyendocrinopathy, enteropathy, and X-linked (IPEX) syndrome[148], this individual required continual intravenous feeding to support growth and was hospitalized numerous times for bloodstream infections. No molecular diagnosis was established despite a comprehensive, multidisciplinary clinical evaluation and locus specific testing for *FOXP3*[149] and *IL2RA*[150]—prime candidates for IPEX syndrome. Prior to obtaining a life sustaining hematopoietic stem cell transplantation, DNA was obtained from the proband and parents for genome sequencing. pVAAST was utilized to interpret variants and prioritize genes using a *de novo* inheritance model. The *de novo* inheritance model and pVAAST identified a single significant candidate gene—*STAT1* (Figure 6.5A).

Using the phenotypes of *hypothyroidism*, *paronychia*, *autoimmunity* and *abnormality of the intestine*, Phevor reaffirmed the highest candidate, *STAT1*. *STAT1* is a gene previously associated with susceptibility to mycobacterial infections[151,152]

A) pVAAST Results



B) Phevor Results

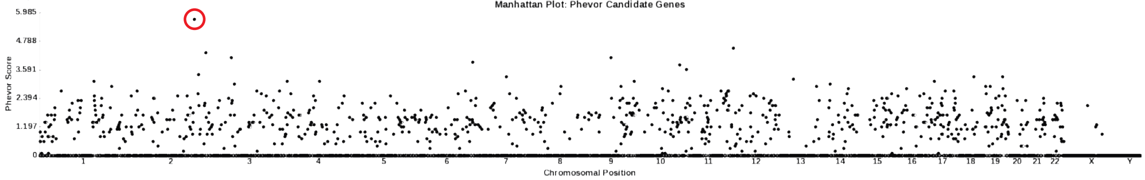


Figure 6.5. pVAAST and Phevor Results for a Proband with Severe Enteropathy

Represented as a Manhattan plot, the pVAAST p-value or Phevor score is plotted against its genomic position. A) pVAAST returns *STAT1* as the single best candidate (circled in red) using a *de novo* inheritance model where it achieved genome wide significance. B) Phevor confirms *STAT1* as the best candidate.

which are phenotypically somewhat dissimilar to the phenotype of the proband. Moreover, ontology information returned by Phevor in its report points to autoimmune disorders connected with *STAT1*.

Gain-of-function mutations in *STAT1* are known to cause immune mediated human disease[153] and *STAT1* is a transcription factor that regulates FoxP3—the foremost candidate gene. A single *de novo* mutation (NM_139266.2:c.1154C>T; p.Thr385Met) was identified by pVAAST and confirmed by Phevor as very likely to be responsible for the child's illness. Supporting this conclusion are the recent reports of this same allele causing chronic mucocutaneous candidiasis and an IPEX-like syndrome[153,154].

This case demonstrates Phevor's ability to identify pathogenic genes even when patients present with an atypical phenotype. This feature is just as important, if not more so, than its ability to make novel phenotype associations (e.g., *NFKB2*). Much of the phenotype data existing for model organisms does not translate to humans.

6.4 Utah Genome Project – Large Pedigrees

The Utah Population Database contains nearly 20 million entries detailing the genealogy and health records of Utah residents and their ancestors[155]. Access to this database makes it possible for Utah Genome Project investigators to identify entire families plagued with particular recurring and likely genetic diseases. Two recent such examples have been diagnosed with the aid of Phevor.

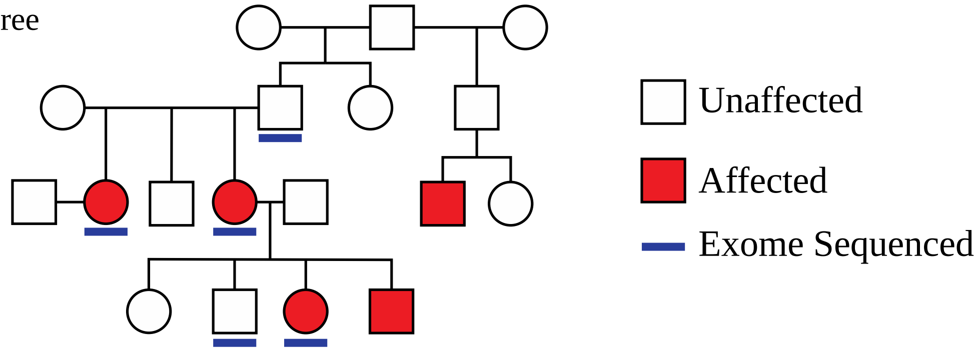
6.4.1 Wolff-Parkinson-White Syndrome

Peter Gruber and colleagues presented an extended pedigree to the UGP having multiple family members diagnosed with Wolff-Parkinson-White syndrome (WPW)[156,157]. The phenotype of WPW includes *heart palpations, dizziness, lightheadedness, chest pain, difficulty breathing* and even *sudden cardiac death.* Wolff-Parkinson-White has only a single gene known to cause the disorder—*PRKAG2*[156], displaying dominant inheritance with incomplete penetrance.

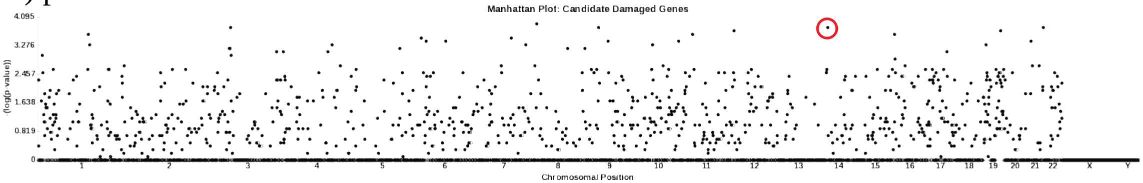
Prior to exome sequencing, all affected family members were tested for causal variants in *PRKAG2* and were found to be negative. This pedigree was selected because of the high number of affected individuals, and the improved chance of finding the causative gene because there is one common male ancestor to all those affected (Figure 6.6A). In total, five members of the family were sequenced, two unaffected and the other three affected. pVAAST was used to analyze the family's exomes. Several genes achieved genome-wide significance (Figure 6.6B), making it difficult to confidently identify the best candidate gene.

The Phevor analysis was performed using the pVAAST prioritized genes and the phenotypes: *prolonged QRS complex*, *shortened PR interval*, *paroxysmal supraventricular tachycardia*, *sudden cardiac death*, *ventricular pre-excitation with multiple accessory pathways*, *paroxysmal atrial fibrillation*, *stroke*, *cardiomyopathy* and *palpations*. The specific phenotypes were used to run Phevor opposed to the WPW disease name, because as previously explained there is only one gene associated with its name. Phevor identified a single strong top candidate (Figure 6.6C). Myosin heavy chain 6 (*MYH6)* is ranked sixth by pVAAST and first by Phevor. A single variant was

A) Pedigree



B) pVAAST Results



C) Phevor Results

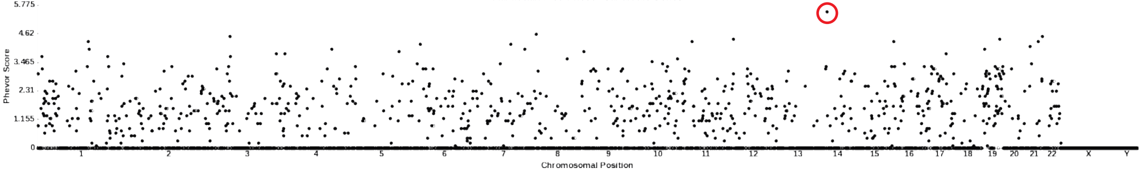


Figure 6.6. pVAAST and Phevor Results for a Family with Wolff-Parkinson-White Syndrome

Manhattan plot with the pVAAST p-value or Phevor score plotted against each gene's genomic location. A) pVAAST returns *MYH6* as the sixth best candidate using a dominant inheritance model. B) The phenotypes *prolonged QRS complex*, *shortened PR interval*, *paroxysmal supraventricular tachycardia*, *sudden cardiac death*, *ventricular pre-excitation with multiple accessory pathways*, *paroxysmal atrial fibrillation*, *stroke*, *cardiomyopathy* and *palpations* together with the pVAAST results shown in panel B were used by Phevor to identify *MYH6* as the top candidate.

found to be shared by all sequenced affected family members (NM_002471.3:c.5652C>T; p.Glu1884Lys), and confirmed by Sanger sequencing.

MYH6 is well known to be expressed in heart tissue and has previously been associated with heart septal defect phenotypes[158]. Although previously associated with several cardiomyopathies[159] and septal defects[158], MYH6 has not been previously associated with WPW. Follow-up genotyping of additionally affected and healthy individuals in the pedigrees demonstrated that the allele segregates in a Mendelian fashion.

Without Phevor, it would have been difficult to come to these conclusions. pVAAST was able to rank the *MYH6* allele properly as the top candidates, but with so many shared alleles, dominant inheritance and incomplete penetrance it would be difficult to connect the phenotype to the disease-gene. Successful molecular diagnosis of this pedigree adds another allele responsible for the WPW syndrome phenotype.

6.4.2 A Very Large Family Plagued with Early Onset Atrial Fibrillation

Figure 6.7A shows two branches of a very extended family plagued with early onset Atrial Fibrillation and other cardiomyopathies. Martin Tristani-Firouzi and colleagues identified these two families using the Utah Population Database. Interestingly, genealogy follow-up studies, again powered by the Utah Population Database have determined that these two families are both descended from of a couple that lived in the early 1800s. Neither family is aware of this fact, as informing them is disallowed, sadly, by the Institutional Review Board (IRB).

The Utah Population Database was able to provide phenotype information; thus

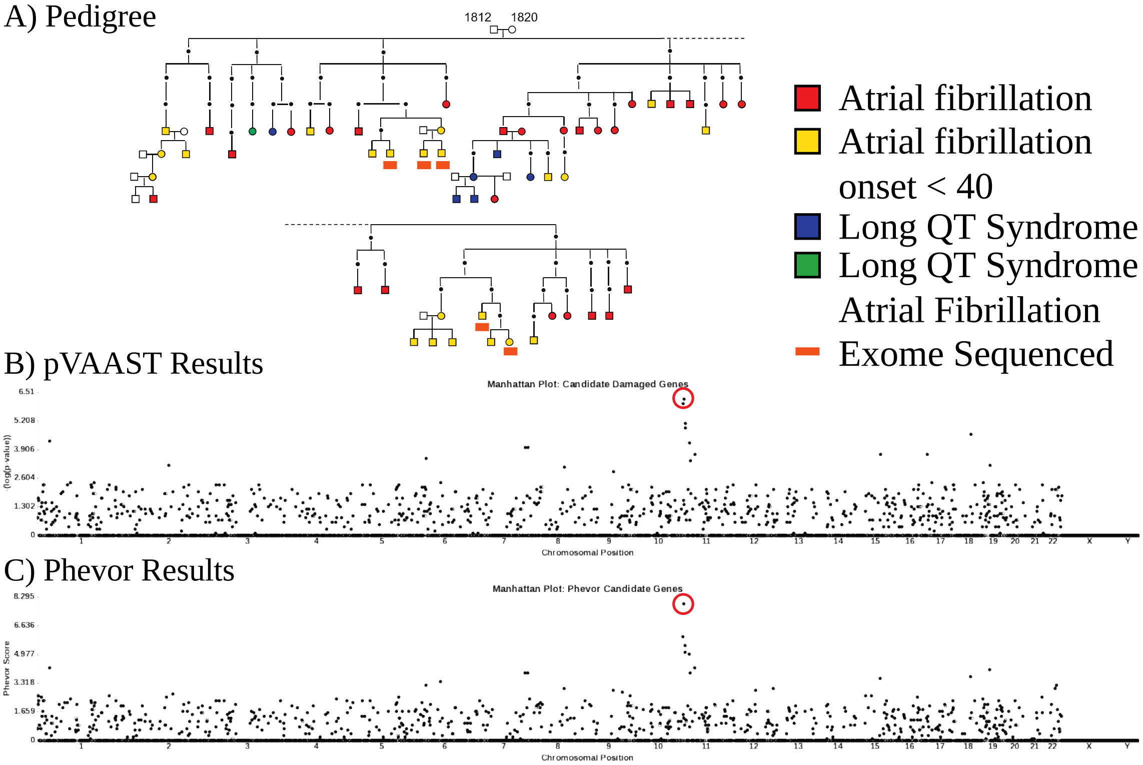


Figure 6.7. pVAAST and Phevor Results from a Pedigree with Atrial Fibrillation Manhattan plot, the pVAAST p-value or Phevor score plotted against each gene’s genomic location. A) pVAAST returns *KCNQ1* as the top candidate using a dominant inheritance model. B Phevor confirms *KCNQ1* as the top candidate.

related many individuals in the family can be identified as affected with: 1) atrial fibrillation, 2) early onset atrial fibrillation, 3) long QT syndrome or 4) a combination of long QT syndrome and atrial fibrillation. This information makes it obvious that atrial fibrillation is an inherited phenotype in this family (Figure 6.7A).

Exome sequencing was performed on five individuals in two portions of the tree, separated by seven generations. Using these data, pVAAST identified a single, very strong candidate. Interestingly, several other candidate genes that received genome-wide significance were found to neighbor the top candidate on the same chromosome (Figure 6.7B).

Utilizing the phenotype *atrial fibrillation* Phevor was used to further delineate the genome-wide significant genes and to identify a single gene responsible for the phenotype: *KCNQ1* (Figure 6.7C). From this pedigree, a molecular diagnosis for atrial fibrillation was established for this variant (NM_000218.2:c.692G>A; p.Arg231His). Sanger sequencing confirmed the variant.

Variants in *KCNQ1* are well known to cause long QT syndrome, short QT syndrome and familial atrial fibrillation[145,160]. Additionally, prior to the completion of this analysis another family reportedly had the Arg231His mutation and atrial fibrillation phenotype[145]. It is possible too that this family has ties to the families in the Utah Population Database. Having the capability through the Utah Population Database to relate these two families together greatly increased the confidence that the disease-gene had been identified. Using Phevor to attach the phenotype connection further strengthened this confidence.

6.5 Conclusion

These results from the Utah Genome Project demonstrate the utility of Phevor for diagnosis. With only a single affected exome, Phevor diagnosed two patients that were misdiagnosed using locus-specific testing (*ABCB11*, *SCN5A*). Phevor was also able to identify a novel allele in a novel gene causing a known phenotype (*NFKB2*). It also identified an allele in a known disease-gene that produces a novel phenotype (*STAT1*). Finally, Phevor aided in the analysis of two large pedigrees (*MYH6*, *KCNQ1*). These results make it clear that Phevor is playing a valuable role in the UGP's analysis pipeline.

CHAPTER 7

WHAT IS NEXT FOR PHEVOR?

The benchmarking and case studies from the Utah Genome Project that I have presented demonstrate the utility of Phevor as a means to identify rare pathogenic alleles of the germline. However, I believe that this use-case scenario only scratches the surface of Phevor's potential. Going forward, obvious application areas for Phevor include 1) cancer; 2) means to bring environmental exposure information to bear on the problem of identification mutagen induced germline and somatic pathogenic alleles; 3) pharmacogenomics applications; 4) and risk assessment for common and complex pathogenic alleles.

7.1 Phevor and Cancer

Cancer is truly a personal disease; somatic and inherited mutations combine to create a mosaic of aberrant genotypes that can evade the host defenses and chemotherapy. Understanding the etiology of an individual's cancer can have a huge impact on their level of care. More and more cancer centers are turning to next generation sequencing to profile individual tumors in hope of targeting therapy. Unfortunately, oncology suffers from the same limitations encountered in clinical genomics. Tumor sequence analysis is being restricted to candidate genes and actionable mutations found in allele databases—missing many novel associations that could play a role in patient care.

The etiology of cancer is a very complex problem because of its progressive, continuously evolving nature. Cancer phenotypes are complex, reflecting the networks of interactions between genes and their products producing them. My results to date have made it clear that Phevor has an unparalleled ability to de-convolute such data for diagnosis. Going forward, I would like to explore Phevor's performance to identify cancer-driving networks and somatic mutations.

I envision two methods for utilizing Phevor to shed light on the personal etiology of cancer—personalizing both diagnosis and treatment. The first of these methods involves integration of cancer specific pathway data into Phevor—a project already underway in collaboration with the Eilbeck Lab and the Utah Genome Project. Even though the somatic mutations driving the tumorigenesis are diverse and personal in nature, their eventual convergence upon cell cycle control offers a common starting point for analysis, and I hope, will prove their Achilles heel.

The second extension to Phevor involves the use of RNA-Seq data for increased statistical power. RNA-Seq data are commonly used to identify up or down regulated genes in tumor samples. It is often not the case, however, that the differentially expressed genes are involved in tumorigenesis or progression; rather, their altered expression is the consequence of a mutation in an upstream regulator. In other words, somatic driver mutations do not necessarily alter the expression of genes they are located in, but do alter the expression of their downstream signaling pathway targets. Thus, the integration of cancer specific pathway data into Phevor will prove essential for this approach too. In this scenario, Phevor would combine somatic DNA variation data with RNA-Seq data in the context of cancer signaling pathways to identify candidate driver mutations in control genes that manifest themselves by altering the expression of their downstream regulatory targets.

7.2 Phevor and the Environment

Environmental exposure to chemicals, pollutants, allergens, even light, is well known to play a role in disease. Better means for making connections between environmental exposures and gene expression, gene function and disease phenotype are clearly needed. Phevor provides an obvious starting point for such a research endeavor. Because Phevor employs networks (pathways, ontologies, etc.) for diagnosis, it is reasonable to think Phevor could connect environmental exposure to genetic consequence, especially as regards interactions between gene products and chemical mutagens. A derivative of Phevor could be developed that relates environmental agents (e.g., pesticides, carcinogen, toxins, pollutants) to genes impacted by exposure. Patient phenotype and genotype data, together even with RNA-Seq and pathway data, could be employed in this application. I recently carried out a proof of concept of this approach as part of a collaboration with the Eilbeck and Camp groups. For this experiment, 15 genes linked to hepatocellular carcinoma and benzene exposure were individually used as seeds for Phevor propagation, in order to identify additional known and novel candidate genes connected to hepatocellular carcinoma and benzene exposure. Using as few as five known genes linked to this exposure related genetic disease, Phevor was able to return all known benzene processing genes ranked in the top 100 candidates genome-wide. These results were without any use of expression or genotypic information. Being able to connect genes that process benzene, leading to developing hepatocellular carcinoma, through Phevor propagation in this manner makes me very optimistic as to the potential of a full Phevor-based application capable of using RNA-Seq and pathway data for environmental exposure diagnostic applications.

7.3 Phevor and Pharmacology

Drug development is an expensive, complicated process of trial and error. Bringing a pharmacological agent to market requires hundreds of thousands of tests, screenings, clinical trials and millions in research dollars. When finally accepted as a working drug, the limitations and prescription use-case scenarios are often so narrow that the drug is only ever prescribed to a subtle subset of patients, across a very narrow spectrum of phenotypes. Connecting genes with their phenotype is Phevor's specialty. Clearly, a derivative of Phevor capable of identifying related phenotypes involving related genes and pathways, which might benefit from the drug, would be immensely useful. Expanding pharmaceutical use-case scenarios in this way might save millions in research and development costs by repurposing drugs already shown to be safe and effective. This Phevor derivative would help many with disorders that do not yet have effective pharmacological therapy.

7.4 Risk Assessment and Phevor

Using phenotype as an input, Phevor can connect genes using their related biological properties with the phenotype they would exhibit if damaged. Combining phenotype with a patient’s genotype, Phevor can accurately identify the pathogenic allele. Reversing this process—using genotype to predict the phenotype—is a logical extension of the Phevor algorithm. In other words, this “reverse” Phevor application would calculate a genetic burden on various biological systems (pathways, gene-families, GO functions, processes, etc.) to deduce their phenotypic manifestations, and the relative risk of these manifestations.

Recently the American College of Medical Genetics (ACMG) has released guidelines on reporting incidental findings from many known disease-genes arising from genome and exome sequencing. ACMG’s intentions are to provide an early warning for various life-threatening phenotypes. Using the spectrum of variants found during exome sequencing to predict a phenotype burden using reverse-Phevor would provide an automated means for this process. For example, a high variant burden in biological systems that regulate heart rhythm could be used to identify patients likely to develop cardiovascular disease as they grow older. Likewise, a high variant burden in biological systems regulating insulin or sugar metabolism could be used for diabetes prognosis and relative risk calculations.

7.5 My Conclusions

My dissertation work has centered upon the exploration of algorithmic methods for exploring the interplay between genotype and phenotype. The work described in this dissertation has moved far beyond traditional heritability-based approaches. I hope that my work has convincingly demonstrated the ability of next generation sequencing, when combined with the right software, to shed real light on these phenomena, and to produce practical applications for translational medicine.

REFERENCES

1. Cumming, D.R., Furber, S.B., and Paul, D.J. (2014). Beyond Moore's law. Philosophical transactions Series A, Mathematical, physical, and engineering sciences 372, 20130376.

2. Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. Nature 470, 198-203.

3. Metzker, M.L. (2010). Sequencing technologies - the next generation. Nature Reviews Genetics 11, 31-46.

4. (N.A), (2010). The human genome at ten. Nature 464, 649-650.

5. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

6. Saunders, C.J., Miller, N.A., Soden, S.E., Dinwiddie, D.L., Noll, A., Alnadi, N.A., Andraws, N., Patterson, M.L., Krivohlavek, L.A., Fellis, J., et al. (2012). Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Science Translational Medicine 4, 154ra135.

7. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. The New England Journal of Medicine 369, 1502-1511.

8. Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94, 441-448.

9. Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature 321, 674-679.

10. Kaufman, S., Max, E.E., and Kang, E.S. (1975). Phenylalanine hydroxylase activity in liver biopsies from hyperphenylalaninemia heterozygotes: deviation from proportionality with gene dosage. Pediatric Research 9, 632-634.

11. Fölling, A. (1934). Über Ausscheidung von Phenylbrenztraubensäure in den Harn

als Stoffwechselanomalie in Verbindung mit Imbezillität. In Hoppe-Seyler´s Zeitschrift für physiologische Chemie. p 169.

12. Afzelius, B.A., Eliasson, R., Johnsen, O., and Lindholmer, C. (1975). Lack of dynein arms in immotile human spermatozoa. The Journal of Cell Biology 66, 225-232.

13. Liechti-Gallati, S., and Kraemer, R. (1995). Cystic fibrosis mutations and immotile cilia syndrome. Clinical Genetics 47, 328-329.

14. (1977). Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Case 26-1977. The New England Journal of Medicine 296, 1519-1526.

15. Bennett, S. (2004). Solexa Ltd. Pharmacogenomics 5, 433-438.

16. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20, 1297-1303.

17. Rehm, H.L., Bale, S.J., Bayrak-Toydemir, P., Berg, J.S., Brown, K.K., Deignan, J.L., Friez, M.J., Funke, B.H., Hegde, M.R., Lyon, E., et al. (2013). ACMG clinical laboratory standards for next-generation sequencing. Genetics in Medicine: Official Journal of the American College of Medical Genetics 15, 733-747.

18. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Human Mutation 21, 577-581.

19. Turner, E.H., Ng, S.B., Nickerson, D.A., and Shendure, J. (2009). Methods for genomic partitioning. Annual Review of Genomics and Human Genetics 10, 263-284.

20. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272-276.

21. Fuentes Fajardo, K.V., Adams, D., Program, N.C.S., Mason, C.E., Sincan, M., Tifft, C., Toro, C., Boerkoel, C.F., Gahl, W., and Markello, T. (2012). Detecting false-positive signals in exome sequencing. Human Mutation 33, 609-613.

22. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Medicine 5, 28.

23. Leonard, H., and Wen, X. (2002). The epidemiology of mental retardation: challenges and opportunities in the new millennium. Mental Retardation and Developmental Disabilities Research Reviews 8, 117-134.

24. Ropers, H.H. (2010). Genetics of early onset cognitive impairment. Annual Review of Genomics and Human Genetics 11, 161-187.

25. Shaffer, L.G., American College of Medical Genetics Professional, P., and Guidelines, C. (2005). American College of Medical Genetics guideline on the cytogenetic evaluation of the individual with developmental delay or mental retardation. Genetics in Medicine: Official Journal of the American College of Medical Genetics 7, 650-654.

26. Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86, 749-764.

27. Petit, C. (1996). Genes responsible for human hereditary deafness: symphony of a thousand. Nature Genetics 14, 385-391.

28. Scharenberg, A.M., Hannibal, M.C., Torgerson, T., Ochs, H.D., and Rawlings, D.J. (1993). Common variable immune deficiency overview. In GeneReviews(R), R.A. Pagon, M.P. Adam, H.H. Ardinger, T.D. Bird, C.R. Dolan, C.T. Fong, R.J.H. Smith, and K. Stephens, eds. (Seattle (WA)).

29. Hauck, F.R., Tanabe, K.O., and Moon, R.Y. (2011). Racial and ethnic disparities in infant mortality. Seminars in Perinatology 35, 209-220.

30. Lynberg, M.C., and Khoury, M.J. (1990). Contribution of birth defects to infant mortality among racial/ethnic minority groups, United States, 1983. MMWR CDC Surveillance Summaries: Morbidity and Mortality Weekly Report CDC Surveillance Summaries / Centers for Disease Control 39, 1-12.

31. Kochanek, K.D., Kirmeyer, S.E., Martin, J.A., Strobino, D.M., and Guyer, B. (2012). Annual summary of vital statistics: 2009. Pediatrics 129, 338-348.

32. McKusick-Nathans Institute of Genetic Medicine, J.H.U. (2014). Online Mendelian Inheritance in Man, OMIM®. (Baltimore, MD).

33. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research 42, D980-985.

34. Consortium, E.P., Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the

human genome. Nature 489, 57-74.

35. Singleton, M.V., Guthery, S.L., Voelkerding, K.V., Chen, K., Kennedy, B., Margraf, R.L., Durtschi, J., Eilbeck, K., Reese, M.G., Jorde, L.B., et al. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet 94, 599-610.

36. Mayo, B.J., Klebe, R.J., Barnett, D.R., Lankford, B.J., and Bowman, B.H. (1980). Somatic cell genetic studies of the cystic fibrosis mucociliary inhibitor. Clinical Genetics 18, 379-386.

37. (N.A), (1994). Population variation of common cystic fibrosis mutations. The Cystic Fibrosis Genetic Analysis Consortium. Human Mutation 4, 167-177.

38. (N.A), (2014). Locus Specific Mutation Database. In. (Carlton, Australia, Human Genome Variance Society.

39. Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Research 11, 863-874.

40. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature Methods 7, 248-249.

41. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38, e164.

42. Hu, H., Huff, C.D., Moore, B., Flygare, S., Reese, M.G., and Yandell, M. (2013). VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genetic Epidemiology 37, 622-634.

43. Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L.B., and Reese, M.G. (2011). A probabilistic disease-gene finder for personal genomes. Genome Research 21, 1529-1542.

44. Clark, M.J., Chen, R., and Snyder, M. (2013). Exome sequencing by targeted enrichment. Current Protocols in Molecular Biology edited by Frederick M Ausubel [et al.] Chapter 7, Unit7 12.

45. Fullwood, M.J., Wei, C.L., Liu, E.T., and Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Research 19, 521-532.

46. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant

call format and VCFtools. Bioinformatics 27, 2156-2158.

47. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Research 42, D756-763.

48. Choi, J.Y., Muallem, D., Kiselyov, K., Lee, M.G., Thomas, P.J., and Muallem, S. (2001). Aberrant CFTR-dependent HCO3- transport in mutations associated with cystic fibrosis. Nature 410, 94-97.

49. Afzelius, B.A. (1976). A human syndrome caused by immotile cilia. Science 193, 317-319.

50. El Zein, L., Omran, H., and Bouvagnet, P. (2003). Lateralization defects and ciliary dyskinesia: lessons from algae. Trends in Genetics: TIG 19, 162-167.

51. Carlen, B., and Stenram, U. (2005). Primary ciliary dyskinesia: a review. Ultrastructural Pathology 29, 217-220.

52. Seidman, J.G., and Seidman, C. (2001). The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. Cell 104, 557-567.

53. Gonzalez, K.D., Li, X., Lu, H.M., Lu, H., Pellegrino, J.E., Miller, R.T., Zeng, W., and Chao, E.C. (2014). Diagnostic exome sequencing and tailored bioinformatics of the parents of a deceased child with cobalamin deficiency suggests digenic inheritance of the MTR and LMBRD1 Genes. JIMD Reports.

54. Whisstock, J.C., and Lesk, A.M. (2003). Prediction of protein function from protein sequence and structure. Quarterly Reviews of Biophysics 36, 307-340.

55. Vingron, M., and Waterman, M.S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. Journal of Molecular Biology 235, 1-12.

56. Wrabl, J.O., and Grishin, N.V. (2004). Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. Proteins 54, 71-87.

57. Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345-351.

58. Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America 89, 10915-10919.

59. Henikoff, S., and Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. Proteins 17, 49-61.

60. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4, 1073-1081.

61. Cooper, G.M., Stone, E.A., Asimenos, G., Program, N.C.S., Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Research 15, 901-913.

62. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research 20, 110-121.

63. Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., et al. (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Research 42, D764-770.

64. Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289-1303.

65. Hawks, J., Wang, E.T., Cochran, G.M., Harpending, H.C., and Moyzis, R.K. (2007). Recent acceleration of human adaptive evolution. Proceedings of the National Academy of Sciences of the United States of America 104, 20753-20758.

66. Huelsenbeck, J.P., and Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276, 227-232.

67. Cai, T., Lin, X., and Carroll, R.J. (2012). Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. Biostatistics 13, 776-790.

68. Pertea, M., Pertea, G.M., and Salzberg, S.L. (2011). Detection of lineage-specific evolutionary changes among primate species. BMC Bioinformatics 12, 274.

69. Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M., et al. (2014). Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet 94, 223-232.

70. Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

71. Norton, N., Robertson, P.D., Rieder, M.J., Zuchner, S., Rampersaud, E., Martin, E., Li, D., Nickerson, D.A., Hershberger, R.E., National Heart, L., et al. (2012). Evaluating pathogenicity of rare variants from dilated cardiomyopathy in the exome era. Circulation Cardiovascular Genetics 5, 167-174.

72. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

73. (N.A), (2014). Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), (Seattle, WA).

74. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823-828.

75. Cutting, G.R. (2014). Annotating DNA variants is the next major goal for human genetics. Am J Hum Genet 94, 5-10.

76. McInerney-Leo, A.M., Marshall, M.S., Gardiner, B., Benn, D.E., McFarlane, J., Robinson, B.G., Brown, M.A., Leo, P.J., Clifton-Bligh, R.J., and Duncan, E.L. (2014). Whole exome sequencing is an efficient and sensitive method for detection of germline mutations in patients with phaeochromcytomas and paragangliomas. Clinical Endocrinology 80, 25-33.

77. de Jesus Perez, V.A., Yuan, K., Lyuksyutova, M.A., Dewey, F., Orcholski, M.E., Shuffle, E.M., Mathur, M., Yancy, L., Jr., Rojas, V., Li, C.G., et al. (2014). Whole Exome Sequencing Reveals Topbp1 as a Novel Gene in Idiopathic Pulmonary Arterial Hypertension. American Journal of Respiratory and Critical Care Medicine.

78. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29, 308-311.

79. Dudbridge, F., and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. Genetic Epidemiology 32, 227-234.

80. Moskvina, V., Schmidt, K.M., Vedernikov, A., Owen, M.J., Craddock, N., Holmans, P., and O'Donovan, M.C. (2012). Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. European Journal of Human Genetics: EJHG 20, 890-896.

81. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 15, 1034-1050.

82. Baird, P.A., Anderson, T.W., Newcombe, H.B., and Lowry, R.B. (1988). Genetic disorders in children and young adults: a population study. Am J Hum Genet 42, 677-693.

83. Rath, A., Olry, A., Dhombres, F., Brandt, M.M., Urbero, B., and Ayme, S. (2012). Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Human Mutation 33, 803-808.

84. Jervell, A., and Lange-Nielsen, F. (1957). Congenital deaf-mutism, functional heart disease with prolongation of the Q-T interval and sudden death. American Heart Journal 54, 59-68.

85. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics 42, 30-35.

86. Labrou, Y., and Finin, T. (1999). Yahoo! as an ontology: using Yahoo! categories to describe documents. In Proceedings of the Eighth International Conference on Information and Knowledge Management. (Kansas City, Missouri, USA, ACM), pp 180-187.

87. Mcguinness, D.L. (2002). Ontologies Come of Age. In Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, eds. MIT Press.

88. Rubin, D.L., Shah, N.H., and Noy, N.F. (2008). Biomedical ontologies: a functional perspective. Briefings in Bioinformatics 9, 75-90.

89. Hayes-Roth, F., Waterman, D.A., and Lenat, D.B. (1983). Building Expert Systems. Addison-Wesley Longman Publishing Co., Inc.

90. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32, D115-119.

91. Gruber, T.R. (1993). A translation approach to portable ontology specifications. Knowl Acquis 5, 199-220.

92. Christofides, N. (1975). Graph Theory: An Algorithmic Approach, Computer Science and Applied Mathematics. (Academic Press, Inc.)

93. Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology 6, R44.

94. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25, 25-29.

95. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg,

L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25, 1251-1255.

96. Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., et al. (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Research 37, D555-559.

97. Kohler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J., et al. (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research 42, D966-974.

98. Gkoutos, G.V., Schofield, P.N., and Hoehndorf, R. (2012). Chapter Four - The Neurobehavior Ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. In International Review of Neurobiology, J.C. Elissa and A.H. Melissa, eds. (Academic Press), pp 69-87.

99. Petri, V., Jayaraman, P., Tutaj, M., Hayman, G.T., Smith, J.R., De Pons, J., Laulederkind, S.J., Lowry, T.F., Nigam, R., Wang, S.J., et al. (2014). The Pathway Ontology - updates and applications. Journal of Biomedical Semantics 5, 7.

100. Smith, C.L., and Eppig, J.T. (2009). The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews Systems Biology and Medicine 1, 390-399.

101. Wang, S.J., Laulederkind, S.J., Hayman, G.T., Smith, J.R., Petri, V., Lowry, T.F., Nigam, R., Dwinell, M.R., Worthey, E.A., Munzenmaier, D.H., et al. (2013). Analysis of disease-associated objects at the Rat Genome Database. Database: the Journal of Biological Databases and Curation 2013, bat046.

102. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research 40, D940-946.

103. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., et al. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Research 41, D456-463.

104. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet 85, 457-464.

105. Green, M.L., and Karp, P.D. (2006). The outcomes of pathway database

computations depend on pathway ontology. Nucleic Acids Research 34, 3687-3697.

106. de Matos, P., Adams, N., Hastings, J., Moreno, P., and Steinbeck, C. (2012). A database for chemical proteomics: ChEBI. Methods in Molecular Biology 803, 273-296.

107. Brown, S.D., and Moore, M.W. (2012). Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. Disease Models & Mechanisms 5, 289-292.

108. Wright, E.S., Yilmaz, L.S., and Noguera, D.R. (2012). DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Applied and Environmental Microbiology 78, 717-725.

109. Plauchu, H., de Chadarevian, J.P., Bideau, A., and Robert, J.M. (1989). Age-related clinical profile of hereditary hemorrhagic telangiectasia in an epidemiologically recruited population. American Journal of Medical Genetics 32, 291-297.

110. Porteous, M.E., Burn, J., and Proctor, S.J. (1992). Hereditary haemorrhagic telangiectasia: a clinical analysis. Journal of Medical Genetics 29, 527-530.

111. Shovlin, C.L., Guttmacher, A.E., Buscarini, E., Faughnan, M.E., Hyland, R.H., Westermann, C.J., Kjeldsen, A.D., and Plauchu, H. (2000). Diagnostic criteria for hereditary hemorrhagic telangiectasia (Rendu-Osler-Weber syndrome). American Journal of Medical Genetics 91, 66-67.

112. McAllister, K.A., Grogg, K.M., Johnson, D.W., Gallione, C.J., Baldwin, M.A., Jackson, C.E., Helmbold, E.A., Markel, D.S., McKinnon, W.C., Murrell, J., et al. (1994). Endoglin, a TGF-beta binding protein of endothelial cells, is the gene for hereditary haemorrhagic telangiectasia type 1. Nature Genetics 8, 345-351.

113. Johnson, D.W., Berg, J.N., Baldwin, M.A., Gallione, C.J., Marondel, I., Yoon, S.J., Stenzel, T.T., Speer, M., Pericak-Vance, M.A., Diamond, A., et al. (1996). Mutations in the activin receptor-like kinase 1 gene in hereditary haemorrhagic telangiectasia type 2. Nature Genetics 13, 189-195.

114. Gallione, C.J., Richards, J.A., Letteboer, T.G., Rushlow, D., Prigoda, N.L., Leedom, T.P., Ganguly, A., Castells, A., Ploos van Amstel, J.K., Westermann, C.J., et al. (2006). SMAD4 mutations found in unselected HHT patients. Journal of Medical Genetics 43, 793-797.

115. Jeffery, S., Saggar-Malik, A.K., Economides, D.L., Blackmore, S.E., and MacDermot, K.D. (1998). Apparent normalisation of fetal renal size in autosomal dominant polycystic kidney disease (PKD1). Clinical Genetics 53, 303-307.

116. Korn-Lubetzki, I., Argov, Z., Raas-Rothschild, A., Wirguin, I., and Steiner, I. (2002). Family with inflammatory demyelinating polyneuropathy and the HNPP 17p12 deletion. American Journal of Medical Genetics 113, 275-278.

117. Giroud, M., Mousson, C., Chalopin, J.M., Rifle, G., and Dumas, R. (1990). Miller-Fisher syndrome and pontine abnormalities on MRI: a case report. Journal of Neurology 237, 489-490.

118. Robinson, P.N., Kohler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. American Journal of Human Genetics 83, 610-615.

119. Smith, C.L., and Eppig, J.T. (2012). The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mamm Genome 23, 653-668.

120. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W.A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research 40, D940-946.

121. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25, 25-29.

122. Kohler, S., Bauer, S., Mungall, C.J., Carletti, G., Smith, C.L., Schofield, P., Gkoutos, G.V., and Robinson, P.N. (2011). Improving ontologies by automatic reasoning and evaluation of logical definitions. BMC Bioinformatics 12, 418.

123. Gelman, A. (2004). Bayesian Data Analysis. (Boca Raton, FL.: Chapman & Hall/CRC)

124. Ohishi, K., Inoue, N., and Kinoshita, T. (2001). PIG-S and PIG-T, essential for GPI anchor attachment to proteins, form a complex with GAA1 and GPI8. The EMBO Journal 20, 4088-4098.

125. Kvarnung, M., Nilsson, D., Lindstrand, A., Korenke, G.C., Chiang, S.C., Blennow, E., Bergmann, M., Stodberg, T., Makitie, O., Anderlid, B.M., et al. (2013). A novel intellectual disability syndrome caused by GPI anchor deficiency due to homozygous mutations in PIGT. Journal of Medical Genetics 50, 521-528.

126. Krawitz, P.M., Hochsmann, B., Murakami, Y., Teubner, B., Kruger, U., Klopocki, E., Neitzel, H., Hoellein, A., Schneider, C., Parkhomchuk, D., et al. (2013). A case of paroxysmal nocturnal hemoglobinuria caused by a germline mutation and a somatic mutation in PIGT. Blood 122, 1312-1315.

127. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Analysis of

genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636-639.

128. Yang, Y., Muzny, D.M., Reid, J.G., Bainbridge, M.N., Willis, A., Ward, P.A., Braxton, A., Beuten, J., Xia, F., Niu, Z., et al. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. The New England Journal of Medicine 369, 1502-1511.

129. Boycott, K.M., Vanstone, M.R., Bulman, D.E., and MacKenzie, A.E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics 14, 681-691.

130. Robinson, P.N., Kohler, S., Oellrich, A., Sanger Mouse Genetics, P., Wang, K., Mungall, C.J., Lewis, S.E., Washington, N., Bauer, S., Seelow, D., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Research 24, 340-348.

131. Cooper, D.N., Ball, E.V., and Krawczak, M. (1998). The human gene mutation database. Nucleic Acids Res 26, 285-287.

132. Saunders, C.J., Miller, N.A., Soden, S.E., Dinwiddie, D.L., Noll, A., Alnadi, N.A., Andraws, N., Patterson, M.L., Krivohlavek, L.A., Fellis, J., et al. (2012). Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med 4, 154ra135.

133. Robinson, P., Kohler, S., Oellrich, A., Wang, K., Mungall, C., Lewis, S.E., Washington, N., Bauer, S., Seelow, D.S., Krawitz, P., et al. (2013). Improved exome prioritization of disease genes through cross species phenotype comparison. Genome Research.

134. Yandell, M., Moore, B., Salas, F., Mungall, C., MacBride, A., White, C., and Reese, M.G. (2008). Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. PLoS Computational Biology 4, e1000218.

135. Albertsen, H.M., Smith, S.A., Mazoyer, S., Fujimoto, E., Stevens, J., Williams, B., Rodriguez, P., Cropp, C.S., Slijepcevic, P., Carlson, M., et al. (1994). A physical map and candidate genes in the BRCA1 region on chromosome 17q12-21. Nature Genetics 7, 472-479.

136. Sanguinetti, M.C. (2014). HERG1 channel agonists and cardiac arrhythmia. Current opinion in Pharmacology 15, 22-27.

137. Elliott, C.G. (2013). Genetics of pulmonary arterial hypertension. Clinics in Chest Medicine 34, 651-663.

138. Frank, T.S., Deffenbaugh, A.M., Reid, J.E., Hulick, M., Ward, B.E., Lingenfelter, B., Gumpper, K.L., Scholl, T., Tavtigian, S.V., Pruss, D.R., et al. (2002).

Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 20, 1480-1490.

139. Coonrod, E.M., Margraf, R.L., Russell, A., Voelkerding, K.V., and Reese, M.G. (2013). Clinical analysis of genome next-generation sequencing data using the Omicia platform. Expert Review of Molecular Diagnostics 13, 529-540.

140. Hu, H., Roach, J.C., Coon, H., Guthery, S.L., Voelkerding, K.V., Margraf, R.L., Durtschi, J.D., Tavtigian, S.V., Shankaracharya, Wu, W., et al. (2014). A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data. Nat Biotech, advance online publication.

141. Alonso, E.M., Snover, D.C., Montag, A., Freese, D.K., and Whitington, P.F. (1994). Histologic pathology of the liver in progressive familial intrahepatic cholestasis. Journal of Pediatric Gastroenterology and Nutrition 18, 128-133.

142. Whitington, P.F., Freese, D.K., Alonso, E.M., Schwarzenberg, S.J., and Sharp, H.L. (1994). Clinical and biochemical findings in progressive familial intrahepatic cholestasis. Journal of Pediatric Gastroenterology and Nutrition 18, 134-141.

143. Strautnieks, S.S., Bull, L.N., Knisely, A.S., Kocoshis, S.A., Dahl, N., Arnell, H., Sokal, E., Dahan, K., Childs, S., Ling, V., et al. (1998). A gene encoding a liver-specific ABC transporter is mutated in progressive familial intrahepatic cholestasis. Nature Genetics 20, 233-238.

144. Benson, D.W., Wang, D.W., Dyment, M., Knilans, T.K., Fish, F.A., Strieper, M.J., Rhodes, T.H., and George, A.L., Jr. (2003). Congenital sick sinus syndrome caused by recessive mutations in the cardiac sodium channel gene (SCN5A). The Journal of Clinical Investigation 112, 1019-1028.

145. Johnson, J.N., Tester, D.J., Perry, J., Salisbury, B.A., Reed, C.R., and Ackerman, M.J. (2008). Prevalence of early-onset atrial fibrillation in congenital long QT syndrome. Heart Rhythm: The Official Journal of the Heart Rhythm Society 5, 704-709.

146. Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., Albrecht, B., Bartholdi, D., Beygo, J., Di Donato, N., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380, 1674-1682.

147. Chen, K., Coonrod, E.M., Kumanovics, A., Franks, Z.F., Durtschi, J.D., Margraf, R.L., Wu, W., Heikal, N.M., Augustine, N.H., Ridge, P.G., et al. (2013). Germline mutations in NFKB2 implicate the noncanonical NF-kappaB pathway in the pathogenesis of common variable immunodeficiency. Am J Hum Genet 93, 812-824.

148. d'Hennezel, E., Bin Dhuban, K., Torgerson, T., and Piccirillo, C.A. (2012). The immunogenetics of immune dysregulation, polyendocrinopathy, enteropathy, X linked (IPEX) syndrome. Journal of Medical Genetics 49, 291-302.

149. Baud, O., Goulet, O., Canioni, D., Le Deist, F., Radford, I., Rieu, D., Dupuis-Girod, S., Cerf-Bensussan, N., Cavazzana-Calvo, M., Brousse, N., et al. (2001). Treatment of the immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) by allogeneic bone marrow transplantation. The New England Journal of Medicine 344, 1758-1762.

150. Sharfe, N., Dadi, H.K., Shahar, M., and Roifman, C.M. (1997). Human immune disorder arising from mutation of the alpha chain of the interleukin-2 receptor. Proceedings of the National Academy of Sciences of the United States of America 94, 3168-3171.

151. Dupuis, S., Dargemont, C., Fieschi, C., Thomassin, N., Rosenzweig, S., Harris, J., Holland, S.M., Schreiber, R.D., and Casanova, J.L. (2001). Impairment of mycobacterial but not viral immunity by a germline human STAT1 mutation. Science 293, 300-303.

152. Liu, L., Okada, S., Kong, X.F., Kreins, A.Y., Cypowyj, S., Abhyankar, A., Toubiana, J., Itan, Y., Audry, M., Nitschke, P., et al. (2011). Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. The Journal of Experimental Medicine 208, 1635-1648.

153. Uzel, G., Sampaio, E.P., Lawrence, M.G., Hsu, A.P., Hackett, M., Dorsey, M.J., Noel, R.J., Verbsky, J.W., Freeman, A.F., Janssen, E., et al. (2013). Dominant gain-of-function STAT1 mutations in FOXP3 wild-type immune dysregulation-polyendocrinopathy-enteropathy-X-linked-like syndrome. The Journal of Allergy and Clinical Immunology 131, 1611-1623.

154. van de Veerdonk, F.L., Plantinga, T.S., Hoischen, A., Smeekens, S.P., Joosten, L.A., Gilissen, C., Arts, P., Rosentul, D.C., Carmichael, A.J., Smits-van der Graaf, C.A., et al. (2011). STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis. The New England Journal of Medicine 365, 54-61.

155. DuVall, S.L., Fraser, A.M., Rowe, K., Thomas, A., and Mineau, G.P. (2012). Evaluation of record linkage between a large healthcare provider and the Utah Population Database. Journal of the American Medical Informatics Association: JAMIA 19, e54-59.

156. Gollob, M.H., Green, M.S., Tang, A.S., Gollob, T., Karibe, A., Ali Hassan, A.S., Ahmad, F., Lozado, R., Shah, G., Fananapazir, L., et al. (2001). Identification of a gene responsible for familial Wolff-Parkinson-White syndrome. The New England Journal of Medicine 344, 1823-1831.

157. Harnischfeger, W.W. (1959). Hereditary occurrence of the pre-excitation (Wolff-

Parkinson-White) syndrome with re-entry mechanism and concealed conduction. Circulation 19, 28-40.

158. Ching, Y.H., Ghosh, T.K., Cross, S.J., Packham, E.A., Honeyman, L., Loughna, S., Robinson, T.E., Dearlove, A.M., Ribas, G., Bonser, A.J., et al. (2005). Mutation in myosin heavy chain 6 causes atrial septal defect. Nature Genetics 37, 423-428.

159. Carniel, E., Taylor, M.R., Sinagra, G., Di Lenarda, A., Ku, L., Fain, P.R., Boucek, M.M., Cavanaugh, J., Miocic, S., Slavov, D., et al. (2005). Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. Circulation 112, 54-59.

160. Chen, Y.H., Xu, S.J., Bendahhou, S., Wang, X.L., Wang, Y., Xu, W.Y., Jin, H.W., Sun, H., Su, X.Y., Zhuang, Q.N., et al. (2003). KCNQ1 gain-of-function mutation in familial atrial fibrillation. Science 299, 251-254.