Computational methods for predicting and validating the causes of Mendelian disease

by

Orion J. Buske

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Computational methods for predicting and validating the causes of Mendelian disease

Orion J. Buske
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2016

We still do not know the genetic basis of roughly half of the estimated 7,000 Mendelian diseases. For some diseases, the responsible variants will not be discovered with standard genetic sequencing. If the variant falls outside of the exome, occurs in a repetitive region, or is a larger structural change, it is unlikely to be found by whole exome sequencing. For other diseases, the variant might be found, but the association with the disease has not yet been discovered. In these cases, making such a discovery requires several steps. Computational approaches are necessary to accurately prioritize harmful variants based on available knowledge. Then, additional information is needed to substantiate the association, such as functional tests, animal models, or identifying unrelated families with the same variant. This thesis presents several contributions to help researchers determine the genetic basis of unsolved Mendelian diseases:

First, a method was developed that improves variant prioritization for a class of variants that are usually ignored by analysis pipelines: synonymous variants. After curating known examples from the literature, machine learning methods were trained to prioritize these variants based on a set of designed features.

Second, finding additional families is a substantial hurdle in rare disease research. By collecting detailed phenotype information, computational methods can be used to find patients with a similar presentation. This leads to improved variant prioritization by combining sequencing data from several similar patients, without ever needing to explicitly define cohorts. The power of matchmaking methods grows exponentially with the size of the database, but simulations suggest that several hundred thousand cases are needed to identify the genetic basis of most Mendelian diseases.

Finally, these matchmaking algorithms are implemented in a web portal, PhenomeCentral, which is used by several consortia and hundreds of clinicians and researchers. While this platform is a repository of several thousand undiagnosed cases, matchmaking between platforms is critical to achieve the numbers of cases predicted to be necessary. Towards this end, the Matchmaker Exchange (MME) was established and an API developed. Case profiles are exchanged within a secure federated network to reduce the time for researchers to validate genetic hypotheses.

Acknowledgements

This thesis owes so much to so many.

Principal among them, Michael Brudno, for his years of guidance, encouragement, and support. Not many can claim to have an advisor so dedicated to his students; pulling all-nighters for his students' conference deadlines, and welcoming the entire lab into his home for events. The seed of every project in this thesis was planted by him. It has been a pleasure and an honor. Along with Anna Goldenberg and Gary Bader, my committee has been more gracious and insightful than I ever could have hoped. They will serve as scientific role models for the rest of my career. Shamil Sunyaev, Quaid Morris, and Stephen Meyn provided valuable feedback and delightful discussions.

This work would not have been possible without the support of the University of Toronto, the province of Ontario, the Hospital for Sick Children's Research Training Centre and Garron Family Cancer Centre, and the Canadian taxpayers that ultimately funded this work.

My path leading to this thesis has been punctuated by a series of truly extraordinary mentors. Among them, Mike Cantlon, Randy James, Stuart Reges, Martin Tompa, and Michael Hoffman. They have been my support vectors, bending my academic trajectory with the gravity of their love for their craft.

While each paper acknowledges the individuals involved, I must elaborate on several of these contributions. While I have had the honor and pleasure of presenting the PhenomeCentral work, it is the result of the efforts of an incredible number of people, including the continuous assistance of the PhenoTips team, the CCM team, and our genetic counselors. Special thanks are due to Damian Smedley and Jules Jacobsen, whose efforts were not properly recognized at the time. Thanks to Peter Robinson for his generosity, both personal and scientific, and his constant support of our efforts. We rely so heavily on his labor of love, the HPO. Finally, in the Matchmaker Exchange, I depended on the assistance and keen insights of Francois Schiettecatte, Ben Hutton, Tudor Groza, and Anthony Brookes, and conversations with Kym Boycott and Anthony Philippakis never failed to provide clarity in times of uncertainty.

I am fortunate to have been joined in this journey by a contingent of such wonderful colleagues and friends, including Marc Fiume, Misko Dzamba, Ladislav Rampasek, Yulia Rubanova, Andrei Turinsky, Lee Zamparo, Joanna Drummond, and many others. Their brilliance and kindness humble me. I am deeply indebted to Marta Girdea and Sergiu Dumitriu, the modest masterminds writing code behind the scenes. Much of this work was started by them, and any successes are attributable to their efforts.

Kym Boycott, Taila Hartley, Chandree Beaulieu, and Kristin Kernohan are the inspiration and deep motivation for almost all of this work. I can only hope it helps them in some small way.

Lastly, my friends and family have kept my spirits up and my feet down as I followed this path. I am nothing without them, and this thesis is theirs as much as it is mine. Mariel, who has filled my life with art and reassured me through the peaks and valleys of this endeavor. My parents, of whom I am merely an eclectic collage. Nature and nurture. I inherited my toes and humour from my father, my curls and spirit from my mother. Their curiosity and lenses of logic and pragmatism alongside all that DNA.

Thank you.

Relationship to published work

Parts of this thesis are adapted from published works written in collaboration with colleagues:

- Chapter 2 is adapted from the SilVA paper (Buske et al., 2013).
- Chapter 3 contains sections adapted from the PhenomeCentral paper (Buske et al., 2015a) and the genomic birthday paradox paper (Krawitz et al., 2015).
- Chapter 4 contains sections adapted from the PhenomeCentral paper (Buske et al., 2015a), the Matchmaker Exchange overview paper (Philippakis et al., 2015), and the Matchmaker Exchange API paper (Buske et al., 2015b).

Contents

Chapter 1

Genomics and health

1.1 Genomics

Genomics is the field of study focused on the characterization and interpretation of the collective genetic material — the genome — found in the cells of all living organisms. The genome of each organism is divided into separate molecules called chromosomes, like volumes in an encyclopedia. Most bacteria have a single, circular chromosome. Human cells, by contrast, have 23 pairs of linear chromosomes, with one copy of 23 chromosomes inherited from each parent.[1] Each chromosome is a long chain of DNA nucleotides, with four basic kinds — Adenine, Cytosine, Guanine, and Thiamine (A, C, G, and T). In total, a human genome is composed of over 3 billion nucleotides of DNA. The specific sequence of these nucleotides provides the blueprint for all the cells in your body.

The genome contains distinct regions, called *genes*, where the DNA sequence provides a template to create other molecules with biological functions, such as RNAs and proteins. The fraction of the genome that encodes for proteins varies widely between organisms, from 88% for the *E. coli* bacterium (Rogozin et al., 2002) and 70% for the yeast *S. cerevisiae* (Alexander et al., 2010), to just 1.2% for humans (Consortium et al., 2012). The regions of the genome that do not encode proteins fall into three main categories: *untranslated* regions at the ends of genes, *intronic* regions within genes, and *intergenic* regions between genes. Introns are rare in the genomes of bacteria and fungi, but account for almost half of the non-coding sequence in humans (Alexander et al., 2010). Human genes alternate between relatively short *exons*, which are included in the final functional molecule, and introns, which are not. Protein-coding genes are *expressed* in the following steps:

transcription: copying the gene sequence into a pre-messenger RNA (pre-mRNA) molecule

splicing: removing all of the introns from the pre-mRNA and joining the remaining exons together to form a mature messenger RNA (mRNA)

translation: creating a polypeptide chain by reading nucleotide triplets from the mRNA, finding the amino acid that uniquely corresponds to that triplet, and appending the amino acid to the growing polypeptide chain

[1] Human sex cells (sperm and eggs) have just one copy of each chromosome, and red blood cells and platelets do not contain any DNA.

folding: folding the polypeptide chain into the three dimensional conformation necessary for its function as a protein

For genes with multiple exons, not all exons are necessarily used to form the final mRNA that serves as the protein template. Some exons are always included, but others are only included in specific cell types or at specific times. By varying which exons are included, the same gene can encode hundreds or thousands of different protein sequences (Schmucker et al., 2000).

1.1.1 Genetic variation

As organisms live and reproduce, their genomes change. Changes that occur very early in embryo development and in sperm and egg cells can be inherited by the next generation. Genetic variation is introduced through several different sources:

- errors during DNA replication
- environmental DNA damage (e.g., exposure to radiation), either directly or through misrepair of a DNA lesion
- horizontal gene transfer (typically between bacteria)
- independent segregation (the random selection of one of each pair of chromosomes to create a sperm or egg)
- crossing over, which swaps similar regions between the two chromosomes in each pair

The Human Genome Project resulted in the creation of a human reference genome, a composite of the genomes of several anonymous people that is used as a point of comparison for new human genomes that are sequenced (Lander et al., 2001). Once the genomic sequences of multiple organisms are known, the differences between them and the collective variation across them can be studied. By comparing representative genomes of many related organisms, one can predict ancestral sequences and estimate the evolutionary constraint at various points in the genome (Henikoff & Henikoff, 1992; Cooper & Shendure, 2011). If a region has less variation than would be expected by chance, it is an indication that the region is functional and changes might be harmful. In fact, evolutionary constraint is one of the most powerful features for predicting how harmful a mutation might be (Cooper & Shendure, 2011; Buske et al., 2013).

There are several different kinds of genetic mutations that can occur, which differ in their biological significance and their ease of detection by different sequencing platforms.

SNVs: Single nucleotide variants (SNVs) are substitutions of one nucleotide for another one. *Coding* SNVs occur within the portion of genes that encode for a protein, and broadly fall into two categories: *non-synonymous* and *synonymous*, with the former resulting in a different expected protein sequence and the latter leaving the protein sequence intact. Non-synonymous mutations are further classified by whether they affect only a single amino acid (a *missense* SNV) or result in the entire protein sequence ending prematurely (a *nonsense* SNV).

In-dels: A sequence of one or more nucleotides can also be inserted into or deleted from a genome. Without knowing the ancestral genotype, it is impossible to determine whether a given difference between two genomes is an insertion in the first or a deletion from the second, so mutations of this

type are typically referred to as *indels* to reflect this ambiguity. Similar to SNVs, indels occurring in coding portions of genes are categorized according to their effect on the expected protein sequence. *Non-frameshift* indels result in an addition or deletion of one or more amino acids, but do not affect the amino acid sequence of the rest of the protein. *Frameshift* indels, however, disrupt the reading frame of the gene and result in a completely different sequence of amino acids after the indel (which often results in the protein sequence ending prematurely).

CNVs: Copy number variants (CNVs) occur when a large (typically >1,000 nucleotides) genomic sequence is deleted or duplicated one or more times. These might disrupt genes, delete genes, or result in additional copies of genes (potentially affecting the concentrations of the encoded proteins). CNVs are a class of *structural variation*, which includes inversions, translocations, chromothrypsis, and other large genomic rearrangements.

If a variant occurs on one of the two copies of a chromosome, it is said to be *heterozygous*, while if the same variant occurs on both copies, it is *homozygous*.

1.1.2 Genetic diseases

Genetic diseases are diseases caused by changes in the genetic sequence of an individual, and are estimated to affect over 8% of live human births (Baird et al., 1988). These changes may either be inherited, or happen *de novo* in a sperm, egg, or embryo. Over 7,000 diseases are so-called "Mendelian" because the traits follow a dominant or recessive pattern of inheritance and are caused by genetic variation in a single locus (Amberger et al., 2015). DNA changes can cause disease either through alteration of the encoded protein sequence, resulting in a protein that is non-functional (a loss of function mutation) or with altered functionality (a gain of function mutation), or through alteration of a regulatory region, resulting in disregulation and altered biological activity.

While genetic diseases can potentially be caused by variants anywhere in the genome, sequencing of the 1% of the genome in protein-coding exons — the *exome*[2] — currently results in a molecular diagnosis in 25–30% of cases of human Mendelian disease (Yang et al., 2014; Chong et al., 2015a; Lazaridis et al., 2016). This is a lower bound on the fraction of cases caused by variants in these regions. Studies such as Yang et al. (2014) exclude cases in which standard clinical tests and single-gene sequencing had already identified the cause. The FORGE consortium was able to identify a genetic cause of 55% of their rare disease cohorts using exome sequencing (Beaulieu et al., 2014). This remarkable enrichment of disease-causing variants within the exome has led to widespread use of targeted sequencing and a focus on developing methods for interpreting variants within this region.

Modes of inheritance

Genetic diseases have different patterns of inheritance depending on the biological mechanism of the disease (see Figure 1.1). Because every person has two copies of each chromosome, they also have two copies of every gene on those chromosomes[3]. For some diseases, a mutation in either one of the two copies of the gene will result in the individual having the condition. These conditions are *dominant*.

[2]The exome usually refers to the complete set of exons, including untranslated regions (UTRs), but many whole exome sequencing platforms to not target these regions (Chilamakuri et al., 2014).

[3]The sex chromosomes, X and Y, are quite different and only share genes in a few regions, called pseudo-autosomal regions. Men have only one copy of X and one copy of Y, and therefore have only one copy of most of the genes on these chromosomes.

Both copies of the gene must be functional for the person to be healthy. If a person only shows symptoms of the disease when both copies of the gene are disrupted, the condition is *recessive*. Having only one functional copy of the gene is enough to be healthy. If the gene is on the X or Y chromosomes, the disease is *sex-linked* and the prevalence will be different in females and males.

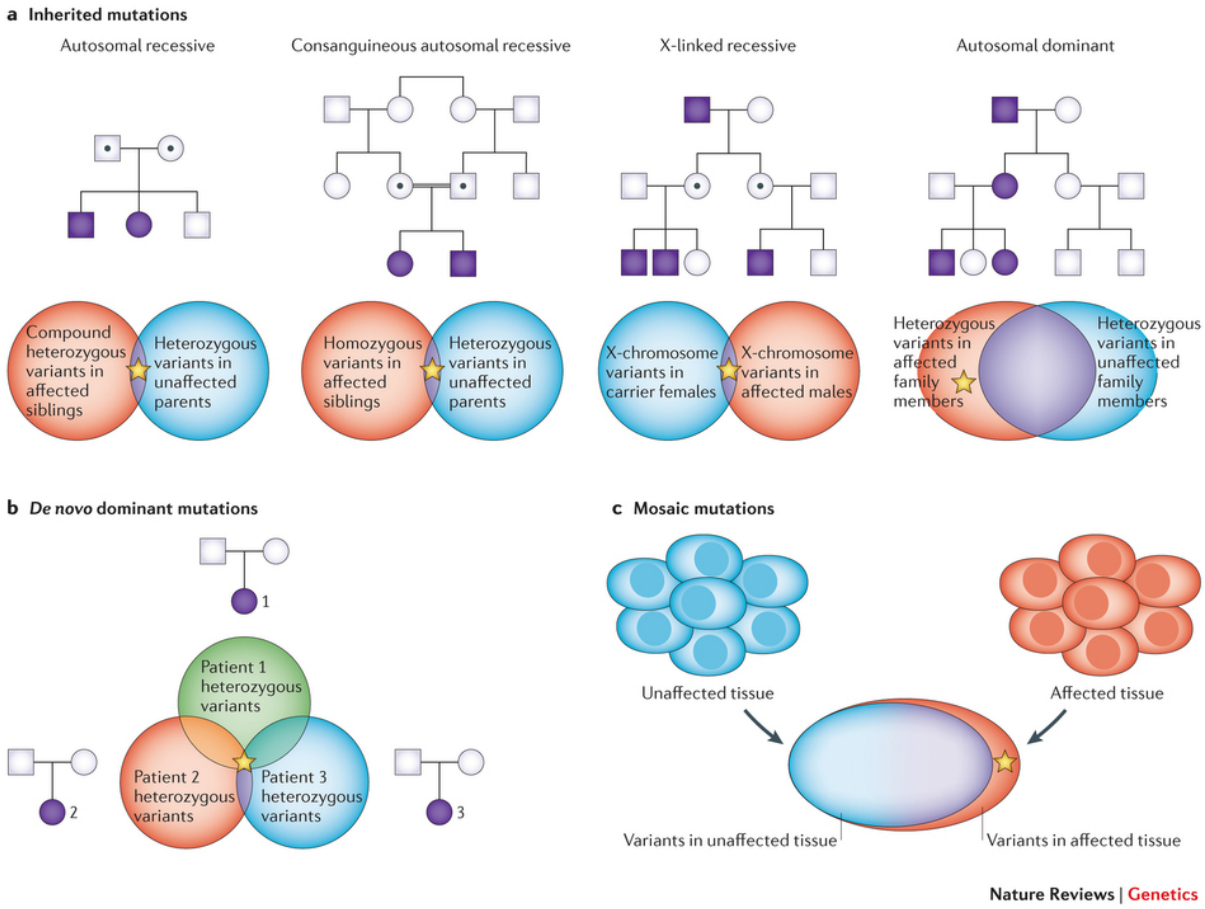

Figure 1.1: Patterns of inheritance for Mendelian diseases, and corresponding study design to determine the genetic cause. (Boycott et al., 2013)

1.2 DNA sequencing

Recent advances in technology for interrogating the DNA sequence of an organism have resulted in a rapid decrease in the cost of sequencing, and corresponding immense growth in the number of organisms whose genomes have been sequenced. Indeed this growth has been super-exponential (see Figure 1.2), with the first complete human genome sequence released in 2003 at a total cost of over $3 billion USD, and the January 2014 release of the HiSeq X Ten that offers whole-genome sequencing at a price point of just $1,000 per genome. This dramatic reduction in cost and increase in scale presents an immense opportunity to advance the understanding of human health and disease.

A major factor contributing to this shift is the switch from low-throughput, high-accuracy Sanger-based sequencing chemistry (at the top-left of Figure 1.2) to high-throughput, lower-accuracy *next-generation* sequencing (NGS) methods (the middle and bottom-right of the figure).

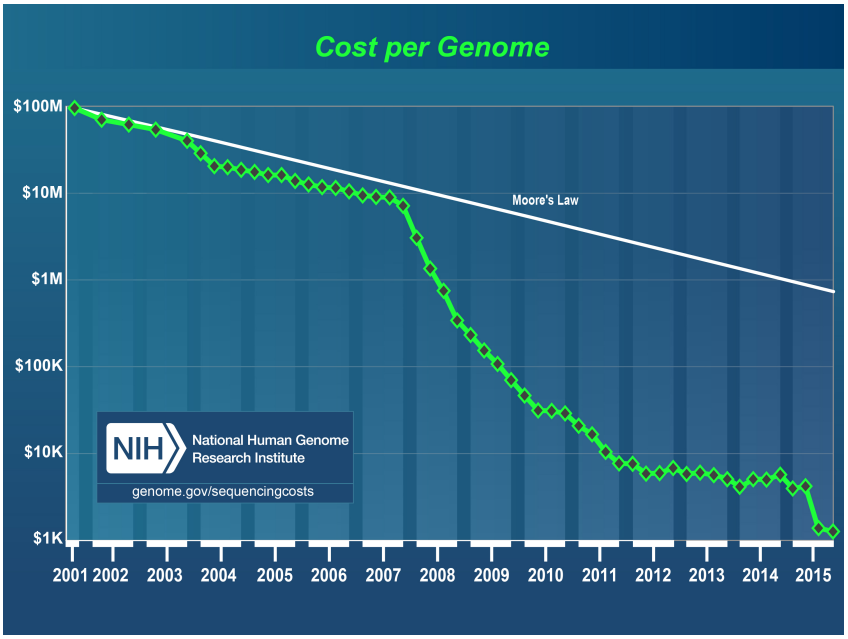

Figure 1.2: The total cost in USD of sequencing a human genome, from 2001 to 2015, versus Moore's Law governing the exponential growth of transistor density in computers. Figure from the NHGRI Genome Sequencing Program, downloaded from genome.gov/sequencingcosts on 29 April 2015.

1.2.1 Sanger sequencing

Sanger sequencing is a method of interrogating the genetic sequence of a single, pre-defined region of the genome less than around 1000 bp in length. The sequence of the flanking regions must be known beforehand so that complementary DNA sequences called *primers* can be synthesized. The region between the primers is then amplified exponentially through a process called Polymerase Chain Reaction (PCR), before being sequenced.

To prepare for sequencing, the PCR-amplified fragments are then replicated using a new mix of nucleotides that includes a low concentration of synthetic nucleotides. The synthetic nucleotides are fluorescently-labeled and designed to terminate a DNA sequence when they are added. When the fragments are replicated, nucleotides are added one at a time, each time presenting a small chance of incorporating a terminating nucleotide. This results in a population of fragments with different lengths (geometrically distributed), each terminating in a fluorescently labeled nucleotide. The fragments are then stratified by length with an electric field until differences in length of a single nucleotide are resolvable. The fluorescence of the stream of molecules is then measured using a laser to determine the nucleotide sequence of the original region.

1.2.2 Next-generation sequencing (NGS)

Rather than interrogating longer sequences of the genome one at a time, NGS technologies interrogate millions or billions of shorter (100–300 bp) DNA fragments in parallel. These fragments, called *reads*, are then aligned back to a reference sequence (or assembled together) to recover the genomic sequence of the sampled organism. A targeting step can be added to enrich for reads that contain sequences complementary to one or more designed sequences, called *probes*. *Whole exome sequencing* uses probes

that tile the sequence of most human exons to enrich the sequencing for these exonic regions. In contrast, *whole genome sequencing* does not include this targeting step.

Next-generation sequencing is particularly well suited for identifying SNVs and small indels, with whole-genome sequencing also enabling the accurate detection of CNVs. On the other hand, structural variants, especially those mediated by large repetitive regions, are much more difficult and often impossible to detect using current high-throughput sequencing technologies. Since most clinical genomic data is currently obtained by whole-exome NGS sequencing, this document will primarily focus on methods involving SNVs and small indels.

1.3 Variant harmfulness prediction

The realization of the medical advantages of the personal genome remains limited by our inability to identify the disease-causing variation from the millions of non-functional (neutral) single nucleotide, structural, and copy number variants which are present in each individual's genome. Despite the successes of using genome sequencing to identify disease-causing mutations in individuals with Mendelian disorders (Majewski et al., 2011b), as well as cohorts of individuals with more common genetic disorders such as autism (O'Roak et al., 2011), the prioritization of variants based on their involvement in disorders remains a significant challenge (Cooper & Shendure, 2011). Methods for identifying disease-causing mutations typically use one of two complementary approaches: statistical association between a variant and a disorder, or the prioritization of all genomic variants found in a genome based on their possible functional effect.

1.3.1 Genotype-phenotype association

In the statistical association approach, individuals with the disorder (cases) are genotyped in parallel with matched controls and statistical tests are then used to identify variants which are overrepresented in cases as compared with controls. These genome-wide association studies (GWAS) have led to the identification of genes associated with many common and complex disorders, including autism (Wang et al., 2009) and type 2 diabetes (Frayling, 2007). To date, over 2,000 GWAS studies have collectively implicated almost 20,000 variants in over 1,400 traits (Welter et al., 2014). However, these tests are not applicable to rare genetic disorders, where cohort sizes are very small and unrelated individuals may all be affected due to different (personal) variants within the same gene or pathway. While approaches like the Cohort Allelic Sums Test (CAST) (Morgenthaler & Thilly, 2007) and the Combined Multivariate and Collapsing (CMC) method (Li & Leal, 2008) aggregate the rare variants seen within a gene or a pathway to mitigate this, the applicability of association-based methods remains extremely limited for small cohorts. The VAAST method (Yandell et al., 2011) extends these approaches and combines SNV prioritization and association testing into a common framework. More recently, Akawi et al. (2015) incorporates phenotypic similarity into an association model to discover four recessive developmental disorders based on rare variant enrichment.

1.3.2 Variant filtration

The alternative approach of prioritizing disease-causing variants based on their predictive features has been extremely effective at identifying causal non-synonymous mutations in a number of Mendelian

disorders, including Charcot–Marie–Tooth neuropathy (Lupski et al., 2010), Hajdu-Cheney syndrome (Majewski et al., 2011a), and Miller syndrome (Ng et al., 2009). In this approach, the variants identified in the genome are filtered to just those with low allele frequencies (common variants are unlikely to cause rare disorders) and are sorted based on a "harmfulness" score, generated by a tool such as PolyPhen, SIFT, or PANTHER (Adzhubei et al., 2010; Ng & Henikoff, 2003; Thomas et al., 2003). While some of these functional variants may not be harmful, functionality is typically used as a proxy for harmfulness within such tools.

Tools for the prioritization of harmful non-synonymous variants typically consider multiple features which may affect the functionality of the protein, including the level of conservation of the changed residue, the severity of the amino acid change (a change from a hydrophobic to a hydrophilic residue is more likely to be harmful than a change within one of these groups), the location of the variant relative to functional regions of the protein, such as active sites, and the likelihood that the mutation will affect protein secondary or tertiary structure. These features are then combined using either heuristic weights (Ramensky et al., 2002) or more rigorous machine learning frameworks (Adzhubei et al., 2010; Thomas et al., 2003) to predict variants likely to have functional effects. All of the features can contribute to the overall success of the prioritization, however, the evolutionary conservation of the modified region has one of the strongest effects, and some argue it may be sufficient on its own (Cooper & Shendure, 2011).

Missense variants are enriched for harmful variants, but they represent a very small fraction (less than 1%) of total human variation. Harmfulness prediction tools have since been developed for additional classes of mutations, including synonymous mutations (Buske et al., 2013), mutations in untranslated regions (Salari et al., 2013), and non-coding SNVs (Khurana et al., 2013). More recently, integrative approaches such as CADD (Kircher et al., 2014) seek to prioritize all variants within a single unified framework.

1.4 Deep phenotyping

Deep phenotyping is the process of precisely describing all of the observable abnormalities in an individual (Robinson, 2012). It has become common to use standardized vocabularies to simplify data sharing and computational analysis. Each aspect of the patient's phenotype is described using a separate term, allowing constellations of symptoms to be unambiguously described. While plaintext descriptions can be easier to collect, standardized vocabularies facilitate exchange and computational analysis of the data. Different clinicians (depending on their specialization) may use completely different words to describe the same set of clinical features, and abbreviations and typographical errors can introduce ambiguities that even specialists cannot always resolve.

1.4.1 Electronic health records

Standardized clinical terminologies have gained popularity in the collection of health care utilization and outcome data in electronic medical records, but these terminologies are inadequate for patient phenotyping in clinical genetics (see review: Robinson, 2014). The SNOMED Clinical Terms (SNOMED-CT) is a large collection of standardized medical terms with defined relationships between these terms, and is widely used in electronic medical records (EMRs) to facilitate data storage and sharing. However, the SNOMED-CT terminology itself is not publicly available, making it unappealing for use in academic research. A complementary terminology, the International Classification of Diseases (ICD) provides a

standardized language for classifying diseases and other health disorders, specifically for health care reporting, international statistics, and epidemiological analysis. Because of its deep integration with health care, ICD is difficult to change and slow to adapt. Further, even ICD-10 and SNOMED-CT terms typically have insufficient granularity and completeness to describe many disorders (Winnenburg & Bodenreider, 2014).

1.4.2 Human Phenotype Ontology

The recording of detailed and standardized phenotypes for patients displaying a broad variety of indications requires a rich vocabulary with clear semantic relationships between the terms, to allow for the identification of similar (yet not identical) indications. Although the London Dysmorphology Database presented one of the first efforts to organize phenotypes typically seen by a clinical geneticist, the Human Phenotype Ontology (HPO) (Köhler et al., 2014) is currently the most complete vocabulary available for recording patient phenotypes for genetic diseases (Winnenburg & Bodenreider, 2014).

The HPO was developed specifically to assist in the detailed characterization and computational interpretation of disorders of human health and is quickly becoming the standard terminology in the field of rare diseases. The HPO provides a standardized, open-source terminology for describing patient phenotypes, with terms organized in a hierarchical, taxonomic structure (see Figure 1.3). The most general terms are at the top, with children always having an "is-a" relationship with their parents (e.g. both "focal seizures" and "tonic/clonic seizure", are a subtype of "seizures", which in turn is a subtype of "neurological abnormality"). Because of this structure, phenotypes can be captured at varying levels of precision without loss of interpretability. Further, this semantic structure enables computational methods that "understand" the similarity between different terms. When clinical features are encoded using an ontology such as the HPO, cases can be compared not just based on the annotated features, but also the corresponding semantic annotations of these features. Finding cohorts of patients with similar (but not identical) phenotypic traits becomes easier when patients are described using terms with defined semantic relationships.

Collectively, these properties make the HPO extremely useful for clinical research, but the broad use of HPO was initially hindered by its size and complexity: the HPO has over 11,000 terms, and only a small fraction of these are relevant for a specific patient. Intuitive user interfaces, such as PhenoTips (Girdea et al., 2013), help clinicians record precise descriptions of their patients and allow for the use of synonyms and variable granularity of presentations.

PhenoTips

PhenoTips provides a simple and easy-to-use interface for entering patient phenotype information using the HPO. With the goal of making digital deep phenotyping as fast or faster than paper, the phenotype entry portion of PhenoTips provides a powerful predictive search function, similar to Google. Typos, synonyms, and acronyms are all accommodated, and the most similar HPO terms are shown immediately in a drop-down (Figure 1.4).

PhenoTips uses the HPO behind the scenes, but makes it transparent so the clinician does not need to consider the structure or the full scope except when beneficial. For example, if a clinician wants to find related terms, PhenoTips exposes the local structure of the ontology and allows the clinician to navigate up and down the hierarchy.

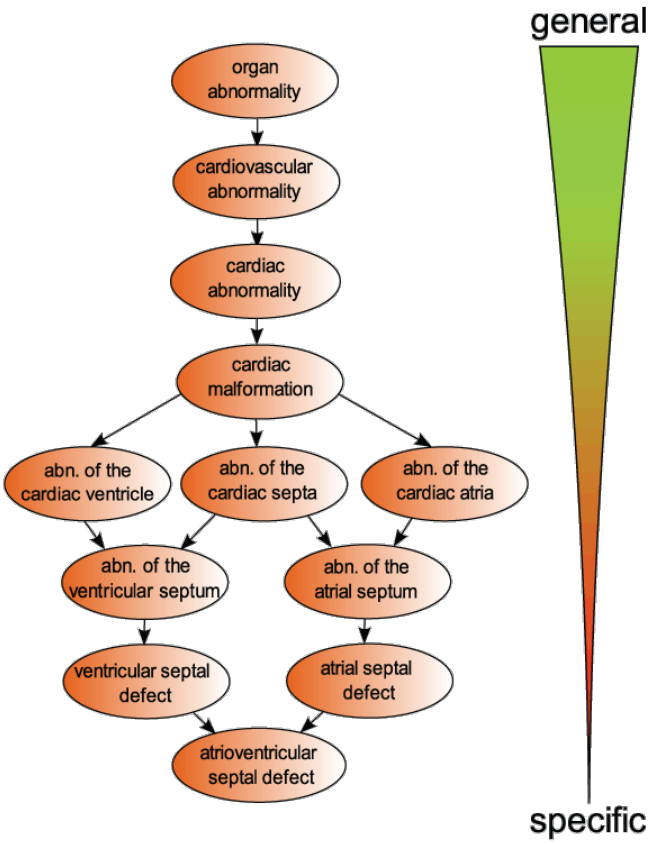

Figure 1.3: The structure of an illustrative portion of the Human Phenotype Ontology, showing the hierarchical nature with more general terms at the top, more specific terms below, and an "is-a" relationship between each node and its parents. (Köhler et al., 2009)

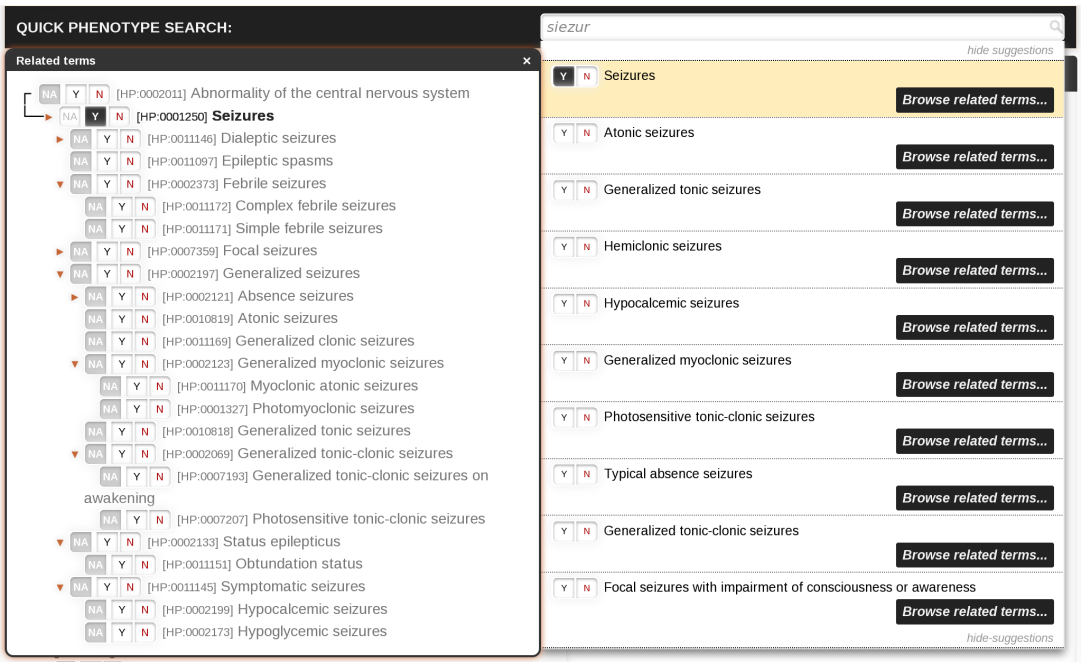

Figure 1.4: A portion of the PhenoTips user interface, showing the hierarchical checkbox interface on the left, and the predictive phenotype quick search box (with results below) on the right.

1.5 Applications to genetic diseases

1.5.1 Phenotypic similarity

After describing a case using HPO terms, similar cases and diseases can be found using a variety of existing similarity measures for comparing terms, or sets of terms, from an ontology. Foundational methods, such as Resnik (1995) and Jiang & Conrath (1997), were developed for lexical analysis and rose to prominence in the field of bioinformatics with their application to the Gene Ontology (GO) (Ashburner et al., 2000). Additional measures, such as simGIC (Pesquita et al., 2007), were developed specifically for use with the GO. Pesquita et al. (2009) provides an excellent review of the most popular similarity measures and their performance on GO-related tasks. The development of the HPO has encouraged the use of semantic similarity measures to predict clinical diagnoses (Köhler et al., 2009; Bauer et al., 2012; Zemojtel et al., 2014), to find representative model organisms for gene prioritization (Hoehndorf et al., 2011; Chen et al., 2012; Smedley et al., 2013), and most recently, to identify similar patients (Buske et al., 2015a; Gottlieb et al., 2015; Akawi et al., 2015).

1.5.2 Diagnosis prediction

Many diseases, especially rare diseases, are associated with a characteristic constellation of symptoms. For example, macrocephaly-capillary malformation (M-CM) is characterized by a prenatal, asymmetric overgrowth of the brain and body, extensive cutaneous capillary malformations, extra or fused fingers or toes, and other skin, joint, and neurological abnormalities (Mirzaa et al., 2012). Several methods have been developed to predict a diagnosis based on a set of HPO terms and the semantic relationships between terms in the ontology.

Köhler et al. (2009) presents a differential diagnosis tool, Phenomizer, which reports the most likely diagnoses and suggests additional terms to consider that will result in specific diagnoses becoming significant (above a p-value threshold). Phenomizer identifies significant diagnoses through two steps. First, it computes a similarity score between the patient phenotypes and the phenotypes associated with each disease in the Online Mendelian Inheritance in Man (OMIM) database. This similarity score is based on the information content (IC) of the most informative common ancestor (MICA) of each pair of terms. The frequency f of a phenotype term is defined as the fraction of diseases in OMIM annotated with that term (or a descendant of that term), and the corresponding information content is $-log(f)$. The similarity of a query patient Q and a disease D is defined by averaging the best matches for each query term:

$$\text{sim}(Q \rightarrow D) = \text{avg}\left[\sum_{t_1 \in Q} \max_{t_2 \in D} \text{IC}(\text{MICA}(t_1, t_2))\right] \tag{1.1}$$

In order to assess whether or not a given similarity score is significant, Phenomizer compares the score to a null distribution for that disease, obtained by randomly sampling the same number of phenotypes 10,000 times.

This approach assumes that the terms annotated for a patient are correct, that no terms are left out, and that all terms are associated with the disease. To better address this uncertainty, Bauer et al. (2012) developed a Bayesian network that incorporates the structure of the HPO and known disease-phenotype associations. The network is augmented to model false positive and false negative terms, and incorporate information about the how frequently certain phenotypes are observed in each disease. This results in a model with greater predictive power than Köhler et al. (2009) on a large simulated dataset.

1.5.3 Gene prioritization

The HPO maintains over 115,000 associations between 11,000 phenotypic terms and 7,000 rare diseases and their associated genes (Köhler et al., 2014), which provides an effective bridge from phenotype to genotype within the rare disease domain. A number of recently published methods have focused on using HPO terms to improve the prioritization of candidate genes from exome sequence data, including PHIVE (Robinson et al., 2014), Phevor (Singleton et al., 2014), Phen-Gen (Javed et al., 2014), and PhenIX (Zemojtel et al., 2014). For example, Zemojtel et al. (2014) focused on rare variants in 2,741 known human disease genes. By combining inheritance-based gene filtering with phenotypic similarity between the patient and the disease, a correct diagnosis was obtained in 28% of the cases. Smedley & Robinson (2015) provides an excellent review of these methods, and several of them are described in more detail in section 3.1.2.

1.5.4 Rare diseases and gene discovery

Rare genetic disorders collectively affect around 350 million people worldwide, but the number of people affected by any one of these disorders can be extremely small. As a result, rare disease research is historically underfunded and effective therapies are relatively rare. By taking advantage of affordable sequencing and rapidly-growing databases of observed human genomic variation, researchers are in a better position than ever to identify the genetic mechanisms of rare diseases, and thus further contribute

to the body of knowledge regarding the relationship between *genotype* — the genomic sequence of an individual — and *phenotype* — the collection of observable traits of an individual.

Rare disease research presents an immense opportunity for improving the quality of life of affected individuals and understanding the relationship between human genetics and health (Boycott et al., 2013). However, siloing of data severely impedes the discovery of genetic causes of these disorders, while directly copying such data across various resources is difficult due to legal and privacy concerns. Individuals with the same disease may be seen by different clinicians and sequenced at different centres, with each individual's data being stored in one of a rapidly growing number of different databases and patient registries. Data sharing efforts are critical for researchers to identify cohorts and validate findings for rare genetic diseases.

The Finding of Rare Disease Genes (FORGE) project addressed some of these problems by collecting and pooling patients with rare or undiagnosed diseases across Canada, and was successful in identifying the molecular etiology of 66 of 100 rare disorders. However, these are still rare diseases that were common enough to find a sufficient number of individuals within Canada to identify the cause. Of the disorders that remained unsolved, around 20 were due to insufficient power to discriminate between the many candidate mutations, and the remainder did not have any good candidates from exome sequencing. The imminent use of diagnostic whole-genome sequencing for rare disease patients demands the ability to predict harmful mutations, and to find unrelated affected individuals worldwide. Further chapters will discuss the progress towards achieving both of these goals, though non-coding mutations, complex multigenic and pathway disorders, and non-germline mutations such as mosaicism continue to pose a challenge.

Chapter 2

Variant harmfulness prediction

High-throughput sequencing results in the hypothesis-free reporting of thousands of potentially-harmful variants per exome, and an order of magnitude more per whole genome. The challenge is then to identify those variants that are most likely to cause the observed phenotype from the large background of non-harmful or irrelevant variants. Among coding variants, both nonsense and frameshift variants are relatively rare and usually deleterious (at least to the protein on that haplotype). Missense variants, however, are much more common and variable in their effect, and exome sequencing is able to identify missense variants more accurately than indels and structural variation. Together, these properties have made missense variants a popular target for harmfulness prediction tools.

2.1 Related work

2.1.1 Missense variant harmfulness prediction

Many tools have been developed to address the problem of missense variant harmfulness prediction, including SIFT, PolyPhen, Panther, and MutationTaster, with most taking the same approach: annotate variants with biologically-relevant sequence or structure features and then train a classifier to distinguish harmful and benign variants. The performance of these methods are usually limited by the amount of training data, which is often minimal, highly biased, or both. Evolutionary conservation measures, such as GERP, are also extremely useful for assessing the potential harmfulness of variants, with some evidence that these measures perform as well or better than traditional harmfulness prediction tools (Cooper & Shendure, 2011).

PolyPhen 2 (Adzhubei et al., 2010), is a missense harmfulness prediction tool commonly used in exome analysis pipelines. The tool incorporates features related to sequence homology, protein structure, and amino acid chemistry. A Naïve Bayes classifier was then trained to discriminate between disease-associated and non-disease-associated variants using the 11 most informative features (after greedy feature selection on 32 initial features). This method provides an effective first-pass prioritization of missense variants, but the high false positive rate (20–30% at true positive rates of 80%) make it inadequate for prioritizing all the variants found from whole-exome sequencing.

Modest improvements have since been made in missense harmfulness prediction, especially by tools that combine the results of multiple existing tools, such as CAROL (Lopes et al., 2012).

2.1.2 Beyond missense variation

Fewer tools exist to prioritize classes of variants other than missense variants. One such method, GECCO, classifies copy-number variants associated with mental retardation with high accuracy (Hehir-Kwa et al., 2010). Other methods focus on variation outside of the coding sequence, including evaluating changes in RNA folding energies and ensembles (Waldispühl & Ponty, 2011; Halvorsen et al., 2010; Salari et al., 2013), and predicting alternative splicing by analyzing exonic splicing enhancers and silencers (Barash et al., 2010b,a) or with sequence models (Xiong et al., 2015).

Most pipelines for identifying disease-causing variants still filter out synonymous SNVs at the earliest stages (Zemojtel et al., 2014, e.g.,). However, there is substantial evidence that synonymous SNVs affect mRNA splicing, mRNA structure, and protein expression, and some of these SNVs contribute to disease (see reviews: Cartegni et al., 2002; Chamary et al., 2006; Sauna & Kimchi-Sarfaty, 2011; Hunt et al., 2014). These mechanisms include introducing cryptic splice sites (Hellwinkel et al., 2001), affecting exon inclusion (Montera et al., 2001; Xiong et al., 2015), altering translational efficiency (Griseri et al., 2011), and disrupting transcription factor binding sites (Stergachis et al., 2013).

Splice changes are perhaps the best-studied effect of functional synonymous SNVs (Cartegni et al., 2002). The creation or modification of a splice donor or acceptor site, or the binding site of a splicing enhancer, silencer, or regulator can lead to intron inclusion or alternative splicing of the exon, and therefore a drastically different protein product (Drögemüller et al., 2011). Synonymous substitutions that change a common codon to a rare one, or vice versa, can also result in a different protein by affecting translational efficiency, as is the case with a mutation in the CFTR gene associated with Cystic Fibrosis (Bartoszewski et al., 2010). Additionally, synonymous mutations have been shown to change the expression (Kudla et al., 2009) and function (Komar et al., 1999; Cortazzo et al., 2002) of proteins in *E. coli*, and play a role in substrate specificity (Kimchi-Sarfaty et al., 2007) and cancer outcomes (Ho et al., 2011) in humans, though the latter claim has been controversial (Renneville et al., 2011).

2.2 SilVA: Silent Variant Analysis using random forests

In around 30% of Mendelian cases, exome sequencing will identify the genetic cause of the condition. The other 70% are a combination of:

1. cases where the variant is not able to be interrogated by exome sequencing, such as CNVs flanked by repetitive regions longer than the read length, and

2. cases where the causal variant is interrogated by exome sequencing but is filtered out, incorrectly prioritized, or for which there is insufficient evidence to associate with the condition.

For these cases in which the standard analysis pipeline fails to identify a promising candidate, we developed the Silent Variant Analyzer (SilVA), a Random Forest-based method for prioritizing synonymous variants in the human genome (Buske et al., 2013). A researcher can then review the top few synonymous variants and follow-up on any promising candidates. To our knowledge, no prior method combines multiple genomic features to identify "silent" genetic variants with functional effects. We use a manually-curated dataset of 33 rare synonymous disease-causing variants in order to train and evaluate the overall efficacy of SilVA, as well as two additional datasets for independent validation, showing that SilVA is able to accurately predict the harmfulness of silent variants in these datasets.

2.2.1 Dataset collection

One of the challenges in investigating synonymous disease-causing variants is the relatively small number of known examples. Perhaps partly because they are so often excluded from analysis, there are very few published examples of harmful synonymous variants. In fact, in many cases, the causal synonymous variants were only identified because the variant was the only one found after Sanger sequencing of all exons in the causal gene. A literature search was conducted using PubMed and Google Scholar based on combinations of the keywords: "harmful", "deleterious", "pathogenic", "causal", "synonymous", "silent", "splicing", "mutation", "variant", "polymorphism". In total, over 70 potentially-pathogenic synonymous variants were identified from the literature.

In many cases, the evidence for pathogenicity was only *in silico* predictions or the absence of a more likely candidate from the sequencing results. Some variants, such as rs1800093 and rs4633, were excluded because the variant was only functional in conjunction with another variant on the same allele (Bartoszewski et al., 2010). Others, such as rs34533956, were statistically associated with the disease phenotype but were not functionally validated (Narendra et al., 2009). For model training, we selected only the 33 variants with experimentally validated functional effect and association with a disease (Table 2.1).

For training and benchmarking negative controls, we used all rare synonymous variants from an individual in the 1000 Genomes Project (NA10851) (Durbin et al., 2010). We identified 758 variants with minor allele frequencies less than 5%. For case studies and validation, we trained SilVA on the NA10851 variants, but used the 746 variants in another 1000 Genomes Project individual (NA07048) during testing. Fifty-nine variants were shared by both NA10851 and NA07048.

After developing and benchmarking the SilVA method, we obtained two further validation datasets (Table 2.2). The first contained seven synonymous variants found in families with Meckel syndrome (OMIM:249000), a rare, lethal ciliopathic associated with kidney, liver, and central nervous system abnormalities (Khaddour et al., 2007). Four of these variants were reported to be novel, of which two (MKS1: E139E, TMEM67: A813A) were suspected to cause a Meckel syndrome phenotype. The other three variants were predicted to be benign polymorphisms with minor allele frequencies of 1–7%. The second dataset contained 12 synonymous mutations encountered by the Molecular Diagnostic Lab at the Hospital for Sick Children (Toronto, Canada). Of these 12 variants, six were determined to be pathogenic by a molecular diagnostician, while the remaining six were believed to be benign polymorphisms. Of the six pathogenic ones, two were already in our training data, while the other four were novel.

2.2.2 Features

We designed a set a features to capture the various mechanisms through which synonymous mutations can be pathogenic. Each variant is annotated with 26 features across 6 categories: 1) conservation, 2) codon usage, 3) sequence features (CpG and relative mRNA position), 4) exon splicing enhancer and suppressor (ESE/ESS) motifs, 5) splice site motifs for both canonical and cryptic splice site detection, and 6) pre-mRNA folding energy (see Table 2.3).

The GERP++ score is used to measure the evolutionary conservation at the mutation position (Davydov et al., 2010). Relative synonymous codon usage (RSCU) (Sharp & Li, 1987) features are calculated using codon frequencies in the Codon Usage Database (Nakamura et al., 2000). Splicing regulatory features include the SR-protein motifs for SF2/ASF, SC35, SRp40, and SRp50, scored using

Table 2.1: Thirty-three (33) examples of rare synonymous variants used as positive examples in model training and validation. All the mutations have experimentally-validated phenotypic effects and are implicated in or causal of a disease (column 4). For all listed mutations, the third codon position is affected. D_{ss} is the number of residues between the mutation position and the nearest splice site. †The following abbreviations are used: FAP: familial adenomatous polyposis; AIS: androgen insensitivity syndrome; A-T: ataxia-telangiectasia; HIGM3: immunodeficiency with hyper-IgM, type 3; CMS: congenital myasthenic syndrome; CF: cystic fibrosis; CTX: cerebrotendinous xanthomatosis; AIP: acute intermittent porphyria; LNS: Lesch-Nyhan Syndrome; GT: Glanzmann thrombasthenia; CESD: cholesteryl ester storage disease; FTDP-17: frontotemporal dementia with parkinsonism, chromosome 17 type; FD: familial dementia with swollen achromatic neurons and corticobasal inclusion bodies; PSP: progressive supranuclear palsy; HNPCC: hereditary nonpolyposis colorectal cancer; NF-1: neurofibromatosis, type 1; SNEM: Leigh's encephelomyelopathy; PKD: pyruvate kinase deficiency; HSCR: Hirschsprung disease; SMA: spinal muscular atrophy; TCS: Treacher Collins Syndrome; XL-SMA: X-linked infantile spinal muscular atrophy; FPCT: familial porphyria cutanea tarda.

Gene	Mutation	D_{ss}	Disease†	Reference	Effect
APC	R623R,G>T	89	FAP	(Montera et al., 2001)	loss of exon 14
AR	S888S,C>T	59	AIS	(Hellwinkel et al., 2001)	activates cryptic 5' splice site in exon 8
ATM	K1192K,G>A	0	A-T	(Gilad et al., 1998)	skipping of exon 26
CD40	T136T,A>T	4	HIGM3	(Ferrari et al., 2001)	loss of exon 5
CFTR	G893G,G>T	21	CF	(Faa et al., 2010)	creates 5' splice site in exon 15
CHRNE	G305G,C>T	2	CMS	(Richard et al., 2007)	creates 5' splice site in exon 9
CYP27A1	G145G,G>T	11	CTX	(Chen et al., 1998)	activates cryptic 5' splice site
F9	V153V,G>A	61	Hemophilia B	(Knobe et al., 2008)	unknown mechanism
FAH	N232N,C>T	10	Tyrosinemia, type 1	(Amstel et al., 1996)	loss of exon 8
FBN1	I2118I,C>T	25	Marfan Syndrome	(Liu et al., 1997)	loss of exon 51
FGFR2	A344A,G>A	52	Crouzon Syndrome	(Gatto & Breathnach, 1995)	activates 5' splice site
HEXA	L190L,G>A	0	Tay-Sachs	(Akli et al., 1990)	loss of exon 5; reduced transcript levels
HMBS	R22R,C>G	21	AIP	(Llewellyn et al., 1996)	loss of exon 2
HPRT1	F199F,C>T	12	LNS	(Steingrimsdottir et al., 1992)	loss of exon 8
ITGB3	T420T,G>A	0	GT	(Jin et al., 1996)	loss of exon 9
LIPA	Q298Q,G>A	0	CESD	(Klima et al., 1993)	aberrant splicing; exon skipping
MAPT	L284L,T>C	29	FTDP-17	(D'Souza et al., 1999)	increases exon 10 inclusion
MAPT	N296N,T>C	27	FD	(Spillantini et al., 2000)	increases exon 10 inclusion
MAPT	S305S,T>C	0	PSP	(Stanford et al., 2000)	increases exon 10 inclusion
MLH1	S577S,G>A	0	HNPCC	(Kohonen-Corish et al., 1996)	loss of exon 15; predicted transcript destabilization
NF1	K354K,G>A	0	NF-1	(Fahsold et al., 2000)	loss of exon 7
PAH	T323T,A>G	0	Phenylketonuria	(Ho et al., 2008)	increases exon 9 inclusion
PAH	V399V,A>T	2	Phenylketonuria	(Chao et al., 2001)	loss of exon 11 from all mRNA
PDHA1	G185G,A>G	44	SNEM	(De Meirleir et al., 1994)	loss of exon 6
PKLR	A423A,G>A	0	PKD	(Kanno et al., 1997)	loss of exon 9
PTS	E81E,G>A	0	PTPS deficiency	(Imamura et al., 1999)	loss of exon 4
RET	I647I,C>T	61	HSCR	(Auricchio et al., 1999)	aberrant splicing
SMN1	F280F,C>T	5	SMA	(Lorson et al., 1999)	loss of exon 7
TCOF1	S1127S,A>C	11	TCS	(Macaya et al., 2009)	loss of exon 22
TP53	T125T,G>A	0	Cancer susceptibility	(Warneford et al., 1992)	retention of intron 4
UBA1	N577N,C>T	10	XL-SMA	(Ramser et al., 2008)	reduced expression; altered methylation pattern of exon 15
UROD	E314E,G>A	0	FPCT	(Mendez et al., 1998)	loss of exon 9
ZFP36	R109R,C>T	284	Cancer progression	(Griseri et al., 2011)	decreases translational efficiency

Table 2.2: Independent validation dataset consisting of seven variants (two putatively pathogenic, five putatively benign) associated with Meckel syndrome and twelve variants (six pathogenic, affecting splicing, and six polymorphic) encountered by the Molecular Diagnostic Lab at the Hospital for Sick Children (Toronto, Canada). Of the eight pathogenic variants, two (those in TP53 and FGFR2) were already included in our training data. We used SilVA to rank these variants relative to all (746) rare putatively-neutral synonymous variants in a 1000 Genomes Project individual not used during model development or training (NA07048). The SilVA method ranked all pathogenic variants higher than all polymorphic variants. Moreover, we ranked 4/6 new pathogenic and putatively pathogenic variants as more harmful than any control variant (a rank of 1). For all listed mutations, the third codon position is affected.

Gene	Mutation	Rank	Score	Description [MAF]
Meckel Syndrome				
TMEM67	A813A,G>A	1	0.737	novel, putatively pathogenic
MKS1	E139E,G>A	1	0.705	novel, putatively pathogenic
MKS1	L557L,G>C	277.5	0.020	polymorphic [0.06]
TMEM67	D799D,T>C	311	0.015	polymorphic [0.01]
TMEM67	C62C,T>C	356	0.011	novel, putatively benign
TMEM67	T964T,A>C	447.5	0.006	polymorphic [0.07]
TMEM67	A984A,A>G	722	0.000	novel, putatively benign
Molecular Diagnostics Lab at the Hospital for Sick Children				
TP53	T125T,G>A	1	0.795	pathogenic, in training data
ACVRL1	P459P,G>C	1	0.794	pathogenic
FGFR2	A344A,G>A	1	0.762	pathogenic, in training data
CFTR	E528E,G>A	1	0.524	pathogenic, exon skipped
PKP2	G828G,C>T	29	0.153	pathogenic, cryptic splicing
IDS	G374G,C>T	73	0.083	pathogenic, cryptic splicing
TP53	L257L,C>T	106	0.065	polymorphic, novel
FGFR2	V232V,A>G	169	0.042	polymorphic [0.18]
CFTR	T854T,T>G	329.5	0.014	polymorphic [0.44]
CDKN1C	E236E,G>A	435	0.006	polymorphic [0.02]
IDS	T146T,C>T	501.5	0.004	polymorphic [0.24]
TP53	P36P,G>A	638.5	0.001	polymorphic [0.01]

Table 2.3: An example of the biological and evolutionary features used by SilVA to predict the harmfulness of synonymous mutations.

Feature	Description
Conservation	
GERP++	Conservation at the mutation position
Codon usage bias	
RSCU	RSCU of new codon
$\|\Delta\text{RSCU}\|$	Change in RSCU caused by mutation
Sequence features	
CpG?	Does the mutation change a CpG?
CpG_{exon}	Observed/expected CpG content of exon
f_{pre}	Relative distance to end of pre-mRNA
f_{post}	Relative distance to end of mature mRNA
Exon splice enhancer/suppressor motifs	
SR-	SR-protein motifs lost
SR+	SR-protein motifs gained
FAS6-	Hexamer splice suppressor motifs lost
FAS6+	Hexamer splice suppressor motifs gained
PESE-	Octamer splice enhancer motifs lost
PESE+	Octamer splice enhancer motifs gained
PESS-	Octamer splice suppressor motifs lost
PESS+	Octamer splice suppressor motifs gained
Splice site motifs	
MES	Max splice site score
$\|\Delta\text{MES}\|$	Max change in splice site score
ΔMES+	Max splice site score increase
ΔMES-	Max splice site score decrease
MES-MC?	Did strongest site change?
MES-CS?	Is a cryptic site now strongest?
MES-KM?	Did a known site change most?
Pre-mRNA folding free energy	
$\Delta\Delta G_{\text{pre},50}$	Folding energy change, pre-mRNA, 50 bp window
$\Delta\Delta G_{\text{post},50}$	Folding energy change, mature mRNA, 50 bp window
$\Delta D_{\text{pre},50}$	Ensemble diversity change, pre-mRNA, 50 bp window
$\Delta D_{\text{post},50}$	Ensemble diversity change, mature mRNA, 50 bp window

ESE Finder 3.0 with default thresholds (Smith et al., 2006), the FAS-hex3 hexamer dataset from FAS-ESS, used for the ESS6 features (Wang et al., 2004), and PESX enhancer and suppressor octamers, used for the pESE and pESS features (Zhang et al., 2005). The splice site motif strength features (MES) are calculated using MaxEntScan (Eng et al., 2004). The effect on the change in free energy from pre-mRNA folding ($\Delta\Delta G$) features are calculated with UNAFold 3.8 (Markham & Zuker, 2008), and the effect on the ensemble diversity (ΔD) features are calculated with ViennaRNA 2.1.1 (Lorenz et al., 2011). Prior to training, we pre-process each of the features to have zero mean and unit variance.

2.2.3 Models

We compared the ability of 5 different methods, the GERP++ score and 4 machine learning models, to prioritize the most-likely harmful synonymous variants from whole-exome analysis. These methods are:

1. Sort by GERP++ conservation score. Mutations at more conserved residues are ranked higher.
2. Fisher's linear discriminant (FLD). The variants are ranked by the one-dimensional projected value.
3. Support vector machine (SVMmap), using the nu-SVR regression mode of the lib-svm toolkit, version 3.11 (Chang & Lin, 2011). We then sort variants by the regression score.
4. Neural network (NNet), with a single, fully connected hidden layer of 5 hidden units, using the PyBrain Python package, version 0.3 (Schaul et al., 2010). We activate the trained network on the test SNVs and use the value at the output node to prioritize them.
5. Random forest (Forest), with 1001 trees and the default number of variables used for each split (the square root of the total number of variables), using the randomForest R package, version 4.6-6. Variants are ranked by the number of votes, with the most popular variants ranked highest.

2.2.4 Performance

The performance of each method was evaluated using stratified 50/50 train/test splits and leave-one-out cross-validation.

Stratified 2-fold cross-validation

For 50/50 splits, we trained each model on half of the positive and negative examples (~17 known deleterious, 379 presumed benign, or control), and then ranked the remaining (16 known deleterious and 379 control variants). In each round, we excluded from the training set any positive examples that occurred within the same gene as any of the positive test mutations. Each method was then evaluated according to the quality of the top-most predictions. We aggregated the results across 50 iterations of training and testing, each time with a new random subset of deleterious and control variants. As shown in Figure 2.1A, the random forest method outperforms the other methods, with more than three times the true positive rate (at a false positive cutoff of 1%) as simply using the GERP++ score.

Leave-one-out cross-validation (*in silico* infection)

To compare the prioritization performance of the five methods in a more realistic scenario, we performed *in silico* "infection" experiments (leave-one-out cross-validation). In each experiment, we held out one

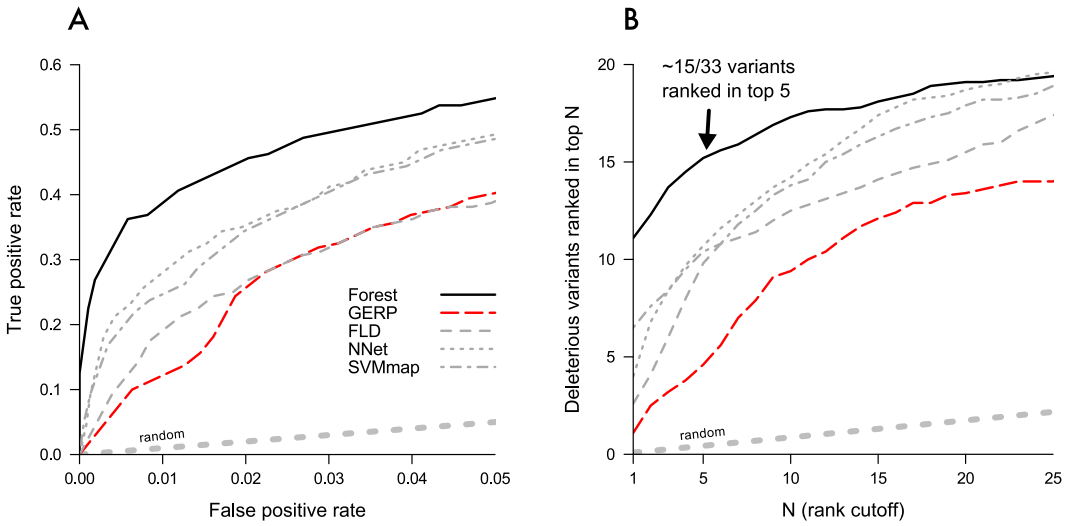

Figure 2.1: Plots comparing the performance of several machine learning methods (Forest: random forest; SVMmap: support vector machine; FLD: Fisher's linear discriminant; NNet: neural network) and the GERP++ score at classifying harmful synonymous variants. **A**) This plot is the bottom-left region on a receiver operating characteristic (ROC) curve. Curves were averaged over 50 training iterations, with half of the positive and negative examples used for testing. The performance of a random ordering appears at the bottom. Random forest is able to rank more of the held-out positive examples highly than any of the other methods at low false-positive thresholds. **B**) Performance on leave-one-out cross-validation experiments using 33 deleterious variants and 758 rare (putatively neutral) variants from 1000 Genomes Project individual NA10851. If we consider the causal variant to have been found if it is ranked in the top 5, random forest succeeds on an average of 15.2 deleterious variants (vs. 10.7 for NNet, 10.4 for SVMmap, 9.8 for FLD, and 4.6 for GERP++).

of the known deleterious variant and half of the variants in a human genome (1000 Genomes Project individual, NA10851) and trained SilVA on the remaining variants. As in the 50/50 split, we excluded from training any positive examples within the same gene as the held-out variant. We repeated this process 10 times with different random subsets of control variants and averaged the rank of the held-out deleterious variant within the test dataset. This leave-one-out cross-validation method allows us to estimate the number of disorders for which our method is able to rank the deleterious variant among the top few variants genome-wide.

The prioritization performance of the five methods are compared in Figure 2.1B. The random forest method achieved the best performance, consistently ranking the deleterious variant in the top 5 most-harmful variants for ~15 of 33 diseases (versus ~5 for the GERP++ conservation score alone).

Feature comparison

To better understand the relative contributions of the features used within SilVA and to explore the relative importance of each category of feature, we compared SilVA's cross-validation performance leaving out different classes of features from the analysis (Figure 2.2B). Removing features related to codon usage, mRNA folding, splicing enhancer and suppressor motifs, and sequence (CpG, relative position in mRNA) does not substantially affect performance. Removing either splice site features or conservation (GERP++), however, causes SilVA's performance to drop substantially, with splice site features

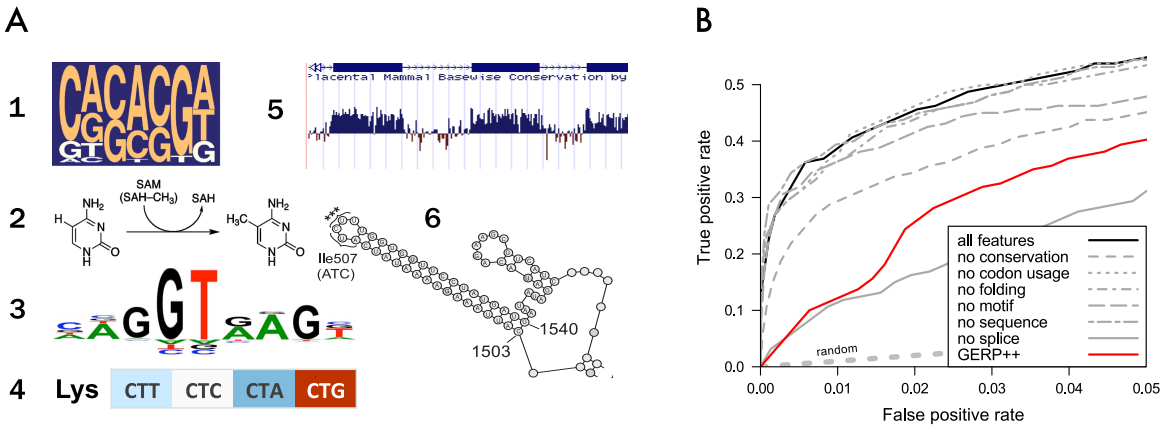

Figure 2.2: **A**) Illustrations of the various feature categories used within SilVA to predict the harmfulness of synonymous variants: 1) exon splicing enhancer/suppressor motifs (SF2/ASF shown), 2) sequence features (CpG and relative position in the mRNA), 3) splice site motifs, 4) codon usage bias, 5) conservation, and 6) RNA folding (image adapted from Bartoszewski et al. 2010). **B**) The bottom-left region of a receiver operating characteristic (ROC) curve comparing the performance of the SilVA method with groups of features removed. Results were aggregated across 50 iterations of stratified 50/50 cross-validation. For comparison, we also show the performances of sorting by the GERP++ score and a random ordering.

appearing to be more informative than conservation for harmfulness prediction. Per-feature forward and backward feature selection was also tried but the redundancy between different features within each category made it difficult to interpret the underlying importance. Several feature importance measures can actually be calculated straight from trained random forests (Archer & Kimes, 2008), but these measures can be biased when the scales of the features are different, as they are in this case (Strobl et al., 2007).

Harmfulness classification

Based on the benchmarking results, the random forest method was selected and all features were kept. The SilVA score was defined as the fraction of trees in the random forest that predict the mutation to be harmful.

To evaluate the ability of the SilVA score to differentiate between harmful and benign variants, we measured the mean SilVA scores for our harmful mutation dataset and common polymorphisms (MAF>5%) which are unlikely to be harmful. We computed the scores of the 33 harmful variants using leave-one-out cross-validation and the scores of all common polymorphisms from the 1000 Genomes Project (May 2011 phase 1 release v2) and found the harmful variants to have a significantly higher mean score (0.322 vs. 0.031, Student's t-test: $p \leq 1.8 \times 10^{-7}$). Further, we still find a significant difference when comparing against rare synonymous variants from a healthy individual (0.322 vs. 0.031, $p \leq 1.9 \times 10^{-7}$, 1000 Genomes Project individual NA07048), and even when we focus on just variants within three residues of a splice site (0.544 vs. 0.153, $p \leq 3.8 \times 10^{-6}$). The SilVA score is thus an effective tool for prioritizing synonymous variants.

To use SilVA as a classifier, we defined score thresholds of 0.27 and 0.485, corresponding to true positive rates of 52% and 33% and false positive rates of less than 1% and 0.1%, respectively. Based on

these thresholds, SilVA classifies variants as likely benign, potentially pathogenic, or likely pathogenic to aid interpretation. Because we expect harmful synonymous variants to be extremely rare, we do not intend SilVA to be used in the same way as typical non-synonymous harmfulness prediction tools and thus focus on ranking variants instead of classifying them, though we also report classification results for each dataset.

When applied to the original dataset of 33 deleterious variants, 11 were classified as likely pathogenic, 6 as potentially pathogenic, and 16 as likely benign. For comparison, of the rare synonymous variants across 82 CEU 1000 Genomes Project individuals, an average of less than 1 variant per genome was classified as likely pathogenic, 7 as potentially pathogenic, and 727 as likely benign. Variants that were mistakenly classified as benign tended to be far from splice sites and disrupt ESE/ESS motifs or translational dynamics. Though we have features that attempt to capture these mechanisms, the machine learning algorithms did not find these specific features to be informative in training.

2.2.5 Application to two datasets

We then applied the final SilVA model to evaluate two smaller, additional datasets: synonymous variants associated with Meckel syndrome, and a collection of synonymous variants clinically observed and stratified by a molecular diagnostics laboratory (Table 2.2).

Meckel syndrome variants

First, we used SilVA to predict the harmfulness of a collection of synonymous SNVs reported by Khaddour et al. (2007) across many cases of Meckel syndrome. Khaddour et al. describe seven synonymous mutations in the MKS1 and TMEM67 (MKS3) genes, of which four are novel and three are known polymorphisms (minor allele frequencies of 1–7%). Two of the novel mutations are suspected of causing Meckel syndrome through the disruption of splice donor motifs. These variants were not included in our training dataset because they did not meet our criterion of experimental validation.

As controls we used all (746) rare synonymous variants in a 1000 Genomes Project individual not used for training or benchmarking (NA07048). In agreement with the literature, SilVA ranks the two suspected harmful mutations (MKS1:E139E,G>A; TMEM67:A813A,G>A) higher than every control variant (a rank of 1) and none of the remaining five mutations within even the top 250 variants.

Variants from HSC's Molecular Diagnostics Laboratory

The Molecular Diagnostics Laboratory at the Hospital for Sick Children (HSC) conducts Sanger sequencing for gene panels in patients with suspected genetic disorders. Each variant is analyzed by a molecular diagnostician, who classifies it as benign or harmful based on an interpretation of its likely molecular effect and a literature review. The Molecular Diagnostic Laboratory provided us with six pathogenic synonymous variants and six benign polymorphisms identified during their analyses, with two of the pathogenic variants already appearing in our training data. Similar to our analysis of the Meckel variants, we implanted the 10 remaining variants in a 1000 Genomes individual (NA07048). SilVA ranks all pathogenic variants higher than all polymorphic variants, with two of the four new pathogenic variants (and both of the ones in the training data) ranking higher than any control variants (a rank of 1).

2.2.6 Assessment of genome-wide synonymous constraint

The rate of synonymous substitutions is widely used as a proxy for the neutral mutation rate, including for the purposes of identifying selection on a gene (*e.g.* McDonald et al., 1991). However, as detailed in the previous sections, synonymous substitutions can exert a phenotypic effect and thus be selected against. Previously, there have been several attempts to understand the fraction of synonymous sites that are under constraint, and the strength of selection at these sites both overall (see review: Chamary et al., 2006) and at specific locations such as exon splicing enhancers (Parmley et al., 2006). The heterogeneity of both the genome and even individual genes, and substantial methodological differences, have resulted in widely-varying estimates, with some suggesting that up to 39% of synonymous substitutions are under selection (Hellmann et al., 2003). Simultaneously all such studies have used comparison of multiple mammalian genomes, and not analysis of human polymorphisms; constraint observable from human polymorphisms would represent generally stronger selection, due to the small human effective population size (Ne $\approx 10^4$) (Tenesa et al., 2007).

More recently, Salari et al. (2013) compared the effects of common human polymorphisms and random mutations on RNA structural ensembles, and found significant evidence of ensemble-stabilizing selection. If a significant fraction of human synonymous sites are under constraint and the SilVA score reflects this, we would expect to see a difference in SilVA scores between common synonymous polymorphisms and a matched set of random mutations. We test this hypothesis by applying SilVA to each synonymous SNP in NA10851 (9596 variants with allele frequencies of 5–95%) and a matched random synonymous SNV within the same gene. The matched random mutation was controlled for: 1) creation or destruction of CpG dinucleotides and 2) splice site proximity (the random mutation created/destroyed a CpG site only if the synonymous variant did, and whether or not the mutation was within three bases of an exon boundary). We then compared the distribution of SilVA scores for the two datasets. While the mean observed scores for polymorphisms and random mutations were similar (0.031 and 0.034, respectively), the difference in the means is highly statistically significant due to the large number of datapoints (Student's t-test, paired, $p \leq 2.4 \times 10^{-6}$). The overall higher scores of random mutations suggest that factors beyond CpG and exon boundaries impose purifying selection at synonymous sites of the human genome that is statistically significant.

To further quantify this constraint, we measured the difference in the number of random mutations and true polymorphisms (Figure 2.3) above a certain SilVA score. This difference can be interpreted as the number of mutations "rejected" during evolution as being unfit, and represents synonymous sites under constraint (Cooper et al., 2005). At a SilVA threshold of 0.005, we observe 626 more random mutations (6531) than true polymorphisms (5905). Thus, we estimate that 6.5% of potential synonymous substitutions (626/9596 SNPs) have been rejected since human divergence due to constraint beyond just CpGs and splice sites. Note that this is a conservative estimate, as there are likely additional functional features in the genome that the SilVA score is not modeling.

2.2.7 Implementation

SilVA is implemented as a collection of shell, Python, and R scripts and is freely available from compbio.cs.toronto.edu/silva. After using the `setup.sh` script to install dependencies and download necessary datasets, input files can be annotated and then scored with the pre-trained model. The input file can be in either VCF format or a custom tab-delimited format. The protein coordinate (`.pcoord`) format is

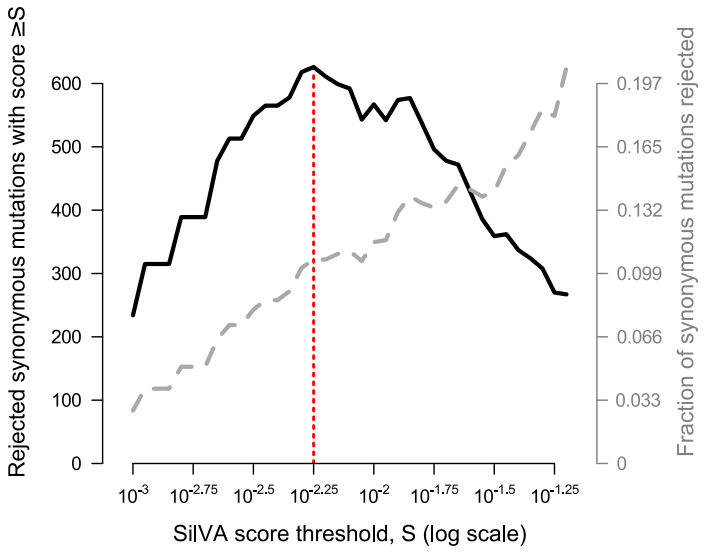

Figure 2.3: The number and fraction of synonymous mutations that are rejected at various SilVA score thresholds. The largest number of rejected mutations occurs at a SilVA threshold of $10^{-2.25}$, marked with a red line, where 626 more random mutations pass the threshold than actual polymorphisms, corresponding to a 10.6% rejection rate at this threshold.

a tab-delimited file, one line per variant, with 4 columns:

1. gene symbol
2. amino acid position
3. reference amino acid single-letter code
4. nucelotide change in HGVS format

Lines starting with # are ignored. For example:

```
#gene  aa pos  aa  mutation  comments
IDS    374     G   C>T       pathogenic, cryptic splicing
```

It is worth noting that the protein coordinate format does not unambiguously describe a genomic mutation, but instead reflects the level of information available in many publications. When run, SilVA attempts to resolve the ambiguity and reports a warning if the variant cannot be mapped to a unique synonymous SNV.

SilVA then filters out non-synonymous variants, common variants, and variants on the Y chromosome (at the time of development, the 1000 Genomes Project did not have high-quality data for the Y chromosome). The remaining synonymous variants are annotated with features and scored using the pre-trained random forest model. The variants are then output in order of descending score, one per line, with the following tab-delimited fields:

1. rank of the variant within the dataset, with ties assigned the average rank

2. SilVA score, a number between 0 and 1, which corresponds to the fraction of classifiers that predict the variant to be harmful
3. SilVA classification, either "likely pathogenic", "potentially pathogenic", or "likely benign", depending on the SilVA score
4. gene symbol, e.g., "PKHD1"
5. RefSeq transcript identifier, e.g., "NM_170724"
6. chromosome of the variant from the VCF file, e.g., "1"
7. position of the variant from the VCF file, 1-indexed, e.g., "51949681"
8. identifier from the VCF file, e.g., "."
9. reference sequence from the VCF file, e.g., "T"
10. alternate sequence from the VCF file, e.g., "C"

2.3 Summary

While current technologies are able to sequence a human genome relatively cheaply and quickly, the key bottleneck is the interpretation of the variants in order to identify those that are most likely to be related to an observed phenotype or a disorder. The automated prioritization of deleterious variants is an important step towards the realization of genomic medicine. Synonymous variants are usually excluded from analysis pipelines wholesale, despite evidence that some of this "silent" variation has important functional roles.

SilVA represents the first method specifically focused on the prioritization of disease-causing synonymous SNVs. We curated 33 high-confidence disease-related variants, and evaluated several machine learning approaches for prioritizing these from amongst a set of rare putatively-neutral synonymous SNVs based on a number of features, including sequence conservation, splice sites, splice-regulatory motifs, codon frequency, CpG content, and RNA secondary structure energy. Our results indicate that splicing information and sequence conservation are currently the two most informative features for identifying deleterious synonymous variants, and the performance degrades without either of these features. The random forest method outperforms other statistical learning methods at prioritizing disease-causing SNVs, and yields variant scores that are significantly higher in known harmful variants than control variants. When a deleterious SNV is added to a human genome, this method ranks the deleterious SNV among the top five candidates for 15 of the 33 diseases and is able to identify harmful variants in independent validation sets. Together, these findings indicate that automated methods for identification of deleterious synonymous SNVs can be useful in parallel with methods that prioritize other types of genomic variation for the analysis of full human genomes.

However, addressing the challenge of translating genotype to clinical phenotype requires not just identifying potentially pathogenic variants, but finding those that are most likely to contribute to the abnormal phenotypes of interest. This challenge requires effective use of animal models, in which controlled perturbation experiments and knockout models can be developed, and accessible databases of human phenotypes and genotypes.

Chapter 3

Deep phenotyping and disease gene discovery

3.1 Phenotypic matching of rare disease patient profiles

3.1.1 Phenotypic similarity measures

Phenotypic similarity measures assess the distance between two phenotypic terms or profiles and are useful for finding cohorts with the same undiagnosed condition, similar model organism experiments, and likely diagnoses. We compared the performance of 13 different semantic and structural similarity measures (11 information content measures and 2 topological measures) by their ability to match patients with the same disease based on annotated HPO terms. See Table 3.1 for the definitions of these terms and Figure 3.1A for additional detailed description of the benchmarking method and a summary of the results. The Resnik, Lin, JC, and Jaccard measures score the semantic similarity between pairs of terms at a time. To extend these measures to compare two sets of terms, P and Q, we employed two methods commonly found in the literature (Pesquita et al., 2009): $sim_{avg,max}$: averaging the score across all pairs of terms, and $sim_{avg,avg}$: averaging the score of the best match for each term in P (it is worth noting that the latter produces an asymmetrical similarity measure). In contrast, the PhenoDigm, UI, and simGIC measures directly score the similarity between two sets of terms.

3.1.2 Comparison on simulated and real data

We compared the performance of the 13 similarity measures using a synthetic dataset of 1,000 patients and a real dataset of 720 deeply-phenotyped patients (annotated with five or more observed HPO terms) from the PhenomeCentral web portal (Buske et al., 2015a). We considered two patients to match if they were submitted as part of the same cohort, diagnosed with the same disease, or annotated with the same gene as a likely candidate or confirmed cause (candidate genes were only used in this case if at most two were specified). These criteria resulted in 225 real cases with at least one match in the database.

Table 3.1: The 13 similarity measures used to find patients with the same rare disease based on the HPO terms annotated for each patient. The Resnik, Lin, JC, and Jaccard measures compare two ontology terms, a and b. To measure the similarity between two patients (i.e., between two sets of ontology terms, P and Q), either the average score (avg) or best score (max) for each term in P is averaged together. The smoothed reciprocal of the JC distance measure was used as a similarity measure. In contrast, the UI, PhenoDigm, and simGIC measures directly score two sets of ontology terms. Three variants of the PhenoDigm score are described in (Smedley et al., 2013), and all three were included in the evaluation. The information content of a term is defined as $IC(t) = \log p(t)$ where $p(t)$ is the fraction of all disease-HPO mappings that involve term t (or a descendant of t). We also compared this to a topological definition: $IC(t) = (|g_t| + 1)/N$, where N is the number of terms in the HPO and g_t is the set of terms including t and all descendants of t. In the table, g^t is the set of terms induced by t (the set of nodes including t and all ancestors of t), and g^P is the set of terms induced by the set of terms in patient P.

Score	Variations	Equation	Reference
Resnik(a, b)	avg, max	$\max_{t \in g^a \cap g^b} \mathrm{IC}(t)$	see the review (Pesquita et al., 2009)
Lin(a, b)	avg, max	$\frac{2 * \mathrm{Resnik}(a,b)}{\mathrm{IC}(a) + \mathrm{IC}(b)}$	"
JC(a, b)	avg, max	$\frac{1}{\mathrm{IC}(a) + \mathrm{IC}(b) - 2 * \mathrm{Resnik}(a,b) + 1}$	"
Jaccard(a, b)	avg, max	$\frac{\lvert g^a \cap g^b \rvert}{\lvert g^a \cup g^b \rvert}$	"
UI(P, Q)		$\frac{\lvert g^P \cap g^Q \rvert}{\lvert g^P \cup g^Q \rvert}$	"
PhenoDigm(P, Q)	avg, max, combined	see reference	(Smedley et al., 2013)
simGIC(P, Q)		$\frac{\sum_{t \in g^P \cap g^Q} \mathrm{IC}(t)}{\sum_{t \in g^P \cup g^Q} \mathrm{IC}(t)}$	(Pesquita et al., 2007)

Simulated patient phenotypes

To benchmark the performance of phenotypic matching and gene prioritization algorithms, we generated 500 pairs of synthetic patients with the same genetic disease, following the standard protocols used in the literature (Robinson et al., 2014; Zemojtel et al., 2014; Javed et al., 2014). Each pair of patients was randomly assigned a disease gene and associated OMIM diagnosis, with 250 pairs assigned an autosomal dominant disease and 250 pairs an autosomal recessive disease. Diseases were sampled uniformly rather than by prevalence, because population prevalence varies by orders of magnitude for these diseases and this prevalence does not necessarily reflect the frequency in a database of undiagnosed patients. The phenotype of each patient was sampled from the set of HPO terms associated with the disease in several ways (HPO version 2014-06-09, disease-phenotype mappings and inheritance mode from the phenotype_annotation.tab file released with the HPO).

Evaluation

We evaluated these measures under several conditions. Initially, all phenotype terms were included. We then introduced noise to model the sources of variability and error in real patient records using three methods from Zemojtel et al. (2014). Clinicians frequently use less precise terms, so we artificially introduced imprecision by randomly replacing every term with a term drawn uniformly from the set of ancestors of that term (including the term, excluding the root: HP:0000118 [Phenotypic abnormality]). We modeled variability in clinical presentation and unrelated clinical features both by sampling a random subset of five terms, and by adding two random terms for every five terms in the patient description (sampled uniformly from the set of disease-phenotype associations).

To assess the sensitivity of the information-content-based measures to the particular corpus used to compute it, we evaluated each measure using three different methods for calculating information content (two different corpora and a topological method). The primary method used the disease-phenotype associations from the HPO, the same corpus used to simulate patients. We compared this to information content computed from the corpus of disease-phenotype associations provided by OMIM, as well as directly from the topology of the graph using the same method as GeneYenta (Gottlieb et al., 2015).

Results

Overall, the two best-performing measures were the PhenoDigm score (Smedley et al., 2013) and the simGIC score (Pesquita et al., 2007) (see Figure 3.1A).

On simulated data, the PhenoDigm score outperformed all other measures and ranked a true match first for 69% of the patients after introducing random phenotypic noise and imprecision, and within the top five patients 90% of the time. The performance was similar when the information content was computed using a smaller corpus and when it was computed using the topology of the HPO (Figure 3.2).

On real data, the PhenoDigm and simGIC scores performed comparably, ranking a true match first for 54–60% of patients, and within the top five matches 70–71% of the time. However, when phenotypic noise and imprecision were added to these real patients, the performance dropped considerably, with respective top one and top five recall rates of 11% and 29% for PhenoDigm, and only 3% and 9% for simGIC.

The PhenoDigm measure is more complex to implement and asymptotically slower than simGIC, $O(n^2)$ and $O(n)$, respectively. In a basic Python implementation, PhenoDigm takes 10 min to perform

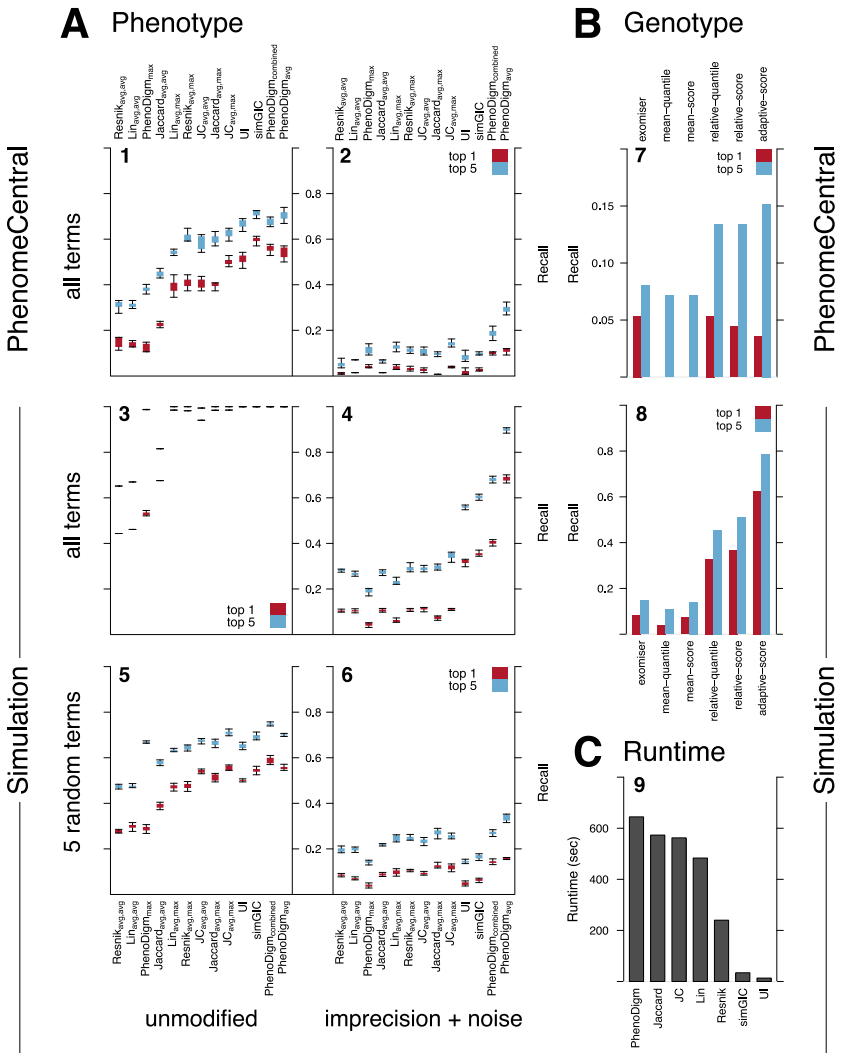

Figure 3.1: **A**) The performance of 13 similarity measures at finding similar patients (panels 1–6). For simulated patients, either all disease-associated phenotype terms were selected (panels 3–4), or five terms were randomly selected (panels 5–6). Noise (adding random phenotype terms) and imprecision (replacing terms with a random ancestor) were then introduced (panels 2, 4, 6). To control for variable cohort size, two cases were randomly selected from each cohort for each of 10 iterations. The performance of each measure is the fraction of cases for which the matching case was ranked within the top one (red/dark) or five (blue/light) most similar cases. **B**) The performance of six methods at prioritizing causal and candidate genes in 112 cases from PhenomeCentral (panel 7) and 1,000 simulated cases with noise and imprecision introduced (panel 8; same parameters as panel 4). As a baseline method, the Exomiser was run on each case individually and genes ordered by their PHIVE score. This was compared with five methods that first identify the most phenotypically similar patents, and then score genes separately for each match. The performance of each method was measured by either the fraction of cases where one of the causal or candidate genes was ranked as the top gene for the most similar patient (red/dark) or among the top five genes (blue/light; either the top gene for one of the four most similar patients or the top gene from the Exomiser directly). **C**) Execution time of each class of similarity measure on the 1,000 simulated cases (N = 499,500 pairwise comparisons).

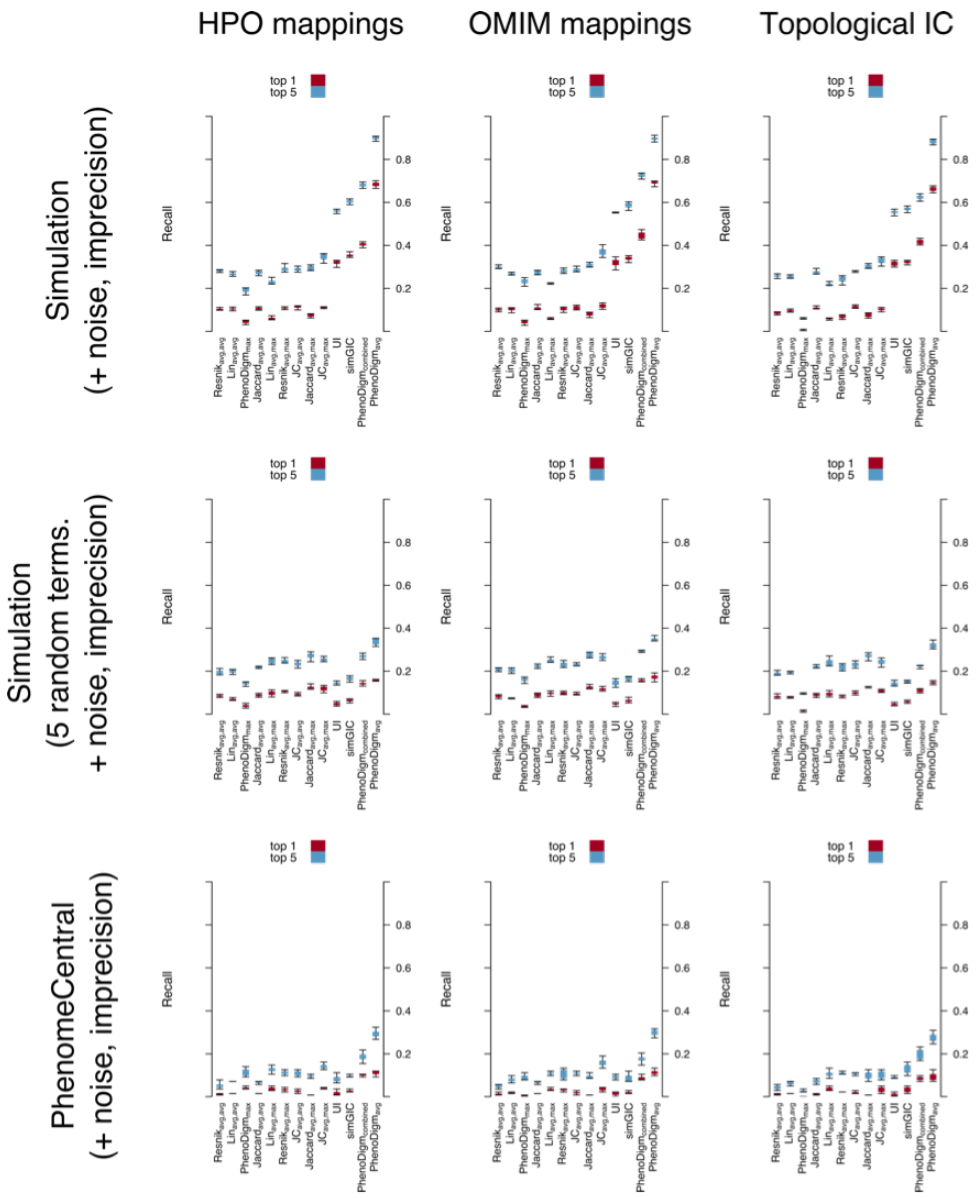

Figure 3.2: The effect of different methods for information content calculation on the performance of each phenotypic similarity measure on simulated patients with noise added (top and middle rows) and real patients with noise added (bottom row). The information content was calculated in three ways: based on the disease-phenotype mappings provided by the HPO (left column), based on the disease-phenotype mappings provided by OMIM (center column), and based only on the topology of the HPO as using the same method as GeneYenta (right column). The overall performance of most measures appear to be robust to these differences.

all pair-wise comparisons of 1,000 simulated cases while simGIC takes just 34 sec (Figure 3.1C). The performance of both could be improved with optimization and caching, such as pre-calculating the similarity between all pairs of HPO terms for PhenoDigm. The relative utility of each score depends on the size of the data set and the amount of noise in the data.

Phenotype matching, though powerful, still has difficulty with extremely atypical presentations and in situations where only some anatomical systems are deeply phenotyped (as can happen when specialists perform less thorough investigations outside of their specialty). Frequently, patients are admitted as cases of novel rare diseases based on their particular constellation of symptoms, only to be diagnosed with an atypical presentation of a known genetic disease after exome sequencing. Yet variant harmfulness prediction methods are not mature enough to accurately identify the causal variants based only on a patient's genotype.

3.2 Improving variant prioritization with deep phenotype data

Combining the massive amounts of exome and whole-genome sequence data now being generated with deep-phenotyping efforts, both in the model organism community and in human clinical research, promises to improve our understanding of the complex relationship between genotype and clinical phenotype. With common diseases, it is easier to find a sufficient number of patients so that case/control studies have statistical power to identify even relatively complex causes, such as interactions between multiple genes in the same pathway, or different causes for different, overlapping sub-phenotypes (Warde-Farley et al., 2012). With rare diseases, heuristic filter-based approaches are often used because a study may have only one or two affected individuals. By exploiting harmfulness prediction tools and allele frequencies in databases of control variants (1000 Genomes Project, Exome Variant Server, HapMap, and dbSNP), the hope is that there are sufficiently few plausible candidates remaining that they can be inspected manually. Yet this is frequently not the case, especially if there is only one family and it is small. The high false positive rate of harmfulness prediction tools combined with the lack of multiple unrelated affected individuals limits one's ability to identify the causal mutations in these difficult cases. However, multiple recent efforts have attempted to address this challenge by leveraging patient phenotypes in novel ways.

3.2.1 Previous work

eXtasy: Adding phenotypic relevance to missense harmfulness prediction

The eXtasy tool (Sifrim et al., 2013) attempts to reduce the false positive rate of existing missense variant harmfulness prediction methods by combining the scores from multiple such methods (SIFT, PolyPhen 2, MutationTaster, and CAROL), in addition to conservation, gene haploinsufficiency scores, and the phenotypic relevance of genes using the Endeavor gene prioritization algorithm (Aerts et al., 2006). The authors compared multiple machine learning methods and found Random Forest to perform the best. Interestingly, they also stratified test disease-associated variants by the year of variant discovery to assess the impact of retrospective bias on their method's performance.

The Endeavor algorithm prioritizes genes in a patient based on a seed set of genes and a combination of expression data, gene associations and interactions, functional annotation, sequence information, orthology, biomedical literature and text mining. In eXtasy, the algorithm is seeded with the set of

genes associated with all OMIM diseases with phenotypes that overlap the HPO terms found in the patient. This use of Endeavor seems to work in eXtasy's published examples, but will likely cause problems if the patient has phenotypes found in many OMIM diseases (such as "intellectual disability").

Exomiser: Applying model organism data to human gene prioritization

The Exomiser tool and web service takes an exome, a mode of inheritance, and an OMIM disorder or collection of HPO terms and ranks the top candidate genes by their PHIVE score (Robinson et al., 2014). This score is the average of the most harmful rare variants in the gene (taking mode of inheritance into account) and the phenotypic relevance of the gene according to gene knockout experiments in mouse models (summarized in Figure 3.3).

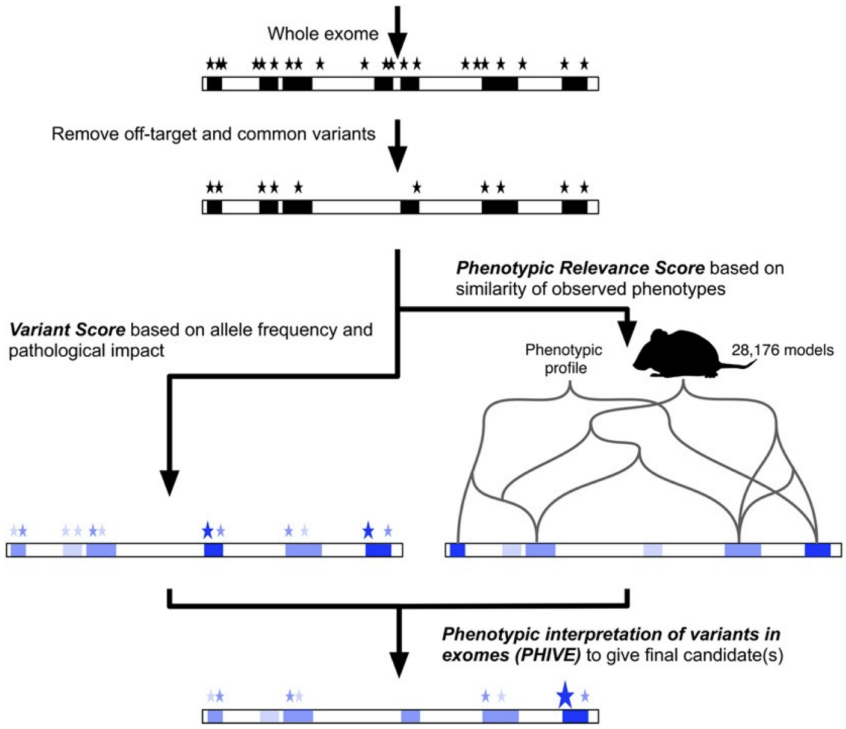

Figure 3.3: A diagram of the Exomiser method showing the separate variant harmfulness prediction and phenotype prioritization, along with the unification into a single PHIVE score.

By relating mouse and human phenotypes and making the reasonable assumption that the mouse phenotypic effects of genomic changes will be correlated with the human phenotypic effects of similar changes, Exomiser makes model organism data directly applicable to discovering the genetic mechanisms of rare human diseases. This is especially important given that almost 25% of human genes have an orthologous phenotyped mouse mutant but no known phenotypic association in humans, and this fraction is expected to rise quickly in the coming years as mouse knockouts for all remaining genes are created and phenotyped.

The Exomiser PHIVE score highly ranks the causal gene in simulation studies where known-harmful variants were spiked into the exomes of healthy individuals from the 1000 Genomes Project, and shows improved ranking performance of the causal gene, especially for autosomal dominant disorders. However, in practice, Exomiser does not help much unless the causal gene has a corresponding mouse model. In-

corporating additional model organisms and protein–protein interaction network propagation to smooth this effect are improvements that have been made to future versions of the tool. An additional concern is that the variant scoring used by Exomiser is rather *ad hoc*, and frequently scores missense mutations higher than nonsense mutations. Missense mutation scores are the maximum of three variant harmfulness scores (each defined in the range of [0, 1]), and nonsense mutations are assigned the fixed score of 0.95. Combining the Exomiser's model organism phenotypic scoring with the CADD variant harmfulness score may be a fruitful direction of improvement.

Phevor: A flexible ontology-based gene prioritization scheme

Singleton et al. (2014) introduces Phevor, a method for ontology propagation for gene prioritization. Given a collection of ontologies where nodes are annotated with genes (such as the HPO or, trivially, the Gene Ontology), patient phenotypes (which are immediately converted into candidate genes), and initial gene scores (such as output by Exomiser, VAAST, or other tools), Phevor performs an iterative weight propagation algorithm to re-score the genes. The eventual gene score is based on the product of the percentile ranks of the initial gene score and the steady-state gene weight after convergence. By exploiting the relationships between genes across different ontologies, Phevor is robust to missing annotations for the causal disease.

Their method is flexible and extensible, but the software is not currently open source and the algorithm provides little justification for the choice of constants. For example, weight is always propagated with a decay factor of 0.5, which is not discussed or compared to other values, and weight is propagated both downwards and upwards in the ontology with this same factor. The true-path rule — that a node implies all its ancestors — suggests that upwards and downward factors might reasonably be different, since they correspond to different semantic shifts. Further, though the performance of the method is evaluated leaving some ontologies out, there is no discussion of weighting the various ontologies as would be found in most network fusion methods.

3.2.2 A cohort-free approach to gene prioritization

While these gene prioritization methods have achieved some accuracy when analyzing singleton cases and families, the power to detect the casual gene in Mendelian diseases can be greatly increased by comparing the genotypes of multiple patients with the same disease. Even without predefined cohorts, incorporating variant prioritization into phenotypic patient matchmaking may help identify plausible genetic mechanisms shared by phenotypically-similar patients. However, to our knowledge, no existing methods attempt to address these two problems simultaneously: finding similar patients and identifying associated genes for those matches. We therefore compared the performance of a state-of-the-art gene prioritization method which scores genes for a single patient individually (the Exomiser; Robinson et al. 2014) to five novel methods that combine these gene scores across patients. These methods were compared on 1,000 simulated cases (as in subsection 3.1.2) and 112 real cases from the PhenomeCentral web portal (Buske et al., 2015a).

Baseline single-patient method

The Exomiser was selected as a baseline method because the source code was readily available and because the PHIVE score (Robinson et al., 2014) uses model organism data instead of human data,

decreasing the likelihood that disease-gene associations (for example, those found in OMIM) published using PhenomeCentral patients were included in the training data for the algorithm. This means that the success rate we achieve on the real data set should be a reasonable estimate of our ability to identify novel genes associated with undiagnosed rare diseases, rather than overfitting to existing knowledge.

The Exomiser (version 7.0.0beta; built from commit a25c26c3) was run with the parameters in Table 3.2.

Table 3.2: Parameters used to run the Exomiser.

Parameter	Value
`min-qual`	30
`max-freq`	1.0
`keep-off-target`	false
`keep-non-pathogenic`	false
`prioritiser`	phive

Pair-wise methods

We compared the baseline single-patient method to the following five pair-wise methods, for a pair of cases P and Q, a gene g, and some other patient $R \notin \{P, Q\}$:

mean-score: The mean of the PHIVE score for each gene appearing in the filtered Exomiser output of both cases.

mean-quantile: The average quantile score of each gene, where the lowest-scoring gene has a quantile score of 0.0, and the highest-scoring gene has a quantile score of 1.0.

relative-score: The mean-score score, scaled by the mean PHIVE score across all other cases in the database. If the gene is not in the filtered Exomiser output for the other patient R, a score of 0.0 is used.

$$\text{relative-score}(P, Q, g) = \frac{\text{mean-score}(P, Q, g)}{\text{mean}_{R \notin \{P,Q\}}(\text{PHIVE}(R, g))} \tag{3.1}$$

relative-quantile: The mean-quantile score, scaled by the mean of the quantile score across all other cases in the database. If the gene is not in the filtered Exomiser output for the other patient R, a quantile of 0.0 is used.

$$\text{relative-quantile}(P, Q, g) = \frac{\text{mean-quantile}(P, Q, g)}{\text{mean}_{R \notin \{P,Q\}}(\text{quantile}(R, g))} \tag{3.2}$$

adaptive-score: An adaptive measure, in which a per-gene variant harmfulness threshold is calculated and the score is adjusted by the phenotypic similarity of any other patients with at least as many harmful variants in the same gene.

$$\text{adaptive-score}(P,Q,g) = \text{mean-score}(P,Q,g) * \prod_R \left(\frac{\text{sim}(R,P)+\text{sim}(R,Q)}{(2*\text{sim}(P,Q))}\right)^{I(P,Q,R,g)}$$
$$I(P,Q,R,g) = \{1 \text{ if } \text{var}(R,g) \geq \min(\text{var}(P,g),\text{var}(Q,g)),\ \text{else } 0\} \tag{3.3}$$

where $\text{sim}(P,Q)$ is the phenotypic similarity between cases P and Q, $\text{var}(P,g)$ is the variant harmfulness score of gene g for patient P, and $I(P,Q,R,g)$ is an indicator variable that is 1 if R's variant harmfulness in gene g is at least that of P or Q and 0 otherwise.

Evaluation

Performance was measured as the fraction of these cases in which a target gene was listed in the top one or top five exome-wide. For the Exomiser method, genes were prioritized for each case in isolation. For the five pair-wise measures described above, the most phenotypically similar patients were first found using the PhenoDigm measure. Gene prioritization was then performed by selecting either: the top gene from the top match; or the combination of the top gene from the top 4 matches and the top gene directly from the Exomiser.

Simulated cases Simulated cases were created as described in subsection 3.1.2. For each patient, we synthesized a corresponding whole-exome VCF file by taking the exome of a healthy control from the 1000 Genomes Project (phase 1 integrated calls; Durbin et al., 2010) and spiking in random pathogenic variants from HGMD (v1.0.3) in the disease-associated gene (one heterozygous variant if the disease was dominant, one homozygous or two heterozygous variants if recessive). HGMD variants were filtered to only non-synonymous variants overlapping RefSeq coding sequences, and variants explicitly labeled with "associated" or "susceptibility" were ignored. Disease-gene associations were taken from OMIM.org (accessed 2015-07-08). We only considered genes associated with a single OMIM disease, a single inheritance mode, and at least five HPO terms, resulting in 156 autosomal dominant and 605 autosomal recessive diseases.

Real cases The real dataset was composed of the 112 patients from PhenomeCentral annotated with five or more observed HPO terms, for which exome sequence data were available, and annotated with between one and five candidate or causal genes. Whole-exome sequence data was present for 692 of the 1027 cases in PhenomeCentral. Of these, 20 were filtered out in quality control: 15 exomes were removed because only SNP calls (and not indel calls) were available; five samples were removed due to abnormally high numbers of exonic variants (four of these were the only samples in the data set sequenced using the AB SOLiD platform, and the other sample was processed using a deprecated pipeline on the NCBI36 assembly).

Results The adaptive scoring method outperforms other methods, increasing seven-fold the number of correctly identified causal genes in simulations with phenotypic noise and imprecision (8% to 63% rank the causal gene first), and doubling the number of real patients with correctly identified genes (from 8% to 15% of patients having a causal or candidate gene ranked within the top five genes) over using the Exomiser separately on each patient (see Figure 3.1B). In the best-scoring approach, we used the

PhenoDigm measure to identify the most phenotypically similar patients and then identified those genes that Exomiser scored highly in both patients and lowly in phenotypically dissimilar cases.

3.3 Gene discovery in matchmaking databases

Once one or more variants of unknown significance are identified in a patient and are hypothesized to cause the disease, additional sources of evidence are necessary to support this hypothesis, such as independent validation in unrelated affected individuals. Finding additional cases provides powerful support for a genotype–phenotype association where the disease is very rare, and *genomic matchmaking* approaches have begun to take hold in the human genetics community in order to accelerate the characterization of the thousands of uncharacterized disease genes. Genomic matchmaking databases (GMDs) allow participants to submit genomic and phenotypic data with the goal of identifying previously uncharacterized disease-associated genes by "matching" to other comparable cases, potentially from different institutions. If a sufficient number of patients in the database are identified with a similar phenotype, this provides strong evidence that mutations in the gene are associated with the disease in question. In light of the enormity of the challenge involved in identifying the several thousand novel disease-associated genes thought to exist, it is important to ask how many patients such databases will need to contain in order to identify a given proportion of these genes. A lot of effort is being put into building databases in human genetics and filling them with detailed patient data. Certainly, these resources become more useful the more data are in them, but it is unclear how much data must be added before a significant fraction of these genes are able to be identified. By making a few simplifying assumptions, we simulated a prototypic GMD in order to estimate this number and provide guidance for policy-making and future developments in the bioinformatics community (Krawitz et al., 2015).

3.3.1 The genomic "birthday paradox"

Before relaxing the assumptions to explore their effect on the matchmaking power of the GMD, we started with a few simplifying assumptions:

- there are a total of 3,000 disease genes to be identified
- all samples submitted to the database have a mutation that is located in one of these genes
- each of the disease genes is associated with a single, delineable disease (i.e., no genetic heterogeneity exists)
- the technical and bioinformatics analysis has functioned correctly and the corresponding variants have been called and annotated such that they can be recognized as potentially pathogenic using automated tools
- the clinical diagnosis was correct in all cases, or that the phenotypic clustering used is able to group together samples with the same disease with complete accuracy
- each of the 3,000 diseases is equally probable

Based on these assumptions, the probability that two particular samples share a mutation in a specific gene of interest is:

$$\frac{1}{3000} \times \frac{1}{3000} = 1.1 \times 10^{-7} \tag{3.4}$$

However, if the goal is to determine whether the database contains some pair of samples representing a disease with a mutation in the same gene, then it becomes apparent that the question strongly resembles the birthday paradox, which refers to the probability that some pair of persons, in a set of n randomly chosen people, will have the same birthday.

It can be shown that if the year has 365 days, and birthdays are equally likely to happen on each of these days, the probability of two of the n people sharing a birthday can be calculated to be slightly over 50% for $n = 23$ and about 99.9% for $n = 70$. Using a standard Poisson approximation of the birthday paradox (DasGupta, 2005), it can be shown that the probability p of a match in a group of r people is:

$$p = 1 - e^{-\frac{r*(r-1)}{2*365}} \tag{3.5}$$

Assuming that the year has 3,000 "days" corresponding to disease genes, and the "persons" correspond to a group of 65 exome- or genome-sequenced samples, the probability of at least two samples having a mutation in the same gene is slightly more than 50%:

$$p = 1 - e^{-\frac{65*64}{2*3000}} = 0.5001 \tag{3.6}$$

By a similar calculation, the probability of finding at least one match exceeds 80% if only 100 patients are registered in the database, 98% if there are 150 patients, and close to 100% if there are 200 patients.

The above calculations give us the probability of finding a single match in the database. Of course, a more important question is how many matches/novel disease genes we can expect to find in a GMD with a certain number of patients, or correspondingly, how big must our database be to be able to identify a certain proportion of the currently uncharacterized disease genes?

3.3.2 Simulations

In order to develop intuition for the challenges involved and to provide a lower bound on the number of patients that GMDs should strive to include, we simulated a genomic matchmaking database and varied a number of factors that influence its ability to identify disease genes. These simulations assume that the GMD first performs an automated search across the entire database that identifies a case group by matching patients with comparable phenotypes, and then performs a rare variant association test using the remaining cases as the control group. The search is conducted over all identifiable diseases in the GMD. A series of parameters (Table 3.3) were varied in order to examine their influence on the total number of novel disease genes that can be identified in this way.

A likelihood ratio test for rare variant association is used to compute gene-wise p-values, based on a chi-square distribution with one degree of freedom (Yandell et al., 2011):

$$-2\log\frac{(\frac{m}{n})^m(1-\frac{m}{n})^{n-m}}{(\frac{a}{n_a})^a(1-\frac{a}{n_a})^{n_a-a}(\frac{u}{n_u})^u(1-\frac{u}{n_u})^{n_u-u}} \tag{3.7}$$

where m is the total number of rare mutations in a given gene that have been detected by the exome sequence and flagged as pathogenic by the bioinformatics analysis. In the case group size n_a, the number of identified mutations is a, whereas the number of predicted pathogenic mutations in the control is u. Both a and u can be viewed as discrete random variables that also depend on the choice of the detection rate, d, the mode of inheritance, and the genetic variability, λ.

The probability distribution of the test statistic D is approximately a chi-squared distribution with

Table 3.3: Parameters investigated that affect the ability of a Genomic Matchmaking Database to identify a certain number of disease genes.

Parameter	Range	Explanation
Mode of inheritance	AR, AD	Autosomal recessive; autosomal dominant
Number of patients n	1–250,000	Number of patient samples entered in GMD
Detection rate d	0.7–1.0	Ability to identify a disease-causing mutation by exome sequencing and bioinformatics analysis
Background rate λ	0–0.05	Likelihood that a control individual will harbor a variant in a disease gene that is called pathogenic by bioinformatics analysis
Prevalence classes	1–4	Number of different classes of disease prevalence in GMD
Total number of "novel" disease genes k	3,000–6,000	Total number of genes mutations associated with unelucidated Mendelian diseases

one degree of freedom. The corresponding p-values were corrected for multiple testing by multiplying by the total number of tested genes (assumed to be 20,000). If the corrected p-value was below a threshold of 0.05, then the gene being tested was considered to be significantly associated with the disease, that is, there was a successful "genomic match."

After exploring the effect of varying each parameter independently DasGupta (2005), the following simulation was performed to estimate the size of GMD necessary to identify a given number of disease genes:

- equal numbers of autosomal recessive and autosomal dominant diseases
- diseases were randomly assigned to one of the four overall prevalence categories
- an overall detection rate of 80%
- a background variation rate of 0.01

With these parameter settings, the GMD would require over 40,000 patients to identify half of the novel disease genes if we assume that there are 3,000 such genes, and well over 100,000 samples to identify half of the novel disease genes if there are 6,000 such genes (Figure 3.4).

3.4 Summary

We compared the performance of a number of popular semantic similarity measures, and found that the PhenoDigm score is best able to accurately and robustly match patients based on their phenotypes in a diagnosis-free manner. However, for situations where speed and simplicity are important, the simGIC score is an easy-to-implement alternative that provides comparable performance on real data while being linear in computational complexity, making it significantly faster than pair-wise metrics such as Resnik's measure and the PhenoDigm score for deeply phenotyped cases.

We also show that combining genetic information across phenotypically similar patients dramatically improves the prioritization of candidate genes beyond only using the data for a single patient. The

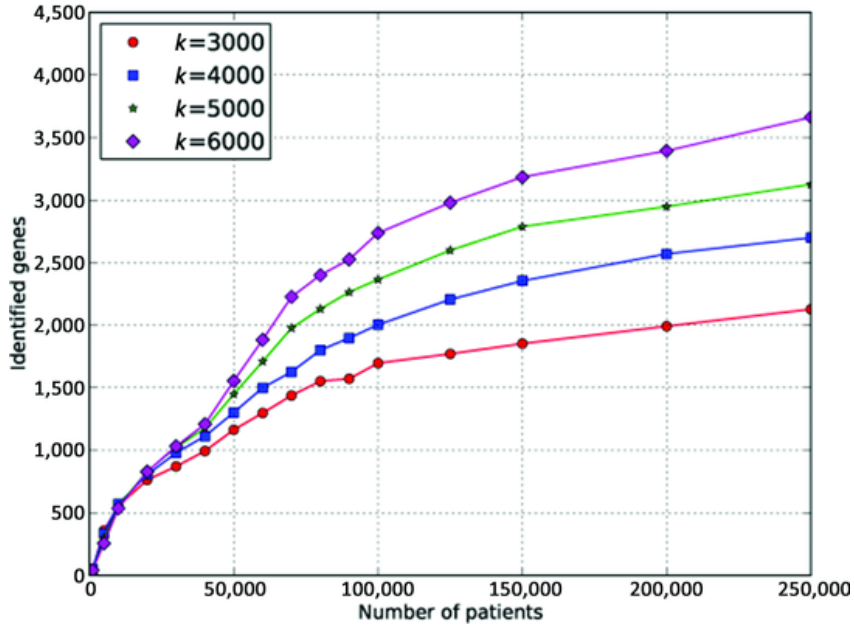

Figure 3.4: Simulations were carried out with "four prevalence classes" (each five times less prevalent than the next more-common category), and varying the total number of novel disease genes from k = 3,000 to k = 6,000.

method first selects a variant harmfulness threshold, and then discounts the average score for every other case with a variant above the threshold, weighted by the phenotypic similarity of the other case. This reduces the score of genes that are ubiquitously prioritized (such as HLA-A and TTN), while preserving high scores if a large cohort of phenotypically similar patients all have deleterious variants in the same gene. Although the performance of the best method is only 15%, this is an intentionally conservative estimate and reflects our performance on novel diseases using clinical exomes without using any human data for gene prioritization. These results highlight our ability to identify similar patients and causal genes using only HPO terms and whole-exome VCF files, demonstrating the feasibility of hypothesis-free matchmaking of novel rare disease cases using computational tools.

Databases of cases with candidate genes present another opportunity for identifying novel disease genes. We simulated a genomic matchmaking database to investigate how the performance of such databases scale with the number of cases. While only 65 undiagnosed cases are necessary before the first "match" is likely to be made, we find that between about 50,000 and 200,000 cases will be required to identify about 2,000 disease genes in a GMD. This demonstrates the reason for the early success of "serendipitous matchmaking," when two clinicians discover that they have patients with a mutation in the same gene simply by talking to each other about their many cases. However, while the first (several) matches are made quickly, a much larger number of cases are needed to discover a significant fraction of all undiagnosed diseases, highlighting the need for broad data sharing through GMDs.

There are several limitations to these findings worth discussing in more detail:

- In the gene prioritization simulations, we selected random pathogenic variants from HGMD to create *in silico* exomes with known causal genes. However, HGMD variant classification does not follow ACMG guidelines (Richards et al., 2015), raising concern that some of these variants are misclassified. While the number of such cases is likely small enough to have little impact

on the presented results, future work should consider using newer resources with higher quality annotations of variant pathogenicity, such as ClinVar (Landrum et al., 2016).

- The GMD simulations made the assumption that all of the cases actually represent one of the diseases whose underlying disease-associated gene is unknown, whereas actually some of the cases are likely to represent previously known diseases that were not recognized.

- We simulated GMDs with detection rates of 0.7–0.9, which may be optimistic given that exome sequencing is not well suited to capturing regulatory mutations, and at least currently an unknown proportion of noncoding exonic mutations (5' and 3' UTR, deep intronic splicing mutations, silent mutations causing exon skipping, etc.) are not identified reliably by bioinformatics pipelines. Therefore, the number of cases needed to discover a certain number of diseases genes in a GMD as estimated by our simulations should probably be considered a lower bound. GMDs should therefore endeavor to recruit large numbers of patients in order to be effective. To the extent possible, efforts should be made to increase the detection rate by comprehensive deep phenotyping of patients (to improve matching), and to reduce the background variation rate by sophisticated bioinformatics analysis (to reduce the number of false-positive calls in persons who do not have the disease in question).

- We assumed the correct gene is identified a certain fraction of the time, but do not address the problem of false positive matches arising from spuriously identified candidate genes. See Akle et al. (2015) for an analysis of this problem.

- Background variation rate is not uniform across loci. Guo et al. (2016) recently used exome sequence data to model gene-specific variation and characterize the sample sizes and study design needed to discover rare disease genes using burden testing. They observed that a gene's background variation rate is linearly correlated with the number of cases necessary to detect it, resulting in substantial (10-fold) differences in the number of number of cases needed to detect the various genes.

Chapter 4

Data sharing approaches to novel disease gene discovery

The content of genetic tests has gradually expanded over the years, with major leaps happening recently with the introduction of exome and genome sequencing. Although the rate of solving Mendelian disorders has increased with the ability to simultaneously sequence all genes, a large fraction of patients still remain without a diagnosis. A portion of these unsolved cases harbor suspicious variants in candidate disease genes. For such cases, finding just a single additional unrelated case with a deleterious variant in the same gene and overlapping phenotype may provide sufficient evidence to implicate the gene and enable a diagnosis for the patient. Methods for identifying these additional cases have evolved over time. From word of mouth between colleagues to sharing published case reports, laboratory diagnosticians and clinicians have worked to uncover connections between patients (Loucks et al., 2015). In a world of rapidly evolving information technologies, however, a more efficient solution is needed that can scale with the exploding growth in genomic sequencing.

4.1 The PhenomeCentral web portal

To address this need, we developed PhenomeCentral (Buske et al., 2015a), a matching network for patients with rare and undiagnosed diseases. The PhenomeCentral portal enables clinicians and researchers to quickly and easily find similar patients submitted by other contributors.

Rather than a traditional database that users query using a sophisticated language or complex filters, users "query" the PhenomeCentral repository simply by contributing a patient record. The critical fields for matchmaking are "Clinical symptoms and physical findings", which allows for the selection (presence or absence) of relevant phenotypic terms from the HPO, and "Genotype information", where genetic variants can be entered and uploaded. The phenotype terms can be selected either using a search box, or through a set of (expandable) check-boxes. PhenomeCentral supports both entering a curated set of candidate genes and uploading a VCF file (with appropriate patient consent). The VCF file is automatically processed using the Exomiser software (sanger.ac.uk/resources/databases/exomiser) to identify an additional, computationally-prioritized set of candidate genes. The set of selected phenotypic terms and candidate genes for each patient are compared to those in all other patient records in the repository, using algorithms described below. This *query-by-example* approach frees the users from the

responsibility of composing the right interrogation, and gives them incentive to contribute data and participate in the growth of PhenomeCentral. The user can immediately see information about the most similar other patients in the database and contact those submitters, in a way that preserves the privacy of the patients (Figure 4.1).

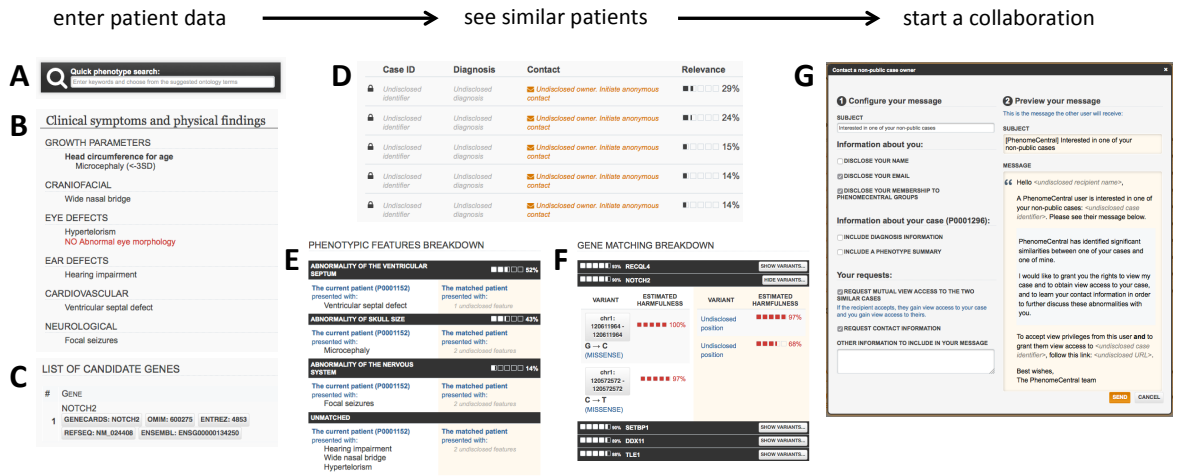

Figure 4.1: Finding similar patients in PhenomeCentral. Patient data can be contributed to PhenomeCentral through the PhenoTips user interface, including **A**) the phenotype search box that enables rapid entry of phenotype terms from the HPO, or selected records can be automatically de-identified and transferred from any PhenoTips instance. The patient record can contain **B**) both present and absent phenotypic features as well as **C**) genetic information, including candidate genes and VCF files. The patient's features are then immediately compared with all other patients in PhenomeCentral, and **D**) the best matches are shown to the user. A detailed breakdown of the **E**) phenotypic and **F**) genotypic similarity is shown for each match, enabling the user to see the underlying reasons for the match and determine whether or not the match is worth following up. **G**) A customizable email template facilitates contacting the (potentially undisclosed) submitter of another patient record.

Submitting a patient record to PhenomeCentral enables high-quality matchmaking against a rapidly growing number of cases. PhenomeCentral has seen a consistent rise in the number of patient records and user accounts since the first collaborators started submitting data in June 2013, with the number of patients tripling since the official launch on Rare Disease Day, February 28, 2014 (Figure 4.2A). As of April 19, 2015, PhenomeCentral contains data from 1,027 clinically phenotyped patients with rare genetic diseases (1,243 records in total, including unaffected relatives) collectively entered by 391 user accounts spanning five continents (Figure 4.2B).

4.1.1 Data entry and user interface

Users can enter de-identified patient data directly into PhenomeCentral through a user-friendly Web interface based on the PhenoTips software (Girdea et al., 2013), by pushing existing patient records from private PhenoTips installations, or by uploading them in bulk (Figure 4.1A-C). The patient record consists of relevant phenotypic terms, genetic information such as exome sequence data or a curated gene list, metadata including age of onset and mode of inheritance, and non-identifiable demographic information such as month and year of birth, sex, ethnicity, and family history. Combined, this functionality enables the recording of all relevant study data within a single portal, and allows research

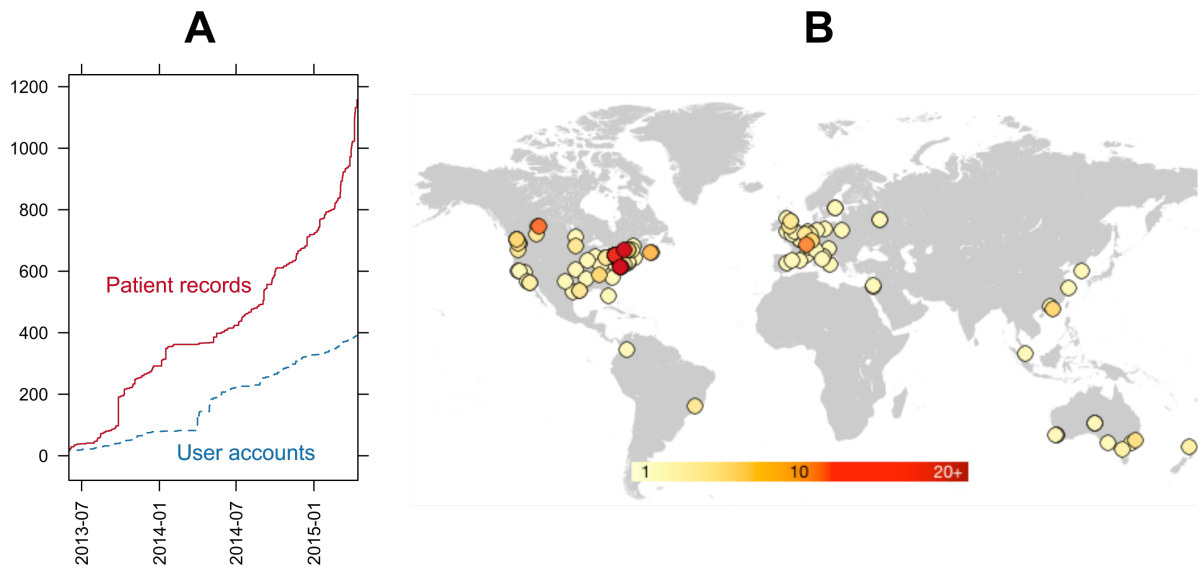

Figure 4.2: **A**) The number of patient records (red solid line) and user accounts (blue dashed line) on PhenomeCentral over time. **B**) The locations of PhenomeCentral users, estimated from the domain name of institutional email addresses associated with user accounts. The approximate region was identified by querying freegeoip.net with the IP address associated with the domain name of each email address. One point is plotted per domain name, with the color corresponding to the number of users with that domain (the darker the color, the more users with email addresses on that domain).

consortia to use PhenomeCentral as their primary data repository. Clinical Genetics has also adopted PhenomeCentral as the preferred repository for depositing structured phenotype data associated with case reports published in the journal.

PhenomeCentral records all phenotypic information as HPO terms, using the PhenoTips software for patient phenotyping. Two resources are available to help ensure users properly and completely enter the patient's phenotype into PhenomeCentral. First, the Monarch Initiative and PhenomeCentral jointly developed a set of annotation guidelines and best practices for clinical phenotyping using PhenoTips and the HPO (phenomecentral.org/annotation-guidelines). Second, a widget is included in PhenomeCentral that displays the Monarch Initiative's annotation sufficiency metric (Washington et al., 2014) as a rating from 1–5 stars. This provides the user with real-time feedback on the specificity of the patient description and encourages the user to enter more terms and more specific terms to phenotype the patient.

Each patient record in PhenomeCentral can be set to one of three different visibility settings:

private the record is visible only to the submitter unless explicitly shared with other users or groups, and does not participate in any matchmaking activity.

matchable the record is not directly visible unless explicitly shared, but similar patients are shown to the submitter, and other contributors with similar patients can discover the existence of the record. The matched phenotypes and genomic variants are obfuscated (phenotypes are made more general and only gene-level information is provided).

public the record is visible to all registered users on PhenomeCentral and participates in matchmaking activity. Contributors of similar patients are shown the submitter's contact details and the matched phenotypes and genomic variants. Whenever a whole exome is provided for a public case, only the

top 10 potential causal variants (ranked by the Exomiser) are shown to other users.

Private records are most useful when a consortium enforces a period of direct data sharing among its scientists before broader sharing is allowed (e.g. in the Neuromics consortium there is a 6 month waiting period before any sharing). Matchable records allow patients that are not yet published or consented for full sharing to participate in matchmaking activities with enhanced patient and submitter privacy. Contacting the submitter of a matchable case is simplified with a customizable message template, allowing the user to quickly and easily choose what patient information to include in the message and add a personal message (Figure 4.1G).

4.1.2 Phenotypic and genotypic matching

Phenotypic similarity is computed using the simGIC score described in section 3.1 for performance reasons. If the two patients share a clinical diagnosis, the score is boosted towards 1.0 proportional to a scaling factor d, currently set to 33%.

To find similar patients quickly as the database scales, a seeding step is added to avoid having to iterate over every case to sort by the simGIC score. First, every patient in the database is indexed with the induced subset of HPO terms. To find matches for a particular patient, the induced subset for that patient is used as the query on the database. The resulting information retrieval score used by Lucene, the document indexing tool used by PhenomeCentral, behaves similarly to the simGIC score. The top 50 patients are efficiently retrieved from the database in this manner, and then they are manually scored using the simGIC measure and the top 10 patients returned. As long as the score is *similar enough* to the Lucene score, the top 10 patients will fall within the top 50 from the seeding step and the results will be the same.

The genotypic similarity is the highest gene score of any gene that is listed as a candidate gene in at least one of the patients. If the candidate gene does not appear in the filtered exome results, the score is 0.9^{N-1}, where N is the number of candidate genes entered for that patient. If the candidate gene does appear in the filtered exome results, the Exomiser gene score E is scaled by a confidence factor c and then boosted towards 1.0, proportional to the number of candidate genes entered: $cE + \frac{1-cE}{N}$. The confidence factor reflects our confidence in the automatically-prioritized exome candidates, and is currently set to 0.5 (meaning the maximum score an exome candidate can receive is half that of a manually-prioritized gene).

Only whole-exome data is currently supported in PhenomeCentral, as whole-genome VCF files are too large to reliably upload through a Web browser. To address this challenge we developed a browser-based large file transfer tool (github.com/sickkids-ccm/dcc-file-transfer, but this has not yet been integrated into PhenomeCentral. When a VCF file is uploaded for a patient record, the Exomiser is automatically run to identify candidate genes. The Exomiser uses the VCF file, the HPO terms annotated in the patient record, and the mode of inheritance of the patient's disease if specified to score genes according to their phenotypic relevance and the estimated harmfulness of variants identified in exome sequencing (Figure 4.3). The highest-ranked genes are then shown to the user (Figure 4.4) and incorporated in patient matchmaking.

PhenomeCentral currently uses an improved scoring measure within the Exomiser (hiPHIVE; Bone et al., 2015; Smedley et al., 2015) for annotating and filtering exome sequence data, and for prioritizing genes by phenotypic relevance. HiPHIVE combines the model organism information of the original

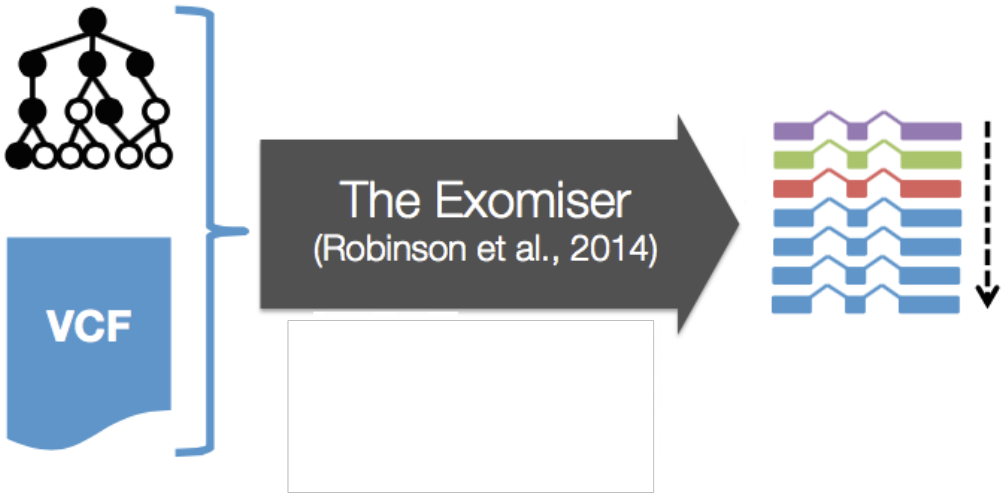

Figure 4.3: A diagram of the use of the Exomiser within PhenomeCentral. The phenotype (HPO terms) and genotype (VCF file) for a case are inputted to the Exomiser, which outputs a list of genes scored by their predicted relevance to the disease.

VARIANTS

2012.07.05.09.38.07_GenomeSub_mcgill_vcf_KB_174_81272.vcf (reference genome GRCh37)

TOP GENES IN VCF

Gene scores are computed from the uploaded VCF file and patient phenotypes using the Exomiser. Each score reflects the phenotypic relevance of the gene and the harmfulness and allele frequency of the variants.

Score	Gene	Variant Position	Variant	Effect
99%	SRCAP	chr16: 30748691 - 30748692	C → T	STOPGAIN
97%	NOTCH2	chr1: 120611964 - 120611965	G → C	MISSENSE
94%	HLA-DQB1	chr6: 32632593 - 32632594	C → T	MISSENSE
		chr6: 32632592 - 32632593	C → G	MISSENSE
90%	SCARF2	chr22: 20779975 - 20779976	- → G	FS_DUPLICATION
		chr22: 20779975 - 20779976	- → G	FS_DUPLICATION

Maximum number of genes: 4 Maximum variants per gene: 2 REFRESH HIDE OPTIONS

Figure 4.4: A screenshot of the highest scoring genes and corresponding predicted pathogenic variants, as ouputted by the Exomiser for a Floating Harbor Syndrome case within PhenomeCentral.

PHIVE score with known human gene/disease associations and protein-protein interaction data. This version was not used for benchmarking in Chapter 2.3 as it could lead to overly optimistic results, but this functionality is included in PhenomeCentral as clinicians are interested in discovering the causal gene, even if it was previously associated with a human disease.

4.1.3 Match visualization

After entering information about a case, PhenomeCentral then displays the most phenotypically similar patients and patients that share a plausible genetic mechanism.

As shown in Figure 4.1E, a breakdown of the phenotypic similarity of the two patients is shown using a greedy clusting method. Specifically, given two patients, P and Q, with corresponding induced subsets of the HPO g^P and g^Q, the ancestor term shared by both patients with the highest information content is selected: $\text{argmax}_{t \in g^P \cap g^Q} \text{IC}(t)$. This ancestor term becomes the root of a new cluster that consists of the descendant terms that were annotated in the two patients. These clustered terms are then removed, and the process repeated until the root of the next cluster is the root of the HPO (HP:0000118 [Phenotypic Abnormality]). Any terms remaining are considered "Unmatched" and displayed at the bottom.

4.1.4 Finding similar diseases

Diagnosis suggestions are also shown based on the symptoms entered into the patient record using the same matching methods as described in section 4.1.1 and the same visual interface as is used for patient matchmaking in section 4.1.2. Unlike the diagnosis suggestions within PhenoTips, this interface then enables the user to explore the breakdown of the phenotypic overlap with the particular disease. This was achieved by creating prototypical records for each OMIM disease, and then performing matchmaking against this set of records separately.

4.1.5 Case studies

There have so far been several successful matches of undiagnosed patients with rare genetic diseases enabled by PhenomeCentral, with two described in detail below. The first match involves an initially undiagnosed patient who was matched with a group of mandibulofacial dysostosis with microcephaly (MFDM) patients. A mutation identified in EFTUD2, the gene responsible for MFDM (Lines et al., 2012), confirmed the diagnosis. The patient record listed a few typical features of MFDM including microcephaly, micrognathia, and developmental delay, however many other common features were absent, including most ear abnormalities: microtia/dysplastic pinna(e) (present in 98% MFDM patients); malformations of the auditory canal, and/or middle ear with associated conductive hearing loss (77% of MFDM patients) (Lines et al., 2014). Furthermore, the patient had abnormalities of the heart and hand, but the specific features were atypical compared to those commonly reported in MFDM patients, as well as subglottic stenosis and vocal cord paralysis, which are not characteristic of MFDM. Despite this atypical presentation, PhenomeCentral matched this patient with a patient previously diagnosed with MFDM, and correctly identified EFTUD2 as the causal gene for the pair of patients.

In the second match, a pair of patients were matched together based on various overlapping abnormal phenotypes in multiple organ systems including myopathy, thrombocytopenia, and peripheral nerve conduction abnormality. Although the patients had overlapping phenotypes, the phenotypes were not

specific enough to confirm a diagnosis in either patient. The differential diagnoses included: Thrombocytopenia, X-linked, with or without dysthropoietic anemia [OMIM:300367], Quebec Platelet Disorder [OMIM:601709], and Platelet Disorder, Familial, with associated myeloid malignancy [OMIM:601399]. However, in addition to these overlapping phenotypic findings, both patients had the same mutation in the STIM1 gene, ranked as the top candidate for the pair of patients by PhenomeCentral. Follow-up studies were completed by the clinicians who contributed these patients into PhenomeCentral, and these patients were diagnosed with York Platelet Syndrome [OMIM:185070], which is characterized by thrombocytopenia, striking ultrastructural platelet abnormalities, and deficiency of platelet Ca(2+) storage in delta granules (Markello et al., 2015; Bone et al., 2015).

4.2 The Matchmaker Exchange

While PhenomeCentral was one of the first such platforms, there are now many platforms that use genotype and phenotype-driven matching algorithms to identify cases with common phenotypes and disrupted genes. However, at the time it was developed no organized system existed to facilitate the interaction between these multiple disconnected projects before the Matchmaker Exchange (MME, matchmakerexchange.org) (Philippakis et al., 2015). To coordinate these efforts and harness the collective patient data across all of the databases, groups representing rare disease repositories collaborated to launch the MME (Figure 4.5). This collaborative effort has developed a federated platform (exchange) to facilitate the identification of cases with similar phenotypic and genotypic profiles (matchmaking) through a standardized application programming interface (API) and procedural conventions. Clinicians and researches can deposit their cases in any of the connected databases and find similar cases in other databases without having to separately query each service, or deposit data in each one.

Siloing of data severely impedes the discovery of the genetic causes of rare disorders, but data sharing presents a number of legal and privacy challenges. Historically, most genetic and genomic data sharing has been accomplished through the aggregation of data in a single centralized site, such as the National Center for Biotechnology Information's (NCBI) Database of Genotypes and Phenotypes (dbGaP; Mailman et al., 2007) or other large data centers such as those employed for the International Cancer Genome Consortium (ICGC; Zhang et al., 2011) and the Cancer Genome Atlas (TCGA; Weinstein et al., 2013). An alternative approach is the use of a federated network in which multiple distributed databases are connected through APIs, whereby each database supports queries of other databases in the network (Figure 4.6). This allows each database to be autonomous with respect to its own data schema, maintain ongoing control of its own data, and continuously innovate at its own pace. This approach allows for easy data analysis given that a data holder is in complete control of the entire dataset; however, a higher regulatory burden must be overcome to allow data to be shared with another entity. In addition, users may only wish to share certain datasets with others and only under certain circumstances that can be better controlled by the use of an API to enable data access. Finally, data annotations such as phenotype are dynamic within a patient, but static in disconnected databases where they can be difficult to capture longitudinally. A federated system makes it easier to support longitudinal connections to patient phenotype and updated genomic interpretations.

It is this latter federated model that was chosen to support the MME, though some data contributors may prefer to deposit data into an existing matchmaker service for participation in the MME instead of setting up their own matchmaker. This initial approach allows each participating matchmaker service

Figure 4.5: Databases and programs that gathered to form the basis for the MME. The MME includes representatives from the founding organizations and databases supporting or intending to support matchmaker services. The MME has been identified as a demonstration project for the Global Alliance for Genomics and Health (GA4GH) and the MME has been leveraging the expertise of the GA4GH working groups for guidance on pertinent aspects of the project.

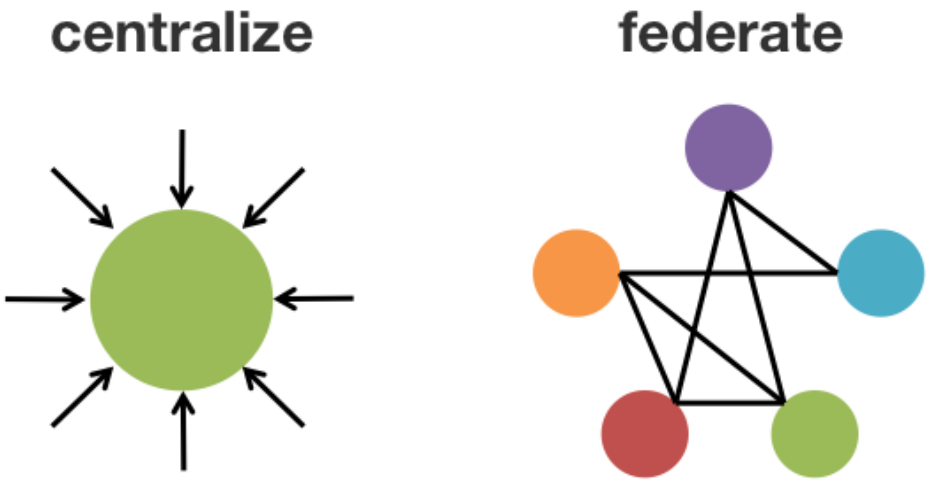

Figure 4.6: Two distinct approaches to bridging silos. The diagram on the left represents collecting datasets from a variety of sources into a centralized repository. The diagram on the right represents a federated network of autonomous services with pairwise connections.

to maintain their autonomy and primary purpose, while contributing valuable data to the MME and the genomics community. Data contributors no longer need to deposit the same datasets into multiple databases in order to find matches, and they will have more options for databases in which to deposit data, including databases in their own jurisdiction if certain regulations prohibit data from leaving a region. Also, data contributors may decide to put some cases into one database and other cases into another database depending on the focus of each database. The decision of where to start may be based upon a variety of factors as described below, including the database's supported content and algorithms for matching. However, in the MME, data contributors are discouraged from depositing the same dataset into multiple databases in order to minimize data duplication and "self-matching".

In the current MME architecture, matchmaker services connect to one another in a pair-wise fashion using an API. This pairwise architecture is sufficient for the time being because 1) there are currently a small number of functional matchmaker services, and 2) each matchmaker service collects different data for different use cases, so each service is likely to only connect with a subset of the other services. If the community continues to grow and converge on a data model, it may be useful to introduce a central exchange server to which each service connects. This hub-and-spokes model would reduce the number of necessary connections from $O(n^2)$ to $O(n)$, enable centralized collection of statistics, and simplify the on-boarding process for new members of the MME by reducing the number of keys that need to be exchanged and, potentially, the number of data transfer agreements that need to be signed. However, this proposal presents logistical challenges. Who will pay for and maintain this server? Where will the server be located? The MME has taken a completely decentralized approach until these questions need to be addressed.

4.3 The MME API

After the user has deposited a case into a matchmaker database, the database can search for matches in connected databases using the MME API (Buske et al., 2015b). Many MME partners perform some form

of internal matchmaking to identify similar patients within their database, but each organization has a different focus, collects different types of data, and stores their data in different formats. The MME API provides a standardized language for exchanging patient profiles in order to enable matchmaking between patient databases. The MME API specifies the format of both the query, which is sent to participating databases (which we call *matchmaker services*), and the response, which contains information about matching individuals in the remote database.

Developing efforts such as the Global Alliance for Genomics and Health (GA4GH) APIs are designed to facilitate the exchange of genetic data between databases; however, these are currently targeting genetic data and hypothesis-driven queries. The initial version of the API follows a *query-by-example* philosophy, in which the request is simply a description of the individual to be matched and the response is a list of the descriptions of similar individuals. Because the API is built around the description of an individual rather than a complex query language, it is easy to understand, straightforward to implement, and provides the various databases the flexibility of experimenting with matching algorithms and regulating the amount of data that is disclosed. Further, because the case is used as the query, more specific and complete case records will return more relevant matches, thus encouraging users to submit the most complete and specific case information possible.

An overview of the match request and response process is shown in Figure 4.7. The user starts by contributing a case to one of the Matchmaker Exchange services. On behalf of the user, the matchmaker service then queries other MME services using the MME API. These other services use the structured patient data in the query to identify and return descriptions of similar cases within their respective databases. They are not permitted to store request data for uses other than analytics and diagnostics (i.e., the data exchanged over the API does not become a part of the data stored by the receiving services). Similar cases found through the API are then reported to the users for evaluation. The users can then follow up with each other on any promising matches using contact information provided with the query and response. It is currently up to each MME service to define the process for alerting their respective users of the match (i.e., step 4 in Figure 4.7).

The MME API was designed to enable automated sharing of case data between patient databases. It started as a collaboration between PhenomeCentral and GeneMatcher, based on an initial PhenomeCentral patient transfer API, and grew from there to support the use cases of additional groups. The overarching principle guiding the design was to create a framework that is flexible enough to support a large number of data types and workflows, as the various members of the Matchmaker Exchange support varying depth of phenotypic and genetic data. The three members initially connected by the API had near-orthogonal data models, visibility settings, and approaches to matchmaking:

- PhenomeCentral collected HPO terms and exome sequencing data, allowed users to discover similar cases but not search them directly, and focused on displaying similar cases to the user within the website.
- GeneMatcher collected candidate genes, kept all entries completely private and unsearchable, and sent emails whenever two users entered the same gene.
- DECIPHER collected HPO terms and variant data with a focus on CNVs, made case data publicly available, and developed user interfaces for filtering and searching within the database.

The development of the MME API was an exercise in compromising these views, harmonizing the

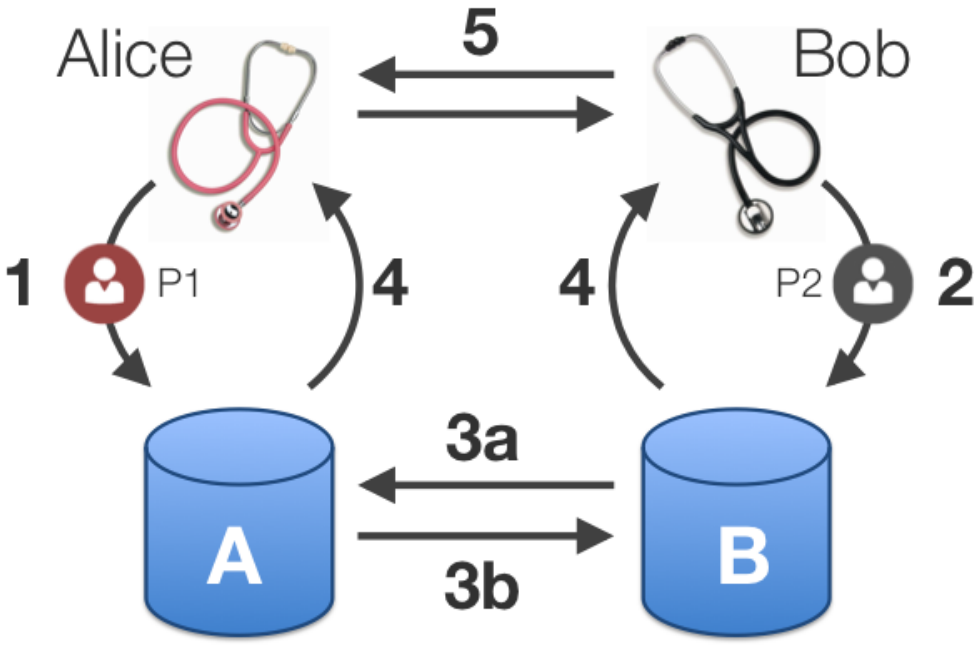

Figure 4.7: Overview of the matchmaking process, in which **1**) Alice deposits case P1 into Matchmaker A; **2**) some time later, Bob deposits a similar case P2 into Matchmaker B; **3a**) Matchmaker B then sends a match request with a description of P2 to Matchmaker A and **3b**) receives a match response with a description of similar patients (including P1) from Matchmaker A; **4**) Matchmaker A informs Alice and Matchmaker B informs Bob of the P1-P2 match; and **5**) Alice and Bob communicate if the match warrants further investigation.

data model where possible, and adding functionality to each service to support a minimum of shared functionality (e.g., candidate genes).

For the initial version of the API, we decided on a hypothesis-free approach in which the patient record defines the query and the receiving site determines how to optimally process the query. This was chosen because it was expected that the receiving database likely has the best understanding of the data available and how to use it to measure patient similarity. One added advantage of this approach is that to obtain optimum matches, the query patient has to be deeply phenotyped, thus encouraging contribution of data into the network. We believe that our approach will have utility beyond the rare disease community, and have contributed our APIs to the Global Alliance for Genomics and Health. Wherever possible, we coordinated field names and data formats with those used by the GA4GH APIs, and will continue to engage in the development of these standards.

4.3.1 Initial version

The API defines a set of data types, each with a corresponding set of properties (e.g. the Disorder type has two properties, "id", which is mandatory, and "label", which is optional). An object is a particular example (instantiation) of a type (an example Disorder object in JSON format is: {"id": "OMIM:269880", "label": "SHORT syndrome"}). The core of the format is a specification of an individual with relevant phenotypic and/or genotypic features (the Patient type, defined in Table 4.1). A match request (see Figure 4.8B) contains a single case in this format, used as the query, and the match response contains a scored list of the most similar cases in the remote system, also in this format. The Patient type is designed to be flexible to facilitate matchmaking between cases with varying degrees of phenotypic and/or genotypic detail. It can contain a list of diagnoses, phenotypic features, and/or genotypic features, along with metadata such as an identifier, sex, and contact information of the submitter of the case (so that promising matches can be followed up on). The API standardizes a small number of required fields, making it easy to implement regardless of the data stored by the matchmaker service, and many optional fields, enabling additional information to be conveyed to improve the accuracy of matchmaking and help users interpret the matches.

Table 4.1: Fields of the MME API. Example values from a patient description in (Hood et al., 2012). *The "Req" column contains a check mark for properties that are mandatory for objects of the given class. †It is preferred to have both the "features" and "genomicFeatures" properties defined for every Patient object; it is mandatory to have at least one of the two.

Type	Property	Req*	Expected Type	Description	Example
MatchRequest	patient	✓	Patient	query patient	see Fig. 2B lines 2–53 and Patient type
Patient	id	✓	string	unique, persistent patient identifier	"F0000011"
	label		string	human-readable identifier, no personally identifiable information	"174_170258"
	contact	✓	Contact	contact details for depositor of patient record	see Fig. 2B, lines 5–9 and Contact type
	species		string	NCBI taxon identifier	"NCBITaxon:9606"
	sex		string	genetic sex ("FEMALE", "MALE", "OTHER")	"FEMALE"
	ageOfOnset		string	age interval at onset of the majority of the symptoms (HPO term identifier)	"HP:0003623"
	inheritanceMode		string	mode of inheritance (HPO term identifier)	"HP:0000006"
	disorders		list of Disorders	list of diagnoses	see Fig. 2B, lines 12–17 and Disorder type
	features	†	list of Features	list of phenotypic traits	see Fig. 2B, lines 18–33 and Feature type
	genomicFeatures	†	list of GenomicFeatures	list of candidate causal genes and variants	see Fig. 2B, lines 34–52 and GenomicFeatures type
Contact	name	✓	string	name of the clinician or organization	"Kym Boycott"
	institution		string	institution of the clinician	"FORGE Canada"
	href	✓	string	contact URL; either public webpage or email address (mailto)	"http://dx.doi.org/10.1016/j.ajhg.2011.12.001"
Disorder	id	✓	string	OMIM or ORDO identifier	"MIM:136140"
	label			human-readable description	"Floating-Harbor Syndrome"
Feature	id	✓	string	HPO term identifier	"HP:0004322"
	label		string	human-readable description	"Short stature"
	observed		string	the feature has been explicitly observed ("yes") or explicitly not observed ("no")	"yes"
	ageOfOnset		string	age interval at onset (HPO term identifier)	"HP:0003577"
GenomicFeature	gene	✓	Gene	candidate gene	see Fig. 2B, lines 36–38 and Gene type
	variant		Variant	candidate variant in gene	see Fig. 2B, lines 39–45 and Variant type
	zygosity		number	allelic dosage (1: heterozygous, 2: homozygous)	1
	type		GenomicFeatureType	cDNA effect of the mutation	see Fig. 2B, lines 47–50; GenomicFeatureType type
Gene	id	✓	string	gene symbol, ensembl gene ID, or entrez gene ID	"SRCAP"
Variant	assembly	✓	string	reference assembly identifier	"GRCh37"
	referenceName	✓	string	chromosome	"16"
	start	✓	number	start position (0-based)	30748691
	end		number	end position (0-based, exclusive)	30748692
	referenceBases		string	VCF-style reference allele of at least one base	"C"
	alternateBases		string	VCF-style alternate allele of at least one base	"T"
GenomicFeatureType	id	✓	string	SO term identifier	"SO:0001587"
	label		string	human-readable description	"STOPGAIN"
MatchResponse	results	✓	list of MatchResults	list of similar/matching patients	see Fig. 2D, lines 2–10 and MatchResults type
MatchResult	score	✓	MatchScore	scoring details for the match	see Fig. 2D, lines 4–6 and MatchScore type
	patient	✓	Patient	matching patient	see Fig. 2D, line 7 and Patient type
MatchScore	patient	✓	number	overall match score (in the range [0, 1], where 0.0 is a poor match and 1.0 is a perfect match)	0.983

The sharing and automated analysis of genetic and phenotypic data has necessitated standardization using a number of ontologies and controlled terminologies. In this API, we use the Sequence Ontology (Eilbeck et al., 2005) to describe the class of the genetic variants (e.g. whether it is insertion, deletion, or SNV; missense or stopgain, etc.) and the Human Phenotype Ontology (HPO) (Köhler et al., 2014) to describe patient phenotypes. The HPO has over 11,000 terms corresponding to phenotypic abnormalities, which are structured from general (e.g. "abnormality of the nervous system") to specific (e.g. "atonic seizures"). Importantly, the HPO has the "true path rule", which states that the presence of a lower-level term implies the presence of all ancestors of the term (a patient with "atonic seizures", by definition, also has "seizures" and has an "abnormality of the nervous system"). This feature makes it possible to "obfuscate" a term by using one of its ancestors instead, and to match distinct but related terms by identifying shared ancestors.

Diagnoses are specified using OMIM (Amberger et al., 2015) or Orphanet (orphadata.org) identifiers. Each phenotypic feature (a Feature object) is specified using a term from the HPO, and can be recorded as either observed (the default) or explicitly absent (it may be important for similarity measures and differential diagnosis to know if particular features or co-morbidities were explicitly checked for but not observed in the individual). To protect privacy, phenotypic features can be intentionally obfuscated in the query or the response by substituting HPO terms with ancestors of those terms. Each genotypic feature (a GenomicFeature object) represents a candidate gene or variant believed to be directly involved in the individual's phenotype. It contains a gene identifier, specified as an HGNC gene symbol, an Ensembl gene identifier, or an Entrez gene identifier, and can include details about the type of variant (specified as a Sequence Ontology term) and/or the specific variant with respect to a reference genome. Extensive additional documentation is available on the GitHub page (github.com/ga4gh/mme-apis).

The match response (see Figure 4.8D and Table 4.1) contains a list of the cases in the database most similar to the case specified in the query, scored according to the particular matchmaker service's matching algorithm. Scores must be a number between 0.0 (a poor match) and 1.0 (an excellent match), but scores are not yet comparable across matchmaker services as matching algorithms vary. Currently, only an overall score for the strength of each match is required, but more detailed scoring of the phenotypic and genotypic aspects of each match will likely be added in future versions.

The MME API is semantically versioned (semver.org), with version numbers taking the form "X.Y", where X is incremented for major releases and Y is incremented for backwards-compatible minor releases. Every request must specify the API version within the HTTP Accept header, and the remote server must provide the API version of the response in the Content-Type header of every response (see Figure 4.8A and C).

The remote server should use HTTP status codes to report any error encountered processing the match request. Table 4.2 contains a list of status codes and their meanings with regards to this API. The error response should include a JSON-formatted body with a human-readable "message" containing further details about the error (see Figure 4.8E). The exact error message is up to the implementer, and additional fields can be provided with further information.

All communication between servers in the Matchmaker Exchange must occur over secure HTTP (HTTPS), and requests are currently authenticated through a simple yet effective protocol. If Matchmaker B wishes to accept match requests from Matchmaker A, Matchmaker B securely sends a secret authentication token to Matchmaker A (e.g. through encrypted email). We recommend the authentication token be a randomly generated SHA1 hexadecimal digest. This authentication token must be

Table 4.2: HTTP status codes and their intended use within the MME API.

HTTP Status Code	Reason Phrase	Description
200	OK	no error
400	Bad Request	missing/invalid data
401	Unauthorized	missing/invalid authentication token
405	Method Not Allowed	invalid method (POST required)
406	Not Acceptable	missing/unsupported API version
415	Unsupported Media Type	missing/invalid content type
422	Unprocessable Entity	missing/invalid request body
500	Internal Server Error	default error

specified as the X-Auth-Token header of all requests that Matchmaker A makes to Matchmaker B (see Figure 4.8A). Matchmaker B will then verify the authentication token and may perform additional checks such as validating the originating IP address of the request (though this is not required). We are currently exploring support for a federated user authentication scheme, such as OAuth 2.0 (oauth.net), in future versions of the API.

4.3.2 Testing

In order to facilitate testing the ability of systems to query, match, and respond to requests, we have compiled a standardized test dataset of 50 de-identified individuals spanning 22 disorders. These cases were selected from publications by the FORGE Canada (Beaulieu et al., 2014) and Care4Rare Canada projects (care4rare.ca), and deliberately include conditions with diverse phenotypes. Some of the conditions involve multiple organ systems (e.g., OMIM:269880 SHORT syndrome; OMIM:182212 Shprintzen-Goldberg Syndrome), while others mainly affect a single system (e.g., OMIM:614665 Meconium ileus; OMIM:243150 Intestinal atresia, multiple). In addition, multiple individuals with variable severity were included for many of the disorders (e.g., OMIM:615960 Cerebellar Dysplasia and Cysts; OMIM:615273 Congenital disorder of glycosylation, type IV), which serve as internal controls for evaluating the performance of matchmaking algorithms. These test cases are available in the MME API JSON format, and are annotated with phenotypic features, the diagnosed disorder (OMIM identifier), and the causal variant(s). New matchmaking organizations can use this dataset internally, to verify that the query and response are formatted correctly and the matching is accurate, or externally, to verify that links to other matchmaker services are functioning properly. In these cases, an additional property of the Patient object, "test", should be set to true. This informs the system being queried that the query is a test, allowing it to respond accordingly. Normally, the system being queried will match against real patient data, return any matches, and notify users of identified matches. With a test query, the system should run the match against test data, return any matches, and suppress any notifications.

4.3.3 Deployment and validation

As a pilot, three databases were connected using the MME API: DECIPHER (Chatzimichali et al., 2015), GeneMatcher (Sobreira et al., 2015), and PhenomeCentral (Buske et al., 2015a). We validated the API through two means. First, through the use of the test data, which recovered all of the expected

matches. Second, as a preliminary test with clinical cases, we used the MME API to find matches for unsolved PhenomeCentral cases within GeneMatcher.

Test dataset

Using an initial test dataset of 50 published patients compiled in collaboration by the Care for Rare Canada Project, we validated that these three connections behaved as expected and matching patients were able to be identified. We also used the MME API to identify matches within GeneMatcher for 60 unsolved PhenomeCentral cases, although this matching was done almost entirely based on candidate genes (GeneMatcher did not support phenotype matching at the time and has few cases with this type of data). Altogether, 10 matching cases were followed up on, of which 2 were real matches, 2 are still being followed up on, and 6 were deemed to be too phenotypically dissimilar to be the same disease.

Pilot dataset

We identified 60 unsolved PhenomeCentral cases submitted by the Care4Rare Canada project, which together included 45 different candidate genes (1–5 candidate genes per record). At least one match was found for 37 out of 60 PhenomeCentral cases, with 33 matching cases returned in total. Of the 33 matches, 16 were duplicate records (entered by the same clinician in both systems) and 2 were excluded because GeneMatcher had many (>= 30) candidate genes per record. We followed up on the 10 matching genes within the remaining 15 matching records, with 6 of the gene matches classified as false positives (i.e. phenotypes of the two patients were not significantly similar after clinician review), 2 of the gene matches still unresolved, and 2 of the gene matches classified as potentially significant hits with additional validation currently underway. Most of the cases do not have phenotypic information and in such cases matching was done using only the gene, which may contribute to the false positive rate of this test.

4.3.4 Implementation within PhenomeCentral

After entering a patient record into PhenomeCentral, the user can opt to have their patient record participate in the Matchmaker Exchange. This allows the case to be matched to and matched by similar cases in other repositories in the Exchange, currently GeneMatcher, DECIPHER, Monarch, and MyGene2. The MME API uses a query-by-example philosophy, where a match request consists primarily of a set of HPO terms and several genetic features (candidate genes or variants). The match request is securely sent to other sites in the MME, and each site instantly responds with a description of the most similar patients in their database and contact details to connect with the submitter of each match. After inspecting the phenotypic and genotypic evidence for each match, the user can follow up on promising matches directly.

At the same time, patient records in PhenomeCentral that participate in the MME can be discovered by users that submit cases into other sites in the MME when those sites send match requests to PhenomeCentral. In these situations, a summary of the phenotypic and genotypic profile of the similar record will be returned, allowing the user of the other site to evaluate the match and follow up if it is promising. To reduce the identifiability of the case, patient details are obfuscated before responding to the match request: phenotypes are replaced by their ancestors and only gene-level genetic data is returned (variant-level details are left out). However, this decision is currently being re-evaluated as it

creates data asymmetry with the other MME services and prevents them from being able to reproduce the similarity scores.

PhenomeCentral was developed with a focus on capturing and utilizing phenotypic information. As discussed in section 4.1.1, matching is phenotype-first, so phenotype information was initially required to perform matching. To enable matching with services in the MME with limited or no phenotype information, this matching approach was extended to support matching between cases with no phenotypic information. Gene information was indexed for each patient, and the scoring metric was altered to have a non-zero score if there is a gene match but no phenotypic similarity.

4.3.5 Reference implementation

An open-source server was developed that implements the API, to both simplify the adoption of the Matchmaker Exchange API for new databases and to augment the API documentation (Figure 4.10). The reference server conforms to current best-practices for open-source software:

- Distribution on GitHub: github.com/MatchmakerExchange/reference-server
- MIT licensed
- Continuous integration and automated builds using Travis CI
- Code coverage measurement and tracking with coverage.py and coveralls.io

Interface

Installation The package uses Python setuptools for installation, allowing the user to install the package with the command:

```
pip install -e .
```

The necessary datasets and mapping files can be downloaded and installed with the command:

```
mme-server quickstart
```

Testing The test suite uses Python unittest framework, and can be run with the command mme-server test. The test suite can also be run via the coverage.py package to calculate code coverage. For testing, the Matchmaker Exchange test dataset of 50 patients is downloaded and imported into the database. The entire testing process is automated and run on every commit and pull request via integration with Travis CI.

Importing data Patient data can be uploaded into the server via either the command-line or Python interfaces. From the command-line, patient data in a file patients.json in MME API JSON format can be indexed with the command:

```
mme-server index patients --filename patients.json
```

The file can be imported from within Python by instantiating the DatastoreConnection class found in mme_server.models, and then calling datastore.patients.index('patients.json').

Running Indexed patient data can be served over the MME API with the command:

```
mme-server start
```

Optional arguments --host and --port can be specified to restrict the host of the server or the port the server listens on, respectively.

4.3.6 Privacy-preserving extensions

The current structure of the API requires sending potentially-sensitive information about a case in order to find matches. This disclosure is not always desirable or possible, depending on consent and privacy concerns. At the BioHackathon 2016 meeting, a prototype of a privacy-preserving MME matchmaking was developed by Hiroki Sudo and Masanobu Jimbo using an inner-product extension of the Lifted-ElGamal library (github.com/cBioLab/secure-innerproduct). The query, a set of HPO terms or gene symbols, is transformed with homomorphic encryption such that fuzzy phenotype matching or exact gene matching can be performed without the query or the response being disclosed to the server. The numerator of the Jaccard score is used as the phenotype similarity score (the total number of phenotypes and ancestors in common). The performance of the implementation is practical, taking just 45 seconds plus 25 seconds per 1000 patients on the server to encrypt the query, perform matching in the encrypted space, and return an encrypted response when run on a 2010 MacBook Pro laptop.

4.4 Summary

PhenomeCentral addresses the increasing need for computational approaches to identify individuals affected by the same or overlapping phenotypes and mutations in the same gene, thereby enabling novel gene discovery. PhenomeCentral is based on the popular PhenoTips software, which makes its user interface familiar to many clinical geneticists, and also allows for the direct transfer of patient records from any other PhenoTips instance to PhenomeCentral. This enables institutional workflows (such as at the NIH Undiagnosed Diseases Program) in which clinicians use the PhenoTips software clinically, store full patient records within the institutional firewall, and then export the de-identified phenotypic records (HPO terms and additional needed demographics) to PhenomeCentral to enable matchmaking.

Since its release, PhenomeCentral has grown rapidly and now contains almost 2,000 deeply phenotyped patients with rare genetic disorders, with accounts for over 750 scientists and clinicians. Most of these patients are undiagnosed, and most have exome sequence data. The coordination with the Matchmaker Exchange also increases the number of potential matches, with the MME API enabling automated querying of other databases (such as GeneMatcher and DECIPHER) for records with same candidate genes and similar phenotype. With additional sites currently implementing the MME API, storing deep phenotype and genotype data for all patients in PhenomeCentral will help ensure the maximum potential for matchmaking for these rare disease patients.

The Matchmaker Exchange has recently expanded to include several new members with live connections using the MME API, and many more participants that are actively developing endpoints. These include multiple databases that collect phenotypic and genomic data directly from patients, and policies governing these cases are currently being discussed. The MME API is also expanding to support exome matching use cases, but the viability of these approaches will remain to be seen, and false positive rates are a significant concern. One proposal for addressing false positives is to allow nodes within the MME

full access to the (potentially limited) data from other nodes, enabling them to more accurately evaluate the significance of matches by having access to the distribution of cases across sites.

A

```
POST /baseURL/match HTTP/1.1
Host: b.org
Accept: application/vnd.ga4gh.matchmaker.v1.0+json
X-Auth-Token: 854a439d278df4283bf5498ab020336cdc416a7d
```

B

```
{
  "patient": {
    "id": "F0000011",
    "label": "174_170258",
    "contact": {
      "name": "Kym Boycott",
      "institution": "FORGE Canada",
      "href": "http://dx.doi.org/10.1016/
              j.ajhg.2011.12.001"
    },
    "sex": "FEMALE",
    "inheritanceMode": "HP:0000006",
    "disorders": [
      {
        "id": "MIM:136140",
        "label": "Floating-Harbor syndrome"
      }
    ],
    "features": [
      {
        "id": "HP:0004322",
        "label": "Short stature"
      },
      {
        "id": "HP:0000878",
        "label": "11 pairs of ribs"
      },
      {
        "id": "HP:0000369",
        "observed": "no",
        "label": "Low-set ears"
      },
      ...
    ],
    "genomicFeatures": [
      {
        "gene": {
          "id": "SRCAP"
        },
        "variant": {
          "assembly": "GRCh37",
          "referenceName": "16",
          "start": 30748691,
          "referenceBases": "C",
          "alternateBases": "T"
        },
        "zygosity": 1,
        "type": {
          "id": "SO:0001587",
          "label": "STOPGAIN"
        }
      }
    ]
  }
}
```

C

```
HTTP/1.1 200 OK
Content-Type: application/vnd.ga4gh.matchmaker.v1.1+json; charset=UTF-8
```

D

```
{
  "results": [
    {
      "score": {
        "patient": 0.94283
      },
      "patient": {...}
    },
    ...
  ]
}
```

E

```
HTTP/1.1 406 Not Acceptable
Content-Type: application/vnd.ga4gh.matchmaker.v2.2+json; charset=UTF-8
```

```
{
  "message" : "unsupported API version",
  "supportedVersions" : [ "2.0", "2.1", "2.2" ]
}
```

Figure 4.8: An example match request and response, based on a patient description in (Hood et al., 2012). **A**) The HTTP header of the POST request to a matchmaker at b.org, serving the API from baseURL. The Accept header specifies that the response should conform to version 1.0 of the MME API. The X-Auth-Token header is set to the secret token that b.org provided the querier to authenticate match requests. **B**) An example request body, describing a particular patient with Floating-Harbor Syndrome (additional features omitted for brevity). **C**) The HTTP header of a successful matchmaking response, indicated by the 200 OK status code. The Content-Type header specifies that the response conforms to version 1.1 of the MME API, which is backwards compatible with the version 1.0 query. **D**) An example response body, containing a list of matching cases and corresponding match scores (patient details and additional matches omitted for brevity). **E**) The HTTP header and body of a failed matchmaking response, in which the server does not support the API version of the query (version 1.0), and responds with an appropriate message, a Content-Type containing the latest API version supported by the server, and a list of all supported API versions (optional).

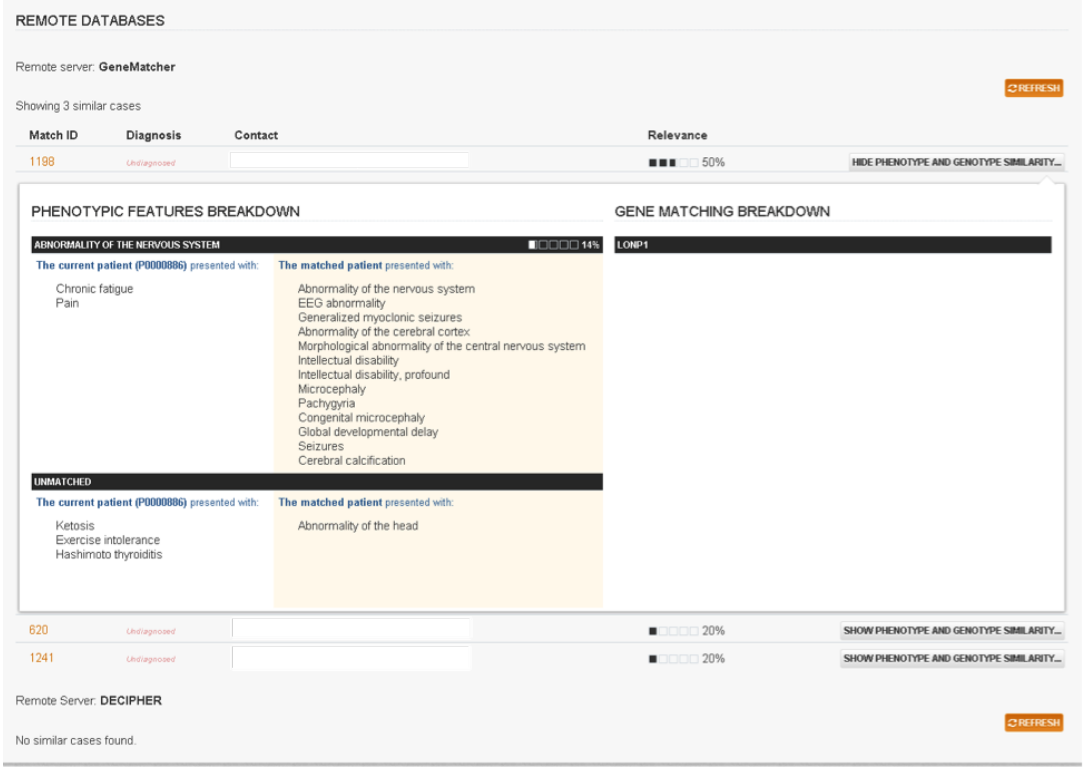

Figure 4.9: The user interface in PhenomeCentral for showing similar patients in remote databases using the Matchmaker Exchange API. Submitter details have been redacted.

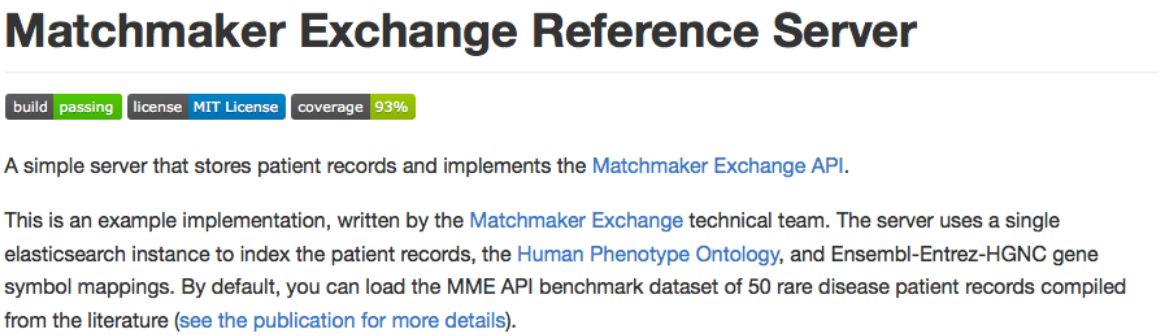

Figure 4.10: The GitHub page for the MME reference server, as shown on github.com/MatchmakerExchange/reference-server on 30 April 2016.

Chapter 5

Concluding thoughts and future work

5.1 Predicting variant harmfulness

5.1.1 Limited dataset of synonymous mutations

Our method's performance is currently limited by the small number of training examples available at the time, but as more examples of pathogenic synonymous variants are found, the model predictions may improve. Since the publication of SilVA, synonymous mutations have gained attention (Hunt et al., 2014) and variant interpretation databases such as ClinVar (Landrum et al., 2016) have become more mature. A manually curated database of deleterious synonymous mutations, dbDSM, was recently created with version 1.1 containing 1,936 entries collected from public databases and literature searches (Wen et al., 2016). Using this larger dataset for training could dramatically improve the performance of SilVA.

Additional features can also be included to address the mechanisms by which synonymous mutations can effect change, such as better measures of exon strength, the effect on RNA folding ensembles, overlaps with known transcription factor binding sites rather than just motif matching (Stergachis et al., 2013), and the scores from differential splicing prediction tools (Xiong et al., 2015).

5.1.2 Distribution of SilVA

While the source code of SilVA was distributed along with installation scripts to download necessary data files, this proved excessively burdensome for potential users. Numerous users sent emails describing difficulties setting up and running the tool on their data, despite the scripts and documentation. Based on the strategies of similar tools, such as CADD (Kircher et al., 2014) and SPANR (Xiong et al., 2015), SilVA could be made available through several additional approaches that encourage use:

1. Make available a website that runs the software on small datasets and display the results.

2. As the SilVA score only applies to synonymous SNVs, it can be pre-computed on all possible mutations. This is significantly simpler (and not much larger) than downloading all the necessary data files to run the software in the first place.

3. Release the software pre-installed within a virtual machine or Docker instance.

5.1.3 Integrated methods for variant harmfulness prediction

As approaches for the stratification of different types of variants are developed, one natural extension of this work and the work on non-synonymous variation would be to design an approach that would consider both types of variants, and attempt to classify deleterious synonymous and non-synonymous variants together.

The Shendure Lab and HudsonAlpha have since developed a unified variant harmfulness prediction framework (Kircher et al., 2014). Their tool, CADD, integrates 63 diverse genomic annotations, many of which are defined across the whole genome, to provide a deleteriousness score (the C score) for any SNV or small indel. As a predictor of variant harmfulness, they trained a support vector machine (SVM) to discriminate between alleles recently fixed in the human population (assumed benign because of their persistence despite selective pressure) and random variants (enriched for deleterious mutations because there is no selection). This innovative approach provided a large dataset with which to train these methods. There is strong correspondence between increasing CADD score and decreasing derived allele frequency, and the different types of variants are automatically stratified by CADD score, and in line with conventional wisdom (see Figure 5.1).

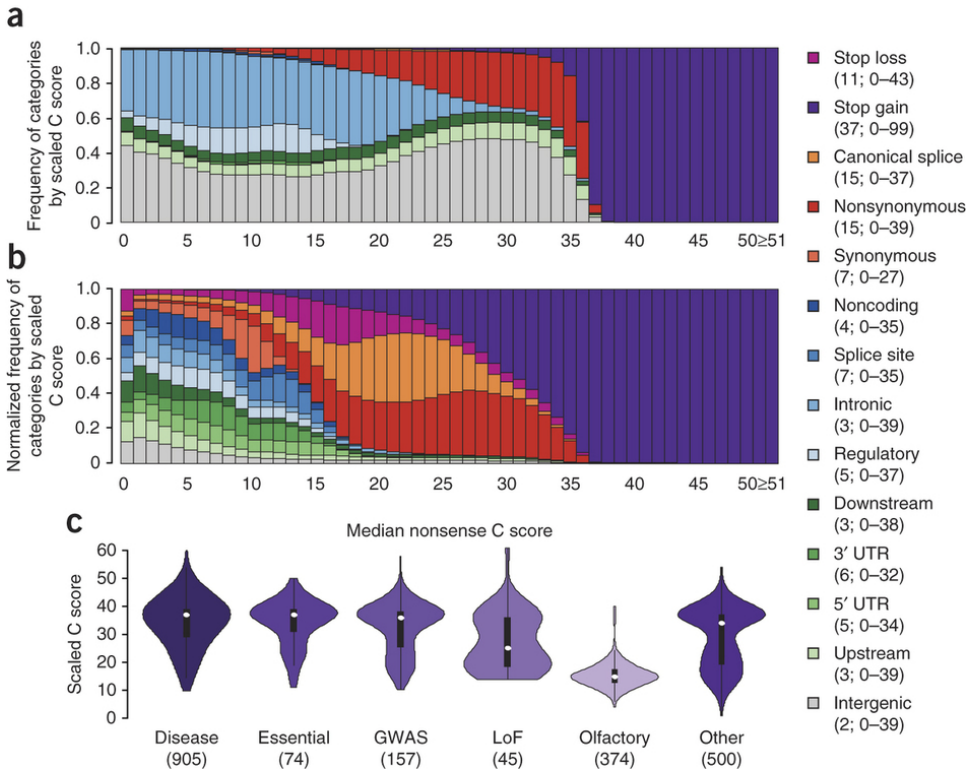

Figure 5.1: (a) the type breakdown of variants within each CADD score bin, (b) the type breakdown of variants within each CADD score bin, normalized by the frequency of each variant type, and (c) violin plots showing the CADD score distribution for various gene sets and variant types.

However, this approach assumes that alleles that have become fixed in the human population are benign, but they are likely enriched for gain-of-function mutations as well. Further, there are around 50 nonsense mutations in the exome of each healthy individual, and the number of deleterious mutations is likely in the hundreds. Even with perfect predictive models of the effect of variants on the function of

individual proteins, this is still far away from understanding the overall effect of the variant on organismal health.

5.2 Phenotypic and genotypic similarity

5.2.1 Capturing per-phenotype severity

Several PhenomeCentral users have requested the ability to indicate the importance of specific phenotypes for matching, a feature already implemented in the GeneYenta system (Gottlieb et al., 2015). In clinical practice, a specific phenotype may be extremely prominent or severe, and any promising match would be expected to display the same phenotype. While PhenomeCentral takes the frequency of a phenotype (in the OMIM corpus) into account, allowing the user to weight importance is likely to improve performance as the links between the HPO and OMIM used to compute information content are incomplete. We find our results to be similar across several different methods and corpora for computing information content, but this incompleteness in mappings can also affect the accuracy of our simulations, as well as simulations by previous authors who utilize these links to sample realistic patients.

5.2.2 Phenotype similarity scoring

There are several improvements that could be made to the phenotypic similarity score employed within PhenomeCentral.

1. The score treats all phenotypes as independent with an additive effect. In reality, co-morbidities affect the relative likelihoods of sets of terms.

2. Subtle differences between very similar diseases are much more important for a diagnosis than similar differences between less-similar diseases. It seems reasonable that the similarity score between two patients is dependent not just on the phenotypes of the two patients, but on the distance between their clinical presentation and the most similar diseases. This suggests that transforming the phenotype space into a manifold where diseases are more equidistant might aid in classification and patient similarity scoring.

3. The existing efforts have focused on pair-wise similarity metrics and gene prioritization. However, there is more power to identify cohorts of similar patients with clustering methods that leverage the distribution of similarity to other patients to control the sensitivity and specificity of matches. This was not implemented in PhenomeCentral because of the desire for a dynamic and interactive user interface that shows matches for a single case immediately. This would be difficult if global clustering needed to be performed any time any patient record changed.

4. In addition to annotating phenotypic traits observed in a patient, PhenoTips and PhenomeCentral enable users to specify absent traits, especially those that might be expected based on comorbidities and were explicitly looked for but not observed. These absent traits are not currently used within the algorithms for patient matchmaking or diagnosis prediction, but the semantic similarity scores could be extended to support this information.

5.2.3 Visualizing phenotype similarity

PhenomeCentral displays phenotypic similarity using a greedy approach that does not align with the way many users think about patient similarity. In particular, there are two situations in which the current method produces poor results.

1. If patients are annotated with phenotype terms that are seldom associated with diseases, they receive a high information content, and are therefore shown at the top of the similarity table. This frequently occurs with terms that are not associated with diseases because they are irrelevant, such as "G-Tube feeding in infancy".

2. Because the matching is greedy, the "Unmatched" terms at the bottom are not a good representation of the phenotypes that occur in one patient and not the other. For example, if two patients have the same term, "Broad thumb," and only one has a related term, "Broad fingertip," the latter will appear as "Unmatched" even though both patients share broad finger phenotypes. Assessing meaningful phenotypic overlap requires matching at an appropriate granularity.

One solution may be to display the phenotypes of each patient, grouped at a particular level of granularity within the HPO such as by organ system (the direct children of the root phenotypic abnormality term). This approach is used to compare phenotypic summaries of rare diseases in a new tool, Phenotate (phenotate.org). Methods for visualizing the phenotypic similarity between two patients, such as the recently developed PhenoBlocks tool (Glueck et al., 2016) shown in Figure 5.2, are also promising direction of future work.

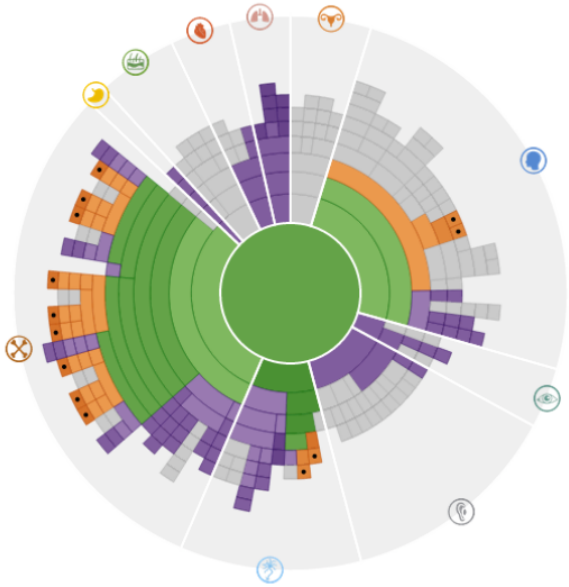

Figure 5.2: A PhenoBlocks visualization of the phenotypic overlap between two patients (image from Glueck et al. 2016). Different organ systems are represented as sectors of the circle, with more general terms in the middle and specific terms at the edges. Terms shared by both patients are colored green and terms in one patient or the other are colored orange and purple.

5.2.4 Recent efforts in cohort-free matching

Since the publication of PhenomeCentral, Akawi et al. (2015) published a probabilistic method for cohort-free rare disease gene discovery. They identified four new recessive diseases from a dataset of over 4,000 families with exome sequencing. Rather than the phenotype-first approach taken by PhenomeCentral, Akawi et al. adopt a gene-first approach to define the potential cohort. After filtering for multiple rare, predicted-pathogenic variants consistent with a recessive mode of inheritance, they are left with an average of just 3.2 variants per proband. For each gene, the probability was calculated of sampling a cohort with the observed phenotypic similarity and variant consequences by chance. These probabilities were combined with Fisher's method and genes with high significance investigated further. This approach to cohort-free evaluation of genotype-phenotype association is incredibly powerful and PhenomeCentral could benefit from attempting these methods, but it will remain to be seen how the performance translates to dominant disorders or higher rates of artifactual variants.

5.3 Extending the MME API

5.3.1 General improvements

While this API has proven successful for the first iteration of matchmaking, we are actively developing extensions to improve the efficacy of the API. These include improvements to the security/privacy configurations and a gradual adoption of hypothesis-driven queries. We believe that two changes could enhance the privacy protections offered by the MME API. First, some MME sites currently apply obfuscation to the provided data before returning it, and require direct communication between the submitting users before showing full patient data. Currently the API does not support reporting when data has been obfuscated; however this information may be useful for the receiving user. Secondly, a centralized identification framework, using a technology such as OpenID, would enable users to have a single sign-on for all of the MME partners, as well as allowing the receiving site to make decisions on what data to show in response to a query based on the user's profile and their membership in the receiving site.

Finally we expect the current hypothesis-free nature of the API to develop into a partially hypothesis-driven approach. Towards this end the API should allow for weighing or requiring of features (e.g. specifying a specific gene or phenotype as "required", suggesting a scoring function to be applied when computing a match score, or filtering the results based on a feature). We have found increasing need for such features as the scoring schemes differ significantly between matchmaker services, making expected results difficult to validate.

5.3.2 Towards a Genome Query Language

The current version of the MME API focuses on supporting 2-sided hypothesis matching, the use case where both the query and the matching patient have a candidate gene that has been identified. However, many databases, including PhenomeCentral, RD-Connect, GENESIS, Broad RDAP, and Geno2MP have cases with exome or whole-genome data for which candidate genes have not been identified. To identify matches in and between these databases, we either need highly accurate automated gene prioritization methods or functionality to dynamically filter exomes for variants in particular genes that meet criteria

for allele frequency, variant effect, and harmfulness prediction. The former option is outside of the scope of the MME, so the latter option has been pursued.

In developing this functionality, it became apparent that the first version of the MME API needed to be updated to be more extensible. New fields could be added to the patient object, but it quickly becomes unclear how these fields interact with other fields for matching purposes. Further, in order to match against exome data, variant filters need to be specified with the query. This use case implies that a user would view the results and update the filters interactively, a significant departure from the query-by-example approach initially taken with the API design. Supporting this use case resulted in a proposal to refactor the various fields and data types in the API into separate, modular components.

A `genome` component is proposed to enable matching on dynamically-filtering genome sequencing data. A query for this component consists of:

- A list of filters for selecting rare, predicted-pathogenic variants in particular genes. If multiple filters are provided, only variants that pass all filters are matched.
- One or more possible modes of inheritance that must be consistent with the number and types of variants in the gene.

Each variant is annotated with a number of attributes, including its position (chromosome, start, end), consequence information (gene, consequence), and population data (alleleFrequency). Each annotation has a data type, which defines which filters are supported for that annotation. The following annotation types are supported:

- integer
  - operator: EQ (default), NEQ, LT, LTE, GT, GTE
  - single value (value)
- float
  - operator: LTE, GTE
  - single value (value)
- nominal/categorical/ontological
  - operator:
  - single value (term):
    - EQ (default): Match must contain the term or a descendant of the term
  - multiple values (terms): LIKE (default), ANY, ALL
    - ANY: For at least one of the terms, the term or a descendant of that term must be present in the match
    - ALL: For every term, the term or a descendant of that term must be present in the match

The following fields are proposed to be added to the API:

inheritanceMode contains a field terms with a list of HPO terms for modes of inheritance (e.g., autosomal dominant, autosomal recessive, X-linked, sporadic). Matches should have genes with variants that are consistent with at least one of the included modes of inheritance. For example, if the inheritanceMode is AR ("HP:0000007"), matching cases must have at least 2 variant alleles that pass the filters (2 heterozygous variants or 1 homozygous variant).

filters a list of filters, each with one or more of the following subfields (depending on the annotation):

annotation

source (used by many annotations)

operator

population (used by alleleFrequency annotation)

value|term|terms (depending on number and type of annotation)

The following annotations are currently included in the proposal for filtering exome data on the receiving site:

referenceName nominal

start integer

end integer

size integer

id nominal; dbSNP identifier (or other)

gene nominal; additional fields:

source Ensembl, RefSeq, Entrez, UCSC

terms list of gene terms

consequence nominal (SO term); additional fields:

source VEP, ANNOVAR, Jannovar

terms list of SO terms

alleleFrequency float; additional fields:

source 1000GP, ESP5600, ExAC, local_db_name

population ALL, CEU, ...

score float; additional fields:

source SIFT, PolyPhen2, MutationTaster, CADD

For example, to match against cases with 2 or more rare (AF < 0.01) harmful (missense or stopgain) variants in NGLY1 or TTN:

```
"query": {
        "components": {
                "genome": {
                        "inheritanceMode": {
                                "terms": [
                                        {"id": "HP:0000007", "label": "Autosomal recessive"}
                                ]
                        }
                        "filters": [
                                {
                                        "annotation": "gene"
                                        "source": "Ensembl"
                                        "operator": "ANY"
                                        "terms": [
                                                {"id": "ENSG...", "label": "EFTUD2"}
                                                {"id": "ENSG...", "label": "TTN"}
                                        ]
                                }
                                {
                                        "annotation": "alleleFrequency"
                                        "source": "ExAC"
                                        "population": "ALL"
                                        "operator": "LT"
                                        "value": 0.01
                                }
                                {
                                        "annotation": "consequence"
                                        "source": "VEP"
                                        "operator": "ANY"
                                        "terms": [
                                                {"id": "SO:...", "label": "Stopgain"}
                                                {"id": "SO:...", "label": "Missense"}
                                        ]
                                }
                        ]
                }
        }
}
```

This capability is currently being added to the MME API, with a pilot deployed at RD-Connect and under development at GENESIS. The API for this functionality was developed in collaboration with the Beacon Network (beacon-network.org), the GA4GH, and Café Variome.

5.4 Next steps: patient-led matchmaking

While PhenomeCentral and the Matchmaker Exchange help clinicians and researchers connect with other specialists with similar cases, they have extremely limited time to pursue these avenues. However,

patients with rare diseases, family members, and caregivers are often extremely motivated to find information related to their disease and follow leads that might result in a diagnosis (Lambertson et al., 2015). In fact, patient-led efforts have already led to the discovery of new rare disease genes (Chong et al., 2015b).

The phenotypic matchmaking approaches used by PhenomeCentral can be repurposed to directly benefit patients and families by helping connect patients to other similar patients. This can help users find a community of similar people, even before they have a diagnosis. Further, by connecting such a system to the MME, it's possible to help clinicians and researchers find cohorts of patients they might not otherwise discover. Several existing databases, including GenomeConnect, MyGene2, and PEER, collect self-reported phenotype data for matchmaking, but none of them use this information to help patients build a support network around their condition.

Towards this end, we created a prototype of a website, PatientKind.org, and invited feedback from an initial set of ~25 users (patients, family members, or caregivers) and ~25 advocacy organizations (Figure 5.3). We are currently redesigning the site and adding functionality based on the feedback we received.

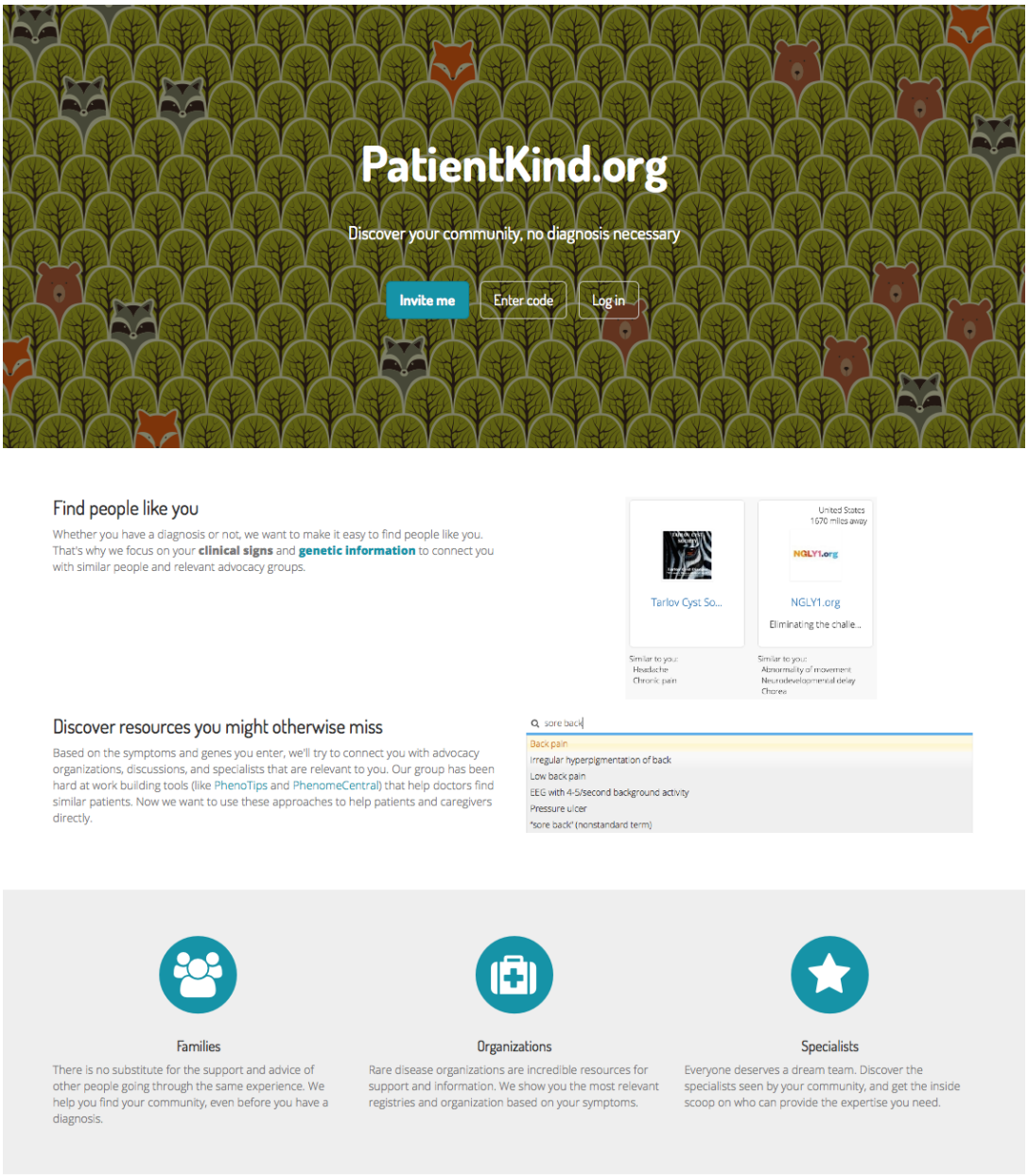

Figure 5.3: A screenshot of the landing page of PatientKind.org, a website for people with rare diseases to find a community based on their symptoms.

Bibliography

Adzhubei, I., Schmidt, S., Peshkin, L., Ramensky, V., Gerasimova, A., Bork, P., Kondrashov, A., & Sunyaev, S. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), 248–249.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.-C., De Moor, B., Marynen, P., Hassan, B., et al. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5), 537–544.

Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A. F., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T. W., et al. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4,125 families. *Nature genetics*, 47(11), 1363–1369.

Akle, S., Chun, S., Jordan, D. M., & Cassa, C. A. (2015). Mitigating false-positive associations in rare disease gene discovery. *Human mutation*, 36(10), 998–1003.

Akli, S., Chelly, J., Mezard, C., Gandy, S., Kahn, A., & Poenaru, L. (1990). A "G" to "A" mutation at position-1 of a 5' splice site in a late infantile form of Tay-Sachs disease. *Journal of Biological Chemistry*, 265(13), 7324.

Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11(8), 559–571.

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1), D789–D798.

Amstel, J., Bergman, A., Beurden, E., Roijers, J., Peelen, T., Berg, I., Poll-The, B., Kvittingen, E., & Berger, R. (1996). Hereditary tyrosinemia type 1: novel missense, nonsense and splice consensus mutations in the human fumarylacetoacetate hydrolase gene; variability of the genotype-phenotype relationship. *Human genetics*, 97(1), 51–59.

Archer, K. J. & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4), 2249–2260.

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. *Nature genetics*, 25(1), 25–29.

Auricchio, A., Griseri, P., Carpentieri, M., Betsos, N., Staiano, A., Tozzi, A., Priolo, M., Thompson, H., Bocciardi, R., Romeo, G., et al. (1999). Double heterozygosity for a RET substitution interfering with splicing and an EDNRB missense mutation in Hirschsprung disease. *American journal of human genetics*, 64(4), 1216.

Baird, P. A., Anderson, T. W., Newcombe, H. B., & Lowry, R. (1988). Genetic disorders in children and young adults: a population study. *American journal of human genetics*, 42(5), 677.

Barash, Y., Blencowe, B. J., & Frey, B. J. (2010a). Model-based detection of alternative splicing signals. *Bioinformatics*, 26(12), i325–i333.

Barash, Y., Calarco, J., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B., & Frey, B. (2010b). Deciphering the splicing code. *Nature*, 465(7294), 53–59.

Bartoszewski, R., Jablonsky, M., Bartoszewska, S., Stevenson, L., Dai, Q., Kappes, J., Collawn, J., & Bebok, Z. (2010). A synonymous single nucleotide polymorphism in δF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *Journal of Biological Chemistry*, 285(37), 28741–28748.

Bauer, S., Köhler, S., Schulz, M. H., & Robinson, P. N. (2012). Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics*, 28(19), 2502–2508.

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., et al. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *The American Journal of Human Genetics*, 94(6), 809–817.

Bone, W. P., Washington, N. L., Buske, O. J., Adams, D. R., Davis, J., Draper, D., Flynn, E. D., Girdea, M., Godfrey, R., Golas, G., et al. (2015). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genetics in Medicine*.

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10), 681–691.

Buske, O. J., Girdea, M., Dumitriu, S., Gallinger, B., Hartley, T., Trang, H., Misyura, A., Friedman, T., Beaulieu, C., Bone, W. P., et al. (2015a). PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Human Mutation*, 36(10), 931–940.

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., & Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. *Bioinformatics*, 29(15), 1843–1850.

Buske, O. J., Schiettecatte, F., Hutton, B., Dumitriu, S., Misyura, A., Huang, L., Hartley, T., Girdea, M., Sobreira, N., Mungall, C., et al. (2015b). The Matchmaker Exchange API: Automating patient matching through the exchange of structured phenotypic and genotypic profiles. *Human mutation*, 36(10), 922–927.

Cartegni, L., Chew, S., & Krainer, A. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature Reviews Genetics*, 3(4), 285–298.

Chamary, J., Parmley, J., & Hurst, L. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, 7(2), 98–108.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 27:1–27:27.

Chao, H., Hsiao, K., & Su, T. (2001). A silent mutation induces exon skipping in the phenylalanine hydroxylase gene in phenylketonuria. *Human genetics*, 108(1), 14–19.

Chatzimichali, E. A., Brent, S., Hutton, B., Perrett, D., Wright, C. F., Bevan, A. P., Hurles, M. E., Firth, H. V., & Swaminathan, G. J. (2015). Facilitating collaboration in rare genetic disorders through effective matchmaking in DECIPHER. *Human mutation*, 36(10), 941–949.

Chen, C.-K., Mungall, C. J., Gkoutos, G. V., Doelken, S. C., Köhler, S., Ruef, B. J., Smith, C., Westerfield, M., Robinson, P. N., Lewis, S. E., et al. (2012). MouseFinder: candidate disease genes from mouse phenotype data. *Human Mutation*, 33(5), 858–866.

Chen, W., Kubota, S., Teramoto, T., Nishimura, Y., Yonemoto, K., & Seyama, Y. (1998). Silent nucleotide substitution in the sterol 27-hydroxylase gene (CYP 27) leads to alternative Pre-mRNA splicing by activating a cryptic 5' splice site at the mutant codon in cerebrotendinous xanthomatosis patients. *Biochemistry*, 37(13), 4420–4428.

Chilamakuri, C. S. R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., & Meza-Zepeda, L. A. (2014). Performance comparison of four exome capture systems for deep sequencing. *BMC genomics*, 15(1), 1.

Chong, J. X., Buckingham, K. J., Jhangiani, S. N., Boehm, C., Sobreira, N., Smith, J. D., Harrell, T. M., McMillin, M. J., Wiszniewski, W., Gambin, T., et al. (2015a). The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *The American Journal of Human Genetics*, 97(2), 199–215.

Chong, J. X., Yu, J.-H., Lorentzen, P., Park, K. M., Jamal, S. M., Tabor, H. K., Rauch, A., Saenz, M. S., Boltshauser, E., Patterson, K. E., et al. (2015b). Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. *Genetics in Medicine*.

Consortium, E. P. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.

Cooper, G. & Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics*, 12(9), 628–640.

Cooper, G., Stone, E., Asimenos, G., Green, E., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome research*, 15(7), 901–913.

Cortazzo, P., Cerveñansky, C., Marín, M., Reiss, C., Ehrlich, R., & Deana, A. (2002). Silent mutations affect in vivo protein folding in Escherichia coli. *Biochemical and biophysical research communications*, 293(1), 537–541.

DasGupta, A. (2005). The matching, birthday and the strong birthday problem: a contemporary review. *Journal of Statistical Planning and Inference*, 130(1), 377–389.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol*, 6(12), e1001025.

De Meirleir, L., Lissens, W., Benelli, C., Ponsot, G., Desguerre, I., Marsac, C., Rodriguez, D., Saudubray, J., Poggi, F., Liebaers, I., et al. (1994). Aberrant splicing of exon 6 in the pyruvate dehydrogenase-E1 alpha mRNA linked to a silent mutation in a large family with Leigh's encephalomyelopathy. *Pediatric research*, 36(6), 707.

Drögemüller, C., Reichart, U., Seuberlich, T., Oevermann, A., Baumgartner, M., Boghenbor, K., Stoffel, M., Syring, C., Meylan, M., Müller, S., et al. (2011). An unusual splice defect in the mitofusin 2 gene (MFN2) is associated with degenerative axonopathy in Tyrolean grey cattle. *PloS one*, 6(4), e18931.

D'Souza, I., Poorkaj, P., Hong, M., Nochlin, D., Lee, V., Bird, T., & Schellenberg, G. (1999). Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proceedings of the National Academy of Sciences*, 96(10), 5598.

Durbin, R., Altshuler, D., Abecasis, G., Bentley, D., Chakravarti, A., Clark, A., Collins, F., Francisco, M., Donnelly, P., Egholm, M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–1073.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5), R44.

Eng, L., Coutinho, G., Nahas, S., Yeo, G., Tanouye, R., Babaei, M., Doerk, T., Burge, C., & Gatti, R. A. (2004). Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: Maximum entropy estimates of splice junction strengths. *Human Mutation*, 23(1), 67–76.

Faa, V., Coiana, A., Incani, F., Costantino, L., Cao, A., & Rosatelli, M. (2010). A synonymous mutation in the CFTR gene causes aberrant splicing in an Italian patient affected by a mild form of cystic fibrosis. *The Journal of Molecular Diagnostics: JMD*, 12(3), 380.

Fahsold, R., Hoffmeyer, S., Mischung, C., Gille, C., Ehlers, C., Kücükceylan, N., Abdel-Nour, M., Gewies, A., Peters, H., Kaufmann, D., et al. (2000). Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. *The American Journal of Human Genetics*, 66(3), 790–818.

Ferrari, S., Giliani, S., Insalaco, A., Al-Ghonaium, A., Soresina, A. R., Loubser, M., Avanzini, M. A., Marconi, M., Badolato, R., Ugazio, A. G., & et al. (2001). Mutations of CD40 gene cause an autosomal recessive form of immunodeficiency with hyper IgM. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22), 12614–12619.

Frayling, T. (2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Reviews Genetics*, 8(9), 657–662.

Gatto, F. & Breathnach, R. (1995). A Crouzon syndrome synonymous mutation activates a 5' splice site within the IIIC exon of the FGFR2 gene. *Genomics*, 27(3).

Gilad, S., Chessa, L., Khosravi, R., Russell, P., Galanty, Y., Piane, M., Gatti, R. A., Jorgensen, T. J., Shiloh, Y., & Bar-Shira, A. (1998). Genotype-phenotype relationships in ataxia-telangiectasia and variants. *The American Journal of Human Genetics*, 62(3), 551–561.

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., Chitayat, D., Faghfoury, H., Meyn, M. S., Ray, P. N., et al. (2013). Phenotips: Patient phenotyping software for clinical and research use. *Human Mutation*, 34(8), 1057–1065.

Glueck, M., Hamilton, P., Chevalier, F., Breslav, S., Khan, A., Wigdor, D., & Brudno, M. (2016). PhenoBlocks: Phenotype comparison visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1), 101–110.

Gottlieb, M. M., Arenillas, D. J., Maithripala, S., Maurer, Z. D., TarailoGraovac, M., Armstrong, L., Patel, M., Karnebeek, C., & Wasserman, W. W. (2015). GeneYenta: A phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation. *Human mutation*, 36(4), 432–438.

Griseri, P., Bourcier, C., Hieblot, C., Essafi-Benkhadir, K., Chamorey, E., Touriol, C., & Pagès, G. (2011). A synonymous polymorphism of the Tristetraprolin (TTP) gene, an AU-rich mRNA-binding protein, affects translation efficiency and response to Herceptin treatment in breast cancer patients. *Human Molecular Genetics*, 20(23), 4556–4568.

Guo, M. H., Dauber, A., Lippincott, M. F., Chan, Y.-M., Salem, R. M., & Hirschhorn, J. N. (2016). Determinants of power in gene-based burden testing for monogenic disorders. *The American Journal of Human Genetics*, 99(3), 527–539.

Halvorsen, M., Martin, J., Broadaway, S., & Laederach, A. (2010). Disease-associated mutations that alter the RNA structural ensemble. *PLoS Genetics*, 6(8), e1001074.

Hehir-Kwa, J., Wieskamp, N., Webber, C., Pfundt, R., Brunner, H., Gilissen, C., de Vries, B., Ponting, C., & Veltman, J. (2010). Accurate distinction of pathogenic from benign CNVs in mental retardation. *PLoS computational biology*, 6(4), e1000752.

Hellmann, I., Zöllner, S., Enard, W., Ebersberger, I., Nickel, B., & Pääbo, S. (2003). Selection on human genes as revealed by comparisons to chimpanzee cdna. *Genome research*, 13(5), 831–837.

Hellwinkel, O. J.-C., Holterhus, P.-M., Struve, D., Marschke, C., Homburg, N., & Hiort, O. (2001). A unique exonic splicing mutation in the human androgen receptor gene indicates a physiologic relevance of regular androgen receptor transcript variants. *Journal of Clinical Endocrinology and Metabolism*, 86(6), 2569–2575.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22), 10915–10919.

Ho, P., Huang, M., Fwu, V., Lin, S., Hsiao, K., & Su, T. (2008). Simultaneous assessment of the effects of exonic mutations on RNA splicing and protein functions. *Biochemical and biophysical research communications*, 373(4), 515–520.

Ho, P., Kuhn, J., Gerbing, R., Pollard, J., Zeng, R., Miller, K., Heerema, N., Raimondi, S., Hirsch, B., Franklin, J., et al. (2011). WT1 synonymous single nucleotide polymorphism rs16754 correlates with higher mRNA expression and predicts significantly improved outcome in favorable-risk pediatric acute myeloid leukemia: a report from the Children's Oncology Group. *Journal of Clinical Oncology*, 29(6), 704.

Hoehndorf, R., Schofield, P. N., & Gkoutos, G. V. (2011). PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39(18), e119–e119.

Hood, R. L., Lines, M. A., Nikkel, S. M., Schwartzentruber, J., Beaulieu, C., Nowaczyk, M. J., Allanson, J., Kim, C. A., Wieczorek, D., Moilanen, J. S., et al. (2012). Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *The American Journal of Human Genetics*, 90(2), 308–313.

Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., & Kimchi-Sarfaty, C. (2014). Exposing synonymous mutations. *Trends in Genetics*, 30(7), 308–321.

Imamura, T., Okano, Y., Shintaku, H., Hase, Y., & Isshiki, G. (1999). Molecular characterization of 6-pyruvoyl-tetrahydropterin synthase deficiency in Japanese patients. *Journal of human genetics*, 44(3), 163–168.

Javed, A., Agrawal, S., & Ng, P. C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. *Nature methods*, 11(9), 935–937.

Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc of 10th International Conference on Research in Computational Linguistics, ROCLINGâĂŹ97*: Citeseer.

Jin, Y., Dietz, H., Montgomery, R., Bell, W., McIntosh, I., Coller, B., & Bray, P. (1996). Glanzmann thrombasthenia. cooperation between sequence variants in cis during splice site selection. *Journal of Clinical Investigation*, 98(8), 1745.

Kanno, H., Fujii, H., Wei, D., Chan, L., Hirono, A., Tsukimoto, I., & Miwa, S. (1997). Frame shift mutation, exon skipping, and a two-codon deletion caused by splice site mutations account for pyruvate kinase deficiency. *Blood*, 89(11), 4213.

Khaddour, R., Smith, U., Baala, L., Martinovic, J., Clavering, D., Shaffiq, R., Ozilou, C., Cullinane, A., Kyttälä, M., Shalev, S., et al. (2007). Spectrum of MKS1 and MKS3 mutations in Meckel syndrome: a genotype-phenotype correlation. *Human Mutation*, 28(5), 523–524.

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., et al. (2013). Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science*, 342(6154), 1235587.

Kimchi-Sarfaty, C., Oh, J., Kim, I., Sauna, Z., Calcagno, A., Ambudkar, S., & Gottesman, M. (2007). A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, 315(5811), 525–528.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315.

Klima, H., Ullrich, K., Aslanidis, C., Fehringer, P., Lackner, K., & Schmitz, G. (1993). A splice junction mutation causes deletion of a 72-base exon from the mRNA for lysosomal acid lipase in a patient with cholesteryl ester storage disease. *Journal of Clinical Investigation*, 92(6), 2713.

Knobe, K. E., Sjörin, E., & Ljung, R. C. R. (2008). Why does the mutation G17736A/Val107Val (silent) in the F9 gene cause mild haemophilia B in five Swedish families? *Haemophilia*, 14(4), 723–728.

Köhler, 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(D1), D966–D974.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., & Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4), 457–464.

Kohonen-Corish, M., Ross, V., Doe, W., Kool, D., Edkins, E., Faragher, I., Wijnen, J., Khan, P., Macrae, F., & St John, D. (1996). RNA-based mutation screening in hereditary nonpolyposis colorectal cancer. *American journal of human genetics*, 59(4), 818.

Komar, A., Lesnik, T., & Reiss, C. (1999). Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *Febs Letters*, 462(3), 387–391.

Krawitz, P., Buske, O., Zhu, N., Brudno, M., & Robinson, P. N. (2015). The genomic birthday paradox: How much is enough? *Human mutation*, 36(10), 989–997.

Kudla, G., Murray, A., Tollervey, D., & Plotkin, J. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *science*, 324(5924), 255–258.

Lambertson, K. F., Damiani, S. A., Might, M., Shelton, R., & Terry, S. F. (2015). Participant-driven matchmaking in the genomic era. *Human Mutation*, 36(10), 965–973.

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(6822), 860–921.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1), D862–D868.

Lazaridis, K. N., Schahl, K. A., Cousin, M. A., Babovic-Vuksanovic, D., Riegert-Johnson, D. L., Gavrilova, R. H., McAllister, T. M., Lindor, N. M., Abraham, R. S., Ackerman, M. J., et al. (2016). Outcome of whole exome sequencing for diagnostic odyssey cases of an individualized medicine clinic: the mayo clinic experience. In *Mayo Clinic Proceedings*, volume 91 (pp. 297–307).: Elsevier.

Li, B. & Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3), 311–321.

Lines, M., Hartley, T., & Boycott, K. (2014). Mandibulofacial dysostosis with microcephaly. In R. A. Pagon, M. P. Adam, H. H. Ardinger, S. E. Wallace, & others (Eds.), *GeneReviews*. University of Washington, Seattle.

Lines, M. A., Huang, L., Schwartzentruber, J., Douglas, S. L., Lynch, D. C., Beaulieu, C., Guion-Almeida, M. L., Zechi-Ceide, R. M., Gener, B., Gillessen-Kaesbach, G., et al. (2012). Haploinsufficiency of a spliceosomal gtpase encoded by eftud2 causes mandibulofacial dysostosis with microcephaly. *The American Journal of Human Genetics*, 90(2), 369–377.

Liu, W., Qian, C., Francke, U., et al. (1997). Silent mutation induces exon skipping of fibrillin-1 gene in Marfan syndrome. *Nature Genetics*, 16(4), 328.

Llewellyn, D., Scobie, G., Urquhart, A., Whatley, S., Roberts, A., Harrison, P., & Elder, G. (1996). Acute intermittent porphyria caused by defective splicing of porphobilinogen deaminase RNA: a synonymous codon mutation at-22 bp from the 5' splice site causes skipping of exon 3. *Journal of medical genetics*, 33(5), 437.

Lopes, M. C., Joyce, C., Ritchie, G. R., John, S. L., Cunningham, F., Asimit, J., & Zeggini, E. (2012). A combined functional annotation score for non-synonymous variants. *Human Heredity*, 73(1), 47–51.

Lorenz, R., Bernhart, S. H., Zu Siederdissen, C. H., Tafer, H., Flamm, C., Stadler, P. F., Hofacker, I. L., et al. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1), 26.

Lorson, C., Hahnen, E., Androphy, E., & Wirth, B. (1999). A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proceedings of the National Academy of Sciences*, 96(11), 6307.

Loucks, C. M., Parboosingh, J. S., Shaheen, R., Bernier, F. P., McLeod, D. R., Seidahmed, M. Z., Puffenberger, E. G., Ober, C., Hegele, R. A., Boycott, K. M., et al. (2015). Matching two independent cohorts validates DPH1 as a gene responsible for autosomal recessive intellectual disability with short stature, craniofacial, and ectodermal anomalies. *Human mutation*, 36(10), 1015–1019.

Lupski, J., Reid, J., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D., et al. (2010). Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *New England Journal of Medicine*, 362(13), 1181–1191.

Macaya, D., Katsanis, S., Hefferon, T., Audlin, S., Mendelsohn, N., Roggenbuck, J., & Cutting, G. (2009). A synonymous mutation in TCOF1 causes Treacher Collins syndrome due to mis-splicing of a constitutive exon. *American Journal of Medical Genetics Part A*, 149(8), 1624–1627.

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature genetics*, 39(10), 1181–1186.

Majewski, J., Schwartzentruber, J., Caqueret, A., Patry, L., Marcadier, J., Fryns, J., Boycott, K., Ste-Marie, L., McKiernan, F., Marik, I., et al. (2011a). Mutations in NOTCH2 in families with Hajdu-Cheney syndrome. *Human Mutation*.

Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A., & Jabado, N. (2011b). What can exome sequencing do for you? *Journal of Medical Genetics*, 48(9), 580–589.

Markello, T., Chen, D., Kwan, J. Y., Horkayne-Szakaly, I., Morrison, A., Simakova, O., Maric, I., Lozier, J., Cullinane, A. R., Kilo, T., et al. (2015). York platelet syndrome is a CRAC channelopathy due to gain-of-function mutations in STIM1. *Molecular genetics and metabolism*, 114(3), 474–482.

Markham, N. R. & Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. *Methods in Molecular Biology*, 453, 3–31.

McDonald, J., Kreitman, M., et al. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, 351(6328), 652–654.

Mendez, M., Sorkin, L., Rossetti, M., Astrin, K., Batlle, A., Parera, V., Aizencang, G., & Desnick, R. (1998). Familial porphyria cutanea tarda: characterization of seven novel uroporphyrinogen decarboxylase mutations and frequency of common hemochromatosis alleles. *The American Journal of Human Genetics*, 63(5), 1363–1375.

Mirzaa, G. M., Conway, R. L., Gripp, K. W., Lerman-Sagie, T., Siegel, D. H., deVries, L. S., Lev, D., Kramer, N., Hopkins, E., Graham, J. M., et al. (2012). Megalencephaly-capillary malformation (mcap) and megalencephaly-polydactyly-polymicrogyria-hydrocephalus (mpph) syndromes: Two closely related disorders of brain overgrowth and abnormal brain and body morphogenesis. *American Journal of Medical Genetics Part A*, 158(2), 269–291.

Montera, M., Piaggio, F., Marchese, C., Gismondi, V., Stella, A., Resta, N., Varesco, L., Guanti, G., & Mareni, C. (2001). A silent mutation in exon 14 of the APC gene is associated with exon skipping in a FAP family. *Journal of Medical Genetics*, 38(12), 863–867.

Morgenthaler, S. & Thilly, W. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2), 28–56.

Nakamura, Y., Gojobori, T., & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1), 292.

Narendra, U., Pauer, G. J., & Hagstrom, S. A. (2009). Genetic analysis of complement factor h related 5, CFHR5, in patients with age-related macular degeneration. *Molecular Vision*, 15, 731.

Ng, P. & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13), 3812–3814.

Ng, S., Buckingham, K., Lee, C., Bigham, A., Tabor, H., Dent, K., Huff, C., Shannon, P., Jabs, E., Nickerson, D., et al. (2009). Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genetics*, 42(1), 30–35.

O'Roak, B., Deriziotis, P., Lee, C., Vives, L., Schwartz, J., Girirajan, S., Karakoc, E., MacKenzie, A., Ng, S., Baker, C., Rieder, M., Nickerson, D., Bernier, R., Fisher, S., Shendure, J., & Eichler, E. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genetics*, 43(6), 585–589.

Parmley, J., Chamary, J., & Hurst, L. (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular biology and evolution*, 23(2), 301–309.

Pesquita, C., Faria, D., Bastos, H., Falcão, A., & Couto, F. (2007). Evaluating GO-based semantic similarity measures. In *Proc. 10th Annual Bio-Ontologies Meeting*, volume 37 (pp.̃38).

Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7), e1000443.

Philippakis, A. A., Azzariti, D. R., Beltran, S., Brookes, A. J., Brownstein, C. A., Brudno, M., Brunner, H. G., Buske, O. J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: A platform for rare disease gene discovery. *Human Mutation*, 36(10), 915–921.

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, 30(17), 3894–3900.

Ramser, J., Ahearn, M., Lenski, C., Yariz, K., Hellebrand, H., Von Rhein, M., Clark, R., Schmutzler, R., Lichtner, P., Hoffman, E., et al. (2008). Rare missense and synonymous variants in UBE1 are associated with X-linked infantile spinal muscular atrophy. *The American Journal of Human Genetics*, 82(1), 188–193.

Renneville, A., Boissel, N., Helevaut, N., Nibourel, O., Terré, C., Pautas, C., Gardin, C., Thomas, X., Turlure, P., Reman, O., et al. (2011). Wilms' tumor 1 single-nucleotide polymorphism rs16754 does not predict clinical outcome in adult acute myeloid leukemia. *Leukemia*.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1* (pp. 448–453).: Morgan Kaufmann Publishers Inc.

Richard, P., Gaudon, K., Fournier, E., Jackson, C., Bauché, S., Haddad, H., Koenig, J., Echenne, B., Hantaï, D., & Eymard, B. (2007). A synonymous CHRNE mutation responsible for an aberrant splicing leading to congenital myasthenic syndrome. *Neuromuscular Disorders*, 17(5), 409–414.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in Medicine*, 17(5), 405–423.

Robinson, P. N. (2012). Deep phenotyping for precision medicine. *Human mutation*, 33(5), 777–780.

Robinson, P. N. (2014). Computational phenotype analysis in human medicine. In *Phenomics*. CRC Press.

Robinson, P. N., Köhler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, 24(2), 340–348.

Rogozin, I. B., Makarova, K. S., Natale, D. A., Spiridonov, A. N., Tatusov, R. L., Wolf, Y. I., Yin, J., & Koonin, E. V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic acids research*, 30(19), 4264–4271.

Salari, R., Kimchi-Sarfaty, C., Gottesman, M. M., & Przytycka, T. M. (2013). Sensitive measurement of single-nucleotide polymorphism-induced changes of RNA conformation: application to disease studies. *Nucleic Acids Research*, 41(1), 44–53.

Sauna, Z. & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics*, 12(10), 683–691.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., Rückstieß, T., & Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*, 11, 743–746.

Schmucker, D., Clemens, J. C., Shu, H., Worby, C. A., Xiao, J., Muda, M., Dixon, J. E., & Zipursky, S. L. (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, 101(6), 671–684.

Sharp, P. M. & Li, W.-H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, 15(3), 1281–1295.

Sifrim, A., Popovic, D., Tranchevent, L.-C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J. R., Aerts, J., De Moor, B., & Moreau, Y. (2013). eXtasy: variant prioritization by genomic data fusion. *Nature Methods*, 10(11), 1083–1084.

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. *The American Journal of Human Genetics*, 94(4), 599–610.

Smedley, D., Jacobsen, J. O. B., Jöger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O., Bone, W. P., Haendel, M. A., & Robinson, P. N. (2015). Next-generation diagnostics and disease gene discovery with the exomiser. *Nature Protocol*, 10, 1304–1351.

Smedley, D., Oellrich, A., Köhler, S., Ruef, B., Westerfield, M., Robinson, P., Lewis, S., Mungall, C., et al. (2013). PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database*, 2013, bat025.

Smedley, D. & Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human mendelian disease genes. *Genome medicine*, 7(1), 1–11.

Smith, P. J., Zhang, C., Wang, J., Chew, S. L., Zhang, M. Q., & Krainer, A. R. (2006). An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Human Molecular Genetics*, 15(16), 2490–2508.

Sobreira, N., Schiettecatte, F., Valle, D., & Hamosh, A. (2015). GeneMatcher: A matching tool for connecting investigators with an interest in the same gene. *Human mutation*, 36(10), 928–930.

Spillantini, M., Yoshida, H., Rizzini, C., Lantos, P., Khan, N., Rossor, M., Goedert, M., & Brown, J. (2000). A novel *tau* mutation (N296N) in familial dementia with swollen achromatic neurons and corticobasal inclusion bodies. *Annals of neurology*, 48(6), 939–943.

Stanford, P., Halliday, G., Brooks, W., Kwok, J., Storey, C., Creasey, H., Morris, J., Fulham, M., & Schofield, P. (2000). Progressive supranuclear palsy pathology caused by a novel silent mutation in exon 10 of the tau gene. *Brain*, 123(5), 880–893.

Steingrimsdottir, H., Rowley, G., Dorado, G., Cole, J., & Lehmann, A. (1992). Mutations which alter splicing in the human hypoxanthine-guanine phosphoribosyltransferase gene. *Nucleic Acids Research*, 20(6), 1201–1208.

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E. M., Akey, J. M., et al. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, 342(6164), 1367–1372.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1), 1.

Tenesa, A., Navarro, P., Hayes, B., Duffy, D., Clarke, G., Goddard, M., & Visscher, P. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome research*, 17(4), 520–526.

Thomas, P., Campbell, M., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research*, 13(9), 2129–2141.

Waldispühl, J. & Ponty, Y. (2011). An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *Journal of Computational Biology*, 18(11), 1465–1479.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J., Abrahams, B., Salyakina, D., Imielinski, M., Bradfield, J., Sleiman, P., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, 459(7246), 528–533.

Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M., & Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6), 831–845.

Warde-Farley, D., Brudno, M., Morris, Q., & Goldenberg, A. (2012). Mixture model for sub-phenotyping in GWAS. In *Pacific Symposium on Biocomputing*. (pp. 363).

Warneford, S., Witton, L., Townsend, M., Rowe, P., Reddel, R., Dalla-Pozza, L., & Symonds, G. (1992). Germ-line splicing mutation of the p53 gene in a cancer-prone family. *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, 3(11), 839.

Washington, N. L., Haendel, M. A., Köhler, S., Lewis, S. E., Robinson, P., Smedley, D., & Mungall, C. J. (2014). How good is your phenotyping? methods for quality assessment. *Proceedings of Phenotype Day*.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), 1113–1120.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*, 42(D1), D1001–D1006.

Wen, P., Xiao, P., & Xia, J. (2016). dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics*, (pp. btw086).

Winnenburg, R. & Bodenreider, O. (2014). Coverage of phenotypes in standard terminologies. *Proceedings of the Joint Bio-Ontologies and BioLINK ISMB*, (pp. 41–44).

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., Hua, Y., Gueroussov, S., Najafabadi, H. S., Hughes, T. R., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 1254806.

Yandell, M., Huff, C., Hu, H., Singleton, M., Moore, B., Xing, J., Jorde, L., & Reese, M. (2011). A probabilistic disease-gene finder for personal genomes. *Genome Research*, 21(9), 1529–1542.

Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *Jama*, 312(18), 1870–1879.

Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science translational medicine*, 6(252), 252ra123–252ra123.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database*, 2011, bar026.

Zhang, X. H. F., Kangsamaksin, T., Chao, M. S. P., Banerjee, J. K., & Chasin, L. A. (2005). Exon inclusion is dependent on predictable exonic splicing enhancers. *Molecular and Cellular Biology*, 25(16), 7323–7332.