Development and Evaluation of Software for Applied Clinical Genomics

by

Casper Shyr

BSc (Computer Science and Biology), The University of British Columbia, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

The University of British Columbia

(Vancouver)

April 2016

Abstract

High-throughput next-generation DNA sequencing has evolved rapidly over the past 20 years. The Human Genome Project published its first draft of the human genome in 2000 at an enormous cost of 3 billion dollars, and was an international collaborative effort that spanned more than a decade. Subsequent technological innovations have decreased that cost by six orders of magnitude down to a thousand dollars, while throughput has increased by over 100 times to a current delivery of gigabase of data per run. In bioinformatics, significant efforts to capitalize on the new capacities have produced software for the identification of deviations from the reference sequence, including single nucleotide variants, short insertions/deletions, and more complex chromosomal characteristics such as copy number variations and translocations. Clinically, hospitals are starting to incorporate sequencing technology as part of exploratory projects to discover underlying causes of diseases with suspected genetic etiology, and to provide personalized clinical decision support based on patients' genetic predispositions. As with any new large-scale data, a need has emerged for mechanisms to translate knowledge from computationally oriented informatics specialists to the clinically oriented users who interact with it.

In the genomics field, the complexity of the data, combined with the gap in perspectives and skills between computational biologists and clinicians, present an unsolved grand challenge for bioinformaticians to translate patient genomic information to facilitate clinical decision-making. This doctoral thesis focuses on a comparative design analysis of clinical decision support systems and prototypes interacting with patient genomes under various sectors of healthcare to ultimately improve the treatment and well-being of patients. Through a combination of usability methodologies across multiple distinct clinical user groups, the thesis highlights reoccurring domain-specific challenges and introduces ways to overcome the roadblocks for translation of next-generation sequencing from research laboratory to a multidisciplinary hospital environment. To improve the interpretation efficiency of patient genomes and informed by the design analysis findings, a novel computational approach to prioritize exome variants based on automated appraisal of patient phenotypes is introduced. Finally, the thesis research incorporates applied genome analysis via clinical collaborations to inform interface design and enable mastery of genome analysis.

Preface

The work described in this thesis is based upon research done by Casper Shyr in Dr. Wyeth W. Wasserman's group at the Centre for Molecular medicine and Therapeutics (CMMT), Child and Family Research Institute (CFRI) at the BC Children's Hospital. Part of the research is done as collaborations with OMICS2TREATID team, led by Dr. Clara van Karnebeek, for which Casper Shyr was granted co-authorship on the publications that came out. Works that have been published in peer-reviewed scientific journals are listed below. Contributions from Casper Shyr and acknowledgements to other members are discussed below.

Work in chapter 2 and 3 were done by Casper Shyr, with support and guidance from Dr. Andre Kushniruk, Dr. Jehannine Austin, Dr. Sohrab Shah, and Dr. Clara van Karnebeek. Dr. Kushniruk and Dr. Wasserman were particularly involved in designing the study. Dr. Kushniruk further assisted with the data analysis. Dr. van Karnebeek facilitated the recruitment process. Cynthia Ye, Alice Chou, and Dr. Ekaterina Nosova helped with the design of tutorial videos. Patrick Tan, Calvin Lefebvre, and David Arenillas supported the usability evaluations by being initial participants. Jonathan Chang and Michael Hockertz provided the equipment necessary to conduct the usability analysis. This work was supported by Canadian Institutes of Health Research (CIHR) grant number MOP-82875, Natural Sciences and Engineering Research Council of Canada (NSERC) grant number RGPIN355532-10, Omics2TreatID, and Genome Canada/Genome BC 174DE (ABC4DE project). Icons and graphic arts incorporated into the figures and tables are modified from open repositories freely available for academic use. These two chapters have received ethical approval from UBC Children's and Women's Research Ethics Board (H12-02738, H13-02034). The work for chapter 2 is in part published as a first-author research article in

C. Shyr, A. Kushniruk, W.W. Wasserman, 2014, "Usability study of clinical exome analysis software: Top lessons learned and recommendations", ***Journal of Biomedical Informatics*****, 51, 129-136.**

The work for chapter 3 was in part published as a first-author research article in

C.Shyr, C.D.M. van Karnebeek, A. Kushniruk, W.W. Wasserman, 2015, "Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists and genetic counselors", ***JAMIA*****, Jun 27. pii: ocv053. doi: 10.1093/jamia/ocv053. [Epub ahead of print]**

The work in chapter 4 was done by Casper Shyr, with data generation and analysis contributions from Jessica Lee and Mike Gottlieb. Jessica was specifically involved with extracting publication records from NCBI, and Mike was involved with the calculations of dN/dS and related gene-level measurements. Dr. Maja Tarailo-Graovac formulated the experiment and oversaw the project collectively with Dr. Wyeth wasserman. Dr. Clara van Karnebeek facilitated the clinical coordination to include a relevant patient case into the manuscript as proof-of-concept. Written informed consent was obtained from the patient's guardian/parent/next of kin for the publication of this report. Additional acknowledgements go towards Drs. J. Wu, J. Rozmus, S. Vercauteren, K. Hildebrand, T. Dewan and A. Garcera for clinical evaluation and management of the patient; Mrs. X. Han for Sanger sequencing; Mr. B. Sayson for data management; Mrs. M. Higginson for DNA extraction, sample handling and technical data; Dr. C. Vilarino-Guell for timely whole exome sequencing; Dr. W. Cheung for MeSHOP support; Mr. D. Arenillas and Mr. M. Hatas for systems support, and Dora Pak for research management support. The work was published as joint-first-authorship with Dr. Tarailo-Graovac in:

C. Shyr, M. Tarailo-Graovac, M. Gottlieb, J.J.Y. Lee, C. van Karnebeek and W.W. Wasserman, 2014, "FLAGS, Frequently Mutated Genes in Public Exomes", ***BMC Medical Genomics*****, 7, 64.**

The work in chapter 5 was done by Casper Shyr, with guidance from Dr. Maja Tarailo-Graovac, Dr. Anthony Mathelier and Dr. Sohrab Shah. Jessica Lee, Wenqiang Shi, Yifeng Li, and David Arenillas provided bioinformatics advice. Mike Gottlieb supplied codes that he wrote from GeneYenta project that were adapted for my own project. I am also grateful to the clinicians involved, including Dr. S. Stockler, Dr. G. Horvath, Dr. S. Sirrs, Dr. K. Boycott, Dr. A. Lehman, Dr. G. Sinclair, Dr. H. Vallance, Dr. R. Salvarinova, and Dr. B. Gibson. Additional contributors include Ms. C. Ye for data analysis; Dr. A. Bhavsar and Dr. Linhua for wetlab experimentation; Ms. X. Han and Dr. B. Drogemoller for Sanger sequencing; Mr. B. Sayson and Ms. K. Townsend for data management; Ms. M. Higginson for DNA extraction, sample handling and technical data; Dr. C. Vilarino-Guell providing data on Ion-torrent based whole exome sequencing; Dr. W. Cheung for MeSHOP support; Mr. M. Hatas for systems support, and Ms. Dora Pak for research management support. The work has not been published at the time of writing.

Chapter 6 and Chapter 7 highlight the collaborative works done with OMICS2TREATID. Casper Shyr provided bioinformatics processing and interpretive analysis on exomes and whole-genomes. Casper also contributed to the methodologies and results write-up in the manuscripts. The work was supported by David Arenillas for software development, Miroslav Hastas for server management, Dr. Virginie Bernard, Dr. Maja Tarailo-Graovac and Jessica Lee for bioinformatics analysis. Wet lab validations were carried out by Dr. Linhua Zhang and Xiaohua Han from Dr. Colin Ross laboratory. For chapter 6, Xiahua Han performed CA5A molecular analyses, M. Higginson and M. Zhou conducted DNA extraction, sample handling, and technical data, and M. Thomas and M. Lafek oversaw consent procedure and data management. K. MacKenzie conducted psychological testing of the siblings, and J. Rilstone helped with text editing. Additional acknowledgements extend towards colleagues in the Departments of Pediatrics and Medical Genetics, University of British Columbia (Canada) for clinical management of patients; Dr. D. Wishart and Dr. Rupa Mandal for NMR analyses (University of Alberta, Canada); Dr. M. Baumgärtner and Dr. P. Burda (Zürich University, Switzerland), Dr. J. Cameron (University of Toronto, Canada), Dr. M. Evans (University of Edinburgh, United Kingdom), Dr. M. Kaczocha and Dr. R. Haltiwänger (Stony Brook University, USA), Dr. G. Lavery (University of Birmingham, United Kingdom), Dr. A. Waheed and Dr. W. Sly (St. Louis University, USA), Dr. S. Corvera (University of Massachusetts Medical School, USA) for in vitro studies on unpublished cases; Mrs. B. Toh for metabolic sample handling; Ms. X. Han for Sanger sequencing; Dr. M. Thomas and Mr. T. Murphy consenting and data management; Ms.

M. Higginson for DNA extraction and sample handling; Mr. D. Arenillas and Mr. M. Hatas for systems support, and Ms. D. Pak, Ms. H. Cheung, Ms. R. Giesbrecht (all affiliated with the University of British Columbia) for research management support.

Chapter 6 was in part published in:

C.D. van Karnebeek, W.S. Sly, C.J. Ross, R. Salvarinova, J. Yaplito-Lee, S. Santra, C. Shyr, et. al., 2014, "Mitochondrial carbonic anhydrase VA deficiency due to CA5A alterations presents with hyperammonemia in early childhood", ***American Journal of Human Genetics*****, 94, 453-461.**

Chapter 7 has been accepted as part of a second author publication and has not been published at the time of writing.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BAM Binary Alignment Map

BP (bp) Base Pair

BWA Burrows-Wheeler Aligner

CDSS Clinical Decision Support System

CNV Copy Number Variation

CPU Central Processing Unit

CTA Cognitive Task Analysis

DNA Deoxyribonucleic Acid

EHR Electronic Health Record

EVS Exome Variant Server

HGMD Human Gene Mutation Database

HPO Human Phenotype Ontology

ID Intellectual Disability/Disorder

IEM Inborn Errors of Metabolism

INDEL (or InDel) Insertion and Deletion

INDELs (or InDels) Insertions and Deletions

LIMS Laboratory Information Management System

MB Megabyte

mRNA messenger Ribonucleic Acid

NGS Next Generation Sequencing

GB Gigabyte

GUI Graphical User Interface

GATK Genome Analysis Toolkit

GWAS Genome-Wide Association Study

OMIM Online Mendelian Inheritance in Man

PCR Polymerase Chain Reaction

RAM Random Access Memory

RNA Ribonucleic Acid

SQL Structured Query Language

SNV Single Nucleotide Variation

SNP Single Nucleotide Polymorphism

SV Structural Variation

UI User Interface

UTR Untranslated Region

VCF Variant Call Format

VPA Variant Prioritization Accelerator

Glossary

CADD A tool for scoring the deleteriousness of single nucleotide variants as well as insertions and deletions in the human genome

causal allele An allele that directly results in an observed phenotype, but may have incomplete penetrance. Causal alleles for severe monogenic disorders and rare, large-effect risk variants in complex disease are highly penetrant

cloud computing A computing infrastructure based on the internet, where the computing and/or storage solutions are provided to the users via third-party data centers

compound heterozygosity The situation where two or more mutations are present within the same gene (generally both heterozygous), resulting in the disease phenotype

dbSNP The database of short genetic variations, including single nucleotide polymorphisms (SNPs), short deletions or insertions, microsatellites and short tandem repeats. Found at http://www.ncbi.nlm.nih.gov/projects/SNP/

ExAC A database of exomes from over 60,000 individuals sequenced as part of various disease-specific and population genetic studies. Found at http://exac.broadinstitute.org/

exome The part of the human genome formed by exons that remain within mature RNA after introns are removed by RNA splicing

exome capture A method that employs probes (e.g. on a micro-array) to hybridize to known coding sections of the human genome and selectively retain them for high-throughput sequencing, while all other fragments are washed away

exome variant server A database of over 6500 exomes from the NHLBI consortium to study the genes and mechanisms contributing to heart, lung and blood disorders. Found at http://evs.gs.washington.edu/EVS/

haplotype A combination of alleles at adjacent locations on the chromosome that are often transmitted together

HGMD The Human Gene mutation Database, representing a curated collection of gene lesions responsible for inherited human diseases. Two versions are available: one is free for academic institutions (but not constantly updated), and the other is the more comprehensive professional version requiring a commercial license fee

index The individual being studied or reported on. It is usually the first affected individual in a family who brings a genetic disorder to the attention of the medical community. A synonymous term often used in biomedical literature is proband

mutation A heritable change in the structure of a gene. It does not infer a deleterious effect

missense variant A single nucleotide change that results in the codon of a protein coding gene to encode for a different amino acid. It is part of the category of non-synonymous variant

NGS Next generation sequencing. Automated Sanger method is considered a “first-generation” technology, and newer methods are referred to as next-generation sequencing that can determine the sequence of DNA at a much higher throughput than Sanger sequencing, orders of magnitude faster, and exponentially cheaper per nucleotide

non-genetic factors Environment and lifestyle are two issues that may influence whether a disease phenotype manifest or not

nonsense variant A type of non-synonymous variant that results in changing a codon to a premature stop codon, resulting in a truncated protein product

penetrance The proportion of individuals carrying a particular genotype that also expresses an associated trait

personalized medicine A model in healthcare that provides medical decisions, practices, and/or products tailored to individual patient. Another term to refer to this is precision medicine

polymorphism A DNA variation that differs from the human reference genome that is observed above a certain level of allelic frequency (typically 1%) in the human population, making it a common variation

proband See definition of “index”

rare mutation A variant that is observed with a frequency less than 1% across the matching population, or none at all

read A portion of the DNA fragment that has been sequenced

risk Presence of faulty genes do not always lead to the expected phenotype

Sanger sequencing The modern version involves a chain termination method using a mix of deoxynucletides and four-color dideoxynucleotides that terminate polymerization when incorporated. DNA sequence is obtained by reading off the fragmentation patterns separated by size on a gel or in a capillary

SNP Common variant observed across the population. Stands for single nucleotide polymorphism. I do not address novel variants as SNPs, but instead refer to them as single nucleotide variants (SNVs), which in my definition, also includes SNPs

structural variant A variant which is 1kb or larger. Copy number variants (CNVs) are included under this category

synonymous variant A substitution of one base for another in the exon of a gene coding for a protein that results in the same amino acid sequence.

variant A change in DNA structure that differs from an accepted standard, typically against the human reference genome

whole genome The complete DNA sequence of an organism’s genome

Acknowledgements

I am extremely grateful to my supervisor Dr. Wyeth W. Wasserman for providing exceptional guidance and support throughout both my undergraduate years as well as my years as the graduate student. I also extend my gratitude to my committee members Dr. Jehannine Austin, Dr. Bruce Carleton, and Dr. Sohrab Shah, whose penetrating questions taught me to question more deeply. I am particularly thankful for my fellow students at UBC, especially members of the Wasserman lab. I owe particular thanks to the exome/whole-genome analysis team: Dr. Maja Tarailo-Graovac, David Arenillas, Dr. Allison Matthews, Dr. Virginie Bernard, Jessica Lee, and Cynthia Ye. Dr. Tarailo-Graovac was especially helpful in formulating the research questions that shaped a significant portion of my thesis. She provided valuable feedback on my experimental approaches, training me to think critically and become a better scientist. She was always available to the students in the lab, and has supported me in both thesis research as well as for non-thesis-related endeavors (e.g. conference presentations, manuscript write-ups). In many ways, she became a co-supervisor in the later stages of my thesis. Her mentorship went beyond scientific research or scientific publications; I highly cherish her wisdom for building a cohesive team in a multidisciplinary project. The importance of maintaining a successful collaborative research environment between scientists from different scientific domains cannot be overstressed and I will be taking what I have learned from her to my future employment. The project was possible due to the efforts of Dr. Clara van Karnebeek, who was an outstanding lead in the TIDEX and OMICS2TREATID pojects. I thank Dr. Elodie Portales-Casamar for enlarging my vision of science. I am indebted to all the participants who took part in my thesis research. I especially thank the TIDEX (www.tidebc.org) team for numerous opportunities of

collaborations, supplying the sequencing data, and advising on the various aspects of clinical DNA sequencing analysis. I am indebted to the patients and families for their participation throughout my thesis. Finally, I am grateful for the doctoral fellowship from Canadian Institute for Health Research (CGSD-GSM), and the additional grants related to funding my collaborations: B.C. Children's Hospital Foundation as "1st Collaborative Area of Innovation" (www.tidebc.org), Genome BC (SOF-195 grant), Genome BC and Genome Canada grants 174CDE (ABC4DE Project), and the Canadian Institutes of Health Research #301221 grant. Special thanks are owed to my wife and my parents, whose have supported me throughout my years of education, both morally and financially.

Dedication

I dedicate this thesis to my wife, Victoria Hu, who has been there since the beginning.

Chapter 1: Introduction

1.1 Biology background

Over a 65-year period, genetics research has moved from Watson and Crick's description of the DNA double helix to the development of high-throughput instruments capable of determining the 6 billion nucleotide sequence of an individual genome in a single day[1]. During this period, clinical genetics research has provided insights into genetic disorders and the types of mutations that arise with clinical phenotypes. The representative mutation classes are single-base substitutions, small insertions/deletions (<30bp), copy number variations (e.g. large deletions and duplications), and structural variations (e.g. translations, inversions). Figure 1-1 and 1-2 show common types of DNA variations and how they can disrupt gene sequences. Between 1980s-90s, dozens of disease genes were discovered, revealing a diverse range of genetic mechanisms and inheritance patterns, from autosomal dominant to X-linked recessive and mitochondrial inheritance (Table 1-1). One historical example was the cloning for the first human disease gene by Dr. Stuart Orkin, revealing a single base substation in X chromosome responsible for chronic granulomatous disease[2].

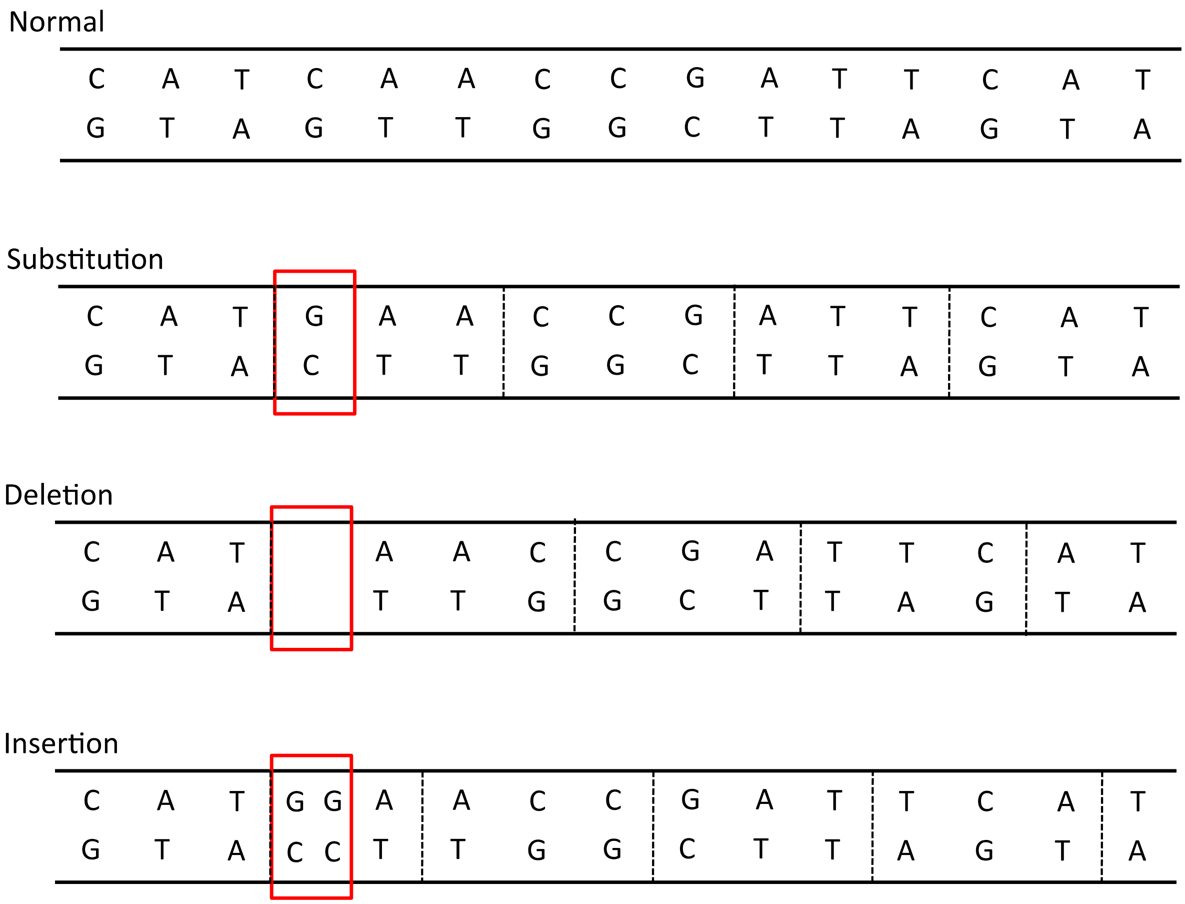

Normal
C A T C A A C C G A T T C A T
G T A G T T G G C T T A G T A
Substitution
C A T G A A C C G A T T C A T
G T A C T T G G C T T A G T A
Deletion
C A T A A C C G A T T C A T
G T A T T G G C T T A G T A
Insertion
C A T G G A A C C G A T T C A T
G T A C C T T G G C T T A G T A

Figure 1-1 Common classes of single nucleotide DNA variants. The dotted line marks the boundary of each codon. The uppermost portion of the figure shows a normal DNA sequence. The lower 3 portions show how codons are changed, depending on the type of mutation introduced. For simplicity, deletion and insertion are illustrated with removal/insertion of a single G-C respectively, but most insertions and deletions observed in human populations span beyond 1bp[3].

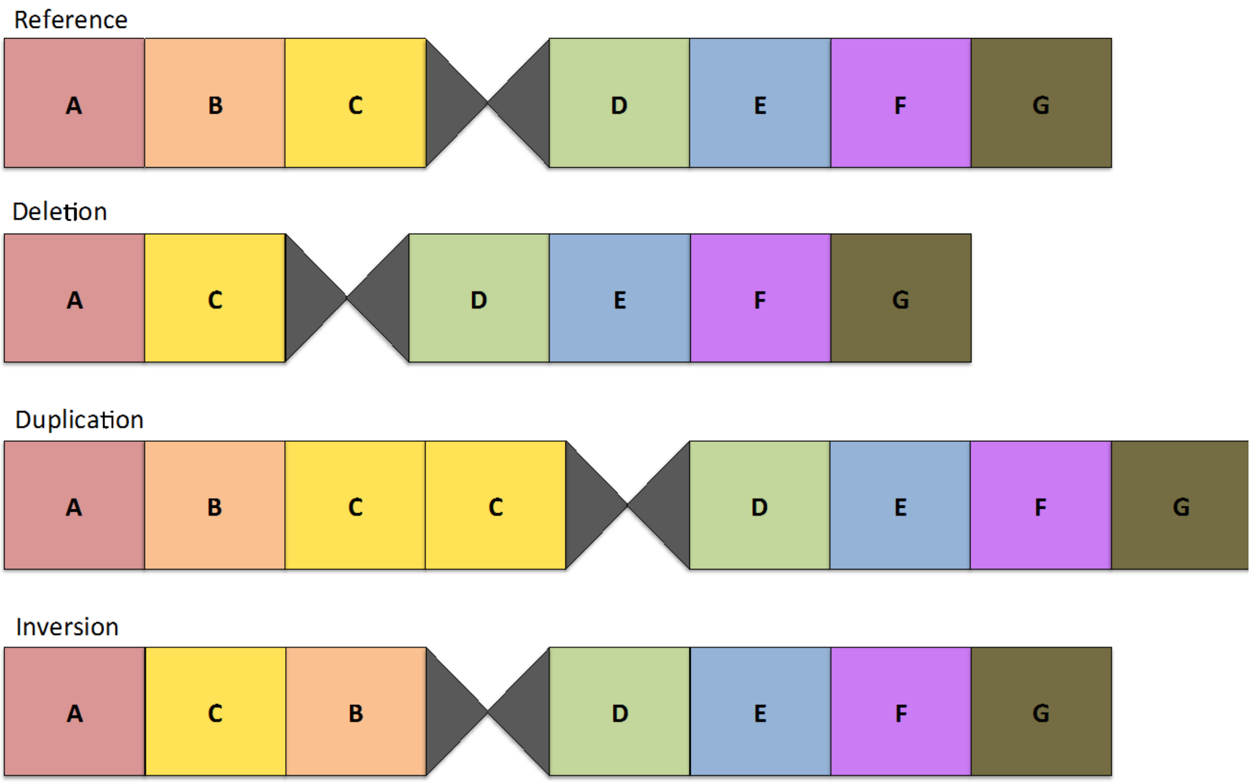


Figure 1-2 Illustrative examples of copy number variations and structural variations. Each letter corresponds to a section in the chromosome. The categories highlighted here are not meant to be exhaustive.

	Example disease	**Example gene**	**Reference**
Autosomal dominant	Hereditary nonpolypopsis colorectal cancer	*MSH2*	[4]
Autosomal recessive	Cystic fibrosis	*CFTR*	[5]
Sex-linked dominant	Chronic granulomatous disease	*PHOX*	[6]
Sex-linked recessive	Duchenne muscular dystrophy	*DMD*	[7]
Mitochondrial-inherited	Lebers hereditary optic neuropathy	*MT-ND6*	[8]

Table 1-1 Common inheritance patterns and a classical disease example for each.

Common disorders such as heart disease, asthma, or diabetes are rarely caused by single gene defects, rather they arise from the combined effects of multiple genes (polygenic) interacting with environmental factors and lifestyle. Although polygenic disorders tend to cluster in families, the pattern of inheritance is difficult to detect, as many do not conform to Mendelian patterns. Importantly, genetic penetrance of phenotype-causing variations varies widely, making interpretation of genomes more challenging.

1.1.1 Technologies for DNA Analysis: 1980s-2015

Molecular genetics advances have been catalyzed by a series of innovations. Over the past 35 years, improvements in DNA analysis approaches and technologies have enabled geneticists to transition from population-driven studies to pursue increasingly specific cases of rare diseases. In the 80s and 90s, the analysis largely focused upon screening for variations that eliminate or create restriction enzyme recognition sites via restriction enzymes, or analysis of varying fragment lengths in polymerase chain reaction (PCR)-amplified products[9]. Highly relevant to the current era are technologies for array-based high-throughput genotyping that can assess thousands of single nucleotide polymorphisms (SNPs) simultaneously throughout the genome. This provides the basis for genome-wide association studies (GWAS), in which cohorts of individuals are profiled across ~10^5 SNPs to reveal those statistically biased between cases and controls[10]. With such high-throughput genotyping technology, the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) determined common allele segregation patterns through linkage analysis across different human populations, providing a haplotype map of the human genome. Using this map, it has been increasingly feasible to link chromosomal regions to complex disorders. Similarly, Array CGH (array-comparative genomic hybridization), a hybridization-based method capable of detecting genomic copy number variations (CNVs) across the whole genome with higher resolution (~10^4bp) than traditional chromosome karyotyping, has been used to study the frequency of CNVs between cases and controls[11, 12].

Sanger’s Nobel prize-winning method for DNA sequencing with terminators, the technology that drove the Human Genome Project, has been accelerated by intermittent innovations to increase throughput and decrease cost[13]. The transition from radioactive terminators to dye-based terminators allowed single tube sequencing reactions, which have subsequently been moved into parallel reactions on a single surface, so called next generation sequencing (NGS). The overall cost per base pair has decreased substantially for the past 25 years (Figure 1-3). The main advantage for NGS is the ability to produce enormous data at a relatively cheap cost in short periods of time[14]. While newer technologies such as single molecule sequencing are under development that may increase read length, sensitivity, accuracy and throughput[15], it is sufficient for this thesis to understand that the capacity now exists to sequence the full human genome for a reasonable cost.

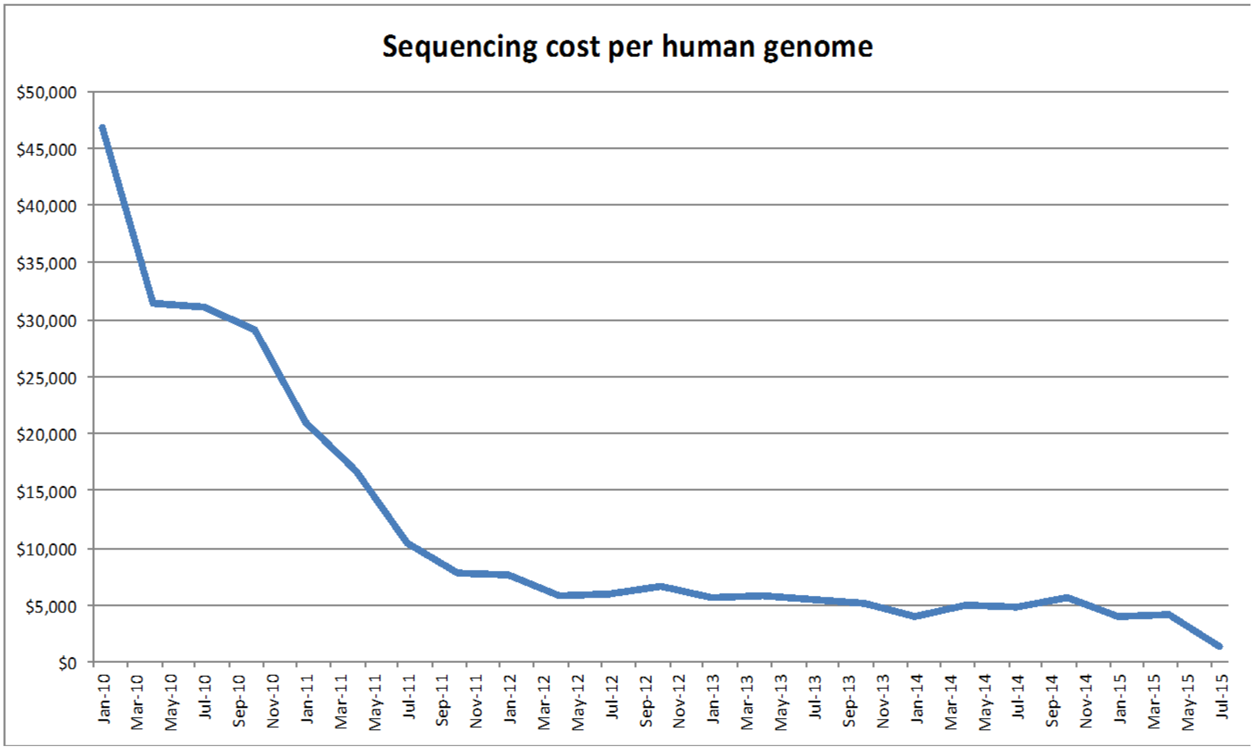


Figure 1-3 The cost of sequencing a human genome. Data from NHGRI Genome Sequencing Program, http://www.genome.gov/sequencingcosts/, on October 30, 2015.

Indeed, researchers are performing genome sequencing at a rapidly accelerating pace. The 1000 Genomes Project (http://www.1000genomes.org, [16]) is an international research effort to capture human genetic variation, but instead of focusing only on common variants like HapMap[17], it reveals rare variants with minor allele frequencies ≤ 1% by sequencing the genomes of upwards of 2000 individuals from various ethnic populations. At present it is the most detailed catalogue of human genetic variation in healthy populations of distinct ancestries, while databases such as ExAC (http://exac.broadinstitute.org) and UK10K (http://www.uk10k.org/data_access.html) are the current most comprehensive resources for allelic frequencies across an aggregated collection of large-scale sequencing projects in various disease-specific population studies. At the time of writing, the UK has announced intentions to sequence 100,000 genomes within its National Health Service, and NGS in Canada is rapidly moving from a research technology to a clinical standard[18].

1.1.1.1 Next-generation sequencing platform

NGS technologies involve three stages of template preparation, massively parallel sequencing with image processing, and informatics processing of reads to identify variation from a reference genome (or perform *de novo* sequence assembly). Due to the dominant role Illumina (California, USA) currently plays within the NGS market, this section illustrates their Genome Analyzer sequencing protocol as an example of the steps involved in next-generation DNA sequencers[19]. After obtaining genomic DNA, the first stage is to fragment the molecules and recover pools of similar lengths. Fragmentation is performed using adaptive focused acoustics technology that focuses acoustic energy to create cavitation events within the DNA sample to disrupt molecular bonds. Oligonucleotide adaptors are added to the ends of the fragments size-filtered by agarose gel electrophoresis. The DNA is PCR-amplified and denatured to single-strands. In a solid-phase amplification process, the strands are hybridized to a template oligonucleotide consisting of immobilized primers affixed to a surface. Subsequently, fluorescent-labeled dideoxynucleotides are added and incorporated one nucleotide at a time. Most Illumina machines use four-color method where each color represents one of the AGCT reversible terminating nucleotides, and the laser-triggered emitted fluorescence is captured by a high-resolution camera. The result is a series of digital images that are converted sequentially to DNA sequences for each discernable position across the surface. For more information, refer to Figure 1-4.

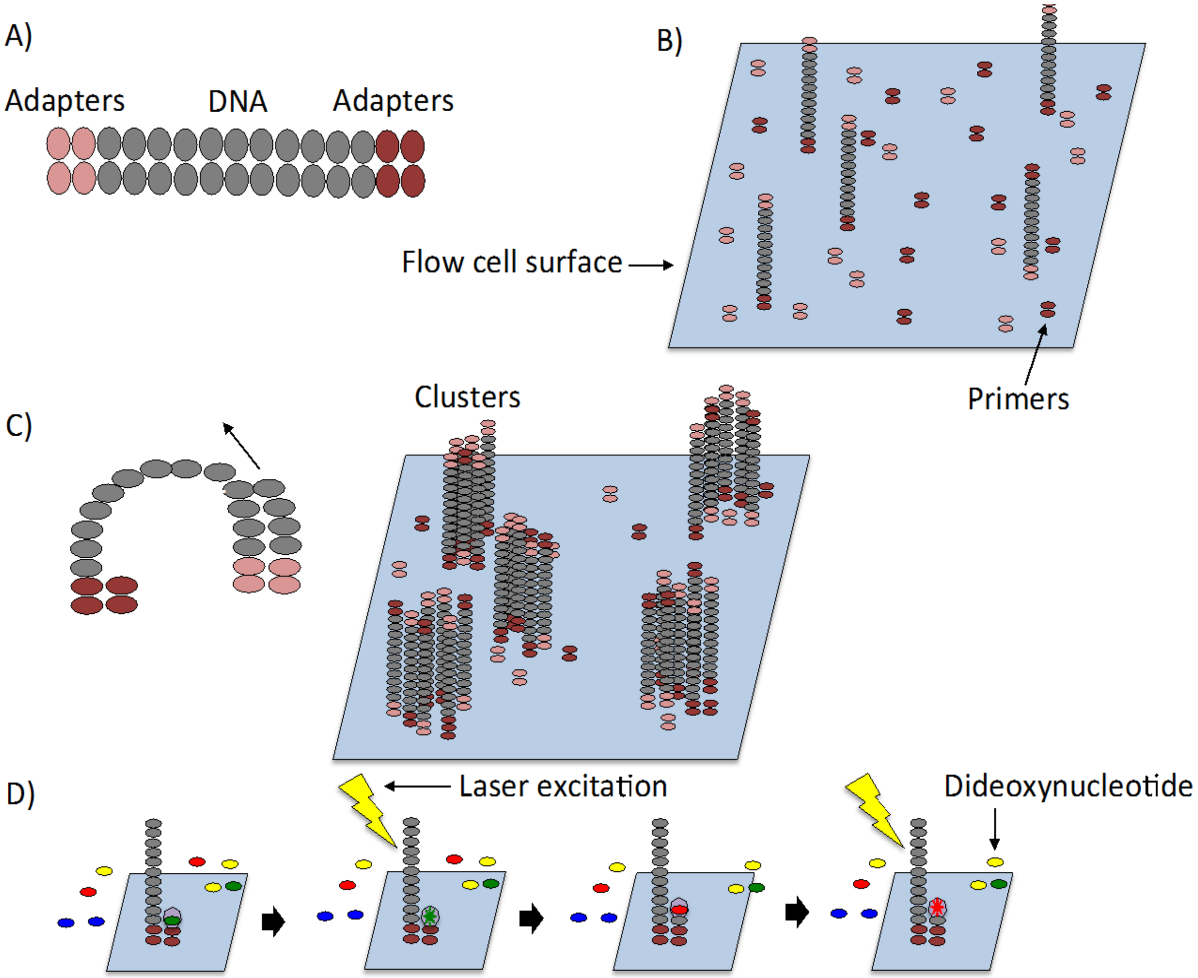


Figure 1-4 An overview of the Illumina sequencing protocol. In part A, input genomic DNA is randomly fragmented, and the fragments are ligated to adapter oligonucleotides on both ends. In part B, the single-stranded DNA-adapter complex hybridizes to the inner surface of the flow cell channels, which is coated with DNA primers. In part C, unlabeled nucleotides and DNA polymerase are added to construct a second "DNA bridge" which is complementary to the first. This is followed by denaturation to break up the bridge, leaving back to single-stranded templates anchored to the flow cell surface. Multiple iterations of this form clusters. Over 100-200 millions of such clusters are generated. This entire process here is called solid-phase bridge amplification. In part D, a four-color method is used where each color represents one of the AGCT nucleotides. Each type of this fluorescent nucleotide, called dideoxynucleotide, is added sequentially with one nucleotide incorporated at a time. A cleavage step removes the inhibiting group and the fluorescent dye. The four colors emitted by the incorporated nucleotides are detected by

a laser that excites the fluorescent molecule at a specific wavelength, and the emission signal is captured by a detector. The result is a series of images that when analyzed sequentially, convert to DNA sequences.

1.1.1.2 Exome sequencing and whole genome sequencing

Recent advances in NGS technology allow researchers to identify rare mutations within small families, reducing the need for large pedigrees or cohorts for causal gene discovery (although proving causality for the family-specific variants can be a challenge, as discussed later in the thesis) [20]. The technology accelerates genome analysis in clinical research for rare disorders, with more than 2400 clinical papers published, ranging from the discovery of genes responsible for Kabuki Syndrome to familial Parkinson's Disease[21, 22]. As sequencing the whole genome was initially cost prohibitive, a DNA filtering procedure has commonly been applied to focus sequence production on protein encoding exons, a process called "exome" sequencing. As protein-altering mutations are most readily interpreted at present, exome analysis has been popular for the discovery of monogenic disease genes[23]. The workflow of exome sequencing is essentially the same as whole genome, with the exception of an exome capture and enrichment stage following library preparation prior to DNA sequencing (Figure 1-5). Enrichment strategies include PCR-based approach, molecular inversion probes, or array-based/solution-based hybridization[24]. Each enrichment method has advantages and pitfalls; PCR has the highest specificity, but performs worse in terms of uniformity in target coverage. Hybridization-based methods hold the advantage of greater throughput, capturing large target regions in a single experiment, but requires expensive hardware and relatively large amounts of input DNA[24]. To illustrate how enrichment works, in aqueous-phase hybridization capture, the randomly shared DNA fragments are hybridized to biotinylated DNA or RNA baits. The hybridized fragments are recovered by biotin-streptavidin-based pull-down, followed by amplification and next-generation sequencing (each protocol is platform dependent).

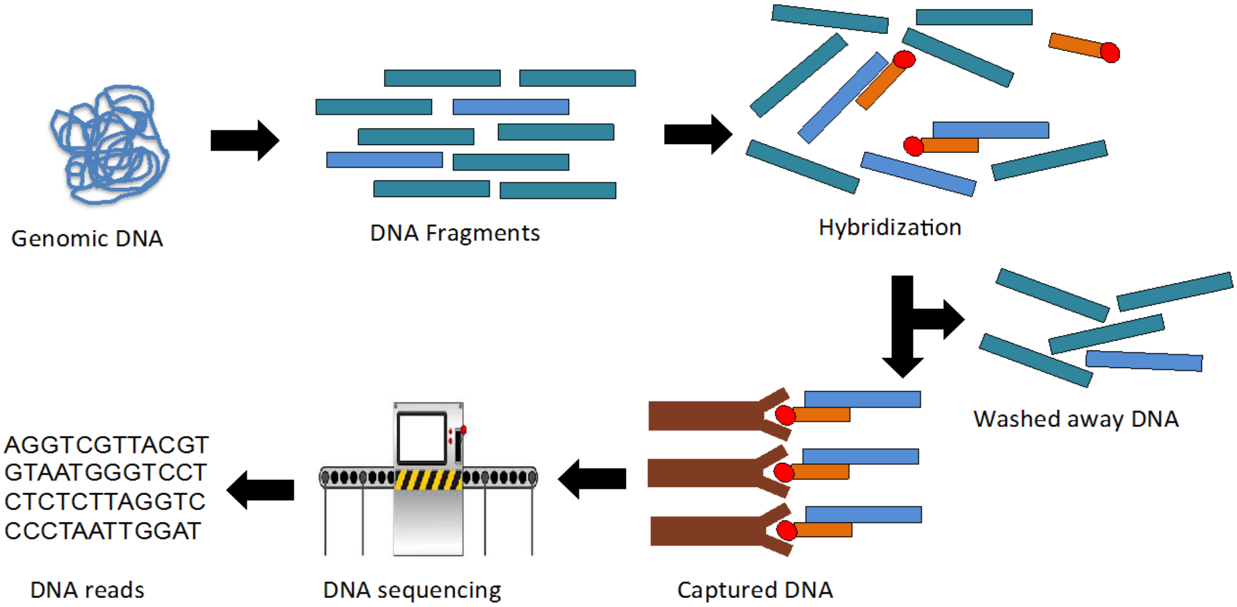


Figure 1-5 The overview workflow of exome sequencing, based upon Shendure et al. [25].

The first major proof-of-principle for exome analysis was published by Shendure *et al.* in which they sequenced 12 exomes: 8 from HapMap individuals, and 4 with Freeman-Sheldon syndrome, an autosomal dominant disorder. They successfully identified variants in *MYH3* known to characterize this syndrome[26]. Miller syndrome is the first genetic disorder with its genetic etiology discovered by exome sequencing. Performed by the same research team, two previously unknown variants shared by four affected individuals in three independent kindreds were identified, affecting the gene *DHODH* that encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway[27]. The gene was confirmed to be disrupted in additional families by Sanger sequencing. While increasing numbers of high penetrance mutations have been discovered outside of protein-coding regions using whole-genome sequencing (WGS)[28], the more limited interpretability of non-coding variants and their high volume remain a challenge for the field. Moreover, the greater financial cost and higher complexity of technological infrastructure required to process, analyze and store whole-genomes versus exomes have thus far limited WGS to select large-scale clinical centers, although at the time of thesis submission a transition from exome sequencing to WGS is clearly in process.

1.1.1.3 Typical NGS bioinformatics pipeline

The informatics workflow for processing genomic data consists of multiple steps. First, initial instrument images are processed to define the sequences. This varies by technology, and is often performed at the instrument level, as image data is too large for long-term retention with current storage technologies. The output from the sequencing machines are called “reads”, which are generally delivered in FASTQ format that consists of a read ID, the instrument-called DNA sequence, and a quality score for each position. Genomic aligners such as Bowtie2[29] and BWA[30] are used to map each read to a corresponding position in a human reference genome. See Table 1-2 for a selected list of alignment software. The typical output from the aligner is a BAM file, containing information about the qualities of sequence reads and the positions of their alignments. Post-alignment processing can be performed by tools such as Picard and Genome Analysis Toolkit (GATK)[31] to correct for possible misalignments, especially at extremity of reads. The diverse software can return different results (Figure 1-6). To call variants and their genotypes, popular software includes SAMtools mpileup[32] and GATK UnifiedGenotyper)[31], which take in a BAM file and return variants that pass a designated mapping criterion and quality score. To attach a functional annotation to the variant, software such as SnpEff[33] and ANOVAR[34] assign labels to each variant, distinguishing coding variants from non-coding variants, and separating coding variants into synonymous, missense, and nonsense categories (see glossary for their distinctions). Computational predictions on allele deleteriousness are generated by software such as SIFT[35], PolyPhen-2[36] and Combined Annotation Dependent Depletion (CADD)[37]. See Table 1-3 for a more complete list of variant callers and damage prediction tools. Evaluation of mapped reads can be manually performed with visualization software such as Integrated Genomics Viewer (IGV)[38].

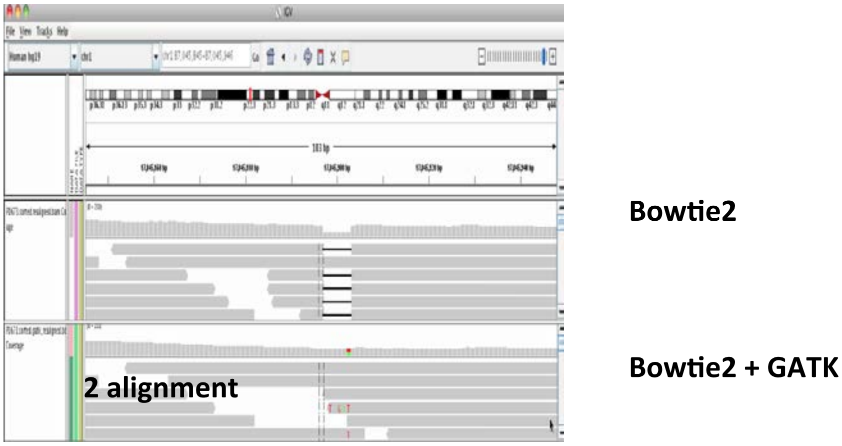


Figure 1-6 Impacts on results from different software. Part A is an example of a deletion which GSNAP aligner is able to detect but Bowite2 pipeline cannot. The alignment here is visualized by IGV. Each horizontal grey bar represents one individual mapped read. Two panels are shown, separated by a horizontal divider. The upper panel shows the alignment with GSNAP, and the lower panel shows the alignment with Bowtie2. In this example, Bowtie2 fails to align the reads properly to allow the detection of the deletion, and instead returns 4 mismatches (TGAT) as shown by the bright red, orange and green colors in one of the reads. GSNAP is able to align the reads correctly, depicting the deletion of the 4-bases deletion while simultaneously avoiding false mismatches. This is only an anecdotal example and does not mean GSNAP is superior to Bowtie2. The accuracy of mapping in regions of insertions/deletions remains an ongoing research problem. Part B shows an example of a successful realignment correction by GATK local realignment function. The figure again is a screenshot of IGV. The upper panel shows the alignment by Bowtie2 without GATK correction. We see 3 of the 5 reads have 2 mismatches: an A at the end, and a

C 4 bases to the right. The bottom panel shows the same alignment after GATK local realignment. GATK correctly extends the 5’ ends of 3 reads with the mismatches so they all now instead show the 4-bases deletion.

Name	**Method**
Bfast	Hash the reference
Bowtie/Bowtie2	FM-index
BWA	FM-index
GSNAP	Hash the reference
MAQ	Hash the reads
Mosaik	Hash the reference
Novoalign	Hash the reference

Table 1-2 The current popular genomic aligners used for mapping short-read DNA sequences.

Name	**Link**
SNVs and short InDels caller	
GATK HaplotypeCaller	https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_haplotypecaller_HaplotypeCaller.php
GATK UnifiedGenotyper	https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_genotyper_UnifiedGenotyper.php
SAMtools	http://samtools.sourceforge.net/
VarScan	http://varscan.sourceforge.net/
SVs caller	
Break Dancer	https://github.com/kenchen/breakdancer
Pindel	http://gmt.genome.wustl.edu/packages/pindel/
SVMerge	http://svmerge.sourceforge.net/
Predicting variant functional impacts	
B-SIFT	http://research-pub.gene.com/bsift/
Name	**Link**
CADD	http://cadd.gs.washington.edu/
MAPP	http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005
PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html
PolyPhen/PolyPhen V2	http://genetics.bwh.harvard.edu/pph2/
SIFT	http://blocks.fhcrc.org/sift/SIFT.html
SNAP	http://www.rostlab.org/services/SNAP
SNAPper/Pedant	http://pedant.gsf.de/snapper

Table 1-3 The current popular variant calling software and variant functional impact prediction tools.

At this time, up to 500,000[1] single nucleotide variants (SNVs) can be reported per exome, with some variance related to ancestry[39]. Filtration and prioritization is necessary to identify causal alleles for Mendelian traits. One approach is to look for novel or rare alleles based on databases like dbSNP, which catalog the observations of SNPs and short insertions/deletions (InDels; typically 1-30bp size range) submitted by large-scale sequencing projects (e.g. 1000 Genomes Project[3]) and other sequencing centers around the world[40]. This builds on the assumption that the filter set(s) contains no alleles actually causing the phenotype being studied. This assumption may not be true because dbSNP holds a subset of cases derived from individuals with unreported phenotypes or unaffected carriers (e.g. variants with incomplete penetrance)[41]. Additional approach for linking variants to phenotypes is to screen for variants present in known disease databases such as Human Gene Mutation Database (HGMD), which is a manual compilation of published human inherited disease mutations. HGMD exists in two formats: an academic version (http://www.hgmd.org/) and a commercial version that is more comprehensive and more frequently updated. The third strategy is to detect relationships between the impacted gene and the observed phenotypes through automated literature analysis using gene-to-disease profile comparisons, such as enabled for MeSHOPs[42]. Tools such as MeSHOPs create a weighted linkage between a particular gene and a disease by extracting keywords (e.g. Medical Subject Headings) from peer-reviewed scientific literature about the gene, and performing statistical over-representation analysis to look for disease keywords that are over-represented in the gene-related literature versus a control set. Candidate alleles can be further stratified based on

[1] This number is dependent on the breadth of exon coverage for the employed exome capture kit, and the specific setups of the bioinformatics pipeline used to generate the data.

the predicted deleteriousness. Mode of inheritance can be used as a filter if exomes from related individuals are available. Parent-child trios are especially important for identifying *de novo* mutations (Figure 1-7). Sequencing and filtering for shared novel variations or genes across multiple unrelated affected individuals is another option, especially if the underlying genetic defect is expected to be similar. Section 1.1.2.1 describes why this approach is helpful, and section 4.1 discusses how this can be difficult to achieve for rare disorders and the implications.

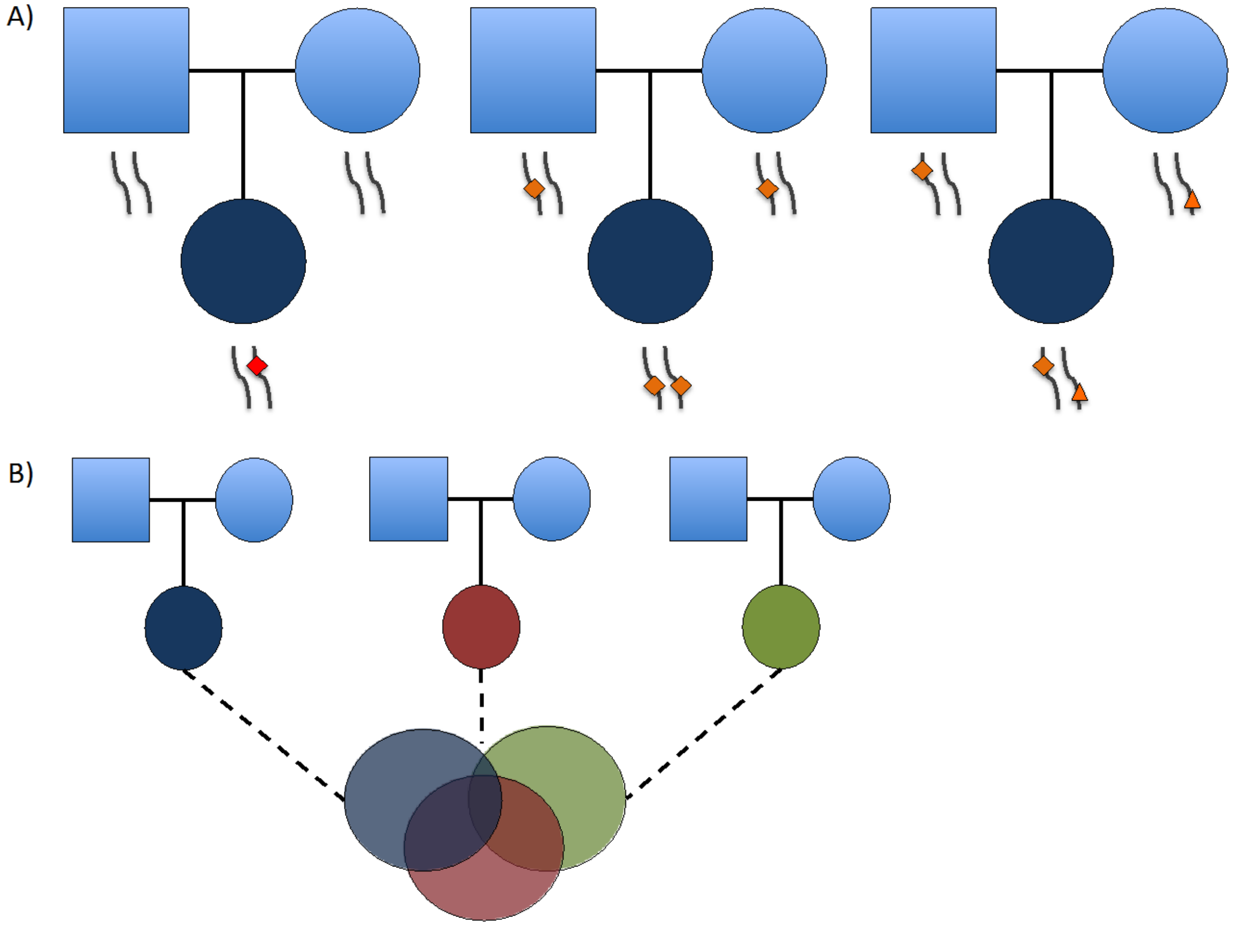


Figure 1-7 Selected filtering strategies for finding disease-causing variants using exome sequencing. Part A shows the 3 example Mendelian models evaluated in trio exome structure (e.g. father-mother-index). The leftmost figure is *de novo* dominant model, the middle one is homozygous recessive model, and the rightmost model is compound heterozygous. Part B shows a distinct strategy where the filtering is done by looking for shared mutations or shared

affected genes across multiple unrelated, affected individuals. Note: the pedigrees in the figure were drawn for illustrative purposes and are not meant to represent the pedigree rules used in clinical literature.

Note 1: as an example of NGS analysis pipeline, Figure 1-8 *describes a snapshot of the Wasserman lab NGS analysis pipeline as of November 2015. This is the pipeline that I helped assemble, and has been used to generate most of the datasets used for my thesis research (section 1.3).*

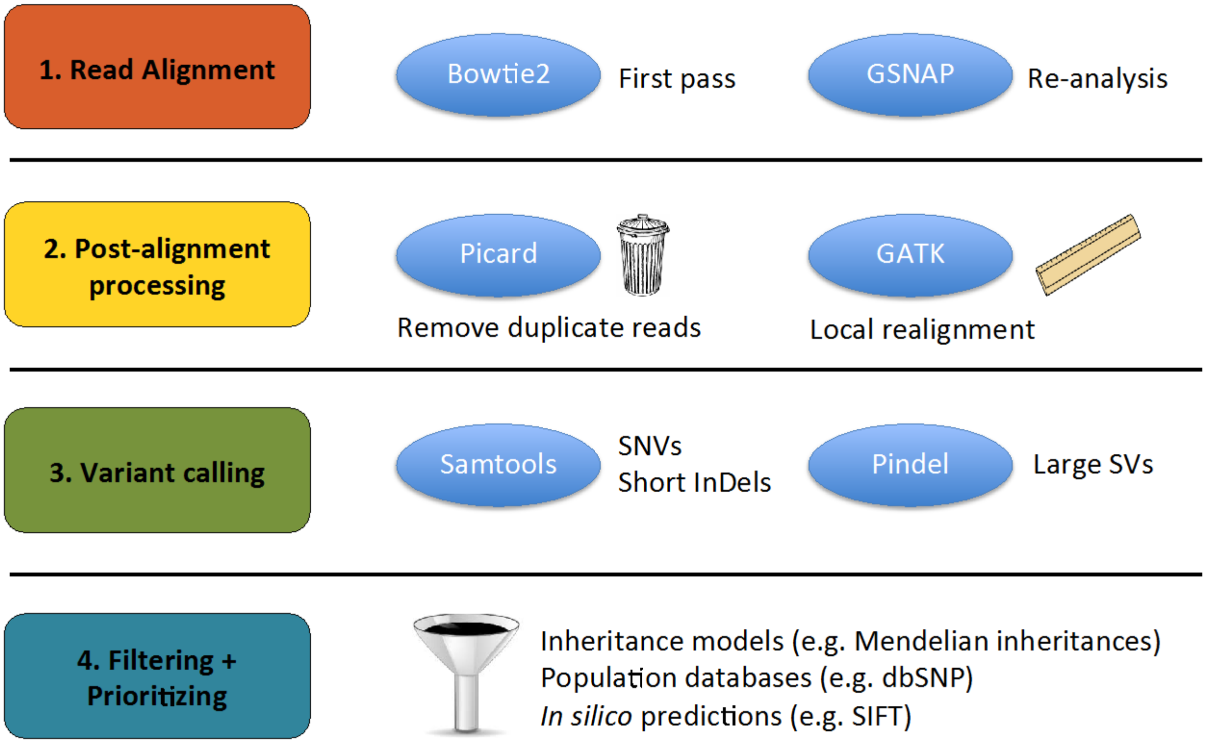

1. Read Alignment
Bowtie2
First pass
GSNAP
Re-analysis
2. Post-alignment processing
Picard
Remove duplicate reads
GATK
Local realignment
3. Variant calling
Samtools
SNVs
Short InDels
Pindel
Large SVs
4. Filtering + Prioritizing
Inheritance models (e.g. Mendelian inheritances)
Population databases (e.g. dbSNP)
In silico predictions (e.g. SIFT)

Figure 1-8 A snapshot of the Wasserman NGS pipeline in November 2015, and the tools involved. We feed the reads across a combination of different genomic aligners that align the reads to the human reference genome (GRCh37/hg19 and GRCh38/hg38 both available). Bowtie2 is used primarily as the first pass through the data, and if re-analysis of the data is needed, we try different aligners such as GSNAP, which is more flexible at detecting larger complex genomic variants (e.g. long insertions/deletions, splicing and gene fusions) than Bowtie2. Picard is next used to remove duplicate reads to correct for PCR amplification bias. GATK local realignment is used to ensure

the ends of the reads are aligned as accurately as possible. Samtools is used to call for SNVs and InDels. Pindel is used for whole-genomes to call for large structural variations (SVs; up to 10kb). The resulting variants are filtered and prioritized by selecting for variants that fit specific inheritance patterns (typically the classical Mendelian models), comparing against healthy and disease population databases, and assigning likelihood of deleteriousness to the remaining variants through in silico prediction software.

1.1.2 Clinical diagnostics

Clinical DNA-based diagnostics enable doctors to characterize an individual's vulnerabilities to inherited diseases and to predict the influence of genetic variation on therapeutic response. DNA-based diagnostics in the past have largely focused on specific genes known to cause disorders (e.g. Sanger-based sequencing on *BRCA1* for breast cancer[43]). When there are numerous genes to screen, such diagnostics are expensive to perform and time consuming to analyze. Presently, WGS is used in select cancer centers to help guide cancer treatment, and in certain medical centers for newborn screening, focusing on known severe childhood diseases. Within the coming decade, it is anticipated that full genome sequencing will be more widely adopted as the first-tier approach in place of specific gene tests[44].

Note 2: while new variants are constantly being reported with associations to specific phenotypes, noise exists in the literature and only a fraction of informative variants are reliable predictors of phenotype[45].

Note 3: The application of genome sequencing brings many ethical questions forward to address. While these issues are critically important, they are beyond the scope of this thesis.

1.1.2.1 Case studies with clinical exomes and whole genomes

The previous sections of this introduction have provided a basis for understanding exome and WGS sequencing, and how such data can be processed. To better demonstrate the clinical relevance of exome/whole genome sequencing and the resulting impact on patient treatment, in this section I highlight my selected co-authored publications of applied exomes and whole-genomes in clinical collaborations. There are two additional purposes to this section: 1) to exemplify my doctoral contributions beyond the chapters included in the thesis, 2) to illustrate the various strategies applied to cases of successful clinical diagnosis with exomes and whole-genomes. The studies below were initiated as part of the Treatable Intellectual Disability Endeavor in British Columbia and approved by the ethics committees of the University of British Columbia (Vancouver, Canada).

Case report 1: Discovery of a rare missense mutation p.Ala458Ser in gene *FAAH2* (OMIM#300654)[46]. During the first round of analysis the team was restricted to a single exome for the index case, which did not highlight a causal model in the midst of 65 candidate mutations. The recruitment of additional family members (parents and siblings) led to a different conclusion because additional data allowed intersection analysis to be performed to narrow the candidate list from 65 to 11. The role of *FAAH2* in the phenotype became more obvious when it was not buried among dozens of candidates. Biochemical and molecular modeling studies confirmed that the mutation resulted in a partial inactivation of *FAAH2,* leading to a disruption of the endocannabinoid signaling pathway. The molecular evidence was sufficient to suggest that phenotypically, this results in the manifestation of autistic features and associated movement abnormalities and learning disabilities as seen in the 2-year old male patient.

Case report 2: *RMND1* (OMIM#614917) deficiency is associated with congenital lactic acidosis, renal failure, deafness, and dysautonomia in neonates[47]. Through trio-exome sequencing on the parents and the index, we identified two heterozygous variants under a compound heterozygous model where each variant was inherited from one parent[47]. One of the two variants had been previously found to be partially disruptive to the gene function by an external lab (data unpublished at the time), leading us to reach out for experimental collaboration. Even though the second of the two variants was predicted to be benign by computational software SIFT and PolyPhenV2, the experimental side found significant reduced levels of *RMND1* in patient fibroblasts, the translation defect in these cells could be rescued with wild-type cDNA. The protein itself was almost undetectable by immunoblot analysis in patient muscle. The experience with this gene highlights the importance for research collaborations (we would not have pursued *RMND1* if we had not known one of the two variants was being studied by a specialized lab) and the potential danger of over-reliance on computational predictions. If a stringent removal of all predicted-benign mutations was applied, *RMND1* would never have made it into the candidate list.

Case report 3: Homozygous missense mutation in *ZFYVE20* (OMIM#609511) in a female patient with intractable seizures, dysostosis, macrocytosis and megalobastoid erythropoiesis[48]. The p.Gly425Arg mutation was identified through trio-sequencing on the parents and index, and later Sanger-confirmed to be absent in unaffected siblings. At the time of analysis, the gene was not known to cause any human disease, and our study was the first to associate the gene to a specific set of human phenotypes. We proposed that the mutation disrupted the endocytosis pathway, and consistent with our model, the mutant allele had a 50% decrease in transferrin accumulation (which was corrected by wild-type allele transfection), and patient’s fibroblasts displayed impaired proliferation rate, cytoskeletal and lysosomal abnormalities. The take-away message here is that by screening for variants beyond the known disease genes, one increases the likelihood for clinical diagnosis. While focusing on known disease genes may simplify the interpretation and accelerate the analysis, it runs the risk to miss novel discoveries for advancing medical genetics.

Case report 4: Second family with *AIMP1* (OMIM#603605) deficiency[49]. Through trio-exome sequencing, we identified a homozygous nonsense variant resulting in a truncated *AIMP1*[49]. Previous studies on the gene had reported a Perlizaeus-Merzbacher-like phenotype, but our female patient instead showed early-onset developmental arrest, intractable epileptic spasms, and a rapid clinical course leading to premature death consistent with a primary neuronal degenerative disorder. Since *AIMP1* has a known critical role in neurofilament assembly, its impairment would expectedly result in neuronal/axonal dysfunction. Our study therefore expanded the phenotype spectrum of a previously characterized disease gene, and shed light to the role of *AIMP1* in differential diagnosis of infantile onset, progressive neurodegenerative disease. The phenotype spectrums are constantly being refined, and this is especially true for rare diseases with limited samples; *AIMP1* would not have been selected if we had been overly stringent looking for complete phenotype matches with previous case reports.

1.1.2.2 Challenges

Several challenges remain for widespread incorporation of exome/whole-genome sequencing into clinical diagnosis and screening. One need is access to comprehensive annotations of variants discovered by diverse sequence collections to facilitate prioritization and filtering processes discussed in 2.2.3 and 2.3.1. Databases linking variants to diseases such as HGMD, CLINVAR, Human Variome Project (http://www.humanvariomeproject.org/) and GEN2PHEN (http://www.gen2phen.org) are efforts to compile comprehensive collections of validated variants, the later aiming to unify human and model organism genetic information to expand Genotype-To-Phenotype(G2P) mapping. Strategies for interpreting variants and guidelines/standards for reporting results from genomic projects at a clinical setting need to be established. Effective means of delivering such genomic information and computational predictions to healthcare professionals have not been sufficiently developed. Tools to empower clinicians to interpret and identify candidate variants are a key need for the clinical implementation of WGS (see section 1.2.2 for selected tools). These translational informatics approaches to convey variant information and facilitate diagnosis or prioritization of candidates is an unmet challenge, with each user community needing specific capacities. These challenges are the elements I address as the main focus of my PhD thesis.

1.2 Computer user-interface evaluation background

As genomic data gets incorporated into patient care, we begin to witness a coalescence of bioinformatics and health informatics. The result is the emergence of a field called biomedical informatics, defined as the "interdisciplinary, scientific field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health" (adapted from slides by EH Shortliffe, AMIA 2009).

A challenge of biomedical informatics within genomic medicine is to integrate and analyze high-throughput data to elucidate the molecular basis of a disease, and translate this knowledge into clinical practice[50]. Genomic medicine holds many applications in healthcare, including personalized medicine based on one's genetics, predictive methods for disease susceptibility and patient outcomes, and stratification. However, the advances in technology and computational methods can lead to a gap between information systems and clinicians, delaying the therapeutic impact of new technologies. Venues such as AMIA Summit on Translational Bioinformatics are formed to bridge this recognized gap. A similar approach is pursued by the eMERGE (electronic Medical Records and Genomics) Network to explore the benefits of coupling DNA repositories to electronic medical record systems to further promote genome science[51]. This consortium focuses primarily on GWAS genotyping data, but as NGS cost continues to decrease, we will need clinical systems specifically targeted for interpreting NGS output for healthcare professionals.

1.2.1 Clinical decision support systems

The genomic output from any NGS pipeline is not clinically useful unless it is properly translated for clinicians or hospital staff working with such data. The translation system has to provide intelligently filtered, patient-specific information and advice(s) to clinicians at the appropriate time. In human computer interaction, computerized clinical decision support systems (CDSS) are designed to improve clinical decision making[52]. The motivation behind CDSS is the acknowledgement that humans cannot follow clinical protocols flawlessly, and many processes can be computerized and automated to minimize errors and speed up decision making time[52]. Automation takes advantage of databases where the vast amount of information stored is beyond what humans can commit to memory. CDSS is most effective when delivered automatically as part of clinician workflow that fits the time and location of decision making, with a computer-generated recommended course of action for clinician consideration[53]. Lobache *et al.* performed a systematic review of clinical decision support systems and found over 90% of CDSS evaluated in randomized controlled trials have significantly improved patient care[54].

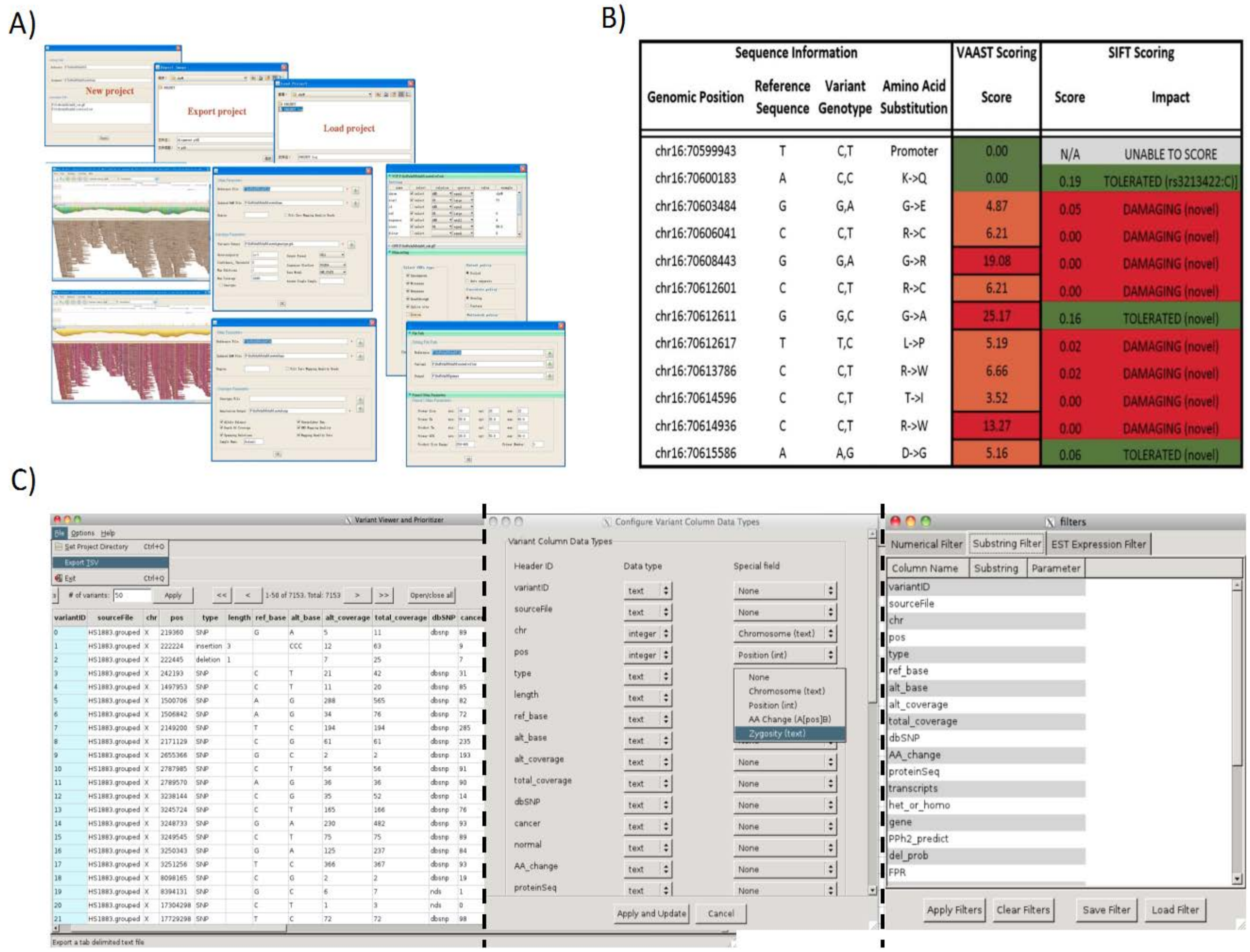
Such support systems have been implemented in numerous medical fields, including medicine, psychiatry, surgery, and pediatrics[55]. For instance, Emery *et al.* developed a decision support system for general practitioners to assess the genetic risk of breast and colorectal cancer by allowing doctors to create family-specific pedigree trees, report qualitative evidence for or against increased risk, and incorporate any rule-based genetic risk guideline. Much of the prior work has focused on gene-specific reporting. With the new genome-scale data comes a new set of challenges for the presentation of data to diverse clinician communities.

1.2.2 Towards visualization and interpretation of sequence variation data

The initial steps in designing genomics interpretation software have focused on users with a strong genetics background, from research biologists to clinical geneticists. Below I highlight a few key selected tools drawn from different software categories.

The first category of tools allows non-computational scientists to process and analyze genomic data in a manner consistent with a typical NGS pipeline (section 1.1.2.3). NextGENe (http://www.softgenetics.com/NextGENe.html) is a self-contained commercialized software package that enables researchers without strong programming and database skills to process data from exome or whole-genome sequencing. Users supply raw sequenced reads and the program handles the mapping, the variant calling, and returns a list of SNVs/InDels discovered, along with annotations for each variant such as the gene that is impacted, the codon change, the genotype, as well as hyperlinks to dbSNP and GenBank. The software has built-in functions for filtering variants to generate specific subsets, such as variants predicted to be damaging, or variants predicted to fit a specific inheritance pattern based on a family pedigree. NextGENe also has a built-in browser to browse the genomic window for a variant and see how the sequenced reads mapped at the location.

While NextGENe is not freely accessible, open source academic software such as MagicViewer[56] also exists, with similar capabilities that allows users without in-depth computational skills to align genomic sequences, perform their own variant calling, attach annotations to the called variants, filter variants based on a set of user-defined criteria, and visualize sequence depth distribution of mapped reads along the reference genome (Figure 1-9). The Galaxy system[57] provides another open-source option for processing FASTQ data, and viewing the annotated sequence variants. Users interact with the whole pipeline via the Galaxy web browser, and unlike the previous two, it requires no local installation. The pipeline starts with quality control on the input dataset, displaying summary information such as per base/sequence quality scores, sequence length distribution, and GC content. This is followed by read mapping, and users have the option to incorporate command-line software such as GATK into Galaxy browser to interact with it graphically instead of via command-line terminal. While Galaxy offers greater flexibilities to incorporate a diversity of open source tools than NextGENe and MagicViewer, Galaxy has a more limited viewing option with less emphasis on filtering capabilities[58].



Sequence Information				VAAST Scoring	SIFT Scoring	
Genomic Position	Reference Sequence	Variant Genotype	Amino Acid Substitution	Score	Score	Impact
chr16:70599943	T	C,T	Promoter	0.00	N/A	UNABLE TO SCORE
chr16:70600183	A	C,C	K->Q	0.00	0.19	TOLERATED (rs3213422:C)]
chr16:70603484	G	G,A	G->E	4.87	0.05	DAMAGING (novel)
chr16:70606041	C	C,T	R->C	6.21	0.00	DAMAGING (novel)
chr16:70608443	G	G,A	G->R	19.08	0.00	DAMAGING (novel)
chr16:70612601	C	C,T	R->C	6.21	0.00	DAMAGING (novel)
chr16:70612611	G	G,C	G->A	25.17	0.16	TOLERATED (novel)
chr16:70612617	T	T,C	L->P	5.19	0.02	DAMAGING (novel)
chr16:70613786	C	C,T	R->W	6.66	0.02	DAMAGING (novel)
chr16:70614596	C	C,T	T->I	3.52	0.00	DAMAGING (novel)
chr16:70614936	C	C,T	R->W	13.27	0.00	DAMAGING (novel)
chr16:70615586	A	A,G	D->G	5.16	0.06	TOLERATED (novel)

Variant Viewer and Prioritizer

File Options Help

Set Project Directory Ctrl+O
Export TSV
Exit Ctrl+Q

# of variants: 50 Apply << < 1-50 of 7153. Total: 7153 > >> Open/close all

variantID	sourceFile	chr	pos	type	length	ref_base	alt_base	alt_coverage	total_coverage	dbSNP	cance
0	HS1883.grouped	X	219360	SNP		G	A	5	11	dbsnp	89
1	HS1883.grouped	X	222224	insertion	3		CCC	12	63		9
2	HS1883.grouped	X	222445	deletion	1			7	25		7
3	HS1883.grouped	X	242193	SNP		C	T	21	42	dbsnp	31
4	HS1883.grouped	X	1497953	SNP		C	T	11	20	dbsnp	85
5	HS1883.grouped	X	1500706	SNP		A	G	288	565	dbsnp	82
6	HS1883.grouped	X	1506842	SNP		A	G	34	76	dbsnp	72
7	HS1883.grouped	X	2149200	SNP		T	C	194	194	dbsnp	285
8	HS1883.grouped	X	2171129	SNP		C	G	61	61	dbsnp	235
9	HS1883.grouped	X	2655366	SNP		G	C	2	2	dbsnp	193
10	HS1883.grouped	X	2787985	SNP		C	T	56	56	dbsnp	91
11	HS1883.grouped	X	2789570	SNP		A	G	36	36	dbsnp	90
12	HS1883.grouped	X	3238144	SNP		C	G	35	52	dbsnp	14
13	HS1883.grouped	X	3245724	SNP		C	T	165	166	dbsnp	76
14	HS1883.grouped	X	3248733	SNP		G	A	230	482	dbsnp	93
15	HS1883.grouped	X	3249545	SNP		C	T	75	75	dbsnp	89
16	HS1883.grouped	X	3250343	SNP		G	A	125	237	dbsnp	84
17	HS1883.grouped	X	3251256	SNP		T	C	366	367	dbsnp	93
18	HS1883.grouped	X	8098165	SNP		C	G	2	2	dbsnp	19
19	HS1883.grouped	X	8394131	SNP		G	C	6	7	nds	1
20	HS1883.grouped	X	17304298	SNP		C	T	1	3	nds	0
21	HS1883.grouped	X	17729298	SNP		T	C	72	72	dbsnp	98

Export a tab delimited text file

Configure Variant Column Data Types

Variant Column Data Types

Header ID	Data type	Special field
variantID	text	None
sourceFile	text	None
chr	integer	Chromosome (text)
pos	integer	Position (int)
type	text	None Chromosome (text) Position (int) AA Change (A[pos]B) Zygosity (text)
length	text	
ref_base	text	
alt_base	text	
alt_coverage	text	None
total_coverage	text	None
dbSNP	text	None
cancer	text	None
normal	text	None
AA_change	text	None
proteinSeq	text	None

Apply and Update Cancel

filters

Numerical Filter Substring Filter EST Expression Filter

Column Name	Substring	Parameter
variantID		
sourceFile		
chr		
pos		
type		
ref_base		
alt_base		
alt_coverage		
total_coverage		
dbSNP		
AA_change		
proteinSeq		
transcripts		
het_or_homo		
gene		
PPh2_predict		
del_prob		
FPR		

Apply Filters Clear Filters Save Filter Load Filter

Figure 1-9 Distinct interface designs for processing, manipulating, and displaying exome variants. Part A shows the interface of MagicViewer for uploading datasets, visualizing the alignment, and filtering for variants of interest. Part B shows an example output from VAAST, which focuses more on variant prioritization instead of visualization. Part C shows three separate views from VVAP, an in-house variant viewer and prioritization software developed by Dr. Jan Friedman’s laboratory at Child Family Research Institute in Vancouver, British Columbia, Canada. The left figure is the variant view of VVAP, displaying variant information in a format similar to Microsoft Excel. The information can be exported as a tab-delimited file. The figure in the middle shows how user can customize many aspects of the input file format by specifying which column in the input file corresponds to which information header. The rightmost figure shows the options for filtering, such as based on specific genomic position, genotype, gene ID, or expression levels.

Not all software is designed to cover the entirety of genome analysis. Some software has been developed with the assumption that the users will perform specific tasks in distinct packages, such as the generation of a list of variants. In such cases, emphasis may be placed on assisting users to interpret the functional roles of the identified variants, and to prioritize them for candidate mutation identification. SVA[59] is a visualization tool specialized for assigning functional associations to each predicted variant and visualizing them in a browser. The method of annotation is primarily done by linking each variant to the gene that it is situated in; once the affected gene name is established, gene functions are compiled by integrating information from Ensembl core, gene ontology, HapMap, 1000 Genomes Project, and KEGG Pathways. As proof of concept, the developers analyzed a patient with metachondromatosis using SVA's filtering function to filter for variants absent in dbSNP and absent in control genomes, and focused on protein-truncating variants. They identified an 11-bp frameshift deletion on *PTPN11* gene known to cause metachondromatosis. Due to the diverse databases that SVA depend on, SVA requires a local memory of over 8.9Gb on the user's machine in order to run, which may not be feasible in a typical clinical computing station.

To bypass the need for high-end computational infrastructure, VarSifter[60] is a light-weight software developed at NIH to allow biological investigators of varying computational skills to quickly sort, filter and sift through sequence variation data. The software takes in VCF files as input, and displays the variants in a GUI with each row representing a variant, and columns represent the annotations such as genotypes, quality scores, and read depth. Once input is loaded, users can setup customized hierarchical set of filtering criteria using logical "AND/OR/XOR" connections. The program supports regular expressions to add flexible text matching. To demonstrate Varsifter's lightweight ability to run on a typical computing station, the authors compared VarSifter to Microsoft Excel and found similar RAM (random access memory) usage, with a faster time to load the data.

The last category of tools presented are those with built-in filtering statistical algorithms to return a list of candidate mutations to users without depending on user specification of filtering criteria. VAAST[61] is a probabilistic variant prioritization program that combines amino acid substitution effect with variant frequency data to identify causative mutations. While useful for identifying variants/genes involved in disease, this tool is not meant for visualizing alignment/variant data as a whole (Figure 1-9).

With so many tools available, it is easy for a user to be overwhelmed and not certain which software is suitable for his/her research needs. Not all tools are designed to handle certain situations that may benefit by having the system generate a specific recommended course of action, or automatically flag information of interest. Furthermore, at the start of thesis research, no proper usability tests had been performed on existing systems to evaluate their usability and impact on workflow efficiency. Certain software, although claimed to address a certain task, may not actually deliver the expected results or may be too complicated for a user to follow. The main exploration of this thesis is the usability of software for applied clinical genome sequence interpretation, including the identification of strengths and weaknesses which impact the speed and accuracy of user decision making, ultimately seeking ways to enable the usage of full genome sequencing in healthcare.

1.2.3 Usability in human-computer interaction

Before deployment, it is essential for the usability of any CDSS to be comprehensively studied to ensure the system adds improvements, and does no harm to the patients (i.e.

technology-induced errors)[62]. Usability is defined as “the capacity of a system to allow the targeted users to carry out their tasks safely, effectively, efficiently and enjoyably”[63]. More specifically, it refers to how quickly the user can attain mastery on the tool, how much benefit and joy it brings to the user, and how error-prone the system is[38]. The vast number of health care information systems setup over the recent years highlights the need for effective ways to evaluate performance and impact to ensure these systems meet the requirements and computational needs for end users and health care organizations[64]. The traditional approaches to assessing information systems are focused on determining how well a system meets its set of pre-defined goals regarding functionality, safety, and impact on outcome measure such as cost and work efficiency. This is frequently performed at the conclusion of the software development cycle, when a complete functional system has been implemented and ideally ready to be shipped for deployment. However, a recent trend in the field of human computer interface has shifted towards the development of evaluation approaches that can be used during the software development stage[39]. The motivation of this approach is that obtaining iterative evaluations while the design is still in progress provides greater flexibility and time to improve the architecture design and deployment of the system (Figure 1-10).

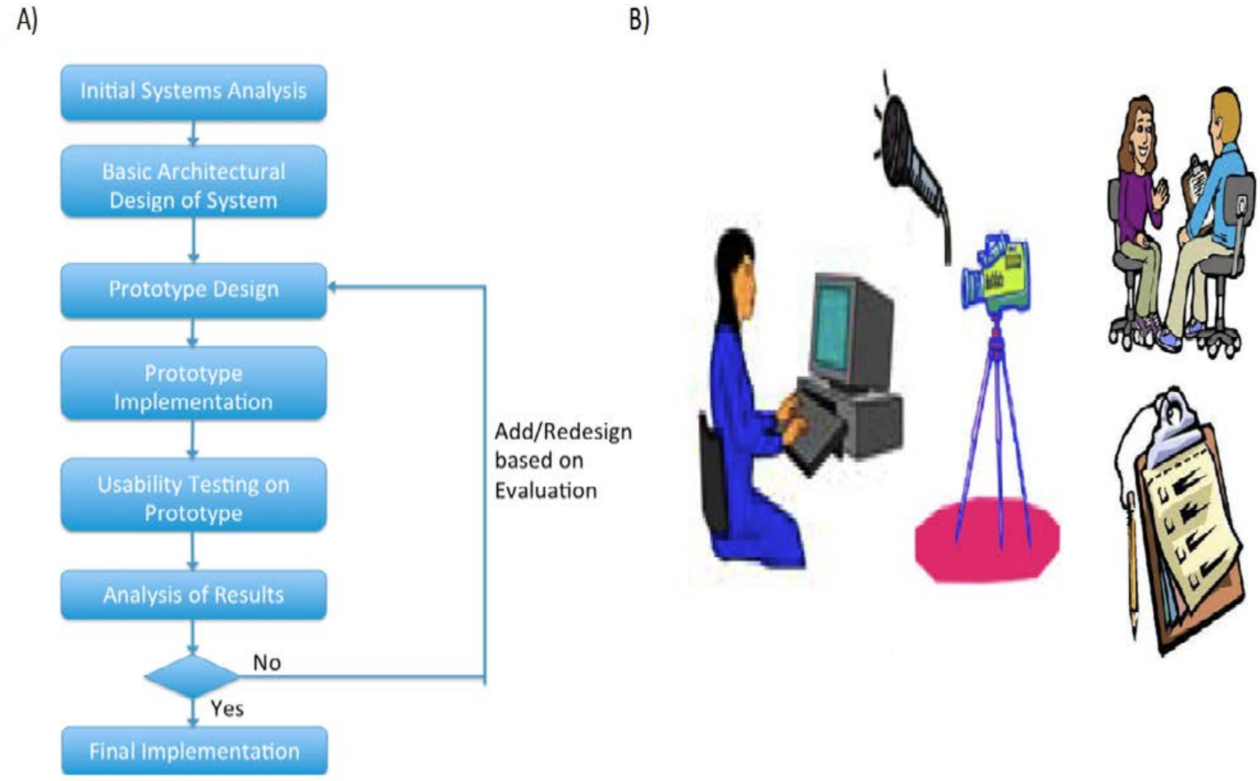


Figure 1-10 Prototype design and evaluation. Part A shows the iterative systems development based on prototyping and iterative usability testing, adapted from [65]. Part B shows the basic setup of a simulation-based evaluation with screen capture and video recordings. Subjects are asked to perform particular tasks using the computer system, and both visual and audio data are captured. The analysis can be complemented by focus-group interviews, and surveys to provide both qualitative and quantitative data.

1.2.3.1 Common methods for evaluating decision support systems

The assessment of health information system involves characterizing a) how easily a user can carry out a task with the system, b) how quickly the user can attain mastery in using the system, c) what effects the system has on work practices and d) what problems exist for the users when interacting with the system[66]. The evaluation outcomes include numerical measurement, which is considered to be more precise, replicable and “objective” compared to subjective measurements[40]. Subjects that are selected to participate in the evaluation should consist of representative target users of the system. Nielsen J. (1993) reports up to 80% of user-interface problems can be detected with as few as 8-12 evaluated individuals[67].

One popular approach to evaluating decision support systems is questionnaire-based surveys[68]. In this scenario, numerical data is gathered from the questions, and simple statistical tests such as a T-test can be used to derive response differences between groups of users. Survey methods hold the advantages of being easy to distribute to large number of users (e.g. via the web), and can be fed into a pipeline for automated analysis of results[68]. Surveys such as Questionnaire for User Interaction Satisfaction (Table 1-4), serve as skeletal examples to test multiple dimension of usability. The main drawback for surveys is the items in the questionnaires are pre-determined and consequently are of limited value in identifying new or emergent issues[65]. Furthermore, it has been reported that people may recall their experience differently when rating a system using the questionnaire versus their experience captured in real-time[65]. Interviews and focus groups allow exploration of more open-ended questions and responses for unexpected aspects can be collected [65]. However, the "recollection problem" still applies when participants are asked to recall their experience with the system during interviews or focus groups. Therefore evaluation performed in real-time while a subject uses software can be particularly valuable.

Examples of Questions	**Corresponding options**
How many operating systems have you worked with?	none, 1, 2, 3-4, 5-6, more than 6
How long have you worked on this system?	less than 1 hour, 1 hour to less than 1 day, 1 day to less than 1 week…etc
On average, how much time do you spend per week on this system?	less than 1 hour, one to less than 4 hours, 4 to less than 10 hours…etc
Rate your reactions to the system	1 to 9 (1=terrible, 9=wonderful)
	1 to 9 (1=rigid, 9=flexible)
Screen layouts were helpful	1 to 9 (1=never, 9=always)
Sequence of screens	1 to 9 (1=confusing, 9=clear)
Use of terminology throughout system	1 to 9 (1=inconsistent, 9=consistent)
System keeps you informed about what it is doing	1 to 9 (1=never, 9=always)
Learning to operate the system	1 to 9 (1=difficult, 9=easy)
System speed	1 to 9 (1=too slow, 9=fast enough)
Technical manuals are	1 to 9 (1=confusing, 9=clear)
Quality of pictures/photographs	1 to 9 (1=bad, 9=good)
Speed of installation	1 to 9 (1=slow, 9=fast)
Customization	1 to 9 (1=difficult, 9=easy)

Table 1-4 A few examples of usability assessment questions manually extracted from QUIS, a questionnaire developed at University of Maryland for assessing user's subjective satisfaction with specific aspects of the human-computer interface, including screen factors, terminology, system feedback, learning factors, system capabilities, software installation, and technical manuals.

1.2.3.2 Evaluations via simulations

Cognitive task analysis (CTA) is a real-time alternative approach to surveys/interviews that particularly addresses decision making and reasoning skills as users perform activities/tasks that involve handling complex information[69]. The first step in CTA is to develop a task hierarchy describing and cataloging the work activities that take place (e.g. physician entering patient data into an electronic health system). Then subjects are asked to perform specified tasks using the system while being recorded (Figure 1-10). The tasks should reflect the typical activities that a normal user would perform in a typical work routine. If the system is still under development, the tasks may be specific to certain aspects of the system or user-system interaction that requires only a partially functioning prototype[69]. Subjects recruited for the simulation may already be experienced with the system, or may be given training on the system prior to or as part of the evaluation. Users ought not be given free-reign to explore the usability of a prototype, but instead be directed to explore specific functionalities or perform a specific task. In cognitive walkthrough, the users systematically step through a list of chosen tasks relevant to the specific system design. An example of a cognitive walkthrough was conducted by Currie *et al.* [70] on a decision support system intended for use in an intensive care setting. The study evaluated the system and assessed the cognitive processes required to prescribe antibiotics to premature infants - a complex clinical task that requires intimate knowledge about multiple patient parameters, yet the decision must be made rapidly and accurately due to high risk of morbidity and mortality associated with sepsis.

Affordable equipment such as screen capture software, cameras, and microphone can record the totality of user-exhibited responses. Subjects can be encouraged to express their thoughts verbally so these can later be transcribed into text files. Such studies can reveal how user characteristics (e.g. differences in experience with computer) relate to usability. CTA can be complemented by interview/questionnaire-style questions at the end of the hands-on session. It can be useful to leave the audio recording running after the question period, as further spontaneous feedback is often offered at this point about some features of the system[66]. To compare between different system prototypes, the design of evaluation studies can be within group, where individuals are asked to try all prototype versions sequentially, or may be performed in time-series in which performance is tracked over time. Between-group testing comparing multiple groups segregated in some manner can also be pursued (e.g. those highly computer literate versus those with little computer experience)[66].

1.2.3.3 Interpretation of usability data

The advantage of video recordings in a human-computer interaction study is it captures the interaction real-time without depending on user's recollections, and the same video recording can be examined and analyzed from multiple perspectives and re-analyzed with a range of methodological approaches. The visual and verbal data collected by simulations are best assigned to different coding categories to facilitate downstream analysis[41]. An example category scheme for categorizing subjects' verbal comments may include: interface consistency, response time, comprehensibility of system messages, help availability, comprehension of graphs and tables, and challenges to entering data[41]. Additional code(s) can be created to address a particular task that fits under a specific analysis. To annotate video-based data, the coding scheme can include interface problems (e.g. data entry, provision of too much or too little information, navigation), content problems (e.g. certain information absent from the database, can user flag a certain field), and slips (errors made by the system, and whether user is able to notice this and correct)[71]. The verbal text files are annotated with the observed problem by a time-stamp that tells the time of occurrence and type of problem (e.g. 01:33 COMMENT:CRITIQUE NAVIGATION)[41].

Once the data is annotated, it can be summarized in a number of ways. The topics to cover generally include task accuracy, user preference, time to complete a task, frequency and classes of problems encountered[72]. One way to present the evaluation is by a summary of types and frequency of problems detected when subjects interact with the system[42]. If the evaluation involves an iterative cycle, this summary should be shared with system designers, then repeated to determine how the new software prototype performs compared to the old. If multiple prototypes are evaluated against each other, statistical tests such as a Chi-squared test can be applied to assess if one system outperforms another under a specific category.

Usability testing method is not limited to recruiting representative end-users, but can also be performed with a usability analyst who notes the problems/cognitive issues as he/she steps through a system[65]. Such findings can then be compared against a pre-established principle of usability and good design. While relatively cost-effective, this type of method requires an analyst trained in human computer interaction (HCI) methodology. For my thesis research, my focus was placed on obtaining feedback directly from clinicians.

1.3 Thesis structure

For my thesis, I set out to apply the principles of software interface design and evaluation to the emerging problem of exome and genome sequence interpretation. The body of my thesis is divided into three themes. The first theme focuses upon software usability, where I evaluated the designs of existing software and prototypes for clinical genome analysis. This work spanned early approaches based on existing tools to concept design by users. Arising from the usability studies, one critical issue highlighted was a lack of clinical decision support tools for variant prioritization. To address the unmet needs, the second theme focuses upon development of a novel bioinformatics solution to improve the accuracy and efficiency in variant prioritization for exomes and whole-genome datasets. The third theme is a culmination of the interface design approaches and developed bioinformatics strategies applied on actual patient data. This work is done in collaboration with a multidisciplinary team of healthcare professionals. Below, I break down to how the thesis is organized and what each chapter constitutes.

1.3.1 Chapter 2 and Chapter 3: Usability evaluation on genomic interfaces

The evaluation of genomic software is presented in two chapters. In chapter 2, cognitive task analysis combined with a think-aloud protocol was performed with clinical geneticists to evaluate the functionalities and usability of existing exome interpretation interfaces. There are two study objectives: 1) To ascertain the key features of successful user interfaces for clinical exome analysis software based on the perspective of expert clinical geneticists, 2) To assess user-system interactions in order to reveal strengths and weaknesses of existing software, inform future design, and accelerate the clinical uptake of exome analysis. My work was the first published application of usability methods to evaluate software interfaces in the context of exome analysis. The results highlight how the study of user responses can lead to identification of usability issues and challenges and reveal software reengineering opportunities for improving clinical next-generation sequencing analysis.

Chapter 3 addresses limitations of the earlier chapter and expands upon it. More specifically, chapter 2 only addressed existing software, thereby constraining user feedback. It was limited to clinical geneticists, failing to address involvement of other healthcare providers, e.g. genetic counselors. Chapter 3 employs a focus group approach, combined with cognitive walkthroughs on prototypes, to ascertain perspectives from healthcare professionals in distinct domains on optimal clinical genomic user interfaces, and digital prototypes that highlighted future software engineering opportunities were translated from users' feedbacks without the constraint of existing software architectures. Each group of users revealed distinct needs and desires.

1.3.2 Chapter 4 and Chapter 5: Novel algorithm for gene-variant prioritization

As rare/novel genetic variants continue to be uncovered, there is a major challenge in distinguishing true pathogenic variants from rare benign mutations. The efficiency and adoptability of exome analysis rests heavily on the ability for software to reliably distinguish pathogenic mutations from rare benign variants within a short amount of time. In chapter 4, through the analysis of public exome datasets, we show that some genes are frequently affected by rare, likely functional variants in general population, and are frequently observed in exome studies analyzing diverse rare phenotypes. We find that the rate at which genes accumulate rare mutations is beneficial information for prioritizing candidates, and propose that clinical reports associating any disease/phenotype to the frequently mutated genes be evaluated with extra caution.

In chapter 5, I expand upon the variant prioritization challenge, presenting a novel method called Variant Prioritization Accelerator (VPA), which utilizes an ensemble machine learning approach trained on variant-level, gene-level and patient-level information for classifying rare variants according to likelihood of pathogenicity. The chapter discusses the advantages to the approach over existing methodologies, and its superior performance to correctly rank novel gene-disease and disease-phenotype associations, and thereby facilitate clinicians to achieve accurate diagnostic from exome data within a given limited amount of time.

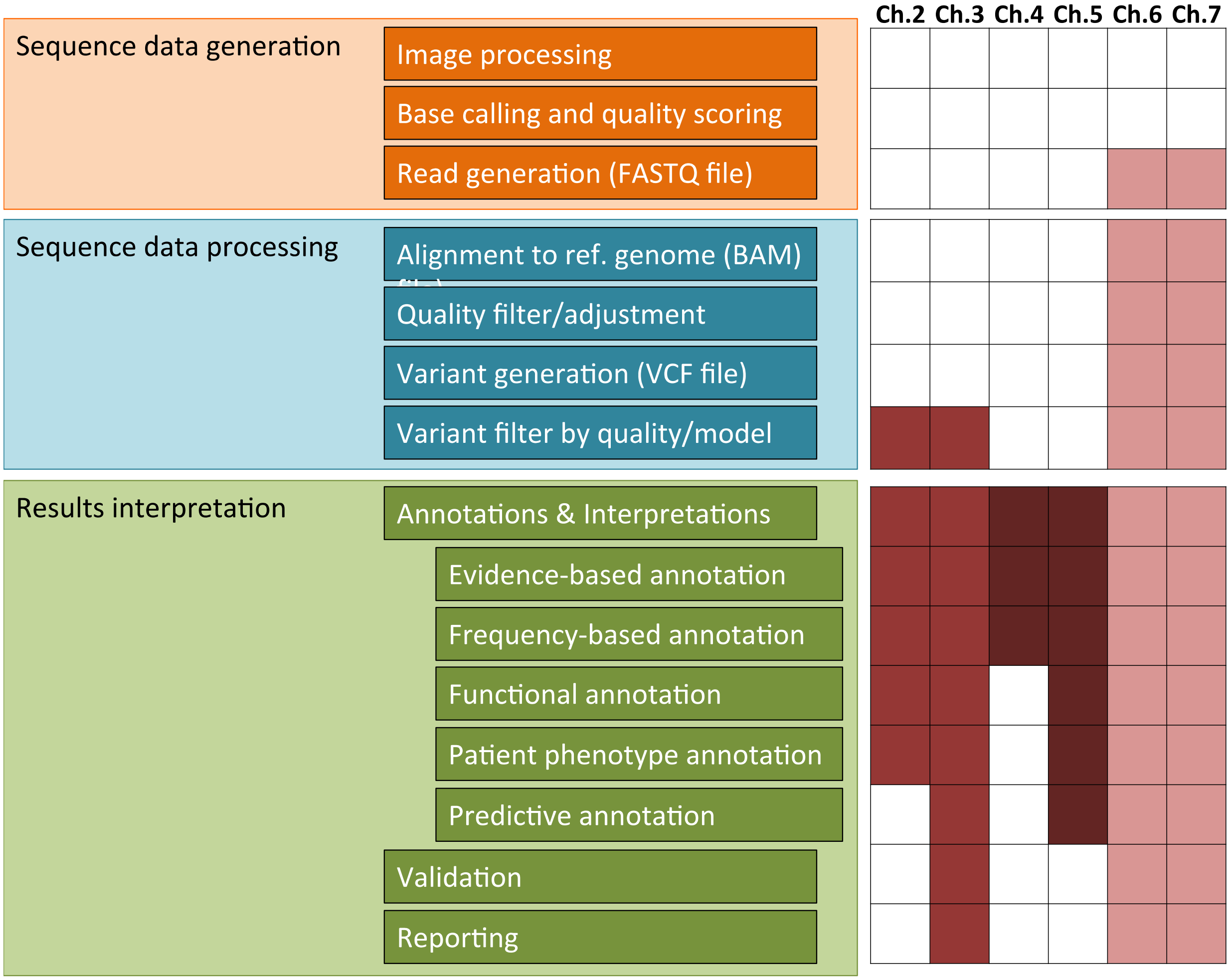
1.3.3 Chapter 6 and Chapter 7: Collaborations in genomic projects

The thesis research culminates in applied collaborative interpretive studies of patient genomes. The collaborations immersed the thesis research in the clinic-facilitated recruitment of subjects, and enabled me to envision system modifications likely to be impactful on users. Below I describe the main collaboration, OMICS2TREATID (http://omics2treatid.org), within which the thesis research was embedded.

Spearheaded by Dr. Clara van Karnebeek, the project focuses upon prevention and treatment of intellectual disability, with specific emphasis on children with treatable genetic conditions called inborn errors of metabolism (IEM). While amenable to treatment[2], many cases of IEM are unfortunately missed in clinical diagnosis due to limitations of karyotyping and arrayCGH, the clinical standard protocols for detection[73]), and a lack of standard protocols and systematic approach for IEM identification. My contribution to OMICS2TREATID includes the bioinformatics processing of patient DNA through the various stages of exome and genome analysis to generate the list of high-confidence variants, applying computational strategies to interpret variants by likelihood to being disease-causing candidates, and providing continuous communications between various domain-specific healthcare providers to decide the best course of action per patient. Chapter 6 describes our first successful discovery of a novel treatable neuro-metabolic disease where we characterized three independent families displaying hyperammonemia with CA5A gene defect. Chapter 7 provides an overview of OMICS2TREATID successes, achieved through a combination of deep clinical phenotyping with exome sequencing analysis via an unbiased semi-automated bio-informatics pipeline. Our

[2] Successful treatment is defined either by improvements in IQ/developmental scores, or biochemical signatures. See [49] for more details.

diagnostic yield and discovery rate exceeded expectation (with 43% exome diagnosis allowed for personalized medicine (or precision medicine, which in this thesis are treated the same), spanning from prevention and tailored symptom management to causal therapy). The results of these studies constitute the evaluative data used in Chapter 5.



Ch.2
Ch.3
Ch.4
Ch.5
Ch.6
Ch.7
Sequence data generation
Image processing
Base calling and quality scoring
Read generation (FASTQ file)
Sequence data processing
Alignment to ref. genome (BAM)
Quality filter/adjustment
Variant generation (VCF file)
Variant filter by quality/model
Results interpretation
Annotations & Interpretations
Evidence-based annotation
Frequency-based annotation
Functional annotation
Patient phenotype annotation
Predictive annotation
Validation
Reporting

Figure 1-11 The process diagram illustrates the major components of a NGS analysis when looking for causal variants for rare disorders. The workflow is represented as 3 main stages: data generation that leads to the generation of sequenced reads in text format, data processing of the reads to detect variants, and results interpretation of the variants, ideally culminating in clinical decision-making. The general framework varies depending on the precise analytical application, and specific details are omitted for discussion later in the thesis, The colored cells to the right

illustrates which part of the workflow I contributed to and their corresponding chapters. The scale of color shows the level of details each chapter touches upon the components in the pipeline (e.g. the darker the shade, the more in-depth the said component is addressed).

1.4 Conclusion

Figure 1-11 provides a schematic overview of the start to end processes involved in determining germline causal variants of rare diseases, with indication on the parts of workflow my contributors relate to and their corresponding chapters. Together the three components of the thesis come together to demonstrate how the design of clinical genome interpretation methods within biomedical informatics emerge from clinical software usability studies, both quantitative and qualitative. The importance of revealing the limitations of early approaches through direct engagement with clinical users, the response to the users through the creation of tools specifically tailored to meet their expressed needs, and the shared mission of bioinformaticians, biologists and clinicians to bring the power of the new technologies into clinical research.

Chapter 2: Usability study of clinical exome analysis software

2.1 Synopsis

Objectives: New DNA sequencing technologies have revolutionized the search for genetic disruptions. Targeted sequencing of all protein coding regions of the genome, called exome analysis, is actively used in research-oriented genetics clinics, with the transition to exomes as a standard procedure underway. This transition is challenging; identification of potentially causal mutation(s) amongst ~10^6 variants requires specialized computation in combination with expert assessment. This study analyzes the usability of user interfaces for clinical exome analysis software. There are two study objectives: 1) To ascertain the key features of successful user interfaces for clinical exome analysis software based on the perspective of expert clinical geneticists, 2) To assess user-system interactions in order to reveal strengths and weaknesses of existing software, inform future design, and accelerate the clinical uptake of exome analysis.

Methods: Surveys, interviews, and cognitive task analysis were performed for the assessment of two next-generation exome sequence analysis software packages. The subjects included ten clinical geneticists who interacted with the software packages using the "think aloud" method. Subjects' interactions with the software were recorded in their clinical office within an urban research and teaching hospital. All major user interface events (from the user interactions with the packages) were time-stamped and annotated with coding categories to identify usability issues in order to characterize desired features and deficiencies in the user experience.

Results: We detected 193 usability issues, the majority of which concern interface layout and navigation, and the resolution of reports. Our study highlights gaps in specific software features typical within exome analysis. The clinicians perform best when the flow of the system is structured into well-defined yet customizable layers for incorporation within the clinical workflow. The results highlight opportunities to dramatically accelerate clinician analysis and interpretation of patient genomic data.

Conclusion: We present the first application of usability methods to evaluate software interfaces in the context of exome analysis. Our results highlight how the study of user responses can lead to identification of usability issues and challenges and reveal software reengineering opportunities for improving clinical next-generation sequencing analysis. While the evaluation focused on two distinctive software tools, the results are general and should inform active and future software development for genome analysis software. As large-scale genome analysis becomes increasingly common in healthcare, it is critical that efficient and effective software interfaces are provided to accelerate clinical adoption of the technology. Implications for improved design of such applications are discussed.

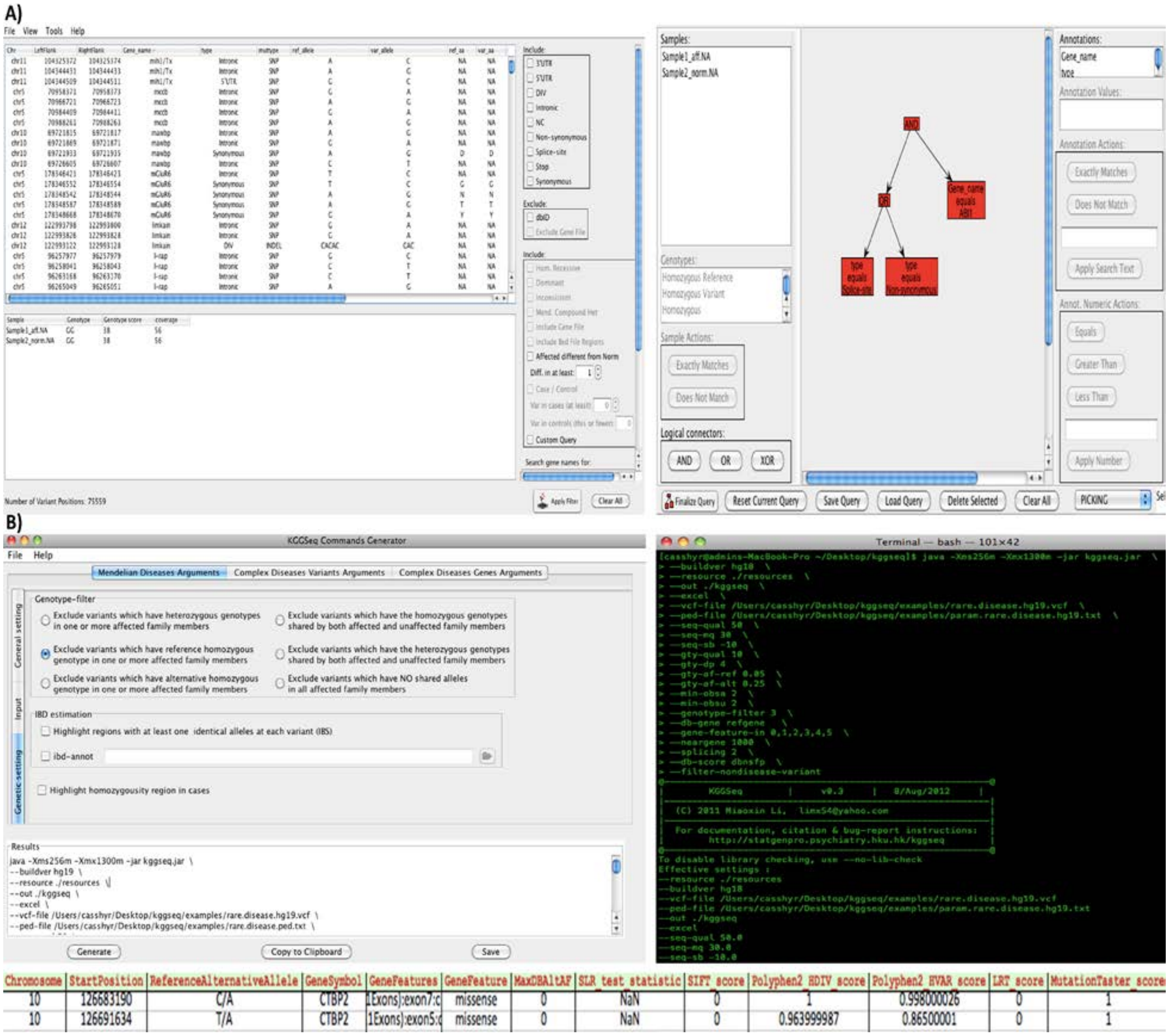
2.2 Introduction

The data output from exome sequencing is immense and computationally complex, and finding relevant sequence variations amongst the hundreds of thousands of variants in each individual remains an ongoing challenge[74-76]. Various software packages have been developed for visualization and interpretation of sequence variation data to address this challenge, but to date no comprehensive usability studies have been reported to identify and investigate user interface features required for efficient clinical work involving exome analysis.

Prior studies illustrate how a lack of systematic consideration of users, the tasks they are involved with, and their work environments can result in poorly designed user interfaces, leading to low adoption rates[77-79]. Such systems are likely to be abandoned[80, 81]. In healthcare systems, poorly designed systems may also jeopardize the quality of patient care, and pose a threat to patient safety and waste precious resources[82-84]. Usability studies in the field of health informatics focus on analyzing user behavior to reveal cognitive and behavioral patterns that may explain such suboptimal outcomes[85, 86], as well as reveal technological considerations that impede clinical translation of patient genomics[18-20].

In the context of usability studies in bioinformatics, Bolchini and colleagues have identified a need for the application of usability analysis to the evaluation of bioinformatics resources and tools[87]. However, there have been few published studies on the usability of such technologies. Usability analysis, involving standard usability testing techniques, have recently been described by Neri and colleagues in the analysis of a user interface for genetic results that are presented to healthcare providers for managing patient genetic profiles. Neri found that usability testing resulted in the identification of problems which were resolvable with simple alterations leading to substantial impact on the quality of user interactions[88].

The framework of our research study is based upon cognitive task analysis (CTA), a cognitive engineering technique that has been successfully applied in informing the design of systems across a variety of clinical domains[89-94]. In this paper, we present the first evaluation of the usability of next-generation sequencing interpretation software, exploring the impacts of different user interface designs on analysis workflows and outcomes. Our methodology builds upon well-established CTA to observe ten clinical geneticists examining two simulated scenario cases using think-aloud protocols[95] to assess end user cognitive behavior. Each subject worked through two hypothetical exome analysis scenarios with two dissimilar exome analysis software interfaces. We highlight the top user desiderata to inform software developers working on the next generation of exome interpretation software, and to inform clinical users who are in the process of choosing a software from this domain. The discussion of this paper addresses recurring usability challenges to overcome and critical features that this class of analysis software should possess. We emphasize that the ultimate goal of the study is not the collection of software-specific usability analysis results, rather the intent is to highlight findings that generalize to all software geared for the clinical interpretation of exome and eventually whole genome sequencing data.



Chr	LeftFlank	RightFlank	Gene_name	type	muttype	ref_allele	var_allele	ref_aa	var_aa
chr11	104325372	104325374	mlh1/Tx	Intronic	SNP	A	C	NA	NA
chr11	104344431	104344433	mlh1/Tx	Intronic	SNP	A	C	NA	NA
chr11	104344509	104344511	mlh1/Tx	5'UTR	SNP	G	C	NA	NA
chr5	70958371	70958373	mccb	Intronic	SNP	G	A	NA	NA
chr5	70966721	70966723	mccb	Intronic	SNP	A	C	NA	NA
chr5	70984409	70984411	mccb	Intronic	SNP	G	A	NA	NA
chr5	70988261	70988263	mccb	Intronic	SNP	A	G	NA	NA
chr10	69721815	69721817	mawbp	Intronic	SNP	A	G	NA	NA
chr10	69721869	69721871	mawbp	Intronic	SNP	G	A	NA	NA
chr10	69721933	69721935	mawbp	Synonymous	SNP	A	G	D	D
chr10	69726605	69726607	mawbp	Intronic	SNP	C	T	NA	NA
chr5	178346421	178346423	mGluR6	Intronic	SNP	T	C	NA	NA
chr5	178346552	178346554	mGluR6	Synonymous	SNP	T	C	G	G
chr5	178348542	178348544	mGluR6	Synonymous	SNP	A	G	N	N
chr5	178348587	178348589	mGluR6	Synonymous	SNP	A	C	T	T
chr5	178348668	178348670	mGluR6	Synonymous	SNP	G	A	Y	Y
chr12	122993798	122993800	lmkain	Intronic	SNP	G	A	NA	NA
chr12	122993826	122993828	lmkain	Intronic	SNP	G	A	NA	NA
chr12	122993122	122993128	lmkain	DIV	INDEL	CACAC	CAC	NA	NA
chr5	96257977	96257979	l-rap	Intronic	SNP	G	C	NA	NA
chr5	96258041	96258043	l-rap	Intronic	SNP	C	T	NA	NA
chr5	96263168	96263170	l-rap	Intronic	SNP	C	T	NA	NA
chr5	96265049	96265051	l-rap	Intronic	SNP	A	G	NA	NA

Sample	Genotype	Genotype score	coverage
Sample1_aff.NA	GC	38	56
Sample2_norm.NA	GC	38	56

Number of Variant Positions: 75559

Samples:
Sample1_aff.NA
Sample2_norm.NA

Genotype-filter

○ Exclude variants which have heterozygous genotypes in one or more affected family members

● Exclude variants which have reference homozygous genotype in one or more affected family members

○ Exclude variants which have alternative homozygous genotype in one or more affected family members

○ Exclude variants which have the homozygous genotypes shared by both affected and unaffected family members

○ Exclude variants which have the heterozygous genotypes shared by both affected and unaffected family members

○ Exclude variants which have NO shared alleles in all affected family members

IBD estimation

□ Highlight regions with at least one identical alleles at each variant (IBS)

□ ibd-annot

□ Highlight homozygousity region in cases

Results

```
java -Xms256m -Xmx1300m -jar kggseq.jar  \
--buildver hg19  \
--resource ./resources  \
--out ./kggseq  \
--excel  \
--vcf-file /Users/casshyr/Desktop/kggseq/examples/rare.disease.hg19.vcf  \
--ped-file /Users/casshyr/Desktop/kggseq/examples/rare.disease.ped.txt  \
```

```
[casshyr@admins-MacBook-Pro ~/Desktop/kggseq]$ java -Xms256m -Xmx1300m -jar kggseq.jar  \
> --buildver hg18  \
> --resource ./resources  \
> --out ./kggseq  \
> --excel  \
> --vcf-file /Users/casshyr/Desktop/kggseq/examples/rare.disease.hg19.vcf  \
> --ped-file /Users/casshyr/Desktop/kggseq/examples/param.rare.disease.hg19.txt  \
> --seq-qual 50  \
> --seq-mq 30  \
> --seq-sb -10  \
> --gty-qual 10  \
> --gty-dp 4  \
> --gty-af-ref 0.05  \
> --gty-af-alt 0.25  \
> --min-obsa 2  \
> --min-obsu 2  \
> --genotype-filter 3  \
> --db-gene refgene   \
> --gene-feature-in 0,1,2,3,4,5  \
> --neargene 1000  \
> --splicing 2  \
> --db-score dbnsfp  \
> --filter-nondisease-variant
@------------------------------------------------------------@
|        KGGSeq         |      v0.3      |   8/Aug/2012      |
|------------------------------------------------------------|
|  (C) 2011 Miaoxin Li,  limx54@yahoo.com                    |
|------------------------------------------------------------|
|  For documentation, citation & bug-report instructions:    |
|        http://statgenpro.psychiatry.hku.hk/kggseq          |
@------------------------------------------------------------@
To disable library checking, use --no-lib-check
Effective settings :
--resource ./resources
--buildver hg18
--vcf-file /Users/casshyr/Desktop/kggseq/examples/rare.disease.hg19.vcf
--ped-file /Users/casshyr/Desktop/kggseq/examples/param.rare.disease.hg19.txt
--out ./kggseq
--excel
--seq-qual 50.0
--seq-mq 30.0
--seq-sb -10.0
```

Chromosome	StartPosition	ReferenceAlternativeAllele	GeneSymbol	GeneFeatures	GeneFeature	MaxDBAltAF	SLR_test_statistic	SIFT_score	Polyphen2_HDIV_score	Polyphen2_HVAR_score	LRT_score	MutationTaster_score
10	126683190	C/A	CTBP2	1Exons):exon7:c	missense	0	NaN	0	1	0.998000026	0	1
10	126691634	T/A	CTBP2	1Exons):exon5:c	missense	0	NaN	0	0.963999987	0.86500001	0	1

Figure 2-1 Overview of the evaluated software interfaces. A) The main layout of Varsifter. The left panel shows the tabular display of variations from the user-supplied input file, and the pre-built filters available as check-boxes on the right. The right panel shows the interface allowing users to design custom queries via graphical icons and logical connectors for designing filters that are not part of the pre-built check-boxes. B) The left panel shows the layout for the GUI command-line generator for KGGSeq which allows the users to specify the parameters visually before copying the text command-line to the terminal. The right panel shows the screenshot of the terminal output as KGGSeq is being executed. The bottom panel shows an example of the final output from KGGSeq, as displayed in Microsoft Excel.

2.3 Methods

2.3.1 Study setting and participants

The study took place within BC Children's Hospital (Vancouver, Canada) from August 2012 to March 2013. Since the goal of the study was to provide both insights and quantitative results addressing the evaluated systems, we sought to recruit as many test users as we could, given the limited numbers of specialists available. Ten clinical geneticists working at this institution with prior exposure to analyzing genomic data were recruited (Appendix A-1). Recruitment was done via email sent by CS at least one month prior to the testing to each of the twelve potential candidates with a response rate of 83%. None of the participating specialists had prior experience with either of the assessed interfaces. The study was approved by the UBC research ethics board. Each participant provided informed consent prior to the study.

2.3.2 Materials

2.3.2.1 Software

We restricted to testing two software due to availability in time allowed from the participants. To decide on which two software to assess, a systematic literature review of available software for exome analysis of Mendelian disorder was conducted with PubMed in April 2012 using a combination of the relevant keywords. We excluded commercial systems to avoid legal/financial complications. Among the nine software that came up from the review, two software, Varsifter and KGGSeq, were chosen for their contrasting design architecture (e.g. GUI versus command-line) and popularity among the clinical research communities (as ranked by their research citations). Varsifter (version 1.5) was downloaded from the NHGRI website.

KGGSeq (version 0.2) was downloaded from developers' website hosted by the University of Hong Kong. Figure 2-1 provides an overview of the interfaces for the two tools.

2.3.2.2 Simulated data

Two sets of simulated data were constructed to represent two clinical scenarios, covering tasks that commonly occur during exome analysis. Each clinical scenario presents a simulated patient suffering from a particular Mendelian disease. The patient's clinical history was constructed based on the typical traits reported in the research literature. Exome results were provided as processed sequence variants as tabulated form and VCF. A bed file consisting of regions of homozygosity (ROH) was constructed to resemble typical ROH data from a 1st-degree consanguineous family. In both scenarios, a disruptive mutation was embedded in the exome to represent the intended causal variant. The mutation was introduced in such a way that it would emerge as a top candidate after prioritizing through a list of specific instructed filters created by CS and approved by WWW (Appendix A-2).

2.3.2.3 Non-simulated data

A list of mitochondrial genes provided to the users was downloaded from the Mitocarta website[96] on 30 June 2012.

2.3.2.4 Interviews and survey instruments

Pre-evaluation semi-structured interviews solicited self-rated computational expertise and perspectives about ongoing challenges faced with sequencing analysis and sentiments towards next-generation sequencing data. The post-evaluation semi-structured interviews addressed specific issues that came up during the evaluation (Appendix A-3). The quantitative questions used in the study are a validated survey called the Software Usability Measurement Inventory (SUMI) version 4.0[97-99].

2.3.3 Data collection equipment

The evaluations were conducted on a 15-inch Macintosh MacBook Pro laptop (with Mac OSX Version 10.6.8, 2.16 GHz Intel Core Duo and 2GB DDR2 SDRAM) with the software pre-installed. All computer screens and the surrounding audio were recorded using QuickTime software Version 10.0.

2.3.4 Experimental procedure

A one-on-one interview was conducted prior to the simulation session. To avoid order bias, for each scenario, we randomly assigned half of the users to utilize KGGSeq before moving on to Varsifter, and the other half to use Varsifter before KGGSeq. Clinicians were instructed to work through the first scenario with the two software packages before proceeding to the second scenario.

At the beginning of each session, an initial 45 minutes were spent familiarizing the subject with the software and the data inputs. Appendix A-3 describes in details the breakdown of this 45-minute period. The 45-minutes introduction to the software packages was deemed sufficient to allow subjects to gain basic background required for interacting with the software (particularly as these domain experts had prior hands-on experience using similar analysis software and practical experience interpreting the results of exome sequencing data).

Following the introduction, the subjects were asked to interact with the simulated cases under a “think-aloud” protocol (Appendix A-3). The duration of these evaluation sessions ranged from 120-150 minutes. Scenarios finished only when either the clinicians found the embedded causal mutation, or if they voiced that the task is too difficult to proceed and they wished to stop. The SUMI survey was given after the simulation. A second one-on-one interview followed the SUMI survey, concluding the evaluation session.

2.3.5 Data annotation and analysis

The audio recordings were manually transcribed into transcripts as Microsoft Word files. Coding categories were assigned to usability issues identified in the transcripts, and further time-stamped as observed from the screen recordings. The usability coding categories and higher-level descriptive themes were developed by CS and AK prior to the analysis of the data, largely drawn from[100-102].

The raw SUMI questionnaire data for each individual was submitted to the Human Factors Research Group, which generated the numerical summaries and statistical evaluations using their SUMICO software. The questionnaires were statistically quantified into software “efficiency”, “affect”, “helpfulness”, “control”, “learnability”, and “global usability”[97, 99]. SUMICO calculates the probability from the chi-squared distribution that the subjects’ responses from the study differ from the expected values based upon the SUMI database (see [97, 99] and Appendix A-3).

Non-parametric Mann-Whitney U test was used to calculate statistical significance for the observed quantifiable differences between the two software packages (for specific details see the results section).

2.4 Results

In this section, the paper describes the quantitative and qualitative results obtained from the study that are specific to the evaluated two tools. Section 2.5 discusses the key findings that reveal broader themes critical for next generation variant interpretation domain.

2.4.1 Overall performance

Table 2-1 shows the number of clinicians able to identify the correct causal mutation in each scenario. In both scenarios, more clinicians were able to identify the causal mutation with Varsifter as compared to KGGSeq. Fewer clinicians were able to identify the causal mutation in the more complex second scenario as compared to the first scenario for both software packages.

		Varsifter	KGGSeq
Clinical scenario #1	Successful completion?	10/10	8/10
Clinical scenario #2	Successful completion?	6/10	5/10

Table 2-1 The proportion of clinicians (n=10) who were able to successfully identify the causal mutation from scenario 1 and scenario 2.

In both scenarios (Appendix A-3.1), the time was notably shorter with Varsifter ($p< 0.05$; 1-tailed Mann-Whitney U-value = 9 for scenario 1 and U-value=1 for scenario 2). For the clinicians who were able to achieve successful completion on both software packages, all performed the work faster with Varsifter (8/8 for scenario 1, 5/5 for scenario 2, see Appendix A-4.1).

Appendix A-3.2 shows how the two tools were perceived differently by the clinical users based on SUMI. Below we highlight the attributes that, according to SUMICO, deviated significantly from the average score of 50 (SUMICO did not provide exact p-values). Varsifter scored the lowest on “efficiency[3]” (score=40) but highest on “affect[4]” (score=58), revealing that despite finding the software difficult to work with, the users nonetheless finished the evaluations with a positive impression. KGGSeq scored the lowest on “helpfulness” (score=43) and “control” (score=40), which refer to the degree to which software is self-explanatory and the extent to which users feel in control of the software respectively.

2.4.2 Descriptive findings from think-aloud content

Table 2-2 show the breakdown of the encountered usability problems captured from think-aloud sessions into usability themes and the frequency of occurrences. Descriptive findings are categorized under twelve usability categories. These codes are further grouped into five major usability themes: (1) Visualization, (2) Information, (3) System response, (4) Functionalities, (5) Overall usability. Example comments that fall into these five categories are shown in Appendix A-4.2. A more in-depth description of the specific usability problems found for each software can be found in Appendix A-6.3.

[3] Efficiency measures the degree to which users feel the software assists them in their work

[4] Affect measures the user’s general emotional reaction to the software

		Varsifter		KGGSeq	
		Positive	**Negative**	**Positive**	**Negative**
VISUALIZATION	**Navigation**	1	21	2	3
	Layout	0	8	1	11
	Operation consistency	0	13	0	1
	Graphics	2	0	0	0
INFORMATION	**Resolution**	6	1	0	17
	Label	0	7	0	19
	System messages	1	3	3	1
SYSTEM RESPONSE	**Response time**	0	9	1	1
	System status	0	2	1	3
FUNCTIONALITIES	**Compatibility**	2	7	1	2
	Scope of functionalities	0	19	1	20
OVERALL USABILITY	**Overall usage**	1	2	0	1
	Total	12	92	10	79

Table 2-2 A breakdown of detected usability issues by categories. We assigned the detected usability problems into 5 main themes that are further subdivided into 12 categories. Example comments can be found in Appendix A-4.2.

2.4.2.1 Visualization

For Varsifter, every clinician complained that text or functions were hidden from view due to scrollbars and/or hidden panels. In one instance, 8/10 participants sought a button that they had observed in the tutorial video, but did not know that it was necessary to use the scrollbar to access the remainder of the options (e.g. “I remember seeing a button to click to exit this window, but I can’t find it”). The feedback about KGGSeq also indicated user concerns with incomplete display in the command-line generator GUI (e.g. “Some of the text descriptions and buttons are not visible on the screen")(Appendix A-4). One clinician commented "the scrolling means I have to remember what is hidden behind this panel and that is a pain". The difficulty with accessing the needed functionalities indicate difficulty in function execution. In some cases, Varsifter's dropdown menus in which similar functions are grouped together were perceived as offering better organization. However, in some cases users had differing views about the logic of the groupings (e.g. "The software should move this function out of 'View'. It is organized very counter-intuitively"), and resulted in navigational difficulties.

2.4.2.2 Information and system response

Clinicians (6/10) indicated that Varsifter responded too slowly to inputs (e.g. "Is the software running? I am clicking this button multiple times but nothing happens"). The actual start up time for the program was a relatively short 7-25 seconds. However, at times the clinicians indicated that they did not know if the program was running, and ended up repeatedly clicking on buttons, which further slowed the program or introduced unwanted errors. As the analysis of large-scale exome data may require more time than is ideal for busy clinicians[42, 43], software should provide an approximate processing time whenever possible and a clear indication of system status. For KGGSeq, there were multiple complaints regarding the use of bioinformatics jargons that the clinicians were not able to comprehend (e.g. "Do I want variants with a high [MutationTaster] score or low score?"). 17/79 comments further criticized the way the information was presented in the output, which the clinicians felt were too overwhelming (e.g. "I don't know what this column means, and I can't find the actual information that I want because there are too many things to look for here").

2.4.2.3 Functionalities

For both software packages, majority of functionality problems (16/19 for Varsifter, 13/20 for KGGSeq) were related to the clinician's inability to execute the software's implemented function. For instance, 4 out of 10 clinicians were unable to filter variations for a particular Mendelian inheritance model in Varsifter (e.g. "The tutorial video showed me how to do it but I don't know how to work it myself", "I can see the button here, but I can't press it. I don't know why it isn't letting me do it, and there is no instruction for how to get it working"). For KGGSeq, the clinicians were unfamiliar with the terminal-style interface (Appendix A-4). The command-line generator interface received some initial praise in the early stages of the tests (n=3; e.g. "I like how all the basic inheritances are already setup so I only have to click it"), but negative comments were subsequently expressed as the subjects realized that only a portion of the functions could be access through the graphical interface (n=20; e.g. "I can't upload this gene file using this software").

2.4.2.4 Frequency of errors

For Varsifter, the clinicians praised the software ability to revert filtered output to a previous unfiltered state (e.g. "I like how I can just uncheck this filter and get my previous result. It allows me to explore different filtering thresholds"). For KGGSeq there was no apparent capacity to revert from one state to a previous state, making it necessary for the user wishing to change a specific threshold for a filtering parameter to restart the entire analysis (e.g. select for variants with population frequency <1% versus <3%). The benefits for this are most apparent when comparing the frequency of errors made by the users when using the system. More mistakes were resolved by Varsifter (12/15) compared to KGGSeq (11/26) because clinicians were able to view the results at each filter step (Appendix A-4.3).

2.5 Discussion

In this chapter, we performed an assessment of clinical genetics exome interpretation software, using a cognitive analysis approach to usability evaluation. The observations specific to clinical genetics and exome/genome analysis are the most important for consideration for future software development. Therefore, we will focus our discussion on the domain-specific lessons learned from the study, providing specific recommendations that could inform the design of future interfaces for clinical exome analysis software and informing clinicians on their choice for software selection. Ultimately the users feedback leads to a clear and concise inventory of features and characteristics desired of clinical exome interpretation software (Table 2-3).

Clinical exome interpretation software user desiderata

- Rich filter functionalities (i.e. variant calls with simple column-based filtering are insufficient)
- Software design structured with focus on genetic models (e.g. Mendelian inheritance)
- User defined workflow management with stepwise reports
- Fast response time with estimates given for wait steps
- Team support to allow multiple clinicians to annotate/review data
- Interoperability with widely used online resources/databases and data formats
- Frequent updating to support emerging tools, data standards and input types

Table 2-3 Implications for new software design

2.5.1 Recurring domain-specific usability challenges need to be addressed

We find that a major impediment for adoption of exome analysis software is a lack of clear presentation (organization), description and help messages for the provided functionalities. Non-computational healthcare professionals will not choose to adopt a software package unless the functionalities are easily executable and can fit into a clinician's workflow. Table 2-4 contains examples of problems frequently encountered in our evaluation and the recommendations from the clinicians to resolve them.

Usability issue	Feedback and recommendation
Navigation difficulties	• Example problem: with so many filtering options available in the system, the user has problem finding the desired option
	• Suggested solution: organize the functionalities into themes as according to the type of analyses they belong to, and visualized as dropdown panels. The groupings should be intuitive to the user. Examples of groupings compiled based on users' feedbacks are given in Table 2-5.
Execution difficulties	• Example problem: user sees a GUI option to filter the variants by compound heterozygous model, but the function is disabled and it is not intuitive how to execute that function
	• Suggested solution: software should always provide an explanation as to why a function does not execute and guidance on how to fix it. This message should be easily accessed (e.g. display on mouse-over), along with links to further instructions.
System logs	• Example problem: when uploading a genome data or filtering multiple exomes, the user is uncertain if the system is in the middle of processing or merely stuck.
	• Suggested solution: system should indicate the current program status and expected run time whenever possible.
Workflow integration	• Example problem: for clinicians working with many families sharing similar inheritance patterns, certain filtering approaches should be automated.
	• Suggested solution: software is best organized into layers, with the ability to develop and save workflows for batch analysis. The layer-structure allows clinicians to go back to previous output and compare the results at each stage of filtering.
Interoperability and data standards	• Example problem: system is unable to take in multiple VCFs (where each VCF contains the data for a distinct subject). Rather, the system forces the user to combine the input files in advance in order to conform to system rigidity.
	• Suggested solution: system needs to be compatible with standard data formats, and be able to integrate with external data resources (Appendix A-5.3). System must also anticipate minor/major updates to the data standards and external resources.

Table 2-4 The top recurring usability problems observed, the features that are desired, and recommendations to developers.

2.5.2 Design software structure with emphasis on genetic models and frequently encountered analytical themes

A key observation from the study is the importance of supporting diverse workflows for the range of potential genetic hypotheses. Specifically, the system should be structured around the commonly used analysis models, such as Mendelian recessive inheritance. Clinicians value such structured approaches, as they are expected to follow standardized protocols in their practice. The ability to develop and save common workflows is key for clinical groups working on many cases over time. There are unique cases, which require unusual analysis approaches. Therefore while the software should be structured around specific standard analysis models, it needs to remain flexible. We compiled a list of frequently employed tasks of clinical exome analysis, organized by the themes of analyses, that the software should be structured to address (Table 2-5).

Themes	**Category**	**Example tasks**
Type of analyses	Population studies	• Look for recurring mutations/genes within a cohort versus control samples • Look for mutations shared by 80% of the affected individuals
	Mendelian inheritances	• Filter the mutations by different classical Mendelian inheritance models • Provide flexibility to work with non-standard family structure (e.g. only exomes for mother and proband, or only exomes for multiple affected individuals)
Area of interest	Genomic coordinates	• Retrieve mutations that fall within regions of homozygosity • Exclude mutations that fall outside of known regulatory regions
	Gene lists	• Retrieve mutations that fall within known mitochondrial genes • Filter for mutations in genes that are abundantly expressed within a specific human tissue
Mutation-level	Conservation	• Sort mutations by their evolutionary conservation score
	Mutation type	• Retrieve all the nonsense, missense and splice-site mutations
	Predicted impact	• Retrieve and rank mutations predicted to be damaging based upon scores from software such as SIFT or PolyPhenV2
	Frequency	• Sort the mutations by their annotated frequencies from dbSNP, and filter out mutations present > 1% frequency.
	Disease databases	• Retrieve mutations that have been reported as disease-causing in HGMD or ClinVar
Technical-level	Coverage	• Retrieve a list of genes that have less than 2 reads covering any exonic regions • Obtain summary statistics on the depth of coverage present in the input dataset
	Collaboration	• Add and share personal annotations to specific variations (e.g. PubMed literature, free text comments)
	Quality thresholds	• Retrieve mutations that have a variant quality score of 30 or greater • Exclude mutations that have less than 2 reads harboring the mutations
	Workflows	• Create a custom workflow to process multiple exome datasets, or to produce incidental findings (e.g. as proposed by American College of Medical Genetics)

Table 2-5 A list of frequently-employed tasks in clinical exome analysis, compiled based on review of PubMed literature and feedbacks captured in the simulations.

2.5.3 Present variants in a tabular format but retain flexibility in layout

Varsifter has a greater emphasis on the GUI while KGGseq is primarily intended for use via the command-line (albeit with an available interface). Our results confirm that clinicians benefited from and appreciated the fuller GUI, both for visualizing the data and performing analyses. Displaying the variations visually in a tabulated form with sortable columns allows the clinician users to browse and prioritize the data, a functionality that KGGSeq lacks. Another advantage of tabular structure is it is highly similar to Excel representation, a program that is frequently used by clinicians. A few clinicians from our study note that they would like the order of the columns to be adjustable so they can customize the type and order of information presented.

2.5.4 Allow customizable filtering pipelines and prioritizing strategies

Users expressed desire for a system to allow them to bifurcate in the workflow, exploring multiple approaches to processing the data at certain steps. While some workflow software platforms such as Taverna, SciTegic's Pipeline Pilot[103, 104] and Galaxy[31] provide this functionality for general informatics work, most specialized exome processing tools have not incorporated the approach in a robust manner. A core component of exome and genome analysis is filtering and variant prioritization. The software should provide an intermediate output to evaluate the effectiveness of a particular filtering step, and the ability to return to the previous result or continue to the next depending on the context of that intermediate output. The iterative design feature not only reduced the amount of slips, but importantly allowed the users to investigate the data under different scenarios (Table 2-5).

2.5.5 Support collaborations and team-based communications

Most exome-sequenced families are examined by multiple clinicians. A consensus opinion about a causal gene candidate may arise from a series of email exchanges, face-to-face meetings and sharing of references such as hyperlinks to scientific abstracts. From this study we learned that most exome analysis software, both free and commercial, do not provide suitable functionalities for facilitating multiple users to collaborate on the same data. Users expressed that an ideal system would allow users to attach notes, links to scholarly articles, as well as comments on individual genes or genetic variations, and that such information be available to multiple users in the same clinical setting. Software that empowers collaborative analysis would be well received.

2.5.6 Maintain high interoperability to data standards

The subjects identified input compatibility as a key factor for exome variant interpretation tools. Many of the filters and prioritization strategies used in exome analysis are built from standard outputs of academic and commercialized bioinformatics pipelines. Being interoperable with the data standards and currency with updates is important for widespread adoption, especially for non-computational clinicians who should not be expected to convert data formats.

2.5.7 Maintain currency with online databases and critical resources

The prioritization of genetic variants can be highly dependent upon accessing external resources such as biological annotations attached to a particular genomic coordinate or to a gene. At present many clinicians manually evaluate each variant by querying online resources (e.g. PubMed, OMIM, HGMD[105], CLINVAR[106]), which was reported to be amongst the most time-consuming steps in the interpretation process. The capacity of software to automate data mining of these resources may accelerate analysis and increase success rates. Appendix A-5.3

shows a list of common data formats for next-generation sequencing data, and the databases and external resources that clinicians indicated a desire to incorporate.

2.6 Conclusions

Software to support exome sequencing is a cost-effective technology increasingly incorporated in clinical genetics[107]. Without a reliable and practical clinical system, complex exome data cannot be processed by most clinicians. In this study we highlighted recurring usability problems, and reported user recommendations and requests for key functionality. Our findings point to the need for changes and/or updates to current exome interfaces. The results should further help clinical users who are choosing what analysis software would suit their needs.

The user desiderata represent a key feature set for future systems to deliver. Our evaluations highlight the many types of filters and prioritization strategies that are needed by the clinicians, and the limitations of simple column-based filtering layouts. In addition, the software can accelerate analysis by reporting findings based on classical genetic models of inheritance where appropriate. The software should retain the ability for the users to define their own custom workflow, providing step-wise reports so the impact of each step can be assessed. As the community moves to whole genome data, the resulting size and complexity will exacerbate concerns about the speed of processing - thus it is critical for the software to provide time estimates whenever a job cannot be completed rapidly (i.e. >10 seconds). Since each case is rarely evaluated by only one specialist, the ability for clinical exome interpretation software to support team collaborations for collective annotation and review of data is desired. The users indicated a need for the software to be compatible with multiple data formats used in the field, as well as providing connectivity to popular online databases and tools.

2.6.1 Limitations of the study

All of the subjects in the study worked within the same academic health research hospital. While this likely introduces bias, our subjects are clinical geneticists with prior experience with exome analysis that by nature are not to be found in a general healthcare facility. The focus on a single healthcare center offered advantages regarding the number of experts we were able to gather, and the time they were able to spend on the study. Most clinical exome analyses are currently performed in similar academic health centers, and therefore we anticipate that the results will have broad relevance to the field. Nonetheless, one future direction from this work would be to perform similar evaluations with clinicians from multiple centers.

Each clinical geneticist was allotted 45 minutes to become acquainted with the software, which is a recognized constraint. However, all of the subjects had performed similar tasks as given in the simulations, and had worked with other exome interpretation software. Furthermore, based upon the nature of the software, and the type of analysis that we asked the clinicians to perform, the 45 minute training period was sufficient for subjects to gain a basic understand the basic functionalities for the purposes of conducting usability testing.

The study was limited to two specific open-source software packages. One could argue for the inclusion of other tools, including commercial packages. We believe the two tested tools present a suitable range of features in order to gain general feedback about software in this specific field. Given the rapidly moving developments in the field, there will always be more new software emerging. We did query the subjects about their experience with other packages throughout the evaluations, such that the user perspectives presented in this study are not restricted to the evaluated tools but also informed by exposure to various commercial and open-source platforms.

As access to low-cost DNA sequencing grows, it is anticipated that whole genome sequence analysis will become a standard diagnostic tool for many fields[108, 109]. The complexity of genome data and annotations will continue to increase as the technologies mature, making it imperative to develop better interfaces that streamline analyses and improve quality.

Chapter 3: Dynamic software design: insights from bioinformaticians, clinical geneticists and genetic counselors

3.1 Synopsis

Objectives With almost no previous research specifically assessing interface designs and functionalities of WES and WGS software tools, we set out to ascertain perspectives from healthcare professionals in distinct domains on optimal clinical genomics user interfaces.

Methods A series of semi-scripted focus groups, structured around professional challenges encountered in clinical WES and WGS, were conducted with bioinformaticians (n=8), clinical geneticists (n=9), genetic counselors (n=5), and general physicians (n=4).

Results Contrary to popular existing system designs, bioinformaticians preferred command line over graphical user interfaces for better software compatibility and customization flexibility. Clinical geneticists and genetic counselors desired an overarching interactive graphical layout to prioritize candidate variants – a 'tiered' system where only functionalities relevant to the user domain are made accessible. They favored a system capable of retrieving consistent representations of external genetic information from third-party sources. To streamline collaboration and patient exchanges, we identified user requirements towards an automated reporting system capable of summarizing key evidence-based clinical findings among the vast array of technical details.

Conclusions Successful adoption of a clinical WES/WGS system is heavily dependent on its ability to address the diverse necessities and predilections among specialists in distinct healthcare domains. Tailored software interfaces suitable for each group is likely more

appropriate than the current popular “one size fits all” generic framework. Our study provides interfaces for future intervention studies and software engineering opportunities.

3.2 Introduction

At this early stage, clinical access to WES/WGS analysis occurs principally on a research basis in academic health research centers where informatics teams are available to assist with data analysis[110]. The prospect of full genome sequencing, compounded by the continual growth in genetic knowledge base, is overwhelming for the health care professional; computerized for interpreting and acting on this information is essential for clinician support and ultimately patient care[111]. For research-focused WES/WGS analysis, distinct software architectures with different engineering emphasis have been introduced, all ultimately sharing the same goal to assist in the identification of key gene(s)/variant(s). The nature of the analysis process includes 5 steps: (1) read mapping of short DNA sequences onto a reference genome, (2) identification of differences between the sample and reference, (3) quality control of candidate variants (including data visualization methods), (4) annotation of the properties of observed variations, (5) prioritization or filtering variations as candidates for the observed phenotype/disorder (reviewed in [110, 112]). Existing software programs address differing portions of the analysis process, with emphasis tending to fall either on categories 1-2, 3-4 or 5 (example software discussed in Appendix B-1).

Many of the early WES/WGS software packages placed greater emphasis on the computationally oriented users, as clinical use was rare[81, 113]. The previous chapter described how we evaluated the usability of exome analysis software based on think-aloud protocols in a study where participants were presented with simulated clinical cases to analyze[114]. While our results highlighted deficiencies of the software for clinical geneticists, such users rarely work in isolation. An interdisciplinary team comprising of informaticians, clinical/biochemical geneticists, subspecialist pediatricians, laboratory scientists and genetic counselors are often involved. This is exemplified by two programs at the National Institutes of Health (Clinical Sequencing Exploratory Research and Clinical Center Genomics Opportunity), seeking to bring together clinicians, genomic researchers, bioinformaticians, and ethicists to tackle challenges in WES/WGS analysis[115, 116]. Despite the expectation of the groups working together, presumably through a shared computational framework, the diversity of perspectives and preferences regarding software design remains undetermined. As the community moves to adoption of WES/WGS as a standard clinical test, it is unclear if the design of analysis software needs to be tailored to domain-specific users.

As far as we are aware, this work represents the first research looking at cognitive insights between distinct domains of medical professionals that most closely interact with genomic data. We surveyed three major groups of specialists that most closely interact with genomic data at the patient-oriented level: data-intensive informatics specialists (a newly emerging clinical role), clinical geneticists, and genetic counselors. In this report, we specifically addressed three key research questions:

1) Are there major cognitive differences and patterns among different user groups?,

2) What do the optimal designs envisioned by informaticians, clinical geneticists, genetic counselors and general physicians look like?,

3) How do the designs desired by the different user groups compare with existing designs?

Bearing in mind of the broad range of clinical applications of genomic data, we focus our research questions primarily in the context of difficult to diagnose germline rare diseases, or diseases with suspected genetic etiology. Through narrative discussions and digital prototypes, we revealed major patterns that distinguish between classes of specialists. We identified properties perceived by users to play a critical role in determining efficacy and efficiency of an analysis software. The results of the study will inform clinical interface design as WES/WGS move into the mainstream.

3.3 Methods

3.3.1 Setting

All focus groups were conducted in the Child and Family Research Institute at BC Children’s Hospital in Vancouver Canada. Sessions were conducted within a conference room with a round table, chairs, a white board with markers, a video recorder (a mounted Sony HandyCamHDR-SR1 + ECM-HW1R Wireless Microphone) and a digital projector connected to a Macbook Pro laptop.

3.3.2 Recruitment

Participants were recruited at least one month prior to the study via email and direct solicitations by CS (or email sent on behalf of CS by contacts reached out by CS) from across various institutions located within the greater Vancouver region. Twenty-six individuals from four different healthcare professions were recruited (crude response rate estimated to range between 20%-45%; Appendix B-2). Each individual was categorized into one of the four user classes: bioinformatician, clinical geneticist, genetic counselor, and non-specialist physician based on their professional and job title (see Appendix B-2). The first three user groups represent the current healthcare professionals that most closely interact with patient genomic data for clinical decision-making in precision medicine (we define precision medicine as ‘the ability to tailor diagnostic and treatment decisions for individual patients’, see [117]). The last group (general physicians) represents the baseline within clinicians that do not have experience working with genomic data.

3.3.3 Focus groups assignment

Each homogenous focus group consisted of participants from the same professional category, and group sizes ranged from four to five. The rationale for the group size was to balance between having enough participants for interactions while not having too many participants in one setting such that not everybody got to express their opinion in the limited timeframe. There was no overlap between the group assignments such that each individual participated only once in a focus group. The participants for each group were randomly assigned. The focus groups were conducted in two rounds: six first round sessions took place between February and June of 2014, six second round sessions took place between September and October of 2014.

3.3.4 Focus group structure

Participants filled out a demographic survey and consented by signing a project participation form at the beginning of each focus group session. Each focus group lasted between 90 to 120 minutes. The sessions were audio-recorded in their entirety and drawings made by participants on a whiteboard were digitally captured. Key matters that were repeatedly referred to in the focus groups were typed on a laptop by the moderator (CS) and projected on a big screen via a projector. Throughout the session, participants had access to drinks and snacks.

The structure of the focus groups was built around the various processing stages of patient exome data (e.g. generation of alignment and variant calls, data annotation and visualization, variant/gene prioritizations). To guide the flow, many of the questions were structured around a hypothetical scenario involving a patient suffering from an undiagnosed rare metabolic disorder (See Appendix B-12 for discussion of study limitations), but participants were encouraged to think and discuss beyond the scenario. Some parts of the focus groups were scripted to raise issues including examining data quality and screening for technical and/or biological abnormalities, filtering exome variant calls at the genetic level, prioritizing mutations at the gene level, and smoothing out the technical challenges when collaborating across multiple researchers, and sharing the clinical findings with patients (See Appendix B-3 for more details).

3.3.5 Analysis

Focus group transcripts were generated from recordings and notes and coded in Microsoft Word. Content analysis was conducted to describe participants' views and perspectives on WES/WGS data[118, 119]. A set of initial codes was formulated based on the research questions and prior studies[65, 114, 120, 121]. Additional emergent themes and codes were identified from the data using an inductive approach[122-124]. The whiteboard drawings were analyzed from the video footages, and were digitally translated using GUI Design Studio Version 4.6. Themes and sub-themes identified from the coded transcripts were used to highlight key features on the digital prototypes. Findings were summarized through tables, figures and narrative discussion.

3.3.6 Approvals

This study was approved by the University of British Columbia Behavioural Research Ethics Board (H13-02034).

3.4 Results

3.4.1 User groups demonstrate dissimilar focus in the analysis pipeline

The diverse WES/WGS analysis software tends to emphasize specific points in the analysis pipeline, in a package specific manner. We sought to understand whether the focal point of the software packages tends to reflect distinct user community desires. To ascertain preferential starting points in WES/WGS analysis, participants were asked to choose between working with raw unaligned sequenced reads, or to work with variants called with an external informatics pipeline. These two choices represent typical options offered by sequencing centers and commercial companies[125, 126]. The preferences from each participant were immediately reflective of the domain they represented, and it was apparent that the same genomic data were treated differently by each of the three user classes (Figure 3-1).

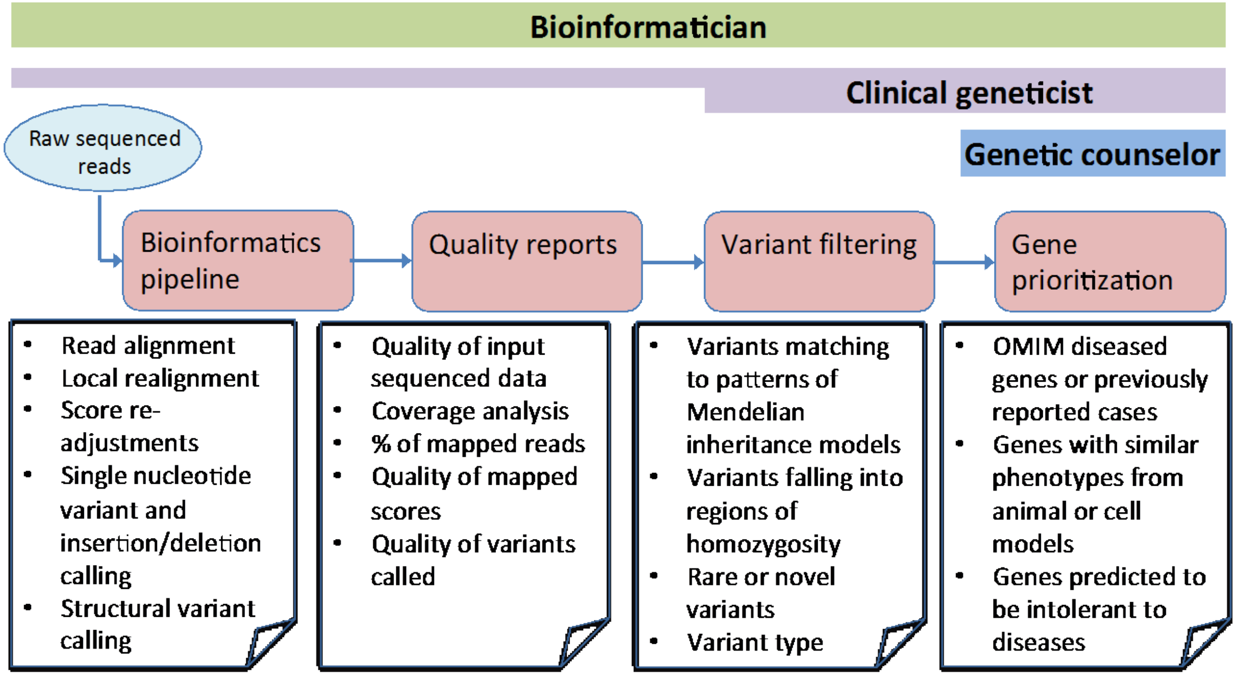

Bioinformatician
Clinical geneticist
Genetic counselor
Raw sequenced reads
Bioinformatics pipeline
Quality reports
Variant filtering
Gene prioritization
• Read alignment
• Local realignment
• Score re-adjustments
• Single nucleotide variant and insertion/deletion calling
• Structural variant calling
• Quality of input sequenced data
• Coverage analysis
• % of mapped reads
• Quality of mapped scores
• Quality of variants called
• Variants matching to patterns of Mendelian inheritance models
• Variants falling into regions of homozygosity
• Rare or novel variants
• Variant type
• OMIM diseased genes or previously reported cases
• Genes with similar phenotypes from animal or cell models
• Genes predicted to be intolerant to diseases

Figure 3-1 Distinct user preferences in genome analysis. Beginning with raw sequenced reads, the exome analysis pipeline can be conceptualized into four distinct compartments: generation of alignments and variant calls, assessment of data quality, filtering of variants based on genetic models, and prioritization of genes based upon

biological functions. The details of the components are annotated largely in the context of genetic diagnosis for rare/complex disorders (refer to Appendix B-12 for discussion on other clinical uses of genomic data). The bars above represent the intensity of user engagement at each step. Bioinformaticians preferred to be involved in every step, with equal attention devoted to all compartments. Clinical geneticist, despite placing heavier emphasis on the final two stages, indicated they would ideally like to be involved in every step too, but they faced difficulties in carrying out the first and second steps (e.g. pipeline execution and quality assessment), which may be attributed to software usability. Genetic counselors (and general physicians, not shown) indicated they would focus on the final output of candidate variants, to which they could apply their domain knowledge to select clinically relevant genes. The text in the lower portion of the figure highlights how the same step in the informatics pipeline (e.g. variant data) can be viewed differently across domain experts.

Bioinformaticians desired to start with raw sequence data, but also indicated that having access to both raw data and externally-provided variant lists would be ideal:

"I prefer to work with raw sequence data because it gives me greater flexibility. If I don't see any interesting candidate from my output, I can re-analyze the data using different thresholds, or try a different genome aligner, or a different variant caller. Having the variant calls is a bonus -- I can go to the variants right away while the pipeline is still processing raw reads. This is especially important when I have multiple whole genomes where the processing time is expected to be long" [Bioinformatician 02]

"Ideally I would like to have both [raw data and variant data]. But having the raw sequence data means I can go back and re-do the analysis as future algorithms improve, or as genome annotations get updated…or if I need to investigate other types of genetic variations like large structural inversions or deletions or duplications" [Bioinformatician 05]

Similarly, clinical geneticists preferred both raw sequence data and variant calls, but with a stronger partiality for working with variant calls over raw reads because they believed working with the already aligned and annotated data gave them a better chance to identify clear causal variants quickly.

"If we are dealing with a recessive disorder, then mosaicism and de novo dominant models are less of a concern. I do not have to worry about twiddling different variant quality scores that is often so important when searching for bona fide heterozygous mutations." [Clinical geneticist 05]

"Starting with only the variant data generally means that the data I am given has already been filtered by some kind of threshold so I am restricted to play within the limit of that threshold. I have yet to find a user-friendly interface that would allow a non-computer savvy clinician such as myself to process an exome data from beginning to end. For now, I am limited to getting only the final sorted list from the bioinformaticians." [Clinical geneticist 03]

Genetic counselors and general physicians expressed no desire for raw sequences, indicating that they did not consider it as part of their professional role (Appendix B-4). There were also differences on the preferred file formats between bioinformaticians versus the geneticists and counselors (Appendix B-5).

3.4.2 Separate interfaces required for data quality assessment

3.4.2.1 Desired statistics

Participants were asked to discuss issues regarding quality examination of WES/WGS dataset(s). Genetic counselors and general physicians stated this entire topic was of no relevance to their line of work.

"I don't think it is up to me to inspect data quality. I don't even know where to begin! That is not what I am trained to do. When I receive the data, I expect it to have already been quality-checked." [Genetic counselor 02]

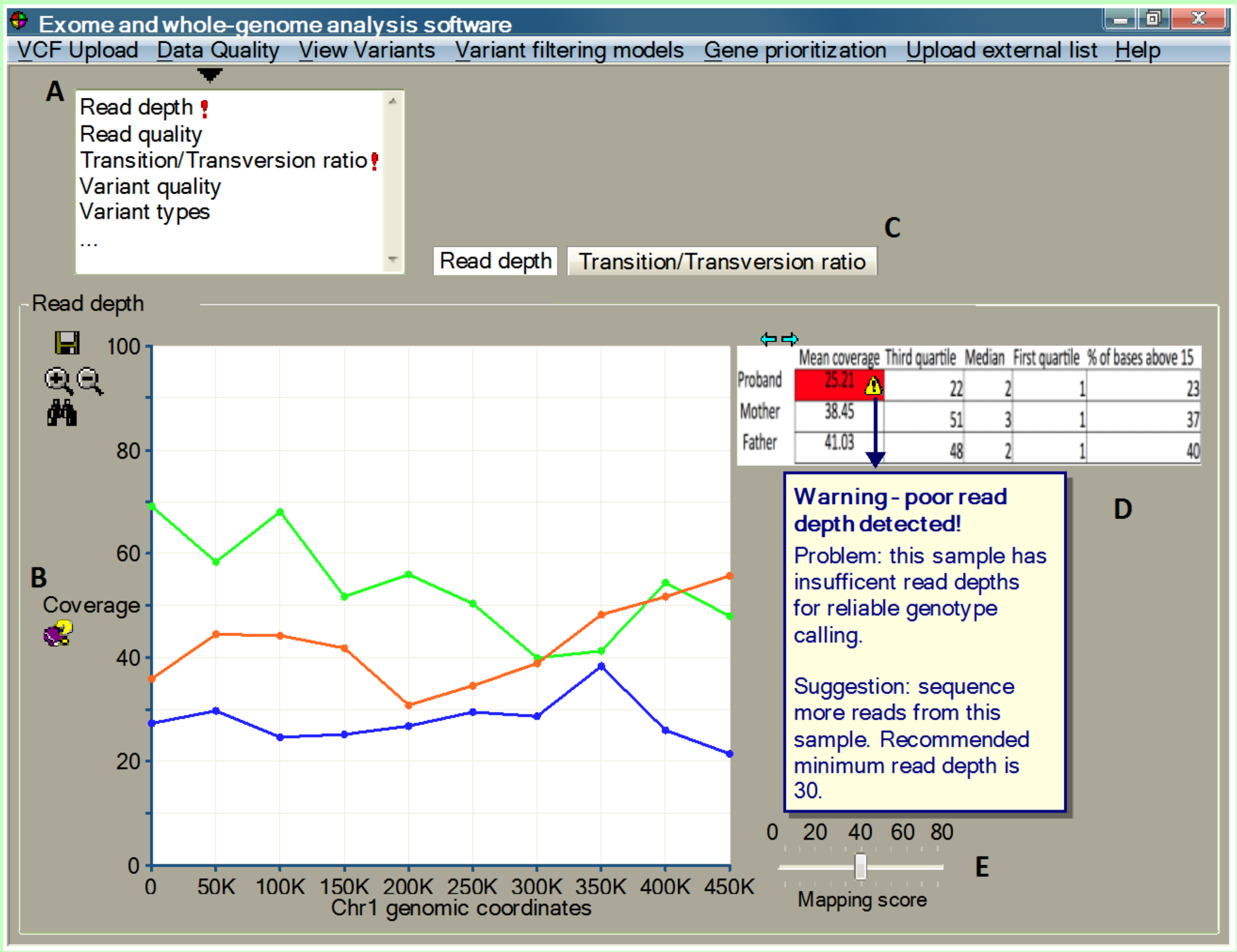
There was a strong overlap between the bioinformaticians and clinical geneticists when commenting on the quality measures desired, and some mentioned quality measurements are not commonly available in current toolkits (Appendix B-6). Both user groups wanted to see a list of genes (or sub-segments of genes) whose exomes were not sufficiently covered, to compare against a list of genes relevant to their study.

"It is important for me to know what genes are included in a capture kit so if there is an insufficient coverage for a set of genes, I can decide if simply re-sequencing the data with the same platform would guarantee more reads at those locations, or if I need to explore alternatives like whole genome sequencing." [Bioinformatician 02]

"I want to know what genes are not sufficiently covered in my exome because currently, all that is given to me is a list of variants. From that list, if I don't see any mutations in those genes, I would be mistaken to think those genes are normal when they could be not." [Clinical geneticist 04]

3.4.2.2 Visual presentation

While the desired metrics and functionalities overlapped highly between the bioinformaticians and clinical geneticists, the preferred methods of presentation differed between them. Figure 3-2 and 3-3 outline the key differences.



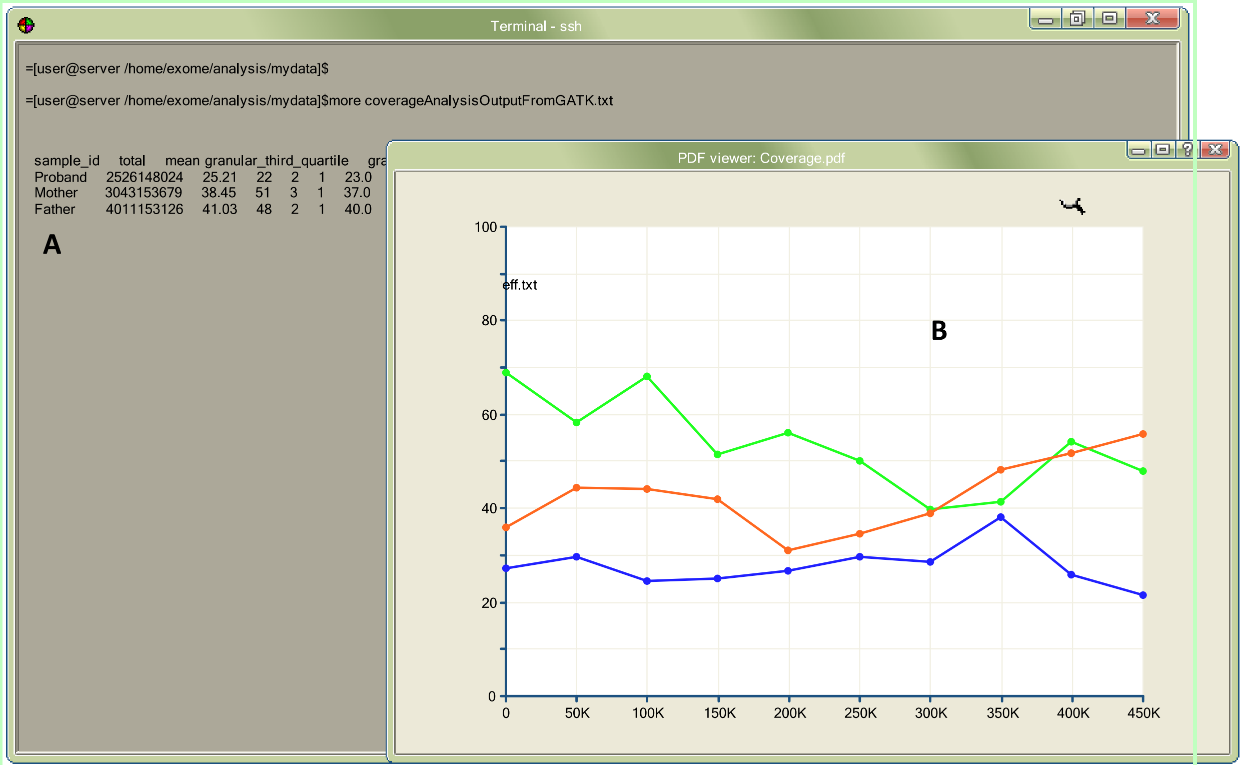
	Mean coverage	Third quartile	Median	First quartile	% of bases above 15
Proband	25.21 ⚠	22	2	1	23
Mother	38.45	51	3	1	37
Father	41.03	48	2	1	40

Warning - poor read depth detected!

Problem: this sample has insufficent read depths for reliable genotype calling.

Suggestion: sequence more reads from this sample. Recommended minimum read depth is 30.

Figure 3-2 A graphical representation of key features desired by clinical geneticists for inspection of data quality. (A) Measurements associated with data quality should be grouped together into a common theme (e.g. a drop-down panel). Quality scores deviating from the norm should be automatically highlighted (e.g. exclamation mark). (B) Computational jargon (e.g. coverage) need to be appropriately explained to a non-computational user. (C) Details on different quality measurements should be displayed separately, but still contained within the same user interface. The example here uses tabs to access different perspective views. (D) Data are best represented both visually (e.g. as a graph) and numerically (e.g. summarized in tabulated form). Simply presenting the quality metrics is not sufficient, software must further describe the nature of the problem, and provide recommendations. (E) The user needs flexibility to explore the distribution of quality scores, and visualize how different thresholds impact the data results. Here, a bar representing the mapping threshold is introduced for the user to dynamically adjust, and the expectation is the interface will update the coverage accordingly.



```
=[user@server /home/exome/analysis/mydata]$

=[user@server /home/exome/analysis/mydata]$more coverageAnalysisOutputFromGATK.txt


 sample_id    total     mean granular_third_quartile    gra
 Proband    2526148024    25.21    22    2    1    23.0
 Mother     3043153679    38.45    51    3    1    37.0
 Father     4011153126    41.03    48    2    1    40.0
```

Figure 3-3 A graphical representation of key features desired by bioinformaticians. A) Terminal interface is the most utilized environment, as it connects with many other command-line software and scripts. Tabulated data quality summaries are displayed directly on the terminal. B) Graphical summaries are also desired, but no intensive graphical user interface app is needed, as bioinformatics users tend to prefer features already available via the terminal display.

3.4.3 Filtrations and prioritizations

3.4.3.1 Variant-level filtration

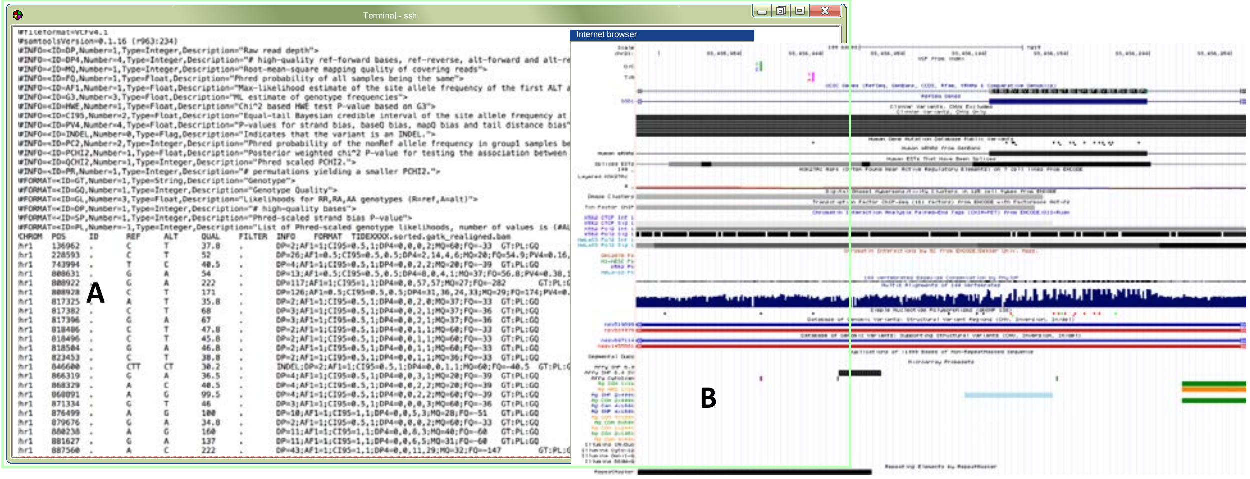
For genetic counselors and general physicians, there were few comments about filtering at the variant level. When the data reached their hands, they expected it to have been filtered based on specified genetic model(s) and allelic frequencies, allowing them to focus on prioritizing candidate genes.

We found a set of filters selected by both bioinformaticians and clinical geneticists, the majority commonly cited in the exome literature (e.g. sort alleles by allelic frequencies, mutation type, and impact prediction[127, 128]). The variants were preferred to be displayed within a table or spreadsheet – a design that is already implemented in many exome analysis systems.

In accordance with how they inspected data quality, bioinformaticians preferred to prioritize variants within the terminal interface (Figure 3-4). Bioinformaticians also displayed the largest diversity in terms of what is desired about each variant (examples discussed in Appendix B-7). The diversity in which bioinformaticians interact with WES/WGS data likely explains why they preferred to work with a command-line rather than to be limited to a graphical tool where the functionalities are by nature more constrained and less flexible to be tailored to context-specific needs.

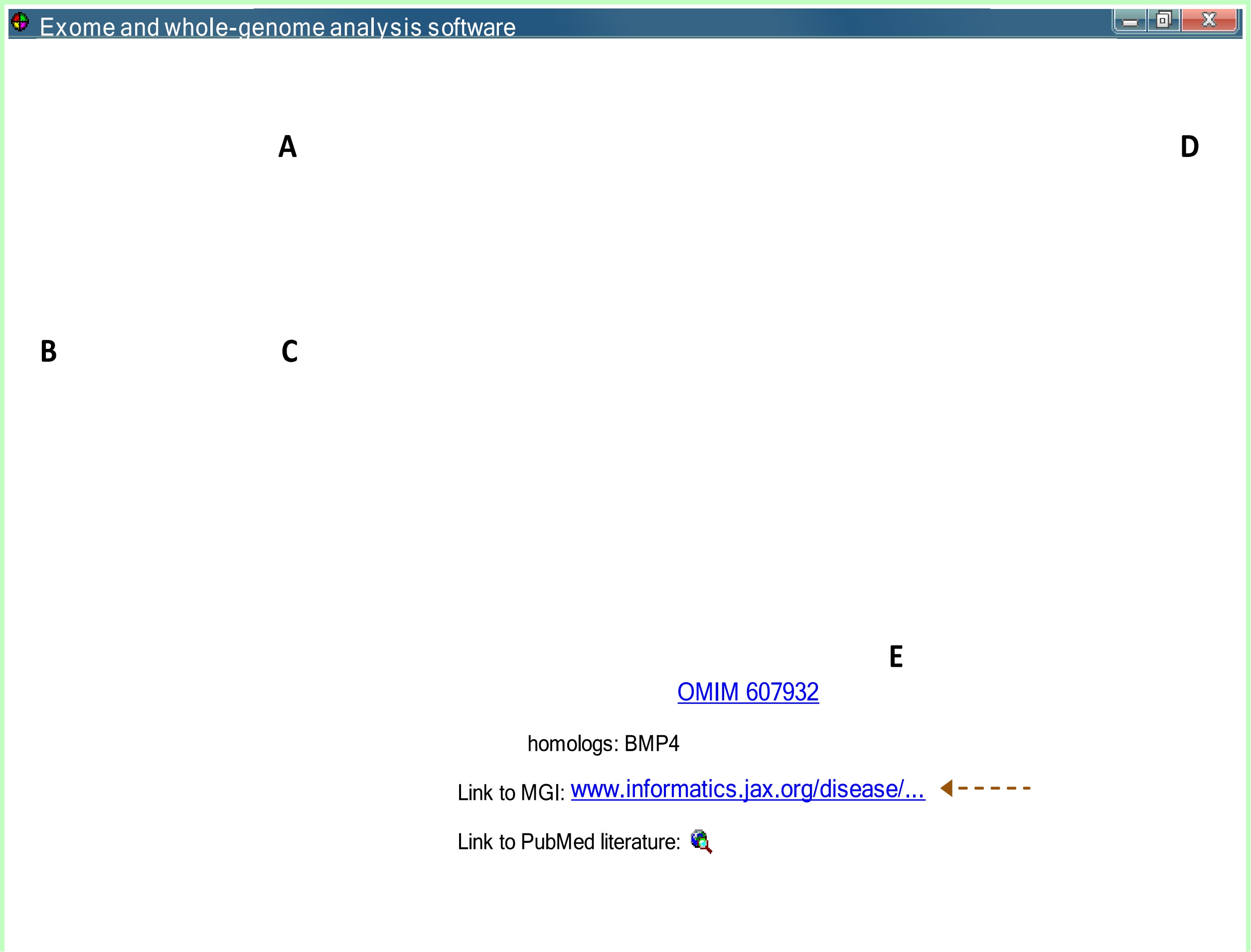
In contrast, clinical geneticists preferred a graphical user interface that is highly dynamic and user-interactive (Figure 3-5). Microsoft Excel spreadsheets were the prevalent choice of clinical geneticists and genetic counselors for viewing variant lists, despite acknowledging it as not being optimized for the purpose.

“The problem with Excel is it starts crashing when I try to feed in more than 65,000 rows of mutation, and that’s just with an exome.” [Clinical geneticist 01]



```
#fileformat=VCFv4.1
#samtoolsVersion=0.1.16 (r963:234)
#INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
#INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-re
#INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
#INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
#INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the site allele frequency of the first ALT a
#INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
#INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
#INFO=<ID=CI95,Number=2,Type=Float,Description="Equal-tail Bayesian credible interval of the site allele frequency at
#INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias"
#INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
#INFO=<ID=PC2,Number=2,Type=Integer,Description="Phred probability of the nonRef allele frequency in group1 samples be
#INFO=<ID=PCHI2,Number=1,Type=Float,Description="Posterior weighted chi^2 P-value for testing the association between
#INFO=<ID=QCHI2,Number=1,Type=Integer,Description="Phred scaled PCHI2.">
#INFO=<ID=PR,Number=1,Type=Integer,Description="# permutations yielding a smaller PCHI2.">
#FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
#FORMAT=<ID=DP,Number=1,Type=Integer,Description="# high-quality bases">
#FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
#FORMAT=<ID=PL,Number=-1,Type=Integer,Description="List of Phred-scaled genotype likelihoods, number of values is (#AL
CHROM  POS     ID   REF   ALT   QUAL  FILTER  INFO    FORMAT  TIDEXXXX.sorted.gatk_realigned.bam
hr1    136962  .    C     T     37.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,0,2;MQ=60;FQ=-33   GT:PL:GQ
hr1    228593  .    C     T     52    .       DP=26;AF1=0.5;CI95=0.5,0.5;DP4=2,14,4,6;MQ=20;FQ=54.9;PV4=0.16,
hr1    743994  .    T     C     40.5  .       DP=4;AF1=1;CI95=0.5,1;DP4=0,0,2,2;MQ=20;FQ=-39   GT:PL:GQ
hr1    808631  .    G     A     54    .       DP=13;AF1=0.5;CI95=0.5,0.5;DP4=8,0,4,1;MQ=37;FQ=56.8;PV4=0.38,1
hr1    808922  .    G     A     222   .       DP=117;AF1=1;CI95=1,1;DP4=0,0,57,57;MQ=27;FQ=-282         GT:PL:G
hr1    808928  .    C     T     171   .       DP=126;AF1=0.5;CI95=0.5,0.5;DP4=31,36,24,33;MQ=29;FQ=174;PV4=0.
hr1    817325  .    A     T     35.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,2,0;MQ=37;FQ=-33   GT:PL:GQ
hr1    817382  .    C     T     68    .       DP=3;AF1=1;CI95=0.5,1;DP4=0,0,2,1;MQ=37;FQ=-36   GT:PL:GQ
hr1    817396  .    G     A     67    .       DP=3;AF1=1;CI95=0.5,1;DP4=0,0,2,1;MQ=37;FQ=-36   GT:PL:GQ
hr1    818486  .    C     T     47.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,1,1;MQ=60;FQ=-33   GT:PL:GQ
hr1    818496  .    C     T     45.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,1,1;MQ=60;FQ=-33   GT:PL:GQ
hr1    818504  .    G     A     46.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,1,1;MQ=60;FQ=-33   GT:PL:GQ
hr1    823453  .    C     T     38.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,1,1;MQ=36;FQ=-33   GT:PL:GQ
hr1    846600  .    CTT   CT    30.2  .       INDEL;DP=2;AF1=1;CI95=0.5,1;DP4=0,0,1,1;MQ=60;FQ=-40.5   GT:PL:G
hr1    866319  .    G     A     36.5  .       DP=4;AF1=1;CI95=0.5,1;DP4=0,0,3,1;MQ=20;FQ=-39   GT:PL:GQ
hr1    868329  .    A     C     40.5  .       DP=4;AF1=1;CI95=0.5,1;DP4=0,0,2,2;MQ=20;FQ=-39   GT:PL:GQ
hr1    868891  .    A     G     99.5  .       DP=4;AF1=1;CI95=0.5,1;DP4=0,0,2,2;MQ=60;FQ=-39   GT:PL:GQ
hr1    871334  .    G     T     46    .       DP=3;AF1=1;CI95=0.5,1;DP4=0,0,0,3;MQ=60;FQ=-36   GT:PL:GQ
hr1    876499  .    A     G     100   .       DP=10;AF1=1;CI95=1,1;DP4=0,0,5,3;MQ=28;FQ=-51    GT:PL:GQ
hr1    879676  .    G     A     34.8  .       DP=2;AF1=1;CI95=0.5,1;DP4=0,0,0,2;MQ=60;FQ=-33   GT:PL:GQ
hr1    880238  .    A     G     160   .       DP=11;AF1=1;CI95=1,1;DP4=0,0,8,3;MQ=40;FQ=-60    GT:PL:GQ
hr1    881627  .    G     A     137   .       DP=11;AF1=1;CI95=1,1;DP4=0,0,6,5;MQ=31;FQ=-60    GT:PL:GQ
hr1    887560  .    A     C     222   .       DP=43;AF1=1;CI95=1,1;DP4=0,0,11,29;MQ=32;FQ=-147          GT:PL:G
```

Figure 3-4 A graphical representation of the key features desired by bioinformaticians when visualizing/filtering variant sets. (A) Analyzing variants within a terminal environment by informaticians allows manipulation of the variant files via custom scripts and/or external command-line programs. (B) Variants are preferred to be visualized within a genome browser (e.g. UCSC Genome Browser[129]) where genomic neighborhood landmarks and any additional relevant biological information (e.g. SNPs, conservation) can be displayed alongside.



OMIM 607932

homologs: BMP4

Link to MGI: www.informatics.jax.org/disease/...

Link to PubMed literature: 🔍

Figure 3-5 A graphical representation of the key features desired by clinical geneticists when performing variant visualization/analysis. Brown dotted arrows point to additional information from specific columns that is available when clicked upon. For instance, clicking the “mutation impact” column would reveal different impact predictions by mainstream prediction software and shows the level of congruency across multiple algorithms. (A) Classical Mendelian models should be built into the system with tabulated summaries automatically available. Outputs from each Mendelian model should be available under separate layouts (e.g. navigated by tabs). (B) Software should provide a quick explanation about the information contained within each column and how to interpret it. (C) The variant table needs to be ranked by evidence (e.g. clinically interesting variants appear at the top of the list). Variants with obvious pathogenic associations need to be automatically highlighted (e.g. flashing red notice). Aside from automated cues, clinical users wanted capabilities to highlight variants that were perceived to be of high interest, to store personal comments for specific variants (e.g. update if a variant is confirmed by Sanger sequencing), or to upload a scientific article related to a particular gene. (D) An integrated pedigree to visualize how the variants are

segregated across a given set of related exomes, and automatically updates the genotypes as users browse across different variants. (E) Hyperlinks that link to external databases should be discouraged. Geneticists complained that the current state of the software relies too much on external references where cross-referencing between different resources on separate interfaces is very distracting. Instead, key clinically relevant information (e.g. the phenotype of a gene knock-out experiment from animal model column) should be computationally compiled and presented within one interface, and only the technical details (e.g. how the experiment was performed) are directed to external sites.

3.4.3.2 Gene-level prioritization

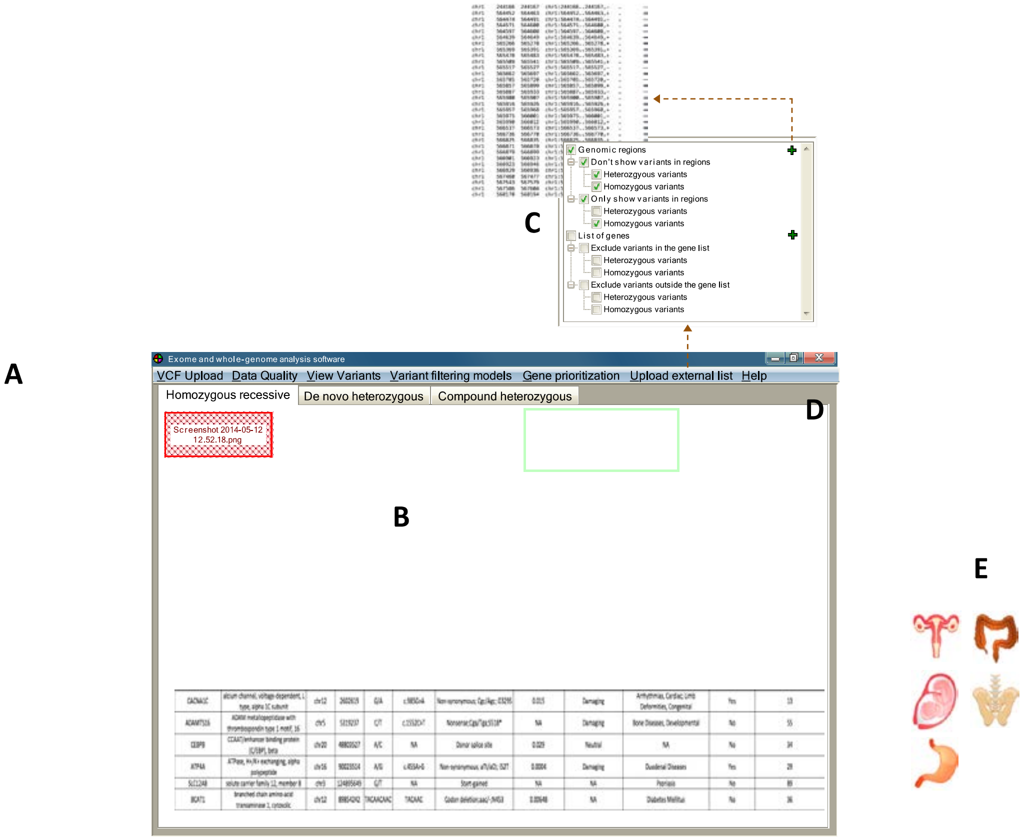
This section discusses the desired prioritization strategies and executions for clinical exomes at the genic level (rather than variant level).

All user groups emphasized a desire for informatics algorithms that conduct automated literature mining or pathway analysis (the overview of such algorithms are introduced in Appendix B-8). The core difference between the user groups is that bioinformaticians wanted such analysis to be integrated with the rest of their command-line based pipeline, while non-computational users wished this functionality to be accessed graphically (Figure 3-6).

Clinicians emphasized that while there are tools that offer online software applications to obtain candidate genes based upon keyword queries (e.g. MeSHOP[42], Genie[130], Ingenuity (http://www.ingenuity.com)), these capabilities are not consistently accessible to integrated WES/WGS analysis software and the output cannot be combined with exome data without additional manipulation. Expanding beyond keywords as input, clinicians further requested graphical search functionalities. One such request is the ability to filter by organ system visually where the user can click on the organ/system of interest in an anatomy diagram (Figure 3-6).

Finally, the clinicians expressed frustration that many gene-ranking software failed to provide the primary literature when returning the results (or it was difficult to retrieve that literature).

"When the program predicts this gene to be related to this particular disease, I want to know how accurate it is. And not just from some kind of confidence score, but I want to see the primary literature. For instance, if the strength of association is based on GWAS literature, then I'm probably not going to treat it seriously." [Clinical geneticist 08]



☑ Genomic regions
- ☑ Don't show variants in regions
  - ☑ Heterozgyous variants
  - ☑ Homozygous variants
- ☑ Only show variants in regions
  - □ Heterozygous variants
  - ☑ Homozygous variants

□ List of genes
- □ Exclude variants in the gene list
  - □ Heterozygous variants
  - □ Homozygous variants
- □ Exclude variants outside the gene list
  - □ Heterozygous variants
  - □ Homozygous variants

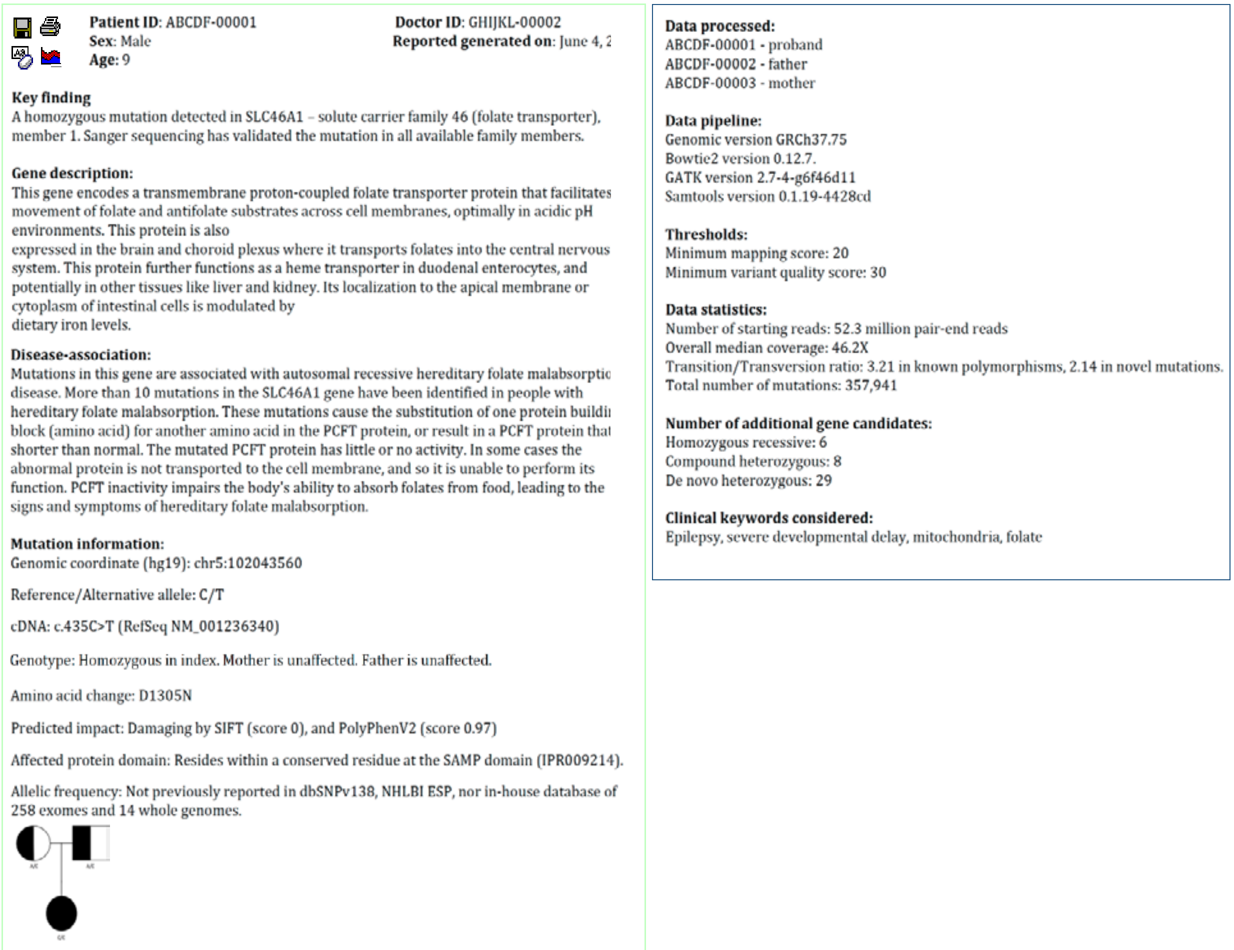
CACNA1C	alcium channel, voltage-dependent, L type, alpha 1C subunit	chr12	2602619	G/A	c.5850G>A	Non-synonymous; Ggc/Agc; G1950	0.015	Damaging	Arrhythmias, Cardiac; Limb Deformities, Congenital	Yes	13
ADAMTS16	ADAM metallopeptidase with thrombospondin type 1 motif, 16	chr5	5319237	C/T	c.1552C>T	Nonsense;Cga/Tga;S518*	NA	Damaging	Bone Diseases, Developmental	No	55
CEBPB	CCAAT/enhancer binding protein (C/EBP), beta	chr20	48800527	A/C	NA	Donor splice site	0.029	Neutral	NA	No	34
ATP4A	ATPase, H+/K+ exchanging, alpha polypeptide	chr16	90023514	A/G	c.455A>G	Non-synonymous; aTt/aCt; I52T	0.0004	Damaging	Duodenal Diseases	Yes	28
SLC12A8	solute carrier family 12, member 8	chr3	124895649	G/T	NA	Start gained	NA	NA	Psoriasis	No	89
BCAT1	branched chain amino acid transaminase 1, cytosolic	chr12	89854242	TACAACAAC	TACAAC	Codon deletion;aac/-;N453	0.00648	NA	Diabetes Mellitus	No	36

Figure 3-6 A graphical representation of the key features desired by clinical geneticists for genetic and genic prioritizations. A) When the user fails to identify any variants of clinical interest, software should provide recommendations on alternative strategies based on what the user has already explored. B) The software should provide easy tracking of the filters currently applied and allow quick adjustments (in this case, via checkboxes to turn a filter on/off). C) Software should allow incorporation of external files containing either genomic coordinates

or list of genes to filter against variant set. D) Software with an embedded dynamic pedigree would allow clinicians to graphically upload multiple exomes (e.g. trio) and assign family memberships via the pedigree. Custom inheritance models could also be setup via the pedigree by specifying expected genotype in a given model. E) Ability to import free-text clinical descriptors, or access terms from a defined ontology (e.g. Human Phenotype Ontology) against which to filter for genes/variants that relate to the specified descriptions. Alternatively, a novel feature emerging from focus groups was the ability to prioritize based on organ systems.

3.4.4 Data sharing with collaborators and patients

A key bottleneck to routine clinical exome analysis was identified to be the preparation of clinical reports for inclusion in medical records and delivery to other physicians. Reports should be concise and automated as much as possible including only clinical information that can be directly extracted from exome data or external databases. Figure 3-7 illustrates an example report separating the clinical genetic findings from technical summaries. Additionally, to streamline exchanges with patients, clinicians wanted the ability to flag genes that have been disclosed by the patients as a set they do not need to be notified about.



Patient ID: ABCDF-00001
Sex: Male
Age: 9

Doctor ID: GHIJKL-00002
Reported generated on: June 4, 2

Key finding
A homozygous mutation detected in SLC46A1 – solute carrier family 46 (folate transporter), member 1. Sanger sequencing has validated the mutation in all available family members.

Gene description:
This gene encodes a transmembrane proton-coupled folate transporter protein that facilitates movement of folate and antifolate substrates across cell membranes, optimally in acidic pH environments. This protein is also
expressed in the brain and choroid plexus where it transports folates into the central nervous system. This protein further functions as a heme transporter in duodenal enterocytes, and potentially in other tissues like liver and kidney. Its localization to the apical membrane or cytoplasm of intestinal cells is modulated by
dietary iron levels.

Disease-association:
Mutations in this gene are associated with autosomal recessive hereditary folate malabsorptic disease. More than 10 mutations in the SLC46A1 gene have been identified in people with hereditary folate malabsorption. These mutations cause the substitution of one protein buildi block (amino acid) for another amino acid in the PCFT protein, or result in a PCFT protein that shorter than normal. The mutated PCFT protein has little or no activity. In some cases the abnormal protein is not transported to the cell membrane, and so it is unable to perform its function. PCFT inactivity impairs the body's ability to absorb folates from food, leading to the signs and symptoms of hereditary folate malabsorption.

Mutation information:
Genomic coordinate (hg19): chr5:102043560

Reference/Alternative allele: C/T

cDNA: c.435C>T (RefSeq NM_001236340)

Genotype: Homozygous in index. Mother is unaffected. Father is unaffected.

Amino acid change: D1305N

Predicted impact: Damaging by SIFT (score 0), and PolyPhenV2 (score 0.97)

Affected protein domain: Resides within a conserved residue at the SAMP domain (IPR009214).

Allelic frequency: Not previously reported in dbSNPv138, NHLBI ESP, nor in-house database of 258 exomes and 14 whole genomes.

Data processed:
ABCDF-00001 - proband
ABCDF-00002 - father
ABCDF-00003 - mother

Data pipeline:
Genomic version GRCh37.75
Bowtie2 version 0.12.7.
GATK version 2.7-4-g6f46d11
Samtools version 0.1.19-4428cd

Thresholds:
Minimum mapping score: 20
Minimum variant quality score: 30

Data statistics:
Number of starting reads: 52.3 million pair-end reads
Overall median coverage: 46.2X
Transition/Transversion ratio: 3.21 in known polymorphisms, 2.14 in novel mutations.
Total number of mutations: 357,941

Number of additional gene candidates:
Homozygous recessive: 6
Compound heterozygous: 8
De novo heterozygous: 29

Clinical keywords considered:
Epilepsy, severe developmental delay, mitochondria, folate

Figure 3-7 An example of automated clinical reporting summarizing the clinical findings from WES/WGS. (A) The system should allow clinicians to save, edit text and insert custom images to the report. The report is designed to be a skeleton for clinicians to build on. (B) Key genetic findings related to the clinical phenotype should be stated right on the front page. These include known clinical relevance about the mutated gene (e.g. what is the biological role of the gene, what phenotype does a person exhibit when the gene is mutated) (C), the type and nature of the mutation (e.g. what is the genomic and transcript coordinate of the mutation, what type of mutation is it, has the mutation been previously reported in clinical literature, what is the allelic frequency, and how is it transmitted across the given family) (D). E) All other information not directly related to the key finding (e.g. the thresholds used by the bioinformatics pipeline that generated the dataset) should be discussed in subsequent pages.

3.5 Discussion

Next-generation WES/WGS sequencing is revolutionizing the study of genetic disorders, with considerable potential for successful application in clinical practice. With large-scale sequencing projects like ClinSeq[131] and Exome Aggregation Consortium (http://exac.broadinstitute.org), and collaborative efforts of sequencing consortiums (e.g. Global Alliance for Genomics and Health, http://genomicsandhealth.org) underway, the global community is in the midst of a multi-year process that will ultimately transition WES/WGS from research labs to clinical labs[132-134]. Despite the continuous flow of new software to assist in the translation process, WES/WGS analysis software for clinical genetic diagnosis is not yet in widespread use[135, 136]. As bioinformaticians have been key processors of WES/WGS sequences in the research setting, they are starting to migrate into emerging clinical laboratory roles. The nature of an interdisciplinary healthcare team necessitates that the software systems and interfaces accommodate the greater diversity of participants to ensure the usability of health information and to provide the requisite utility to diverse clinical users[137, 138].

This report initiates the comparative study of cognitive patterns between healthcare professionals that closely interact with genomic data from multiple domains. Excluding the general physicians included in this study as a control group, the specialist groups represent the three classes of healthcare professionals that currently most closely interact with patient genomic data at the clinical level. While previous focus groups have studied preferences within a general population for results delivery from WES/WGS[139, 140], in this study, we interviewed bioinformaticians, clinical geneticists, genetic counselors, and general physicians to study how domain knowledge influences the cognitive patterns for the analysis of WES/WGS data, and the consequent meaning for software design.

Through a series of scenario-driven focus groups, we found that despite a common goal, the discovery of a causal candidate variant/gene, the user groups exhibit clear differences and divergent patterns among user behaviors. Table 3-1 summarizes and distinguishes the software requirements from each user group.

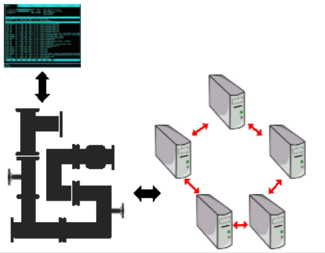
	Properties for analytical interface	Properties for reporting interface	Multidisciplinary collaboration
Bioinformaticians	□Prefer terminal-based interfaces due to superior flexibility for customizing analyses, and compatibility with distributed computing for data processing 	□Desire a digital synchronous collaborative environment for multidisciplinary team interactions	□Suggest new software capacity to foster collaborations with geneticists and counsellors, including secure method of sharing genomic information and personal annotations. □Suggest a focus on visually-friendly representations of complex data to inform clinical users
Clinical geneticists	□Desire capacity to participate in data processing □Prefer a graphical user interface for navigation and execution □Seek support for incorporation of patient phenotype via clinical text-mining and links to biological databases for gene-disease associations □Desire a visually dynamic way to explore the genomic dataset, across varying statistical thresholds	□Desire intelligent system that will automatically highlight abnormal data qualities and/or clinically-relevant information (e.g. in silico prediction of clinically-relevant variants) □Prefer a system that consolidates key information from diverse resources into one interface, rather than distributing across multiple panels	□Necessitates software that connects to collaborative networks to discover similarities of patient cases within institutions and globally
Genetic counsellors and general physicians	□Prefer a streamlined, simple interface related strictly to their domain. Graphical interface should exclude functionalities not related to variant/gene-level prioritization	□Benefit from automated generation of clinical report that prioritizes the clinically relevant variants and masks the clinically uncertain results	□Achieve better workflow efficiency when software maintains an environment to collaboratively review and annotate the variant data with clinical geneticists and bioinformaticians

Table 3-1 An overall summary of the desired software features and design architectures across bioinformaticians, clinical geneticists, genetic counselors and physicians.

It is our interpretation that no single interface will adequately address the needs of all users, necessitating the capacity of future WES/WGS systems to provide interface options to best meet the needs and expectations of the diverse users. The existing academic and commercial software (Appendix B-1) place emphasis upon graphical user interface that are viewed by bioinformaticians as too rigid and not customizable for distributed network analysis. While some tools may be designed to create user-friendly workflows, the lack of design-focus on clinical target users (i.e. geneticists and genetic counselors) impede their adoption in clinical settings (Appendix B-1). The importance of user-centric themes is consistent with emerging models of care and medical decision-making support systems, such as observed for breast cancer diagnosis and management[141, 142], early recognition of sepsis[143], antibiotics prescriptions[144, 145] and interpretation of medical images[146, 147] where extensive evaluations on physicians' and nurses' interactions in work practices reveal similar concepts surrounding issues of sharing information across collaborative settings, and tensions between integration and standardization.

Given the complexities involved, software which attempts to address all possible tasks that arise in clinical genomics is less likely to be incorporated into practice than software specific to exome/whole-genome analytical tasks. To be successful, a medical decision support system should be compatible to an existing clinical workflow[148, 149], and actionable outputs intelligently filtered and presented at appropriate times[150, 151]. In WES/WGS, we found this workflow scope includes a system's capacity to incorporate clinical keywords and genetic hypotheses pertinent to each unique patient (also cited in [152, 153]), and results delivered at specific workflow stages with respect to the disparate foci of counselors, physicians, geneticists, and bioinformaticians (e.g. Figure 3-1). Clinical geneticists expressed desire for an encompassing graphical design that gives them more control over the technical aspects of the pipeline, integrating genomic information with patient history but at the same time removes them from the realm of scripting and the command-line. Meanwhile, genetic counselors (and general physicians) wished to solely focus on gene prioritization and efficient delivery of final results without distraction by functionalities irrelevant to their work processes. The results highlight a need for systems to facilitate the generation of clinical reports, including the appropriate distribution of technical versus clinical details, sharing of notes between clinical staff about specific variants, overview of genes not covered by WES, and the family structure. The format of the prioritized report (Figure 3-5) mirrors the precedent of prioritized information in other modes of clinical reports, e.g. a radiologist’s X-ray report separating clinical impressions from descriptive details of radiographic appearance of specific organs[154].

Strong community observations should be noted by system developers. Our study confirms that an ultimate clinical WES/WGS systems will need to be well connected to online resources, such as animal model phenotypes[155, 156], biological system annotations[157, 158], and disease-focused databases[106, 159, 160]. This is concordant with earlier work that demonstrated the importance of rich access to external resources and databases[114, 161, 162]. The integration of metadata and diverse biological annotations to patient electronic health records will require strict compliance to standards (examples discussed in Appendix B-11). Our study further highlighted the need to integrate access within a single system, sparing users from mastering diverse interfaces.

Our results suggest future software should provide separate interfaces for each target user group. One can envision ‘purpose-driven’ interface options, allowing users to focus on the aspect of the analysis and interpretation relevant to their duties. While the tailored software is fitted to individual domains, it must at the same time facilitate collaboration, as increasingly diverse expertise is key requirement for WES/WGS interpretation. The informatics specialists may be charged with reporting on data linking candidate genes to specific biological processes, clinical geneticists will evaluate specific mutations for a causal role in disease/phenotype, and genetic counselors will indicate the variations that need to be conveyed. These activities are interactive and may require cycles of expert attention. Insights to overcome socio-technical challenges can be drawn from research in Computer-Supported Cooperative Work (CSCW)[163], including themes surrounding information credibility[164], coping with narrative and numeric data[165], scalable methods for managing increasingly large data sets[166], and caution surrounding interpretation of automated systems[167] (discussed further in Appendix B-10). As WES/WGS analysis software matures it will empower clinicians with more automated procedures, which we anticipate will decrease dependency on bioinformaticians for data processing. These experts will continue to be closely involved, developing and applying new approaches for the discovery and interpretation of additional genetic alterations. Advances over the coming years will result in new requirements for collaborative interactions, for instance as the current focus on alterations in protein coding sequences expands to include regulatory sequence alterations. Expansion of the cooperative capacity of the software will assist the diverse users as the field matures.

3.5.1 Conclusions

As high-throughput WES/WGS technologies continue to mature, healthcare providers need efficient software to facilitate interpretation for clinical decision-making. By conducting multiple focus groups of diverse healthcare classes active in clinical genetics, our present study reveals there are distinct types of WES/WGS analysis needs for different classes of domain specialists. The results presented illustrate the cognitive processes and tentative designs envisioned by the range of clinical professionals key to the process. A natural follow-up for future work is to implement the features into a prototype software package and conduct intervention trials to evaluate effectiveness and performance within clinic sites. The limitations to this study are discussed in Appendix B-12.

Chapter 4: FLAGS, frequently mutated genes in public exomes

4.1 Synopsis

Background Dramatic improvements in DNA-sequencing technologies and computational analyses have led to wide use of whole exome sequencing (WES) to identify the genetic basis of Mendelian disorders. More than 180 novel rare-disease-causing genes with Mendelian inheritance patterns have been discovered through sequencing the exomes of just a few unrelated individuals or family members. As rare/novel genetic variants continue to be uncovered, there is a major challenge in distinguishing true pathogenic variants from rare benign mutations.

Methods We used publicly available exome cohorts, together with the dbSNP database, to derive a list of genes (n = 100) that most frequently exhibit rare (<1%) non-synonymous/splice-site variants in general populations. We termed these genes FLAGS for FrequentLy mutAted GeneS and analyzed their properties.

Results Analysis of FLAGS revealed that these genes have significantly longer protein coding sequences, a greater number of paralogs and display less evolutionarily selective pressure than expected. FLAGS are more frequently reported in PubMed clinical literature and more frequently associated with diseased phenotypes compared to the set of human protein-coding genes. We demonstrated an overlap between FLAGS and the rare-disease causing genes recently discovered through WES studies (n = 10) and the need for replication studies and rigorous statistical and biological analyses when associating FLAGS to rare disease. Finally, we showed how FLAGS are applied in disease-causing variant prioritization approach on exome data from a family affected by an unknown rare genetic disorder.

Conclusions We showed that some genes are frequently affected by rare, likely functional variants in general population, and are frequently observed in WES studies analyzing diverse rare phenotypes. We found that the rate at which genes accumulate rare mutations is beneficial information for prioritizing candidates. We provided a ranking system based on the mutation accumulation rates for prioritizing exome-captured human genes, and propose that clinical reports associating any disease/phenotype to FLAGS be evaluated with extra caution.

4.2 Introduction

Rare Mendelian diseases are caused by altered function of single genes and individually have a low prevalence (fewer than 200,000 people in the United States, or fewer than 1 in 2,000 people in Europe)[168] but collectively these affect millions of individuals worldwide[169]. The current best estimate on the number of rare genetic disorders is between 6,000 to 7,000 based on OMIM[170], and a comprehensive reference portal for rare diseases (Orphanet)[171]; however, taking into consideration that the human phenome is far from fully characterized[172] together with higher estimates on rare-disease-causing genes based on human mutation rate and the number of essential genes[173], the number of rare genetic disorders is likely higher.

With the increasing rate of the discovery of rare genetic variants, whole exome sequencing (WES) has the potential to identify the majority of the remaining rare-disease-causing genes in the near future. A major challenge in identification of the true pathogenic variants lies in the differentiation between a large number of non-pathogenic functional variants and disease-causing sequence variants in a studied family (in this study, the term “functional variant” is restricted to missense/nonsense and splice site variants). Current WES analyses of rare genetic disorders use similar approaches[113] to filter the observed variants to enrich for potential causal genes. However, it is well established that a significant proportion of coding variants in each individual represent rare variants (absent from dbSNP or observed with frequency of ≤1%)[40], and that genomes of healthy individuals contain an average of ~100 loss-of-function variants[174]. The analyst must further consider the possibility that non-coding variations (e.g. regulatory alterations) could be involved, thus the filtered results may not contain the causal gene. Thus, for many rare disorders, it is still challenging to separate the real disease-causing variant from the prioritized set of rare, likely functional variants that are not accountable for the investigated phenotype.

There are broadly used tools such as SIFT[175] and PolyPhen-2[176] that provide an interpretation of mutation impacts. Many of these tools focus on the individual variants. In the variant-focused studies, it has been noted that variants tend to arise more frequently in long genes (e.g. *TTN* and *MUC16*). In considering that researchers often focus their interpretation of exome data on the genic level initially, it might be advantageous to have methods and ranking systems that integrate the individual variants at the genic level more systematically to inform variant prioritization. While there are long-standing methods for ranking a set of genes based on their annotations[177], there has been limited work on rankings based on their characteristics from NGS studies such as WES. One ranking system based on the genic level is RVIS[178]. RVIS generates a score based on the frequencies of observed common coding variants compared to the total number of observed variants in the same gene.

To further help in identification of disease-causing variants from families affected by rare Mendelian disorders, we expanded the current, common prioritization parameters that focus mainly on frequency at which variants themselves are seen in normal population, to include the frequency at which genes are found to be affected by rare, likely functional variants. Using rare variations from dbSNP and EVS, we introduced the concept of FLAGS (FLAGS for FrequentLy mutAted GeneS). We showed that these genes possess characteristics that make them less likely to be critical for disease development, but are more likely to be assigned causality for diseases than expected for protein-coding genes in general. We further demonstrated FLAGS' utility via a case study as well as literature review, and application in our in-house database. Finally, we provided a ranking system from FLAGS to assist in the prioritization of genes from exome/whole-genome clinical studies.

4.3 Methods

4.3.1 Terminologies used in this study

In this study, the term "functional variants" refers to variants that are missense, nonsense or fall within a splice site window (see below for specifics). The length of a gene is defined to be the longest open reading frame (ORF) of the gene, thus excluding promoters, untranslated regions and introns. All genes are referred to by their HGNC (HUGO Gene Nomenclature Committee)[179] official gene symbol.

4.3.2 Datasets

In the following sections, we provide detailed descriptions of how the datasets were obtained or generated. Table 4-1 lists the size and descriptive nature of the datasets used in this study. Each gene list referred to in this report can be found online at BMC where the paper is published.

Name of datasets	**Size**	**Description**
FLAGS	100	The top 100 of FrequentLy mutAted GeneS with rare (<1% allelic frequency) functional variants from dbSNPv138 and ESP6500
OMIM	3099	The list of protein-coding genes associated with human diseases from Online Mendelian Inheritance in Man
HGMD	2691	The list of protein-coding genes with damaging mutations (<1% allelic frequency) from Human Gene Mutation Database[160].
WES	300	Downloaded from Boycott *et al.* (2013)[180] - a list of novel genes implicated in human disorders based on whole exome sequencing studies, or novel/known pathogenic mutations discovered by whole-exome sequencing.
Background	18580	The entire set of human protein-coding genes that have complete start and end translation annotations with a specified dN/dS ratio

Table 4-1 Description of the datasets used in this study

4.3.3 FrequentLy mutAted GeneS (FLAGS)

Variations from EVS hosted on the NHLBI Exome Sequencing Project (ESP6500) were downloaded on February 2014. The criteria used to generate the variations are available online (http://evs.gs.washington.edu/EVS/). Variations from dbSNPv138 were downloaded from the NCBI website (version date 20130806). Genomic annotations were assigned to each variation using SnpEff v3.5g with the parameter –SpliceSiteSize 7 and human genome version GRCh37.75. Variants were filtered for allelic frequency <1% according to dbSNP's overall frequency and EVS's combined population frequency. Where a discrepancy in the reported frequency arose between the two resources, we took the higher frequency. Variants were further filtered for "functional" coding mutations that result in a change in the amino acid sequence (i.e. missense/nonsense), or mutations that reside within a putative splice site junction (with a window size of 7, as supplied in the parameter for SnpEff). The remaining mutations were excluded if they were observed more than 10 times within our in-house database consisting of 150 exomes and 13 whole genomes. This last step was included because we noticed it is common to see polymorphic mutations from dbSNPv138 without an allelic frequency attached;

filtering against an in-house pipeline allowed us to remove polymorphic variants that do not have an annotated frequency. Among these remaining mutations, for each gene, we counted the number of mutations observed per gene. Only protein-coding genes with a fully annotated translation start and end, and a valid dN/dS ratio are included for consideration (see Methodology section "Gene length and dN/dS ratio"). From this ranked list, we selected the top 100 genes (0.5% of the 19818 genes overlapping between dbSNP and EVS) with the most observed mutations as a focus for this study. This set will be referred throughout the manuscript as "FLAGS". The entire ranked list is available online at BMC.

4.3.4 Disease genes datasets

To obtain a list of reliable disease-associated genes, we drew from multiple resources. The first list of disease-associated genes was downloaded from OMIM website on March 2014 using the provided file "morbidmap". This list will be referred throughout the manuscript as "OMIM genes". A second list contains pathogenic variations downloaded from the HGMG professional version (file date 20130927)[160]. To focus on likely high-penetrance pathogenic alleles, we filtered the variations in this file by the same frequency criteria as we performed for obtaining FLAGS (see Methodology section "FrequentLy mutated GeneS"), and limited to only the mutations annotated as "DM" (damaging mutations). The affected genes from those remaining variations are compiled, and will be referred throughout this manuscript as "HGMD genes". A third disease set was downloaded from the Supplemental file published by Boycott *et al.* (2013)[180], which provided a compiled list of novel genes and/or novel phenotypes associated with known disease-genes discovered through exome sequencing. For all three disease-associated-gene lists, we mapped the gene symbols to their official HGNC gene symbol (and discarded the ones that could not be mapped), retained only protein-coding genes with a fully annotated translation start and end, and a valid dN/dS ratio. OMIM and HGMD (Human Gene Mutation database) overlap with the top 100 FLAGS by 42 and 37 genes respectively.

4.3.5 Background dataset

The complete list of human-coding genes was downloaded from Ensembl[181] Biomart on March 2014 using version Ensembl Genes 75 with genome version GRCh37.p13. Protein-coding genes without HGNC gene symbol, a proper translation start and translation end annotation according to this genome version were discarded. Genes without a valid dN/dS ratio were removed (i.e. without any observed synonymous polymorphisms according to dbSNPv138 and EVS). This last step was done for two reasons: 1) to ensure there is no bias when evaluating dN/dS ratio in our results, 2) to ensure the genes selected in this study have been covered in NGS studies, since any gene without at least one observed synonymous mutation is presumably not sufficiently captured in either exome or whole-genome studies. The Background set overlaps FLAGS completely. The comparison analyses in the Results section are done without removing the overlap between the gene datasets.

4.3.6 Gene length and dN/dS ratio

We calculated the selection pressures acting on genes by comparing non-synonymous substitution per non-synonymous site (dN) to the synonymous substitutions per synonymous site (dS). This ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site (dN/dS) was calculated using the formula

$$\frac{\dfrac{\#of\ observed\ non-synonymous\ substitutions}{\#of\ \ possible\ non-synonymous\ site}}{\dfrac{\#of\ observed\ synonymous\ substitutions}{\#of\ \ possible\ synonymous\ substitutions}}$$ [182].

The number of possible synonymous and non-synonymous mutations was derived by examining the longest annotated coding transcript per gene (transcript length based upon Ensembl Biomart described above). Only transcripts with annotated start and end positions were considered. The number of observed synonymous and non-synonymous mutations was calculated from the same dbSNPv138 and EVS datasets as described above. We verified that our methodology provides a comparable dN/dS ratios to the ratios reported previously[182]. Gene length was derived by converting the same transcript that was used to calculate the dN/dS ratio into amino acid sequences. In this study, the term “gene length” is defined to be the ORF of the gene, thus excluding promoters, untranslated regions and introns.

4.3.7 Paralogs

The paralogous relationships for human genes were derived from the Ensembl Comparative Genomics API using version Ensembl Genes 75, GRCh37.p13. A custom Perl script was written to extract the paralogs for every gene.

4.3.8 Gene-to-disease phenotypic terms

We used MeSHOP software[42] to identify over-represented disease terms associated with each gene. MeSHOP returns a list of MeSH (Medical Subject Heading) terms for each gene with a p-value for each term. Each p-value was calculated by an over-representation (compared to control) of the MeSH terms assigned to the set of articles within PubMed that are associated with the gene (based on relationships defined in gene2pubmed; articles considered include up to March 2013). From this output, for each gene, the non-disease related MeSH terms were filtered out, and the remaining MeSH terms were selected for significance (using the Bonferroni correction and a significance threshold of 0.05). To derive gene-to-disease relationships with an independent source, we extracted phenotypic diseased terms per gene from Human Phenotype Ontology website[183] by downloading the file “genes_to_diseases.txt” (version April 2014).

4.3.9 Publication record analysis

For our publication analysis on the relationship between a gene and its frequency of citation(s) within biomedical literature, we used Gene Reference into Function (GeneRIF), a manually curated list of experimentally validated gene functions available as part of NCBI’s EntrezGene database. Each entry in GeneRIF contains a short description of a gene function and a PubMed identifier for the publication documenting the evidence of the described function. Therefore, we were able to count the number of papers published on a gene’s functionality by counting the number of PubMed records associated to the gene. The following are the detailed steps of our publication calculation. First, two flat files necessary for our analysis were downloaded via FTP from NCBI Gene on April 2014: GeneRIF (available at ftp://ftp.ncbi.nih.gov/gene/GeneRIF/generifs_basic.gz) and EntrezGene entries for human (ftp://ftp.ncbi.nih.gov/gene/DATA/Homo_sapiens.gene_info.gz). Second, because GeneRIF refers to each gene by its EntrezGene ID, we mapped the gene symbol of all genes on our lists (FLAGS, OMIM, HGMD, Background) to EntrezGene ID using EntrezGene entries downloaded in the previous step. Third, for each gene of interest, we counted the number of PubMed IDs (PMIDs) associated with its EntrezGene ID in GeneRIF. Because GeneRIF does not guarantee one-to-one relationship between a GeneRIF entry and a PMID (http://www.ncbi.nlm.nih.gov/books/NBK3840/#genefaq.Why_does_the_number_of_GeneRIFs) , we filtered out duplicates in the list of PMIDs linked to a gene. Last, to filter the PMIDs by their publication date, we collected the publication date of each PMID via queries into PubMed using the ESummary query provided within the Entrez Programming Utilities (E-utilities).

4.3.10 Statistical analyses

Unless stated otherwise, all statistical analyses and plots were carried out in R (https://www.r-project.org) version 2.15.3. Non-parametric Mann–Whitney U one-tailed test was executed by wilcox.test function with parameter exact = TRUE. Violin plots were generated with Vioplot package.

4.3.11 Mutation Detection using WES – a case study

A 3-year old female patient, born as an only child to non-consanguineous parents of Turkish descent after an uncomplicated pregnancy and delivery, presented with profound early-onset developmental delay, microcephaly, seizures, dysmorphic features, myopia, bone marrow dysplasia with lymphopenia, neutropenia, aplastic anemia and combined immunodeficiency (B and T cell) was enrolled into the TIDEX gene discovery project, approved by the Ethics Board of the Faculty of Medicine of the University of British Columbia (H12-00067).

Extensive clinical investigations were performed according to the TIDE diagnostic protocol[184] to determine the etiology of patient’s condition. These included: chromosome micro array analysis for copy number variants (CNVs) (Affymetrix Genome-Wide Human SNP Array 6.0); telomere length analysis; CT and MRI scans and comprehensive metabolic testing.

Genomic DNA was isolated from the peripheral blood of the patient as well as parents using standard techniques. Whole exome sequencing was performed for the index patient and her unaffected parents using the Ion AmpliSeq™ Exome Kit and Ion Proton™ System from Life Technologies (Next Generation Sequencing Services, UBC, Vancouver, Canada) at 120X coverage. An in-house designed bioinformatics pipeline (Appendix C-3) was used to align the reads to the human reference genome version hg19 and to identify and assess rare variants for their potential to disrupt protein function. The candidate variants were further confirmed using Sanger re-sequencing in all the family members. Primer sequences and PCR conditions are available on request. Deleteriousness of the candidate variants was assessed using Combined Annotation–Dependent Depletion (CADD) scores[37].

4.4 Results

4.4.1 FLAGS: genes frequently affected by rare, likely-functional variants in public exomes

It has been previously reported that *TTN* and *MUC16* appear in multiple exome analyses due to their length[185, 186]; researchers are aware of these genes and are cautious when encountering rare likely functional (missense, nonsense, splice site) variants in WES analyses[187, 188]. In a study of 53 independent families suffering from distinct rare inborn errors of metabolism (comprising of 150 whole exomes and 13 whole genomes; http://www.tidebc.org; Appendix C-4), we confirmed that rare/novel, likely functional variants affecting *TTN* and *MUC16* repeatedly passed all the prioritization steps of our pipeline and appeared in ~5% of our candidate disease-gene lists. However, other genes were repeatedly observed in multiple families affected with different phenotypes (e.g. *DST*). This motivated us to compile a set of FLAGS (FrequentLy mutAted GeneS) to understand their properties and facilitate better interpretation of phenotypes associated with these variants. The FLAGS list was generated by ranking genes based on number of rare (<1%) functional variants affecting these genes in general populations (NHLBI Exome Sequencing Project (ESP6500) and dbSNPv138). As expected, *TTN* and *MUC16* are the top two genes based on the number of rare functional variants; however, other genes that were frequently affected by rare, likely functional variants in multiple TIDE families with unrelated phenotypes were also observed to be frequently mutated in general population. To explore the properties of these frequently mutated genes, we focused our analysis on the top 100 from this ranked list, which we hereafter refer to as FLAGS (Figure 4-1).

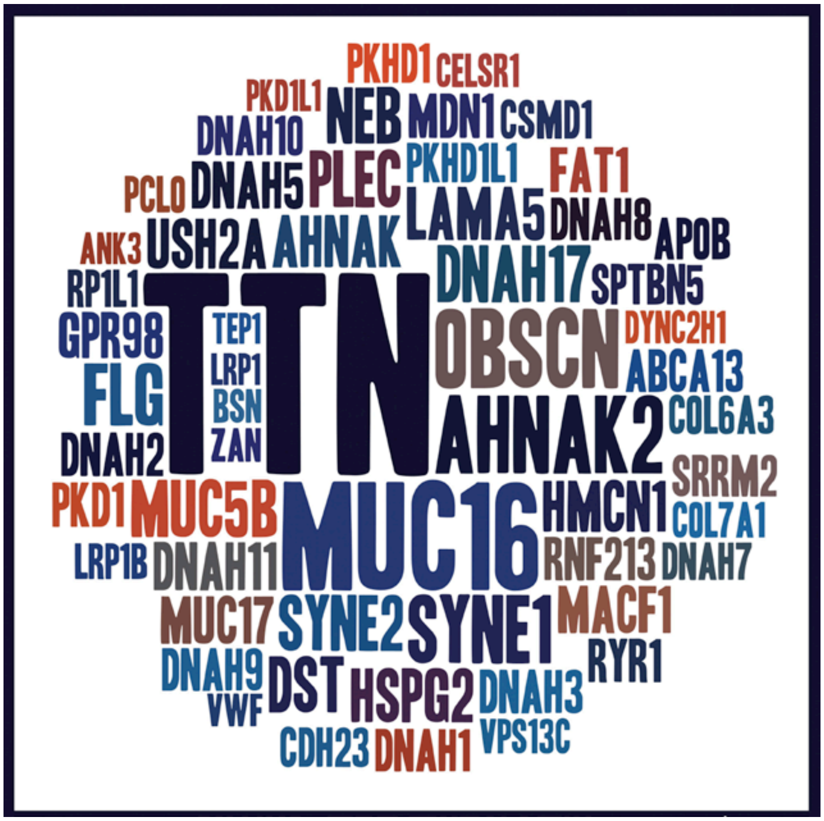


Figure 4-1 The word cloud of FLAGS. A text file was created using a custom Perl script to reflect the frequency of mutation per gene in FLAGS. The Tagxedo (http://www.tagxedo.com/) was then used to generate the word cloud. The size of the words reflects how frequently they are found to bear rare, likely functional variants in the general population. As expected TTN and MUC16 are the top two genes.

4.4.2 FLAGS tend to have longer ORFs

In this study, the assignment of gene length refers to the longest open reading frame. Genes with longer ORFs are expected to have more mutations than shorter genes. To confirm this, we determined the distribution of gene lengths based on the longest annotated open reading frame for each gene. FLAGS have an average length of 4653 ± 3605 aa (amino acids). The high variance is due to two genes (*TTN* and *MUC16*) having extremely long lengths (35992 and 14508 aa respectively) compared to the rest of the protein coding genes. Excluding the 2 outlying genes, the remaining FLAGS genes (n = 98) have an average ORF length of 4233 ± 1399 aa. Figure 4-2A shows the distribution of ORF lengths across different evaluated datasets (with outliers removed to show the distribution clearer). The entire FLAGS have overall much higher ORF length than HGMD, OMIM and Background (HGMD, OMIM comparisons each yield a p-value <2.2e−16, Background comparison yields a p-value of 0.00027). This is aligned with our expectation that FLAGS are frequently mutated from exome analysis because they correspond to genes with long coding regions.

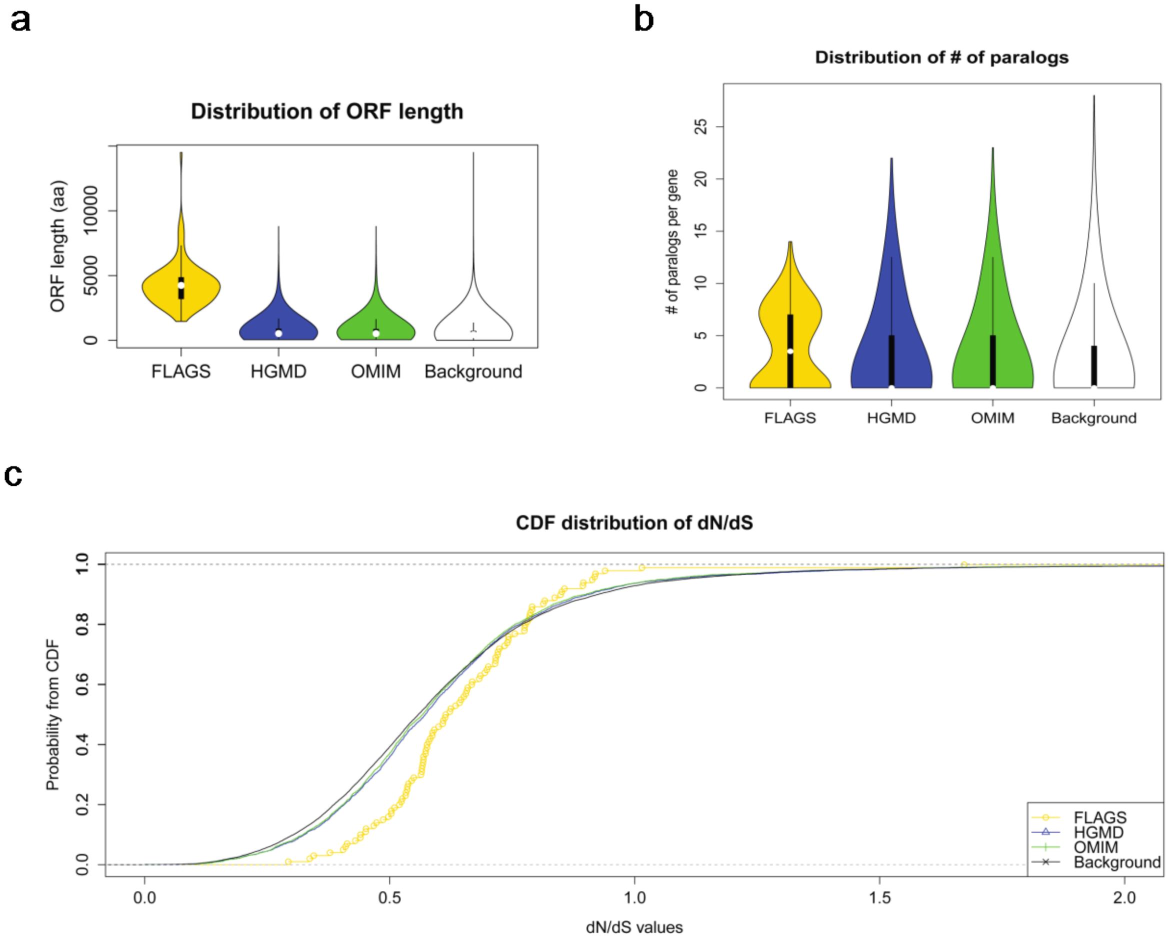


Figure 4-2 Properties of FLAGS. (a) Violin distribution of open reading frame lengths across the evaluated gene sets. Y-axis shows the length defined in terms of amino acids for the longest annotated transcript per gene. Outliers are excluded from the plot. (b) Distribution of number of paralogs per gene across the evaluated gene sets. Y-axis

shows the violin distribution of paralogs based on Ensembl Compara database. Outliers are excluded from the plot. (c) Cumulative distribution of dN/dS ratio across the evaluated gene sets. X-axis is limited from 0 to 2, and Y-axis plots the corresponding probability according to the cumulative distribution function.

4.4.3 FLAGS tend to have paralogs

The presence of paralogs may increase tolerance for otherwise phenotype-inducing functional variations due to functional compensation[189]. We calculated the number of paralogs per gene reported by the Ensembl Compara database[181], and compared this property between different gene sets. FLAGS overall have an average of 4 paralogs per gene. Figure 4-2B shows the distribution of the number of paralogs across the different gene sets. Aligned with our expectation, FLAGS have more paralogs than genes from OMIM, HGMD and Background (OMIM p-value =7.2e−05, HGMD p-value =7.4e−05, Background p-value =8.1e−09). While the existence of paralogs may cause read mapping challenges that leads to an increased frequency of false variant predictions, most of these technical errors will be eliminated by a filter for variant frequency, as they will arise recurrently.

4.4.4 FLAGS tend to have higher dN/dS ratios

Genes which exhibit many functional genetic variations (missense/nonsense/splice site) may have a higher tolerance for variations and thus a reduced likelihood of phenotypes subject to negative selection. For each gene, we calculated the dN/dS ratio as a proxy indicator of the amount of selective pressure acting on protein-coding genes. FLAGS have an average dN/dS ratio of 0.65 ± 0.18. Overall these genes have significantly higher ratio compared to genes from HGMD, OMIM, and Background (each individual comparison yields a p-value <0.005). Figure

4-2C shows the relative densities from cumulative distribution functions for each gene set. The trend indicates that frequently mutated genes have higher dN/dS ratio on average than expected.

4.4.5 Variants detected in FLAGS tend to be predicted as less deleterious

We explored the possibility that the FLAGS genes are affected by less deleterious rare variants compared to other genes. If the variants in FLAGS were less likely to be involved in diseases, then we would expect the variants to have lower predicted damage scores. To calculate this, we used the Phred-scaled Combined Annotation Dependent Depletion (CADD) score developed by Kircher *et al.* (2014) to rank the deleteriousness of each single nucleotide variant[37]. The method objectively integrates diverse annotations into a single measurement for each variant by training upon ~15 million genetic variants separating humans from chimpanzees against a simulated set of variants not exposed to selection. This method was chosen over other variant prediction tools because of its superior performance[37] and its ability to quantify the severity of a variant by a ranking system. This ranking system compares the candidate variant against other possible variants in the genome and assigns it a score based on this comparison; other variant prediction tools do not take into account other possible mutations in the genome[190]. Also, the CADD method includes ranking of nonsense and splice site variants, while other tools only handle missense[37]. For each gene, we calculated the proportion of variants with CADD Phred-scaled score <10, between 10 and 20, and above 20. We found that FLAGS are more enriched for variants with low scores, compared to OMIM and HGMD (Figure 4-3A; p-values =2.6e−11, 2.9e−12 respectively). Likewise, OMIM and HGMD are more enriched for variants with high impact score (>20) than FLAGS (Figure 4-3B; p-values =2.4e−09, and 1.2e−10 respectively). These results are aligned with our expectation. We additionally analyzed the genic tolerance of FLAGS to functional genetic variants, using residual variation intolerance score (RVIS) published by Petrovski *et al.* (2013)[178] and observed trends in the same direction (Appendix C-2).

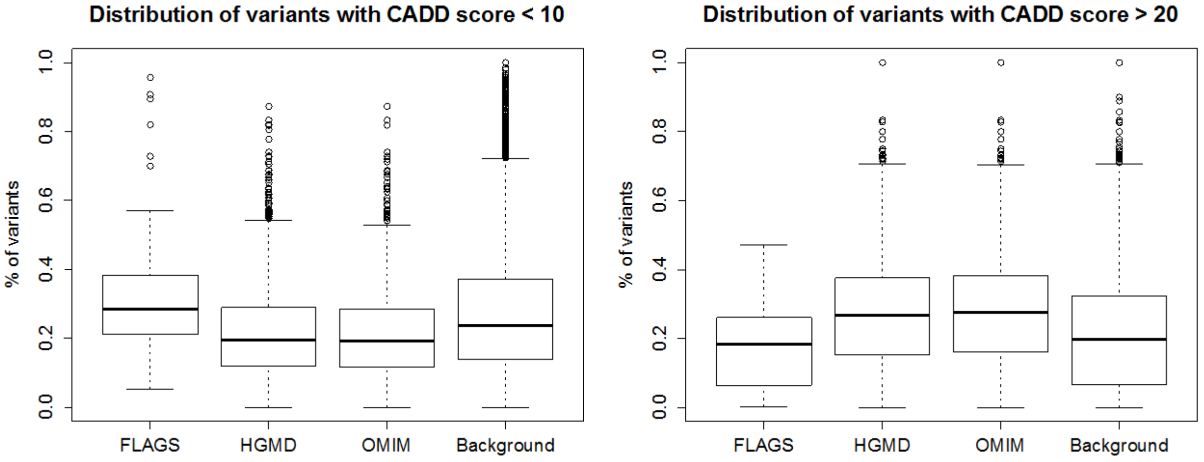


Figure 4-3 FLAGS genes are affected by rare variants predicted to be less deleterious than the variants affecting known disease-genes. Left: A boxplot distribution of proportion of variants with CADD score <10. The Y-axis plots the proportion of variants within each gene set having a Phred-scaled CADD score of <10. The proportion was calculated per individual gene. Right: A boxplot distribution of proportion of variants with CADD score >20. The Y-axis plots the proportion of variants within each gene set having a Phred-scaled CADD score of >20. The proportion was calculated per individual gene.

4.4.6 FLAGS tend to be reported in PubMed and associated with disease phenotypes

We sought to determine if there is a publication bias for pathogenic mutations in the frequently mutated genes. For each gene, we calculated the number of publications related to human diseases and biological functions using GeneRIF annotations (Figure 4-4). FLAGS have an average of 51 articles per gene, which is lower than for genes from HGMD and OMIM

(OMIM p-value =0.00087, HGMD p-value =0.0035). However, FLAGS have more publications than the Background set (p-value =6.3e−12).

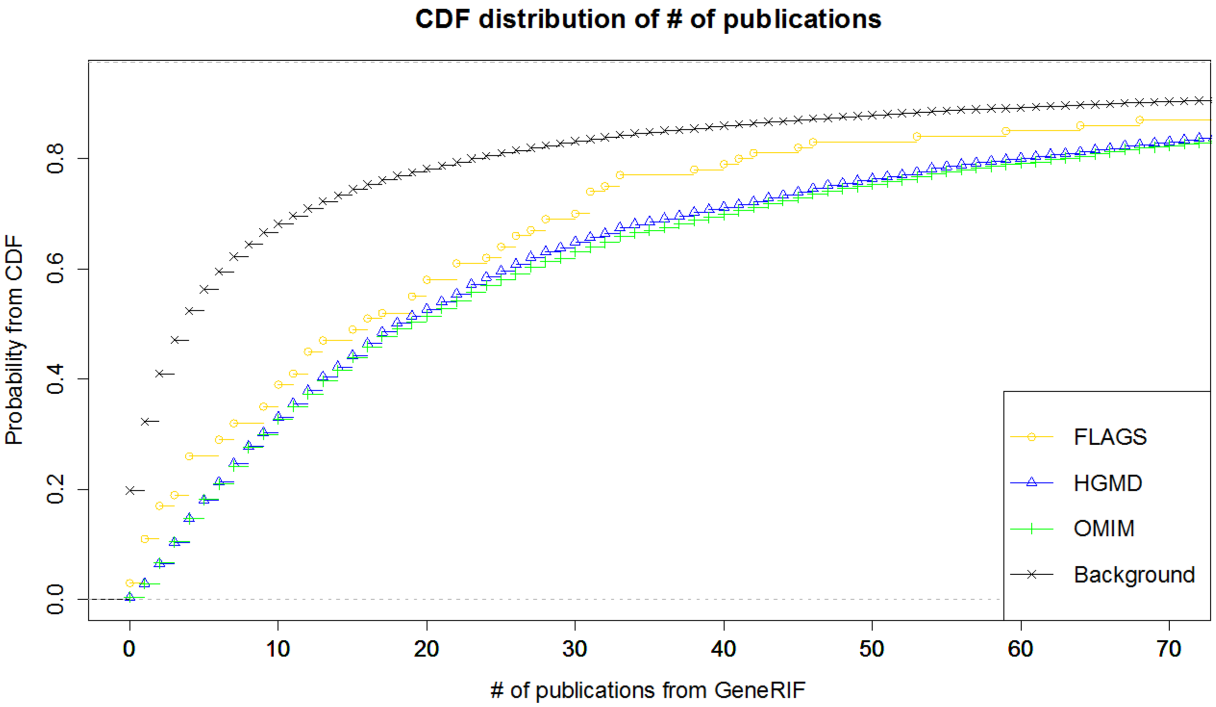


Figure 4-4 Cumulative distribution of the number of publications per gene across the evaluated gene sets. X-axis plots the number of publications from GeneRIF per gene, and Y-axis plots the corresponding probability according to the cumulative distribution function.

We next considered if the frequently mutated genes are associated with greater diversity of disease phenotypes compared to disease-associated genes. Our expectation is that if the frequently seen genes are arising as candidates in more studies, and are less likely to be truly pathogenic, then they could be associated to a wider range of phenotypes in the literature (we recognize the association could also be due to pleiotropy[191], see Limitations). To analyze if FLAGS have been frequently correlated to human diseases, we used two different computational resources (MeSHOP[42], HPO[183]) to extract known significant relationship(s) between genes and human disease phenotypes based on published scientific articles. Figures 4-5A and B show the distribution of the number of disease terms from HPO and MeSHOP per gene within gene sets. From MeSHOP results, we see that FLAGS have slightly fewer MeSH diseased terms per gene than genes from OMIM (mean 8.1 vs. 10.2; p-value =0.013), and significantly fewer terms per gene than HGMD genes (mean 8.1 vs. 9.5; p-value =2.3e−12). FLAGS have more MeSH terms than Background genes (mean 8.1 vs. 3.1; p-value =1.3e−15). These observations are consistent with the results based on HPO annotations, where we again see that while FLAGS have fewer disease phenotypic terms than genes from OMIM and HGMD (mean 2.1 vs. 3.7 and 3.8 respectively; p-values <0.0001), FLAGS exhibit more terms than the Background (mean 2.1 vs. 0.6; p-value =3.7e−14). To adjust for the potential bias that genes with more articles are likely to have more MeSH and HPO terms attached, we repeated the analysis by normalizing the MeSH and HPO terms to the number of publications in GeneRIF. The normalized observations are consistent with the results if no normalization was applied (Appendix C-5).

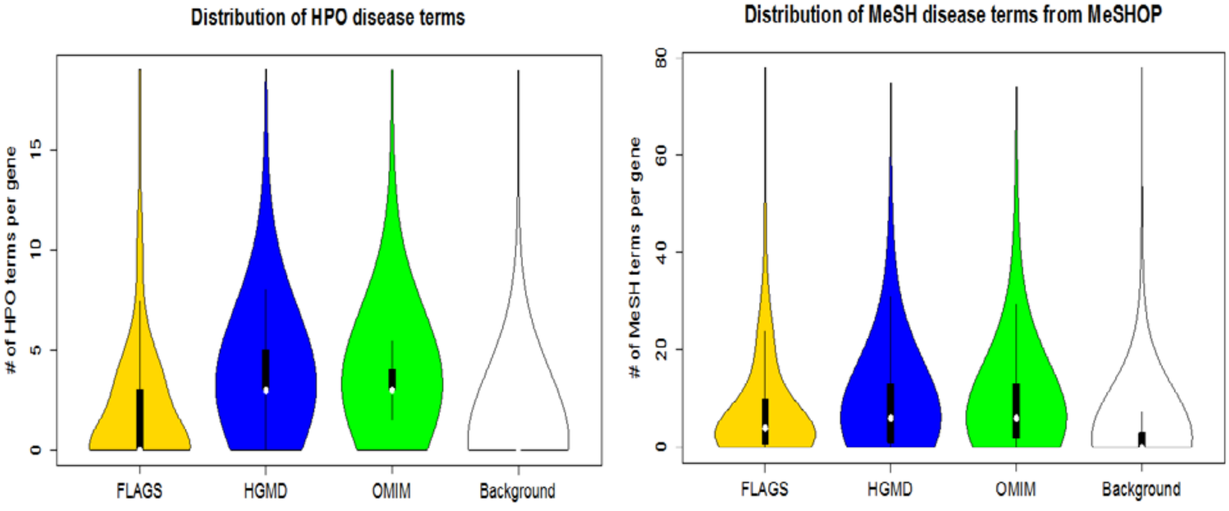


Figure 4-5 FLAGS tend to be associated with disease phenotypes. Left: Violin distribution of number of HPO disease terms across the evaluated gene sets. Y-axis is the violin distribution showing the number of HPO terms per

gene. Outliers are excluded from the plot. Right: Violin distribution of number of MeSH disease terms from program MeSHOP across the evaluated gene sets. Y-axis is the violin distribution showing the number of MeSH terms per gene. Outliers are excluded from the plot.

4.4.7 FLAGS recently implicated in rare-Mendelian disorders

We sought to determine which FLAGS have been reported with pathogenic mutations in NGS clinical studies. Boycott *et al.* (2013) provided a compilation of 178 novel genes discovered to be disease-associated through exome sequencing[180], of which three overlapped with FLAGS (KMT2D/MLL2, HERC2, and DST). To explore the properties of those 3 genes, we analyzed the ratio between number of rare variants and gene length, as well as presence of putative essential protein domains by assessing the distribution of rare variants across the gene. We found that among the FLAGS, KMT2D and HERC2 have the lowest ratios of number of rare variants compared to gene length, while DST is one of the three genes among the FLAGS set with significant non-uniform distribution of rare variants across the gene (p-value =1.2e-04; the other two are EPPK1 and HRNR; see Appendix C-1 for more details on methodology and rationale). If we were to expand this 178 novel-rare-disease gene list from Boycott *et al.* (2013) to include the exome studies reporting on already-known disease-associated genes with known/novel pathogenic mutations, then this expanded set (n = 300) overlapped FLAGS by an additional 7 genes (*TTN, RYR1, PKHD1, RP1L1, ASPM, SACS, ABCA4*). In the discussion we provide our thoughts and literature analysis on why these genes have been reported as disease-associated despite being among the frequent genes to harbor rare functional variants.

4.4.8 Applying FLAGS to prioritize candidate variants: *Case study*

To demonstrate a disease-causing variant prioritization approach using FLAGS and whole exome sequencing data, we selected one family from our TIDE cohort affected by an unknown rare genetic disorder. Through WES performed for the index and her unaffected parents (Methodology - Mutation Detection using WES – a case study), rare variants were identified and assessed for their potential to disrupt protein function. Only those variants predicted to be functional (missense, nonsense and frameshift changes, as well as in-frame deletions and splice-site effects) were subsequently screened under a series of inheritance models. In total, we identified six rare "functional" homozygous, and eight rare "functional" compound heterozygous candidates. Of those, only two genes affected by missense variants were considered functional candidates:

(1) *VPS13B* gene (OMIM 607817) had been found to bear homozygous or compound heterozygous mutations in patients with Cohen syndrome (OMIM 216550). Cohen syndrome is characterized by developmental delay/intellectual disability, facial dysmorphism, microcephaly, neutropenia, and weak muscle tone (hypotonia). The features of Cohen syndrome vary widely in presence and severity among affected individuals. Additional features, perhaps patient-specific, appear in the reports; myopia and small hands and feet are observed in our patient. In our WES analysis, we identified two rare variants affecting this gene in the index, suggesting compound heterozygous inheritance. Neither of the variants was found in more than 160 in-house exomes; one of the variants was predicted to be deleterious using the CADD scores[37] with a score higher than 20, while the second variant was given the score of less than 5. Sanger re-sequencing confirmed that mother is a carrier of one variant, while the father is the carrier of the second

variant and the index is compound heterozygous making the *VPS13B* gene a candidate disease-gene in this family.

(2) *SENP1* gene (OMIM 612157) product is one of the desumoylating enzymes[192] which is important for proper development and survival in mice. *SENP1* was found to regulate expression of *GATA1* in mice and subsequent erythropoiesis[193]. Furthermore, *SENP1* was found to be essential for the development of early T and B cells through regulation of *STAT5* activation[194]. To date, germline mutations in SENP1 had not been described in any human diseases. Our WES analysis identified a rare missense homozygous variant in the index. The variant was not found in more than 160 in-house exomes and was predicted to be the most deleterious of all homozygous variants using the CADD scores[37]. The Sanger re-sequencing of the genomic DNA confirmed that index is homozygous for the variant, while both parents are carriers.

To further prioritize between these two genes, we consider a FLAGS-based approach. The *VPS13B* gene is one of the FLAGS (top 100, rank 67) and is frequently seen to be affected by rare, likely functional variants in general population. On the other hand, SENP1 is rarely affected by functional variants in the general population (rank 11,947). In addition, *VPS13B* is a frequently seen in the TIDE cohort of patients, 22 of 160 individuals have rare, likely functional alleles in the *VPS13B* gene that pass our prioritization filters. In contrast, the family reported here is the only family from the TIDEX cohort of patients with a rare, likely functional variant affecting the *SENP1*. In none of the other 160 exomes did the variants in *SENP1* pass our prioritization filters for rare, likely functional variants. Together with the fact that *VPS13B* does not fit well to her severe hematologic findings and bone marrow dysplasia, FLAGS helped us select *SENP1* as candidate gene for our experimental validation studies. The case report will be published separately. We further applied prioritization of FLAGS on an in-house WES/WGS database and illustrated how trio-based exome families have Mendelian recessive and dominant candidates overlapping with the FLAGS. The FLAGS ranking can be fed into the candidate identification process and highlight genes that should be considered as high-risk candidates for false positives.

4.5 Discussion

WES/WGS studies can identify hundreds to thousands of rare protein-coding mutations per individual. Genes vary in their frequency of appearance; genes that are more likely to harbor rare-coding variants by chance are less likely to be involved in human diseases, especially in the context of rare Mendelian disorders. Previous studies have reported that *TTN* and *MUC16*, the two longest genes in the human genome, should be interpreted with care due to their long lengths[185, 186]. Similar observations have been made in oncology, and methods have been developed for interpreting somatic variants across a population of tumours at the gene level, correcting for confounding co-variants that would lead to higher or lower background mutation rates[195, 196]. While the underlying principle is similar to our study, their utilities are limited for rare Mendelian disorders where it is not possible to obtain sufficiently large cohorts of germline mutations from individuals with the same disorder. In this study, we compiled a list of frequently mutated genes (FLAGS) based upon analysis of rare coding mutations from dbSNP and Exome Variant Server ESP6500. We compared the biological properties of FLAGS against genes from disease databases (HGMD, OMIM) that represent the currently best reliable curated resources for disease-associated genes. We further demonstrated the clinical utilities of FLAGS as a gene prioritization tool. The discussion will illustrate additional clinical benefits of FLAGS, and conclude with ideas for future directions and project limitations.

4.5.1 FLAGS are less likely to be disease-associated

Consistent with our expectations, FLAGS have significantly longer coding lengths, higher average dN/dS ratios, and more paralogs than genes from OMIM and HGMD. Paralogs have been cited as capable to partially compensate for the loss of gene function[189], so the greater frequency of paralogs could mean that mutations are less likely to have a critical impact on phenotype. In the examination of the research literature for FLAGS, we observed fewer disease annotations compared to disease genes, but elevated rates compared to background genes, suggesting that FLAGS have been associated to human disease more frequently than the rest of the protein-coding genes.

4.5.2 Clinical utilization of FLAGS for prioritization

Prioritizing candidates in rare disease studies is important; as it takes substantial time of experts to review each gene[197], getting better specificity without loss of sensitivity has real value. We demonstrated the utility of FLAGS as a prioritization tool by overlapping FLAGS against candidates from clinical exomes in TIDE, without loss of ultimately identified causal genes. We further illustrated with a single clinical case how when multiple equally attractive candidates are under consideration, FLAGS provide a way for clinicians and researchers to decide which gene to focus on first.

4.5.3 Cautionary indicator

While we are not claiming every gene in FLAGS is non-pathogenic, we do wish to make it clear that greater biological evidence is required when interpreting the functional impacts of rare variants in frequently mutated genes. Among the 300 genes with putative pathogenic mutations identified via exome sequencing compiled by Boycott *et al.* (2013)[180], ten genes intersected with FLAGS. We evaluated the gene-level and variant-level evidence for causality based upon the guideline for investigation of causality published by MacArthur *et al.* (2014)[174]. We found that many results are derived based upon single-gene sequencing, rather than taking the less biased exome or whole-genome approach[198, 199]. In addition, many studies reported the mutations as pathogenic simply due to segregation pattern within the family, rare allelic frequency and bioinformatics impact predictions[200, 201], thus lacking experimental validation at both the variant and gene levels. The screen for rare alleles is further complicated when some of the studies look at minor ethnic populations that are not well represented in the population databases[202, 203]. The evidence behind missense variants is especially doubtful when many missense variants are predicted by CADD[37] to be benign with a lower impact rank than the rare mutations observed from dbSNP and ESP6500. Altogether, these observations could explain why these genes harbor frequent rare functional variations despite being reported in diseases. To avoid false-positive reports of causality, especially for FLAGS, it will be very important for reports to follow the recently published guidelines[204] when assigning pathogenicity to new variants identified as well as additional variants identified in genes previously linked to a particular disease. An example of a good paper would be the one where the variant is identified in a genome-wide screening approach with statistical methods applied to compare the distribution of variants in patients against a large matched control cohorts, where the evidence is assessed at both the candidate gene and candidate variant levels, and where the authors recognize the importance of combining both computational comparative approaches and experimental assays for validating the impact of the variant.

4.5.4 Going beyond the top 100 and what the future entails

Genes with frequent rare variants need to be appropriately ranked in order to reduce false associations and streamline clinical analysis. Our current results are limited to the top 100 frequently mutated genes. While it may be insightful to study the characteristics of the genes at the other end of the spectrum (the bottom 100 or alternatively sets of genes with low mutation rates and gene-focused publications to exclude genes with poor coverage in exome capture kits), we perceive the greatest long-term utility to be in the incorporation of the complete set of rankings into the exome interpretation process. To make our prioritization ranking accessible to the broad research community, we provide the FLAGS ranking for the genes represented in both dbSNP and EVS.

The novelty that we bring forth is a ranking that utilizes public control exomes/genomes, which clinicians can readily apply to their clinical cases. As discussed above, the ranking is correlated with gene length, evolutionary constraint, and paralogous gene counts.

The high accumulation rate of mutations can be interpreted partially as genes being under less selective constraint. A utility of the FLAGS ranking is that it provides, albeit indirectly, a gene-level indication of the selective constraint upon a gene, while most existing metrics such as phastCons[205] or PhyloP[206] provide a position-specific value. While the FLAGS ranking is not a substitute for the more direct measures, the genic level information complements them.

Current prioritization tools lack the ability to evaluate at both genic and variant level simultaneously. Ultimately, a scoring mechanism integrating biological and technological features at both the genic and variant level should be developed. A future direction is to improve upon methodologies like RVIS[178] and expand beyond the rate of mutation by employing statistical machine learning techniques to incorporate the genic and allelic features as highlighted in this study and previous works to summarize them into a single computational score. Such a new quantitative measurement should improve the ranking of pathogenicity for each gene, and highlight skeptical candidates to accelerate the clinical translation of genomic research findings. The mechanism itself (e.g. the weights of features) would also shed light on the exact nature of the causes of excess mutation rates and facilitate better biological understanding.

In the long-term, the accumulation of more exomes and whole genomes will provide an increasingly rich body of data for the generation of FLAGS rankings.

4.5.5 Limitations

In the study we relied upon manually-curated GeneRIFs to extract the publications for each gene. One could argue for more sophisticated PubMed queries in combination with semantic rules to increase the sensitivity for assigning human-disease related publications[207, 208]. We also recognize that neither MeSHOP nor HPO capture gene-to-disease terms perfectly. A possible direction is to explore other gene-disease databases such as HuGE Navigator[209]. We further acknowledge that the interpretation of MesHOP and HPO could be influenced by pleiotropic genes. Similarly, we used Ensembl for extracting the paralogous relationships for each gene, but there are other available extraction algorithms and databases for inferring paralogy[210-212]. Additionally, our present study is restricted to genes with both an HGNC symbol and a fully annotated translation start and end. We recognize that not all protein-coding genes fit these criteria, and we are excluding non-coding genes (as well as 5′ and 3′ UTRs of coding genes) from this analysis.

4.6 Conclusion

While most complex disorders generally can confirm the strength of their findings by comparing against a matched background cohort, the nature of studying rare monogenic disorders mean that there is often insufficient sample size to conduct a rigorous statistical analysis on the strength of the finding. In this study, we extracted a list of frequently mutated genes based on rare variants from dbSNP and Exome Variant Server. Our results revealed the biological properties of these genes that could explain why they are frequently mutated, and why extra discretion in statistical and biological interpretation needs to be taken when trying to relate these genes to clinical phenotypes. We propose that the ranking of how frequent a gene is mutated in next-generation sequencing studies is useful for the prioritization of candidate genes.

Chapter 5: An ensemble approach integrating variant and gene information and patient phenotype for prioritizing variants in exomes

5.1 Synopsis

While applications of exomes in clinical research for rare diseases have been broadly successful, clinical geneticists are frequently challenged with the identification of pathogenic mutations from amongst ~10^5 observed variations. Using now established strategies for filtering rare protein-coding mutations fitting specific Mendelian inheritance models, it is still common to observe ~20 to 100+ candidate variants requiring laborious manual review. We present a novel method called Variant Prioritization Accelerator (VPA), which utilizes an ensemble machine learning approach trained on variant-level, gene-level and patient-level information for classifying rare variants according to likelihood of pathogenicity. VPA prioritizes more diverse variant types than other methods, including splice sites and insertion/deletions. Additionally, VPA permits clinicians to describe patient phenotypes in either Human Phenotype Ontology (HPO) vocabulary, or in free-text format without confinement to strict standard vocabularies. Furthermore, VPA allows clinicians to rank each patient symptom to distinguish primary phenotypes from secondary observations. Finally, we compared VPA against published methods on simulated data and clinical cases, and demonstrated how the integration of disparate biological domains and patient phenotypic features resulted in better performance to detect novel gene-disease, disease-phenotype associations, and polygenic traits.

5.2 Introduction

A single exome sequence can reveal over 350,000 variations[5]. Despite the common filtering strategies discussed in earlier chapters, the nature of rare diseases means clinicians typically work with single affected individuals (e.g. singletons) or small nuclear families, and it is still common to emerge from the automated procedures with 20-100+ variants remaining[213]. This creates a bottleneck in the analytical workflow, where a team of geneticists, biochemists and bioinformaticians may be engaged to laboriously review and derive a small list of prioritized candidates (e.g. one to three) based upon prior knowledge from literature[214].

Recently, a series of tools have been developed that incorporate patient phenotype to assist in the identification of disease-causing mutations in exomes. Below we briefly outline the methods of exemplar systems. eXtasy[215] prioritizes non-synonymous mutations by integrating variant deleteriousness scores from existing databases, gene haploinsufficiency scores, and disease genes related to Human Phenotype Ontology[216] (HPO) terms via Phenomizer[217]. Exomiser[218] allows prioritization of all coding mutations by similarly producing a variant deleteriousness score like eXtasy, and considers additional features including allelic frequency, and phenotypic similarity between patient and known diseases and animal models. Phevor[219] allows prioritization of all coding mutations by integrating knowledge from HPO, Mammalian Phenotype Ontology[220], Gene Ontology[158] and Disease Ontology[221].

Amid the plethora of such tools, there remain unresolved challenges, some of which we highlight below. Firstly, existing phenotype-driven tools typically do not score variants outside

[5] This number is dependent on the type of sequencing technology and the type of capture kit employed. Future exome capture kits that offer greater breadth of genome coverage and longer read lengths that allow mapping to previously unmappable regions would likely reveal more variants.

of protein-coding regions, such as splice site variants, and short insertions or deletions (InDels). Since these classes of mutations can be present in exome results[222], software unable to prioritize these types may be excluding important candidates[223, 224], and thus not fully realizing the clinical potential of the technology. The importance of annotating non-protein-coding variants must be addressed, as whole-genome sequencing is gradually replacing exomes[225]. Secondly, existing prediction algorithms for coding variants tend to consider evolutionary conservation as a key measure, but a rapidly growing body of data has emerged providing functional annotation (e.g. data provided by the Encyclopedia of DNA Elements (ENCODE[226]) and FANTOM[227] consortiums) for such features as alternative start sites and alternative splice sites. Similarly, previous studies[178, 228] have shown the utility of accounting for genes that more frequently harbor rare benign protein-coding mutations, but such information is not explicitly incorporated into current variant prioritization methods. Fourthly, all the previously cited tools are constrained to a fixed ontology (largely HPO) as phenotypic inputs. In clinical practice, this is not ideal because it contrasts against how clinicians describe their patients, which is largely in free-text format (often derived by oral dictation)[229, 230]. An argument can be made that physician procedures should change, but in the near term it is our opinion that it is more viable to adapt the software. Finally, the existing methods consider all the patient phenotypes as equivalent. This is not reflective of actual clinical scenarios, where often some described traits are considered to be defining phenotypes while others are secondary[231, 232]. Accounting for the varying importance of each phenotype has potential to improve the diagnostic process, as recently demonstrated in the context of patient phenotype matchmaking software[233].

In this chapter, I present a novel method called Variant Prioritization Accelerator (VPA) which utilizes an ensemble machine learning approach trained on variant-level, gene-level and patient-level information (see Methodology section 2 for more details) for prioritizing rare variants in exomes. To assist the clinical interpretation for non-computational geneticists, VPA outputs the classification results according to clinical terminologies set by American College of Medical Genetics[234]: “Pathogenic”, “Likely Pathogenic”, “Variant of Unknown Significance (VUS)”, and “Benign”. Our method allows the prioritization of diverse classes of rare variants detected in exomes, including splice sites and InDels. Our method permits clinicians to input patient symptoms in either HPO terms or as free text. It allows clinicians to rank the importance of each input phenotype to separate distinguishing phenotypic features from secondary observations. We compared VPA against existing algorithms on simulated exomes and clinical cases, and demonstrated its capacity to better predict novel disease associations.

5.3 Methodology

This section includes: 1) a description of the datasets used for training and testing, 2) an explanation of the types of features considered and the methods used to incorporate them into the model, and 3) the performance evaluation procedure.

5.3.1 Datasets

5.3.1.1 Training set

ClinVar[106] VCF (version 20140502) for human reference hg19 was downloaded from the ClinVar FTP server. To focus on rare diseases, only variants of germline origin, based on the “Origin” annotation, were considered in the downstream analysis. SnpEff[33](version 4.0) was used to annotate each variant using GRCh37.75 as reference, with a splice site window of 7. Only variants marked by SnpEff as ‘HIGH’ or ‘MODERATE’ in the impact field were kept. A custom Perl script was written to categorize variants into ‘pathogenic’, ‘likely pathogenic’, ‘variant of unknown significance (VUS)’, and ‘benign’ based on ClinVar’s CLNSIG annotations (Appendix D-1). The final training set included 2343 pathogenic mutations (in 1501 genes), 1044 likely pathogenic mutations (in 877 genes), 5240 VUS (2399 genes), and 2203 benign variations (in 983 genes).

5.3.1.2 Simulated test set

Mutations annotated as “DM” or “DM?” in HGMD[160] Professional 2015 version 1 were classified as pathogenic mutations. SnpEff was used to annotate the variants on GRCh37.75, and only coding mutations labeled as ‘HIGH’ or ‘MODERATE’ impact were kept. Mutations that overlapped with ClinVar training set were discarded, leaving 589 variants, corresponding to 374 genes. These mutations were embedded in simulated exomes comprising of data from 1000 Genome datasets[16] and NHLBI ESP6500 (http://evs.gs.washington.edu/EVS/). To simulate technical noise present in high-throughput sequencing, and in order to build simulations that were comparable to real exomes under singleton and trio structures, we randomly inserted rare coding mutations (missense/nonsense, <0.01% allelic frequency) to the simulated dataset until the number of rare coding variants in each simulation was comparable to actual clinical exomes (Appendix D-15). The randomly inserted mutations were drawn from an independent pool of 583 Ion Torrent^{TM} exomes provided by Dr. Carles Vilarino-Guell (University of British Columbia). A total of 750 simulated cases were generated for singletons and trios respectively, and broken down as follows: 300 out of the 750 cases were embedded with missense pathogenic mutations (equally divided to represent homozygous recessive model, compound heterozygous model, and *de novo* heterozygous model respectively), 150 cases were embedded with nonsense pathogenic mutations (equally divided to represent the three Mendelian models); the remaining 300 cases were embedded with randomly drawn pathogenic mutations from the entire pool without discrimination between missense versus nonsense classification (100 for homozygous recessive model, 100 for compound heterozygous model, and 100 for *de novo* model).

5.3.1.3 Clinical test set

The patient cohorts were drawn primarily from the Omics2TreatID project [Tarailo-Graovac *et al.*, manuscript accepted] as part of the TIDE-BC initiative (http://www.tidebc.org), and also included clinical cases from TIDEX, Care4Rare and FORGE Canada. Omics2TreatID focuses on patients with intellectual disability plus rare unexplained metabolic phenotypes. The study was approved by the Ethics Board of the Faculty of Medicine of the University of British Columbia (UBC IRB approval H12-00067); each family provided informed consent for publication of results. We selected families for which prior analyses of exomes led to genetic diagnosis, as identified by a team of overseeing clinical genetics, molecular biochemists and genetic counselors. 53 families were included in this test set (Appendix D-16). All the pathogenic[6] variations from these families had been Sanger validated in index and available family members, and shown to be consistent with a determined mode of inheritance (including

[6] At the time of writing, the pathogenicity of these variants, if classified according to latest ACMG criteria (accessed July 2015), range from VUS, Likely Pathogenic, and Pathogenic, each with varying level of confidence. We addressed this issue in the Discussion section.

de novo mutations). In the majority of the cases the families had been informed of the diagnosis but the results were not yet published at the time of the analysis.

5.3.2 Features

In this section, we introduce the three hierarchical sets of features and describe the feature selection and model training process. Table 1 lists the features included in VPA after feature selection (Appendix D-2). Appendix D-17 provides the complete list of features considered.

Type	Granularity	Features	Source
V	1	Frequency of seeing any mutation(s) at the given genomic position	In-house database of 370+ exomes and whole-genomes
V	1	Frequency of seeing mutations within 15bp window	In-house database of 370+ exomes and whole-genomes
V	1	Number of overall homozygotes in ExAC	ExAC: http://exac.broadinstitute.org
V	1	Afr Freq	ExAC: http://exac.broadinstitute.org
V	2	Condel score	CONDEL[235]
V	2	FATHMM rankscore	dbNSFP[236]
V	2	Reliability index	dbNSFP[236]
V	2	VEST3 score	dbNSFP[236]
V	2	CADD raw	CADD[37]
G	3	dN/dS	FLAGS[228]
G	3	Gene length	Ensembl[237]
G	3	Paralogs	FLAGS[228]
G	3	Counts in literature	FLAGS[228]
G	3	Mutation counts	FLAGS[228]
G	4	RVIS	RVIS[178]
P	3	Free-text PubMed	Presented in paper
P	3	Ontology-search HPO	Presented in paper

Table 5-1 The list of features selected in the final model after hierarchical sampling plus hierarchical feature selection. In the first column, V=variant-level, G=gene-level, and P=patient-level. The second column corresponds to the level of hierarchy that the feature belongs to for constructing hierarchical tree during hierarchical feature

selection (see Methods). The third column contains the descriptive information about the feature, and the last column contains the source that was used to derive the feature.

5.3.2.1 Variant-level features

A feature was considered as “variant-level” if it refers to a specific DNA location. Broadly speaking, these features corresponded to: 1) allelic frequency in public genomic databases, 2) variant functional prediction score, and 3) evolutionary conservation score. Whenever possible, this information was extracted from dbNSFP[236] (version at the time of writing is 3.0b2c), but if a newer annotation was available in one of the source data collections of dbNFSP, then custom scripts and local MySQL databases were used to extract the information from the source.

5.3.2.2 Gene-level features

A feature was considered as “gene-level” if the information pertains to a gene rather than a specific position. These features corresponded to two main categories: 1) a report on a biological characteristic of a gene (e.g. gene length), or 2) a computational score computed for that gene (e.g. haploinsufficiency score). As for variant-level features, data were drawn from dbNSFP, with updated information from direct source databases.

5.3.2.3 Patient-level features

A feature was considered as “patient-level” if it describes the clinically reported phenotype(s) exhibited by the patient. Two features were considered in this category: an ontology-based matching score, and a free-text-based matching score. Both scoring systems take in patient phenotypic terms as input and provide a similarity ranking of all protein coding genes (Appendix D-3).

5.3.2.3.1 Phenotype match by ontology

We adapted the GeneYenta[233] phenotype-matching algorithm to data mine patient phenotypes. In brief, GeneYenta is an online patient-matching system that allows clinicians to specify phenotypes of undiagnosed patients to match against patient cases stored in the database. In this study, rather than comparing patient phenotypes, we compared the index patient phenotypes against phenotype terms associated with genes. The probability of a gene given index phenotype was approximated as:

$$P(gene|phenotype) \approx \sum_{diseases} P(gene|disease) \times P(disease|phenotype)$$

The first probability was obtained from DisGenet score in file "ALL gene-disease association" provided by DisGenet[238]. The score was extracted and normalized for each gene. The second probability was based on an adapted algorithm from GeneYenta using HPO annotations. The equation is as follow:

$$match(pat, dis) = \frac{\sum_{t \in T_{pat}} R_t \times max_{t' \in T_{dis}} sim(t, t')}{\sum_{t \in T_{pat}} R_t \times I_t} \times 100$$

where pat = patient, dis = disease in the HPO database, and I = information content. The information content score is based on a negative log function of the number of descendants in the HPO tree for a given HPO node plus 1 divided by the total number of unique HPO terms. The Sim(t,t') is the similarity score between the two phenotype terms, derived from information content, and R_t is an importance ranking (Appendix D-4).

5.3.2.3.2 Phenotype match by free-text

We used MedlineRanker[239] to derive a list of ranked publications based on phenotypic terms in free-text as input. This tool was chosen for three reasons: 1) to provide clinical free-text integration into the model as a proof-of-concept, 2) its API capacity for command-line incorporation, and 3) its convenient provision of a p-value for each ranked entry in the output list. Custom Perl script was used to connect to MedlineRanker's SOAP API (version beta2). Each phenotypic descriptor was executed on MedlineRanker individually as a PubMed query, with the background set equal to be the entire MEDLINE database. The gene(s) associated to each article was/were derived from the MeSH (Medical Subject Headings) vocabulary attached to the article. If no known gene symbol were present in the MeSH list, the gene(s) was/were derived from the article abstract. The final probability score used in this feature is as follows:

$$P(gene|phenotypes) \approx \max R_t \times P(literature|phenotype\ t)$$

where R_t is the weight assigned by user for the phenotype term, and P(literature|phenotype) is 1 minus the p-value derived from MedlineRanker. By default, R_t is assigned to value of 3 unless otherwise specified by the user. A similar approach was designed using PubMed API query to evaluate performance fluctuations when using a different gene-extraction tool (Appendix D-5).

5.3.2.4 Feature selection and model training

The variant-level and gene-level features reflect properties at different levels of resolution. Furthermore, some features were composite scores that were informed by other features in the collection (i.e. dependent upon them). Therefore we selected a procedure that could account for these data properties. We took the pseudocode framework described in Moor *et al.* [240] and developed a combination of hierarchical sampling and hierarchical feature selection using random forest (Appendix D-2). A total of 1000 trees were used to train the random forest based upon multiple training sets derived from hierarchical sampling to assure that no gene-level information would be over-represented in a single tree, and each tree was built based on a hierarchical structure using the hie-ran-forest package in R 3.2.0 to assure the compositions of scores were accounted for. A 10-fold cross-validation on the training set was applied to produce the final model.

5.3.3 Performance evaluation

The input to each evaluation was a VCF (Variant Call Format, version 4.1) file already annotated with SnpEff (GRCh37.75) and restricted to rare (≤ 1% in dbSNPv142) protein-coding mutations (see [228] for a more in-depth description on the bioinformatics pipeline). We compared VPA against Exomiser version 6.0.0, eXtasy0.1, and CADD version 1.2. Based on benchmarking studies[218], Exomiser and eXtasy were selected for comparison as they had been reported as the top command-line accessible tools that incorporate patient phenotype information for gene variant prioritization. CADD scores (‘All possible SNVs’) were downloaded and stored locally. Appendix D-6 describes how we extracted the predicted rank for the pathogenic mutations from each model.

All ranking assessments were performed with the union of candidate variants from all genetic models for each subject. To quantify the performance, we defined a prediction as successful if the causal embedded mutation was among the top 3 predicted candidates per simulated/clinical case. While arbitrary, the focus on 3 was based on feedback from our clinical partners that they have limited time and would be unlikely to review more than 3 candidates in depth[114].

5.3.4 Web access

A web access version of the final model is in progress. A downloadable version is in development and is planned for a future release.

5.4 Results

In section 1, we first describe VPA's performance on simulated test sets and compared it against existing algorithms. In section 2, we extend the comparison to clinical cases. In section 3, we draw results from simulations and clinical examples to demonstrate the advantages of VPA over existing approaches.

5.4.1 Evaluation on simulated test sets

5.4.1.1 Characteristics of test sets and model classifications

VPA classified each variant as either "Pathogenic", "Likely pathogenic", "Variant of unknown significance (VUS)", or "Benign". Figure 5-1 shows the average proportion of variations under each category on 750 simulated cases for trios and singletons respectively. On average, 6 mutations were predicted to be pathogenic from the trios, and 9 mutations were predicted to be pathogenic from the singletons. The "Likely pathogenic" class consistently had more variations than "Pathogenic" class. VUS made up the largest category in most of the cases, closely followed by "Benign" class. VPA was able to assign the embedded pathogenic variant as "Pathogenic" in 93% and 85% of the simulated exomes for trios and singletons respectively.

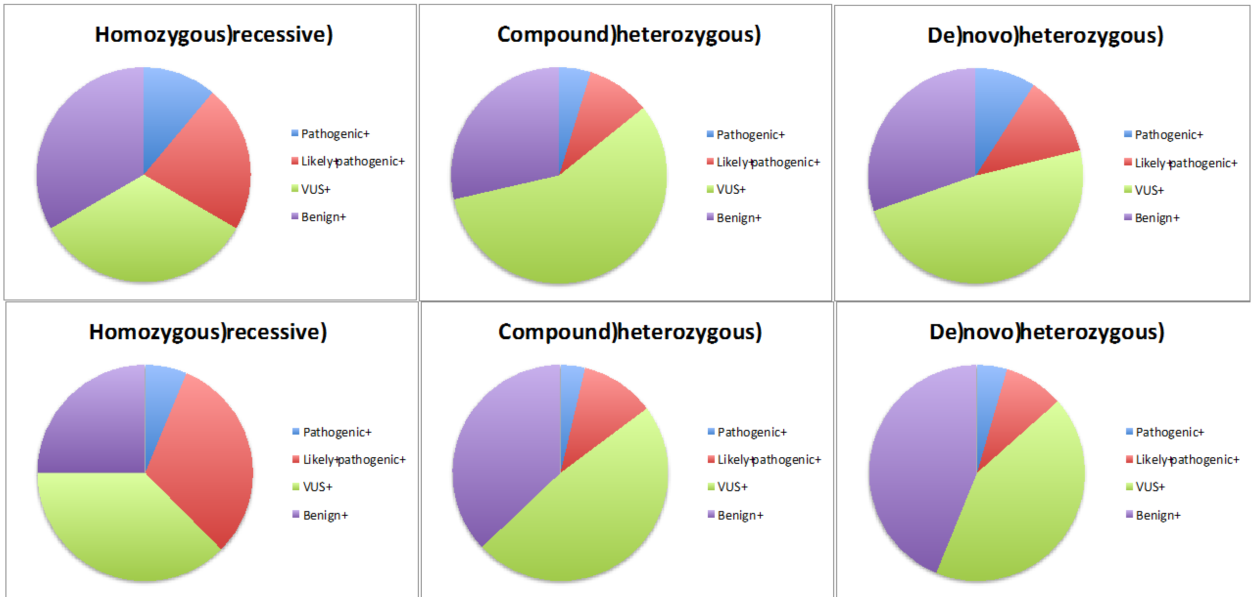


Figure 5-1 A summary of the classifications by VPA on simulated datasets, separated by family structures and genetic models. The top half shows the performance for trios, and the bottom half shows the performance for singletons. Each pie chart shows the proportion of each variant class with respect to the size of the starting input VCF (to see how many variants were present at the start of each simulated case, refer to Appendix D-15). Three genetic models were considered: homozygous recessive, compound heterozygous and *de novo* heterozygous. Throughout this paper, the category "Homozygous recessive" includes both homozygous recessive and hemizygous recessive models. In this study, we did not distinguish between autosomal mutations and mutations on the sex chromosomes, hence X-linked or other sub-types of Mendelian inheritance patterns were not explicitly discussed but instead incorporated as part of the broader types of Mendelian inheritance.

5.4.1.2 Incorporating non-variant level features improves predictive performance

We assessed VPA in comparison to CADD scores using simulated cases, to determine if the inclusion of additional features could boost the performance over CADD alone, and which features were most informative. The performance was evaluated either separately for missense variants and nonsense variants or collectively across distinct Mendelian models, and in different family structures. As expected (since VPA incorporates CADD scores), VPA outperformed

CADD across all considered scenarios, achieving an average successful prediction in 83% for trios and 71% for singletons, versus 42% for trios and 25% for singletons for CADD (Appendix D-14). The performance difference between VPA and CADD was more prominent for singletons than for trios. Both VPA and CADD achieved highest performance for cases reflecting homozygous recessive events, followed by compound heterozygous, and lowest for *de novo* events. Embedded causal nonsense mutations were ranked higher than cases involving missense mutations. To determine the most important features and their relative value, we retrained VPA without CADD scores and re-evaluated performance (Appendix D-7). In the absence of CADD scores, VPA dropped in overall performance by 19% in trios and 21% in singletons when considering missense and nonsense mutations, still higher than CADD's performance.

5.4.1.3 Comparisons against phenotype-based prioritization algorithms

We compared VPA against Exomiser and eXtasy. eXtasy was restricted to evaluation of non-synonymous mutations due to its inability to process nonsense mutations. All phenotype-informed algorithms outperformed CADD (Figure 5-2). Exomiser and eXtasy exhibited performance patterns similar to previous observations with respect to the drop in percentage of successful predictions between trios versus singletons, and decreased predictive power in detecting pathogenic variants under the dominant model versus recessive Mendelian models. VPA and Exomiser achieved comparable performance (83% versus 82% in trios, and 71% versus 69% in singletons); VPA performed marginally better on missense mutations (85% versus 81%), and both models achieved near-identical performance when applied to nonsense mutations (98% versus 97%). VPA and Exomiser outperformed eXtasy, which had an overall missense performance of 67% in trios. To evaluate if Exomiser and eXtasy scores were providing additional information not captured in VPA, we re-trained our model with Exomiser and eXtasy scores as additional features and found no substantial improvement (Appendix D-8).

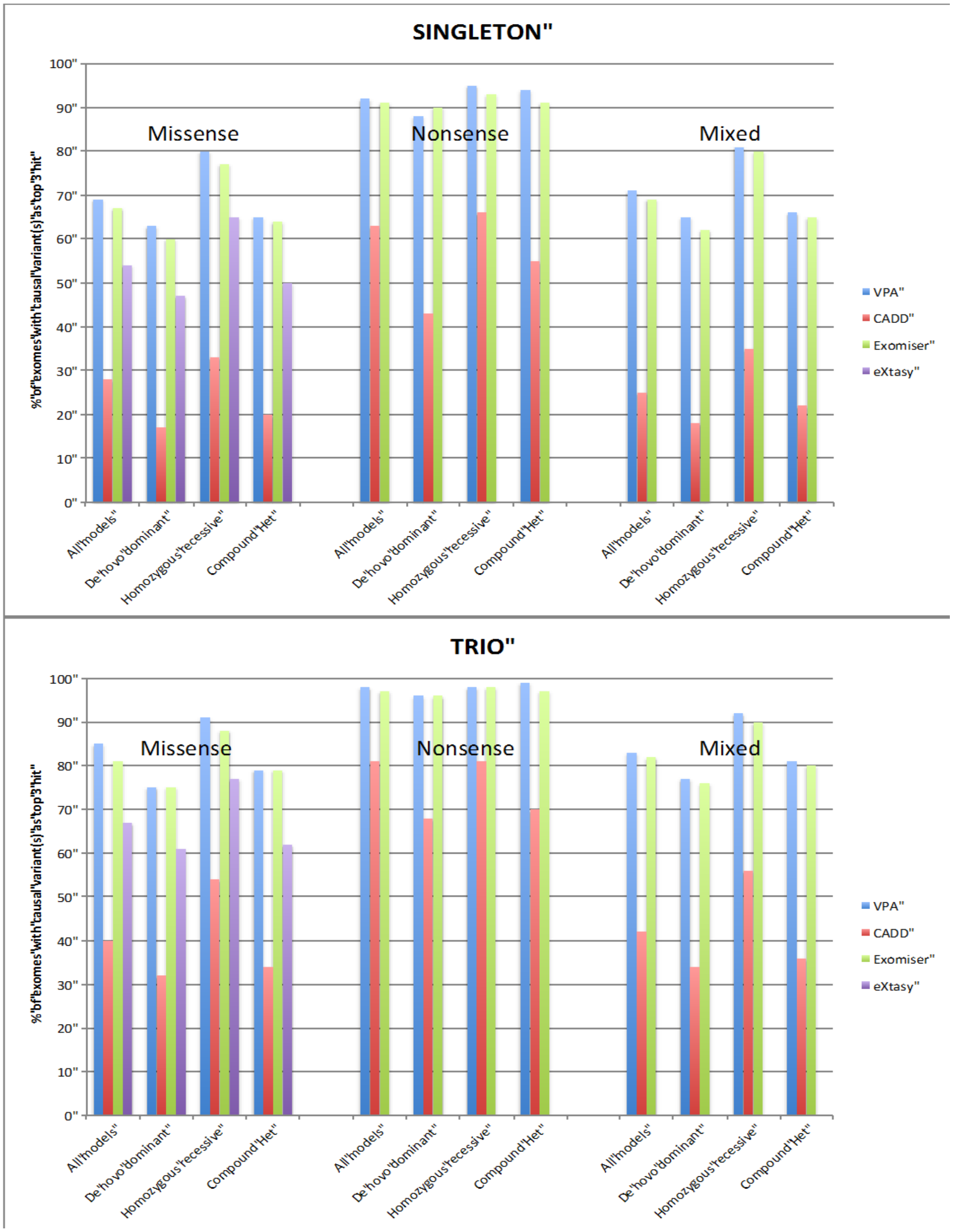


Figure 5-2 Performance on simulated exomes for VPA, CADD, Exomiser and eXtasy. The upper plot shows the performance on singleton cases, and the lower plot corresponds to the trio cases. Performance was separated into the

type of pathogenic mutations that were embedded (missense versus nonsense, or mixed), and the types of genetic models that corresponded to the pathogenic mutations. Y-axis corresponds to the % of simulations with a successful prediction, defined as being able to rank the pathogenic mutation(s) as among the top 3 candidates. The "All models" bars represent the average across the inheritance models in their respective mutation type.

5.4.2 Evaluation on clinical cases

The simulated cases were all derived from previously reported pathogenic mutations in clinical literature. To gather perspectives on novel mutations in patient cases whose phenotypes were not yet compiled in clinical and disease databases, we evaluated VPA and previous algorithms on a cohort of clinical exomes (n=53 families). Using the same performance criteria as defined earlier, VPA achieved successful prediction in 35 families (66%), compared to 29 (55%) for Exomiser, 18 (36%) for eXtasy, and 13 (25%) for CADD. eXtasy could not be evaluated on 3 families because the pathogenic mutations were nonsense changes. The observed performance agreed with preceding simulation-based results, where all phenotype-driven models performed the best for homozygous recessive model (VPA achieved 81% success on families with homozygous recessive pathogenic variants), and achieved similar performances between *de novo* heterozygous and compound heterozygous models (VPA had 60% and 62% respectively). Due to the limited types of mutation eXtasy could classify, the remaining comparative analysis were centered on VPA, Exomiser and CADD.

5.4.3 Benefits of incorporating variant-level, gene-level and patient-level information

In this section, we highlight scenarios where VPA demonstrated superior performance over existing methodologies. We present results from clinical families as illustrative evidence, complemented with simulations when applicable.

5.4.3.1 Importance to include mutations beyond non-synonymous and nonsense categories

Existing algorithms that incorporated patient phenotype information were restricted largely to non-synonymous and nonsense mutations. There were families that were not included in the previous section because the pathogenic mutations are located in splice site regions outside of exons, or are short InDels resulting in coding transcript frameshift that could not be scored by Exomiser or eXtasy. Illustrative clinical example of a novel 13bp heterozygous deletion is discussed in Appendix D-9.

5.4.3.2 Unmappable clinical terms

Among the 53 families with genetic diagnoses from exomes, 24 (45%) had one or more phenotypic terms that could not be mapped to Human Phenotype Ontology (HPO). These clinical descriptors were drawn from primary clinicians who referred the patients to the Omics2TreatID project and were assigned prior to this study. We excluded any unmappable clinical descriptors that represented secondary physical traits and only considered the keywords representing primary symptoms. To examine how much information was lost for clinical terms not mapped to HPO, we compared the rank of each pathogenic mutation in these families between VPA and Exomiser. VPA was found to provide a more accurate rank for the pathogenic mutation in 15 out of these 24 (63%) families compared to Exomiser. The two models had equal performance in 3 families, and Exomiser performed better in 6 families. Due to limited number of samples, we evaluated this property in simulated cases by randomly removing half of the HPO terms per input case. Comparable to performances from clinical test set, VPA demonstrated better resistance when faced with unmappable clinical terms (Figure 5-3), with an average performance of 78% and 63% in trios and singletons respectively, versus 72% and 49% for Exomiser.

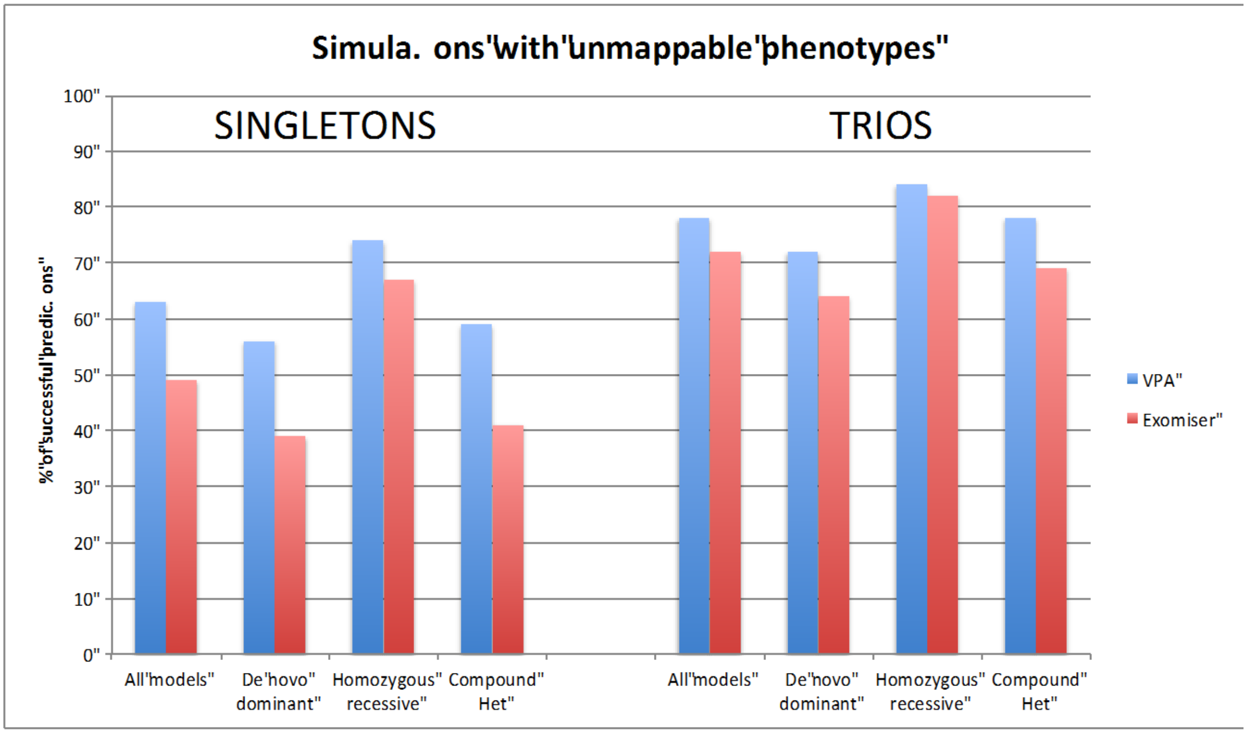


Figure 5-3 Simulations with unmappable phenotypes. To construct simulated cases in which clinical descriptors could not be mapped to HPO but are present in text descriptions, we took the same simulated dataset from earlier analysis and randomly masked supplied HPO terms (with the number of masked terms being 0.5 * the number of supplied HPO terms for each case). The remaining unmasked terms were then supplied to both models for phenotype-based matching via ontology. Y-axis plots the percentage of cases with the embedded pathogenic mutations as among the top 3 predictions per model. The figure shows the performance for singletons (left) and trios (right). A mixed of missense and nonsense pathogenic mutations were considered without discrimination. CADD

performance was not shown because it is not impacted in any way by this experiment. eXtasy was excluded because of its inability to handle nonsense mutations.

5.4.3.3 Novel phenotypes

From the 53 clinical families, we derived a subset of cases (n=16) where the patients harbored pathogenic mutations in previously characterized disease genes but exhibited novel phenotype(s) not yet reported in clinical literature. Out of these 16 families with novel phenotype associations, VPA achieved higher performance than existing algorithms in 8 families based on a comparison of the predicted ranks for pathogenic mutations. Exomiser performed better in one family and drew level with VPA on the remaining seven families. For a more rigorous assessment, we took the same simulated cases from earlier but added random HPO terms to the inputs acting as novel associations (Appendix D-10). While both tools dropped in performance, VPA was more resistant than Exomiser to the added noise (Figure 5-4). In the consideration of missense and nonsense mutations, VPA had a 6% drop versus 16% for Exomiser in trios, and 11% versus 19% in singletons. The higher tolerance reflects a better capacity to connect to the causal gene with diverse clinical observations. In a similar exercise, we demonstrated that VPA was able to predict novel gene associations better than existing algorithms, discussed in Appendix D-11.

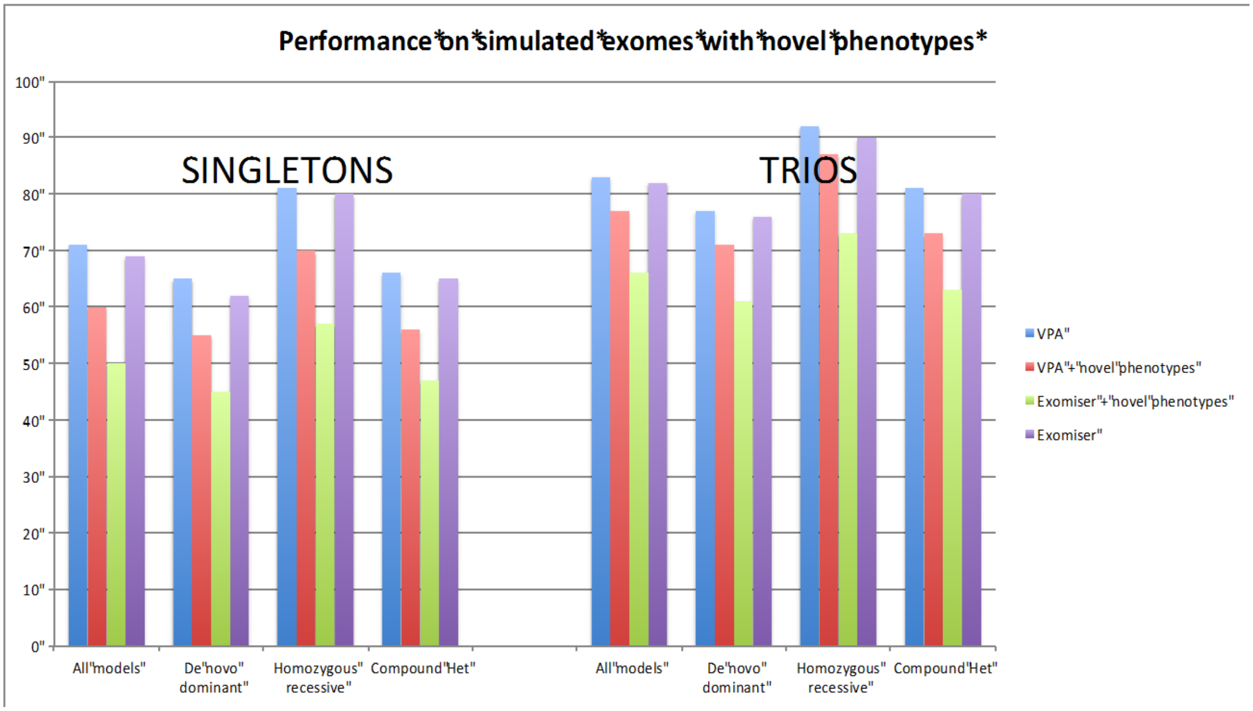


Figure 5-4 Performance on simulated exomes with random HPOs added to represent novel phenotypes. The introduced pathogenic mutations could be either missense or nonsense without discrimination. The left portion corresponds to singleton cases, and the right portion corresponds to trio cases. The blue bars and purple bars represent original performance without novel phenotypes introduced for VPA and Exomiser respectively, and the red bars and green bars represent the performance if novel phenotypes were fed as input. The Y-axis is the percentage of cases with the pathogenic variant(s) predicted to be among the top 3 candidates.

5.4.3.4 Polygenic and oligogenic phenotypes

Previous clinical reports have described polygenic/oligogenic diseases where multiple genes impacted by rare mutations can collectively contribute in the same patient; each affected gene is responsible for triggering a subset of the observed phenotypes. Among the 53 clinical families, 7 families displayed polygenic phenotypes. Among these cases, VPA was able to successfully identify one of the pathogenic mutations as among the top 3 candidates in 4 families, and the second pathogenic mutation was predicted with an average rank of 5 (Appendix

D-19). Exomiser was able to predict the first pathogenic mutation in 3 families, but the second mutation was ranked lower with an average rank of 9.

To complement clinical observations, we constructed 150 simulated cases where we embedded two causal mutations from two different genes (Appendix-12). Performances were evaluated on two-tiers: the ability to identify one of the two embedded pathogenic mutations as among the top 3 candidates, and the ability to identify the second pathogenic mutation as among the top 5 candidates. Across both singleton and trio structures, our model was able to rank these gene candidates higher than Exomiser (Figure 5-5). In trios, VPA achieved an average successful prediction in 73% for the first mutation (versus 69% for Exomiser and 42% for CADD), and 63% for the second mutation (versus 42% for Exomiser and 38% for CADD). The performance distinction between VPA and existing software was more prominent when prioritizing the second mutation. While we observed that CADD was able to rank the two pathogenic mutations in close vicinity, since it was not influenced by the phenotypic diversity, both VPA and Exomiser outperformed CADD.

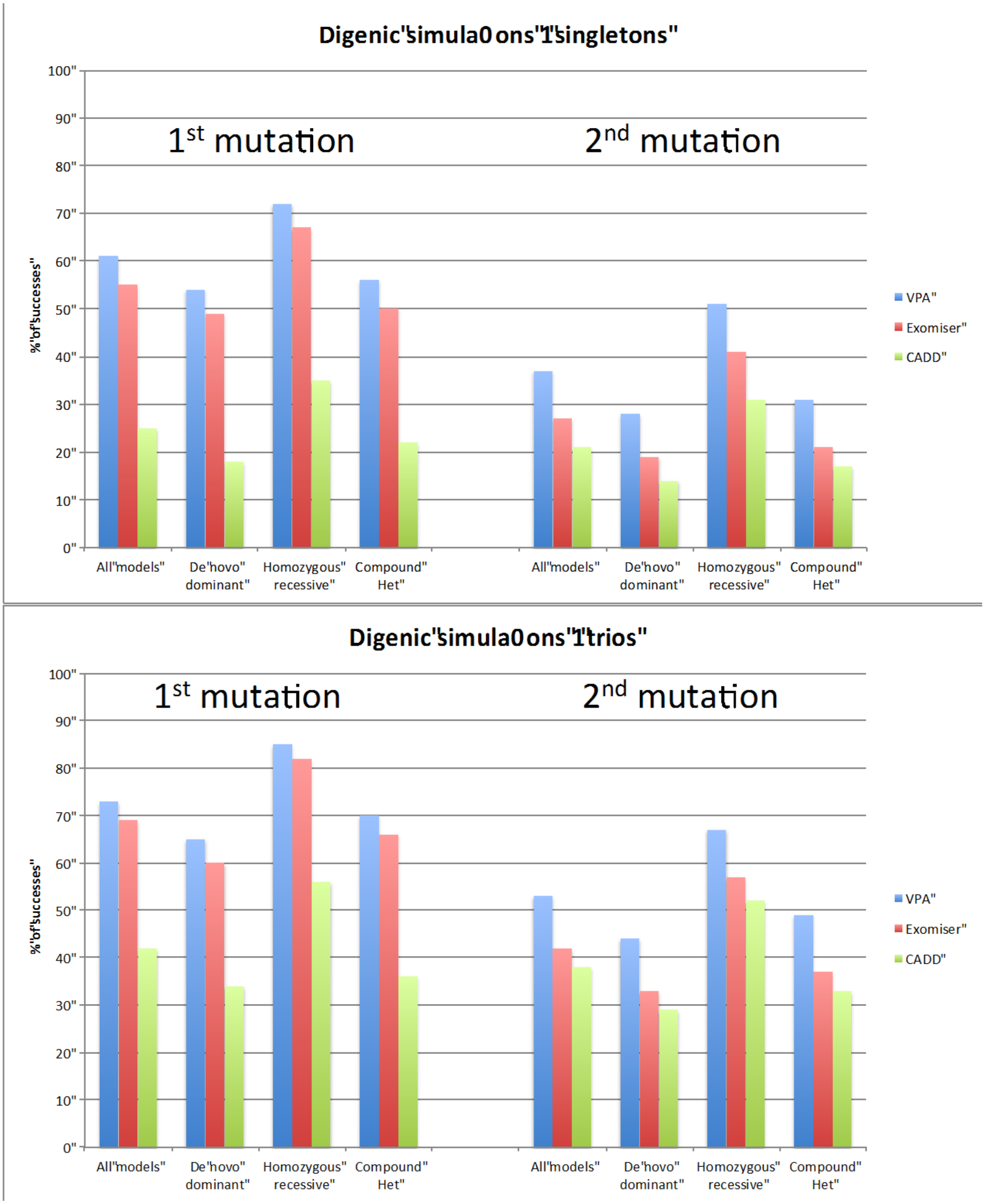


Figure 5-5 Performance on digenic simulations. A mixed of missense and nonsense pathogenic mutations were embedded per case without bias. The plot on the left corresponded to singletons, and the plot on the right corresponded to trios. The performance for the first mutation, shown in the left portion for each plot, was calculated

based upon the % of cases where the evaluated model was able to predict one of the two pathogenic mutations to be among the top 3 candidates. The performance for the second mutation, shown in the right portion for each plot, was calculated based upon the % of cases where the model was able to predict the second pathogenic mutation to be among the top 5 candidates.

5.4.4 Phenotype ranking

The previous performance results were produced under the naïve assumption that all reported phenotypic features for each patient case were equally important and equally clinically distinguishing. VPA allows user the ability to quantitatively rank physical traits from “most important” to “least important”. To assess the performance when such user-supplied ranking scheme is considered, we evaluated on clinical cases where keywords reflecting primary phenotypes were given a score of “5”, and keywords reflecting secondary physical traits were given a score of “1” (Appendix D-13). The assignments of primary versus secondary were provided by the clinicians. Out of 53 families, VPA with weighting-scheme incorporated performed better on 23 families versus VPA with default weights assigned. An overall 72% (n=38) success prediction of picking pathogenic mutation among the top 3 was achieved when phenotypes were clinically ranked, versus the previously reported 66% (n=35) without phenotype ranking.

5.5 Discussion

Clinical exome and whole genome sequencing are becoming preferred methods for clinical genetics, leading to a demand for improved automation for candidate variant prioritization[241]. In this report we introduce VPA, a new ensemble variant prioritization tool based on a hierarchical sampling and feature selection procedure from variant-level, gene-level and patient-level features. The ensemble approach handles a more diverse range of variant types detected in exome data compared to existing methods. Using both simulated and real cases, we demonstrate that VPA outperforms existing tools in most contexts, and at least as well in all. When confronted with noise (unrelated/novel phenotypes and genes), VPA outperforms other patient phenotype-informed methods, and, relatedly, shows greater capacity to detect multiple contributing genes in oligogenic cases. VPA includes a novel clinician phenotype-weighting feature that leads to further performance improvement.

A clinical tool needs to be compatible to existing clinical procedures to improve its usability[242]. VPA assigns clinical labels ("Pathogenic", "Likely Pathogenic", "Variant of Unknown Significance (VUS)", or "Benign") consistent with emerging practice in clinical genetics[243]. Furthermore, VPA's inclusion of clinical free text provides better compatibility to current forms of clinical information, such as provided by dictation. While HPO and other controlled vocabularies are probably preferred in the long run, in our opinion software should reflect actual clinical workflow and not be restricted to the ideal. Due to the high level of noise and complexity in clinical observations, software should be able to handle novel phenotypic associations and limited availability/quality of annotations in disease databases. VPA is able to capture the continuously expanding phenotypic diversities in diseases, and displayed greater resistance to phenotypic noise.

The applied performance evaluation has specific characteristics. Firstly, in addition to sampling from public genomic datasets, we introduced rare coding mutations to our simulated cases that include both true variants and technical noise. Genomic data obtained from projects such as the 1000 Genomes Project have removed most of the noise and therefore do not necessarily reflect realistic test cases. Spiking in rare coding mutations also had the advantage of simulating a “true complete” exome, whereas older exomes cited in prior studies were often derived from outdated capture kits that covered a smaller percentage of the exons, and thereby reported fewer variants in the input sets. Second, we defined a successful prediction as a causal gene-variant pair appearing amongst the top 3 candidates. Previous studies have varied, ranging from top 1 to top 10[215, 218, 219]. Third, we do not restrict to a specific genetic model for inheritance. Published reports sometimes cited performance under the consideration of only particular inheritance model(s) and ignored the variants in other models[244].

There are limitations to this study that might be addressed in future work. Since the ranking of phenotype weights were provided post hoc by the overseeing clinicians, there may have been unintentional bias introduced. Secondly, because result for the detection of novel genes is based upon limited cohort size and is not readily evaluated with simulated cases, additional validation would have to be pursued using a wider range of clinical cohorts. Moreover, we recognize there are more sophisticated literature-mining algorithms that directly extract over-represented genes from a trained list of articles[130, 153, 245, 246], and this could improve performance. Additionally, due to the ongoing experimental validations in many of the clinical cases used as test sets in this study, we recognized the possibility that some of the called pathogenic mutations may in turn be ruled out, depending on the outcome of those validations. Finally, for comparative purposes, in this paper we restricted our analysis primarily on non-synonymous and nonsense mutations. Future exploration could focus on the performance analysis for splice site mutations and InDels, along with the consideration of variants in regulatory regions such as known promoters and enhancers[227]. In anticipation for the replacement of exomes with whole-genomes, we plan to broaden the types of information considered from external non-coding variant scoring systems such as splice site prediction methodologies[247, 248] and regulatory variants prioritization software[249].

In summary, we describe VPA, a novel method that allows the prioritization of rare variants detected in exomes. VPA incorporates variant-level, gene-level and patient-level features to classify variants according to established clinical semantics. VPA allows both clinical free text as well as ontological keywords to describe patient symptoms, and takes advantage of clinicians’ assessment of the relative importance of each phenotype. VPA performs better than existing methods in most cases, and at least as well in all, for detecting novel gene-phenotype associations in prioritization of exome candidates in rare diseases.

Chapter 6: Mitochondrial carbonic anhydrase VA deficiency resulting from *CA5A* alterations

6.1 Synopsis

Four children in three unrelated families (one consanguineous) presented with lethargy, hyperlactatemia, and hyperammonemia of unexplained origin during the neonatal period and early childhood. We identified and validated three different CA5A alterations, including a homozygous missense mutation (c.697T>C) in two siblings, a homozygous splice site mutation (c.555G>A) leading to skipping of exon 4, and a homozygous 4 kb deletion of exon 6. The deleterious nature of the homozygous mutation c.697T>C (p.Ser233Pro) was demonstrated by reduced enzymatic activity and increased temperature sensitivity. Carbonic anhydrase VA (CA-VA) was absent in liver in the child with the homozygous exon 6 deletion. The metabolite profiles in the affected individuals fit CA-VA deficiency, showing evidence of impaired provision of bicarbonate to the four enzymes that participate in key pathways in intermediary metabolism: carbamoylphosphate synthetase 1 (urea cycle), pyruvate carboxylase (anaplerosis, gluconeogenesis), propionyl-CoA carboxylase, and 3-methylcrotonyl-CoA carboxylase (branched chain amino acids catabolism). In the three children who were administered carglumic acid, hyperammonemia resolved. CA-VA deficiency should therefore be added to urea cycle defects, organic acidurias, and pyruvate carboxylase deficiency as a treatable condition in the differential diagnosis of hyperammonemia in the neonate and young child.

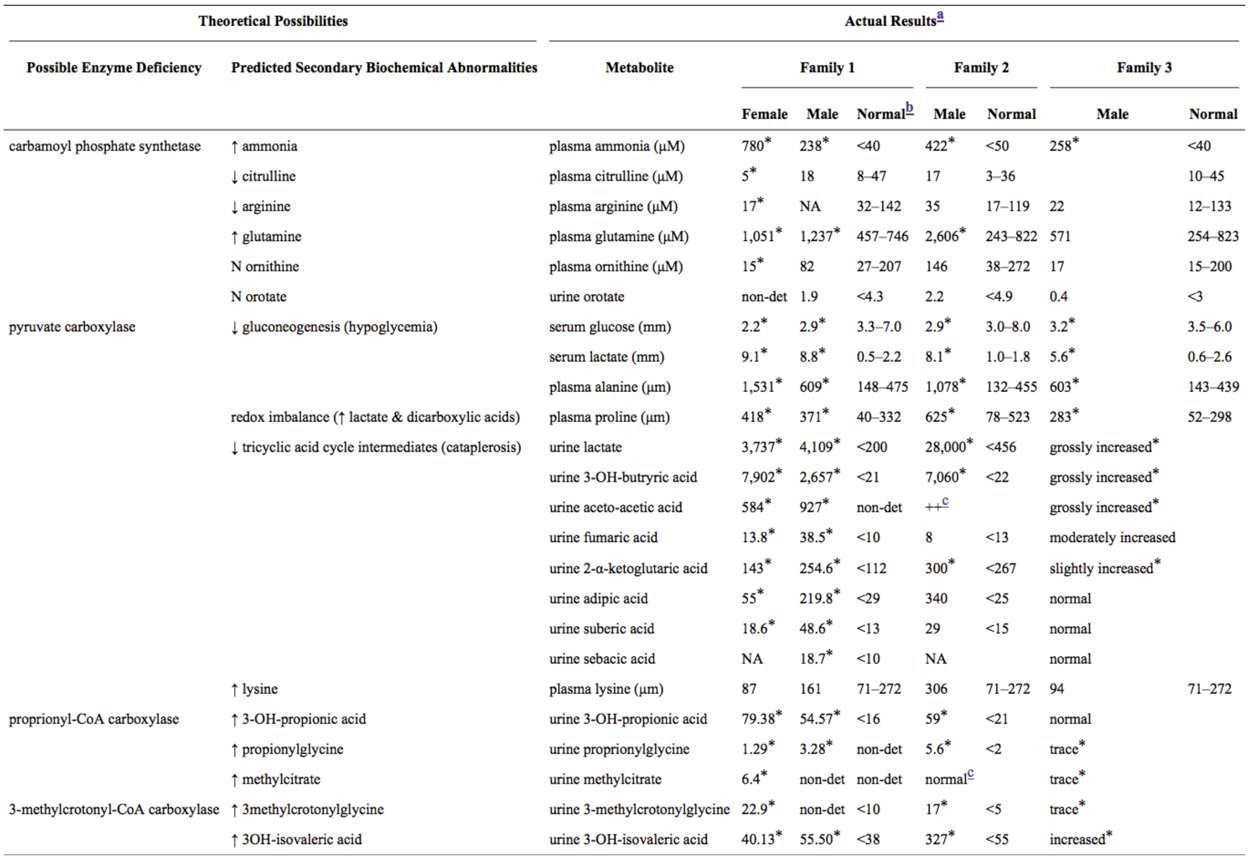
6.2 Introduction

Hyperammonemia is a medical emergency that requires immediate and targeted treatment. Correct diagnosis is therefore essential, but it is challenging given heterogeneous etiologies, including genetic (inborn errors of metabolism), developmental (transient neonatal hyperammonemia), and environmental (infectious hepatitis, medication) causes[250]. Current practice for treating hyperammonemia consists of reducing catabolism and promoting anabolism by a protein-restriction diet and parenteral lipids administration after exclusion of fatty acid oxidation disorder. Ammonia scavenging drugs (sodium benzoate, sodium phenylbutyrate, and arginine hydrochloride) are at present considered the first-line drugs for the treatment of neonatal hyperammonemia[251]. Carglumic acid, a synthetic analog of N-acetylglutamate, is another first-line drug that is able to activate the enzyme of the first and rate-limiting step of the urea cycle. If the foregoing therapies fail to produce any appreciable change in blood ammonia level within a few hours, continuous venovenous hemofiltration is required[251]. We present four children from three unrelated families with infantile hyperammonemic encephalopathy and hyperlactatemia. The underlying cause in each of these children was deficiency of carbonic anhydrase VA (CA-VA) (*CA5A* [MIM 114671]), an inborn error of metabolism broadening the differential diagnosis for hyperammonemia. In this chapter, clinical details of the patients and biochemical interpretations of the validation assays were left out due to not being part of the thesis's focus. Readers interested to read about those detail aspects should go to the published manuscript[252].

This study was initiated as part of the Treatable Intellectual Disability Endeavor in British Columbia and approved by the institutional review boards of BC Children's Hospital and the University of British Columbia. Parents provided written informed consent.

6.3 Family 1

In family 1, the female index (II-1 in Figure 6-3), her younger affected brother (II-2), and her unaffected sister (II-3) were born to healthy nonconsanguineous parents of Belgian-Scottish descent after uneventful pregnancies and deliveries. The index and her male sibling developed lethargy, tachypnea, hypoglycemia, hyperlactatemia, hypernatremia and hyperammonemia with respiratory alkalosis within the first days of life (Figure 6-1). Known urea cycle defects and primary causes of hyperlactatemia were excluded by sequencing and deletion/duplication analysis of N-acetylglutamate synthase (*NAGS* [MIM 608300]), carbamoylphosphate synthetase (*CPS1* [MIM 608307), ATPase deficiency (*TMEM70* [MIM 612418]), and pyruvate carboxylase (*PC* [MIM 608786]). Chromosomal microarray analysis (AffymetrixCytoscan HD) was unremarkable, and homozygosity analysis did not reveal evidence of consanguinity or uniparental disomy.



Theoretical Possibilities		Actual Results[a]							
Possible Enzyme Deficiency	Predicted Secondary Biochemical Abnormalities	Metabolite	Family 1			Family 2		Family 3	
			Female	Male	Normal[b]	Male	Normal	Male	Normal
carbamoyl phosphate synthetase	↑ ammonia	plasma ammonia (μM)	780*	238*	<40	422*	<50	258*	<40
	↓ citrulline	plasma citrulline (μM)	5*	18	8–47	17	3–36		10–45
	↓ arginine	plasma arginine (μM)	17*	NA	32–142	35	17–119	22	12–133
	↑ glutamine	plasma glutamine (μM)	1,051*	1,237*	457–746	2,606*	243–822	571	254–823
	N ornithine	plasma ornithine (μM)	15*	82	27–207	146	38–272	17	15–200
	N orotate	urine orotate	non-det	1.9	<4.3	2.2	<4.9	0.4	<3
pyruvate carboxylase	↓ gluconeogenesis (hypoglycemia)	serum glucose (mm)	2.2*	2.9*	3.3–7.0	2.9*	3.0–8.0	3.2*	3.5–6.0
		serum lactate (mm)	9.1*	8.8*	0.5–2.2	8.1*	1.0–1.8	5.6*	0.6–2.6
		plasma alanine (μm)	1,531*	609*	148–475	1,078*	132–455	603*	143–439
	redox imbalance (↑ lactate & dicarboxylic acids)	plasma proline (μm)	418*	371*	40–332	625*	78–523	283*	52–298
	↓ tricyclic acid cycle intermediates (cataplerosis)	urine lactate	3,737*	4,109*	<200	28,000*	<456	grossly increased*	
		urine 3-OH-butryric acid	7,902*	2,657*	<21	7,060*	<22	grossly increased*	
		urine aceto-acetic acid	584*	927*	non-det	++[c]		grossly increased*	
		urine fumaric acid	13.8*	38.5*	<10	8	<13	moderately increased	
		urine 2-α-ketoglutaric acid	143*	254.6*	<112	300*	<267	slightly increased*	
		urine adipic acid	55*	219.8*	<29	340	<25	normal	
		urine suberic acid	18.6*	48.6*	<13	29	<15	normal	
		urine sebacic acid	NA	18.7*	<10	NA		normal	
	↑ lysine	plasma lysine (μm)	87	161	71–272	306	71–272	94	71–272
proprionyl-CoA carboxylase	↑ 3-OH-propionic acid	urine 3-OH-propionic acid	79.38*	54.57*	<16	59*	<21	normal	
	↑ propionylglycine	urine proprionylglycine	1.29*	3.28*	non-det	5.6*	<2	trace*	
	↑ methylcitrate	urine methylcitrate	6.4*	non-det	non-det	normal[c]		trace*	
3-methylcrotonyl-CoA carboxylase	↑ 3methylcrotonylglycine	urine 3-methylcrotonylglycine	22.9*	non-det	<10	17*	<5	trace*	
	↑ 3OH-isovaleric acid	urine 3-OH-isovaleric acid	40.13*	55.50*	<38	327*	<55	increased*	

Figure 6-1 Overview of Biochemical Abnormalities Resulting from CA-VA Deficiency: In Theory and for the Index Cases for Families 1, 2, and 3. All values in urine are expressed as μmol/mmol of creatinine. Abbreviations are as follows: N, normal (within reference range); NA, not available; non-det, nondetectable. [a]In each individual, the value with maximal deviation from normal during crisis is provided. Asterisks (*) indicate abnormal values. [b]Normal values differ for each family studied because values were measured in different laboratories. [c]Qualitative assessment.

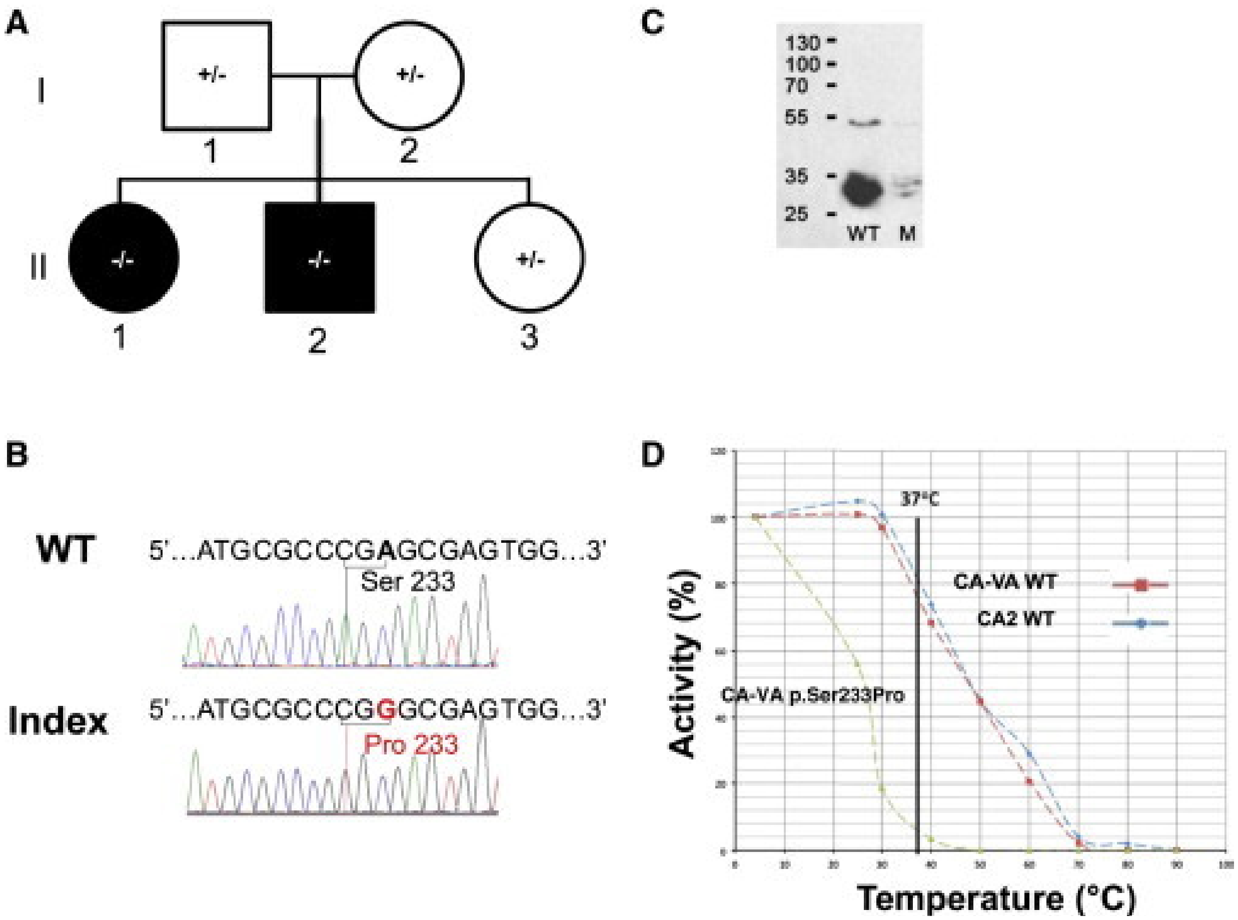


Figure 6-2 Family 1 with p.Ser233Pro Missense Variant. (A) Pedigree (black fill indicates clinically affected individuals, II-1 and II-2). (B) Sanger sequence of *CA5A* from index (II-1) and control (wild-type sequence; WT) subjects; the variant nucleotide position and the corresponding codon alteration (p.Ser233Pro) are indicated. (C) Immunoblot analyses by SDS-PAGE (ImageJ software) of WT and p.Ser233Pro (mutant; M) CA-VA protein levels in COS-7 cell lysates; the molecular weights (kDa) of protein standards are indicated on the left. (D) Thermal stability profiles for WT (red) and p.Ser233Pro mutant (green) CA-VA enzymes. Carbonic anhydrase II (CA2; blue) was used as a control.

Clinical and metabolic findings normalized in both siblings with the administration of intravenous dextrose and bicarbonate, as well as enteral carglumic acid (Carbaglu). In our institution, carglumic acid is used to resolve hyperammonemia of unknown origin[251]. WES was performed for the two affected siblings and their unaffected parents via the Agilent

SureSelect kit and Illumina HiSeq 2000 (Perkin-Elmer). Rare variants were assessed for their potential to disrupt protein function and screened under a series of genetic models—primarily the Mendelian recessive mode of inheritance given the rarity of the phenotype and the pattern of inheritance of most IEMs. Approximately 99% of the observed variations were classified as common (Appendix E Figure 1). Eight rare candidate variants fit the autosomal-recessive model of homozygous (*CA5A*, *ARSB* [OMIM 611542], *CDKN2B* [OMIM 600431], *CDKN2A* [OMIM 600160], *OGG1* [OMIM 601982]) or compound heterozygous (*GRK4* [OMIM 137026], *SPG11* [OMIM 610844], *EPHX2* [OMIM 132811]) variants in the affected siblings. The final set of rare variants (mean average frequency < 1%) was assessed for the potential to disrupt protein function via the Sift and PolyPhen2 software systems. Of these, only one variant (c.697T>C; RefSeq accession number NM_001739.1) in *CA5A* on chromosome 16 was considered a functional candidate; this variant was not reported in dbSNP (version 137), NHLBI ESP, or our in-house genome database (comprising 100 exomes and 10 whole genomes; anno December 2013). Integrative Genomics Viewer 2.0.34 was used to visualize the read alignment and assess variant quality prior to Sanger validation. Given the existence of a related pseudogene[253], the *CA5A* variant was confirmed by targeted Sanger sequencing via carefully designed primers to avoid amplification of the pseudogene sequences; this was achieved by review of paired-end sequence data and selection via a BLAST search of appropriate regions with least similarity, especially close to the 3′ end. Affected siblings are confirmed homozygous, whereas unaffected parents and the unaffected youngest female are heterozygous carriers (Figure 6-2). This variant corresponds to a Ser to Pro substitution at position 233 that is predicted to disrupt structure around the conserved Thr235 residue that forms part of the substrate-binding region of the enzyme (Figure 6-3). Indeed, the p.Ser233 residue is highly conserved evolutionarily across species and in all 12 active human carbonic anhydrase isoforms. Studies of human CA-II have demonstrated that mutations in this hydrophobic patch in the active site destabilize the structure around the substrate-binding region and dramatically reduce the activity of the mutant CA-II[254].

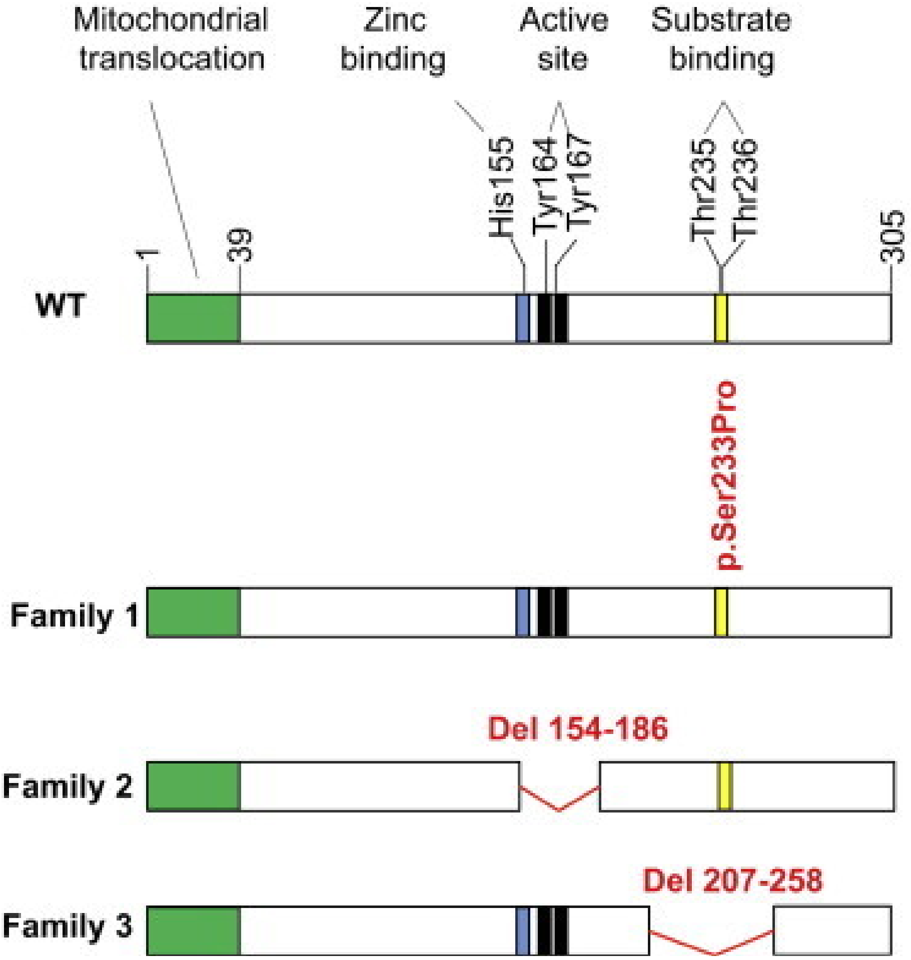


Figure 6-3 Effect of Genetic Variants Identified in CA-VA. Shown is a schematic diagram of the 305 amino acid wild-type (WT) CA-VA. Residues 1–39 encode a mitochondrial translocation signal (green). Homology predictions indicate that histidine 155 binds a zinc ion (blue), tyrosines 164 and 167 are active site residues (black), and threonines 235 and 236 comprise a substrate-binding region (yellow). Shown in red are the deduced CA-VA variants identified in this study. The index and affected brother of family 1 has a nonsynonymous Ser to Pro mutation at residue 233, adjacent to the substrate-binding region. The index of family 2 has a deletion of residues 154–186

(exon 4), thereby missing the metal-binding and active-site residues. The index of family 3 has a deletion of residues 207–258 (exon 6; the substrate-binding region), which results in absent protein.

A marked reduction was observed in the steady-state levels of CA-VA p.Ser233Pro compared with wild-type protein despite similar transfection efficiencies (Figure 6-2C). CA-VA p.Ser233Pro-specific activity in total cell lysates is reduced to 20% of wild-type protein activity, whereas activities of the cotransfected marker enzyme β-glucuronidase were comparable. Thermal stability of mutant recombinant human CA-VA was compared with the WT CA-VA and recombinant human Carbonic Anhydrase II as an additional control (Figure 6-2D). After a 30 min preincubation, the mutant enzyme had lost 80% of its activity at 30°C and almost all its activity at 40°C. By contrast, both WT CA-VA and human CA-II were much more stable at 30°C and 40°C, retaining approximately 100% and 70% residual activity, respectively.

6.4 Family 2

In family 2, a male child (II-1 in Figure 6-4) was born spontaneously at gestational age 36+2 weeks to nonconsanguineous Russian parents. On day 4 of life, he presented with lethargy, weight loss, jaundice, and tachypnea. Initial investigations showed hyperammonemia, hyperlactatemia, mild hypoglycaemia, metabolic acidosis, and ketonuria. Carglumic acid and biotin were initiated, along with protein-free formula and intravenous lipids; 12 hr later, the metabolic acidosis and hyperammonemia resolved.

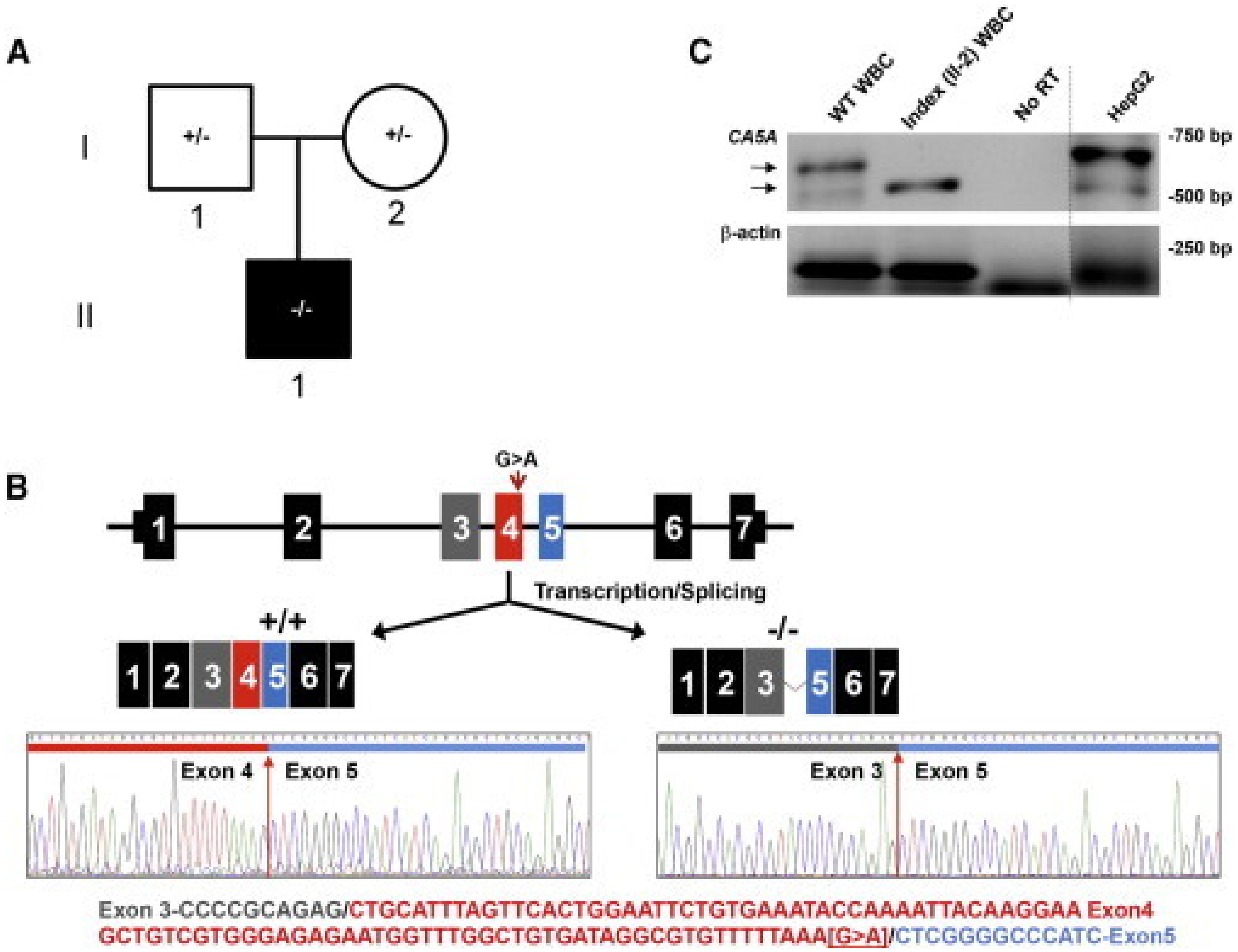


Figure 6-4 Family 2 with Exon 4 Splice Deletion. (A) Pedigree (black fill indicates clinically affected individual). (B) Sanger sequencing of RT-PCR products generated in (B) with exons denoted by colors. Top: The *CA5A* structure (not to scale) and a schematic of the observed *CA5A* transcripts produced in a control subject and the index. Bottom: Sanger sequence of transcripts at exon 4 boundary in a control subject (+/+) and the index (−/−), along with WT *CA5A* cDNA sequence, color-coded as in the top panel. (C) RT-PCR of *CA5A* mRNA from white blood cells (WBCs) or cultured liver cells (HepG2). Arrows indicate the products of differing size amplified from control subject (WT sequence) and index (II-1) WBCs. As controls, reverse transcriptase was omitted from the reaction (no RT) and a control gene (β-actin) was amplified in separate lane on a different cell type (denoted by the line).

Sanger sequencing of all seven exons of *CA5A* in the index identified a synonymous c.555G>A transition (RefSeq NM_001739.1) at the final base of exon 4 (Figure 6-4B). Given that guanine is the most common nucleotide found at this end of an exon in vertebrate genes, RT-PCR was undertaken to demonstrate an effect on mRNA splicing[255]. RT-PCR via primers designed to amplify exons 2–7 of CA5A generated distinctly different product sizes (approximately 550 bp and 650 bp, respectively). Sanger sequencing of the 550 bp band revealed an in-frame deletion of exon 4 from the index RNA (Figure 6-4C). Homology with carbonic anhydrase isoforms identifies three critical residues in the deleted CA-VA transcript (residues 154–185): His155, which binds to a catalytically essential zinc molecule, and Tyr164 and Tyr167, which form part of the active site of the CA-VA enzyme[256]. Thus, this deletion is predicted to significantly impair CA-VA enzyme activity, if not lead to protein misfolding and degradation.

6.5 Family 3

In family 3, a male child (II-5 in Figure 6-5) was born at term by Caesarian section (because of placenta previa) as the youngest of five children to first-cousin consanguineous Pakistani parents. At admission, he was encephalopathic with hyperammonemia and hyperlactatemia with a compensated metabolic acidosis. Urea cycle defects (*OTC* [OMIM 311250], *CPS1* [OMIM 237300], *NAGS* [OMIM 237310] deficiencies) and *PC* (OMIM 266150), citrin (OMIM 605814), and biotinidase (OMIM 253260) deficiencies were excluded by molecular or enzymatic analyses.

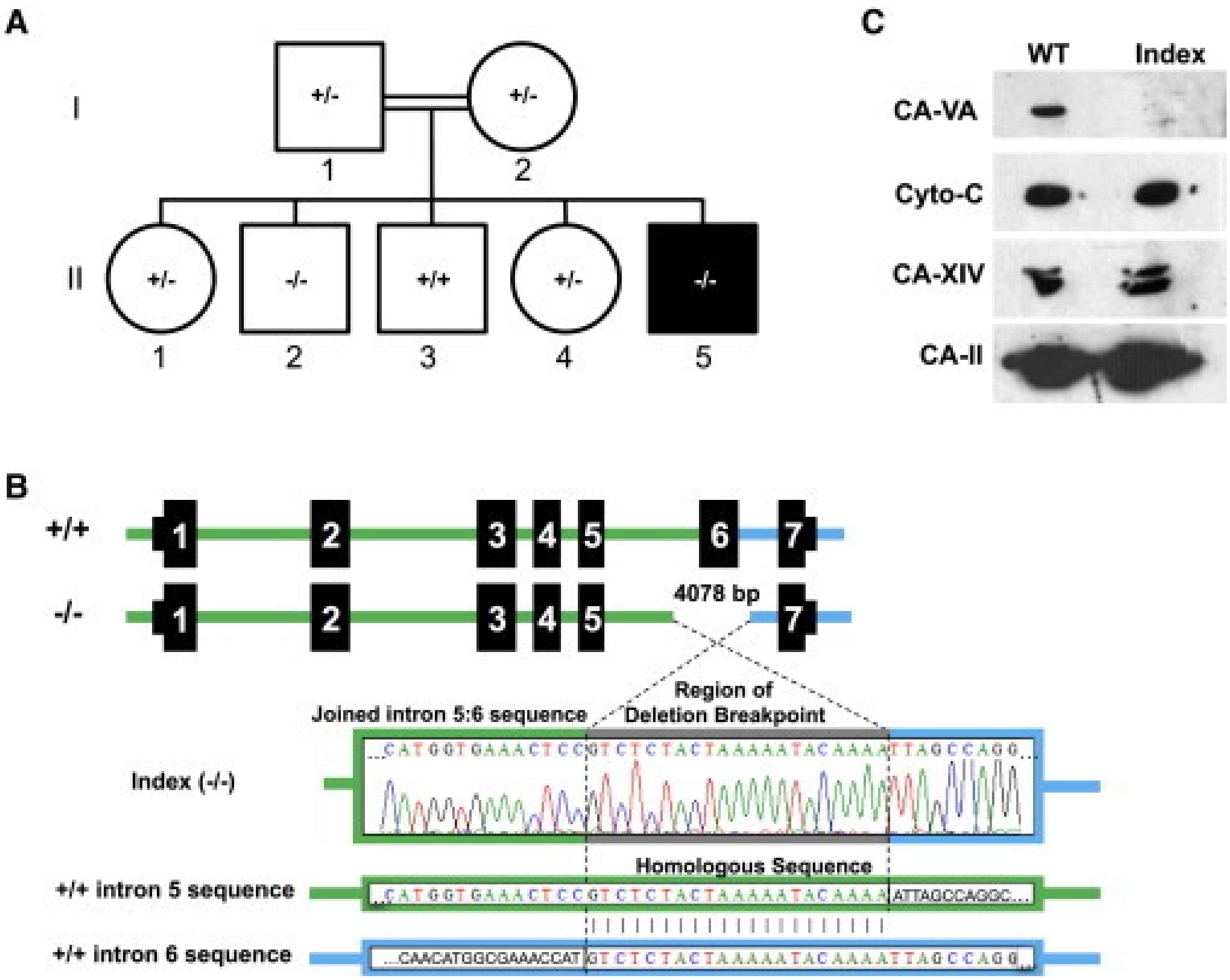


Figure 6-5 Family 3 with Exon 6 Deletion. (A) Pedigree (black fill indicates clinically affected individual, II-5). (B) Top: Schematic representation of the 4,078 bp deletion that encompasses exon 6 of *CA5A*. Bottom: Sanger sequencing of PCR products generated from genomic DNA of the index (II-5) with a 21 bp repeated sequence (gray) at the breakpoint found in both intron 5 (green) and intron 6 (blue). (C) Immunoblot analyses of control subject (WT) and index (II-5) liver homogenates.

Sanger sequencing of the *CA5A* exons for the index revealed a deletion of 4 kb encompassing exon 6 (Figure 6-5B). Absence of CA-VA protein was confirmed by immunoblot in existing liver biopsy tissue (Figure 6-5C).

CA-VA deficiency is a human inborn error of metabolism presenting with hyperammonemic encephalopathy in early life. Initial evidence of causality for the identified CA5A alterations is provided by their similar biochemical phenotypes during metabolic crises. Findings are consistent with dysfunction of all four enzymes to which CA-VA provides bicarbonate as substrate in mitochondria (*CPS1* and three biotin-dependent carboxylases: propionyl-CoA [*PCC*], 3-methylcrotonyl-CoA [3MCC], and pyruvate carboxylase [*PC*]).

Outside of acute events, biochemical parameters remained normal in all affected children except for mildly elevated blood lactate and/or ketonuria. We propose several explanations for the relatively benign clinical course in these individuals and lack of apparent phenotype in the oldest male sibling in family 3. First, overlapping function of CA-VB may help prevent deleterious sequelae of reduced CA-VA activity[257]. In the mouse, Car5A is mainly localized in liver and its deficiency results in profound hyperammonemia. Car5B, though almost undetectable in liver, is predominant in mitochondria of many other tissues. Nonetheless, Car5B deficiency alone has no obvious phenotype. However, when superimposed on Car5A deficiency in the doubly deficient mouse (Car5Adl1Sws/Car5Bdl1Sws), Car5B deficiency aggravated the hyperammonemia and hypoglycemia and shortened survival.18 Thus, Car5B does contribute to handling the metabolic load, though its action is evident only in the absence of Car5A. Second, although carbonic anhydrases accelerate the conversion of CO2 to HCO3− by 1,000-fold or greater, some bicarbonate is produced via the nonenzymatic reaction, even in the absence of carbonic anhydrases[258]. Sufficiency for the product of this limited nonenzymatic source of bicarbonate may differ for the four different bicarbonate-requiring enzymes depending on their individual Kms.

6.6 Conclusion

Thus, CA-VA deficiency should be considered among urea cycle defects, organic acidurias, and *PC* deficiency in the differential diagnosis for hyperammonemia and hyperlactatemia in the neonate and young child. CA-VA deficiency expands the list of treatable inborn errors of metabolism potentially causing intellectual disability[73]. Effective therapy in the affected individuals comprised (1) preventive sick-day management during intercurrent illnesses, including a high-caloric, lipid-rich formula restricted in protein but normal in carbohydrates; and possibly (2) carglumic acid to enhance the activity of the first step in the urea cycle as a treatment for the hyperammonemia.

Chapter 7: Translational value of whole exome sequencing in intellectual developmental disorder patients with unexplained metabolic phenotypes

7.1 Synopsis

Background: Whole exome sequencing (WES) has transformed rare disease-gene discovery and diagnosis. Translation into disease-modifying treatments is challenging, particularly for intellectual developmental disorders (IDD). Inborn errors of metabolism (IEMs) are the exception however; late in 2014, 89 were known to be responsive to causal therapy, i.e. targeting pathophysiology at molecular or cellular level.

Methods: To uncover the genetic basis of potentially treatable IEMs, we combined deep clinical phenotyping with WES analysis via an unbiased semi-automated bio-informatics pipeline, in consecutively enrolled patients with IDD and unexplained metabolic phenotypes.

Results: WES analysis was completed in 59 IDD patients (from 47 families); 8 patients were excluded due to other identified etiologies. The remaining 51 patients in 42 families were predominantly single cases born to non-consanguineous Caucasian parents. (Likely) Pathogenic variants were identified in probands of 38 families, in 43 different genes: 13 genes not previously linked to a human disease phenotype, 21 disease genes with novel patient phenotypes and 9 genes with expected phenotypes. In 7 families, complex phenotypes were explained by two monogenic conditions. In 18 families the diagnosis significantly impacted management beyond genetic counseling, including the discovery of 5 novel IEMs potentially amenable to causal therapy.

Conclusions: Our diagnostic yield and discovery rate exceeded expectation, likely due to enrichment of our cohort for new IEMs and phenotypes, semi-automated bio-informatics pipeline and close collaboration between families, clinicians and scientists. In 43%, WES diagnosis allowed for precision medicine, varying from prevention and tailored symptom management, to causal therapy.

7.2 Introduction

Significant yield of whole exome sequencing (WES) is documented in patients with an unexplained intellectual developmental disorder (IDD)[259], a frequent and heterogeneous condition affecting an estimated 2.5 - 3% of the population worldwide ($3x10^6$ newborns per year)[260]. With co-morbidities ranging from epilepsy, psychiatric/behavioral disturbances, movement disorders, sensory deficits, to other organ dysfunction, IDD poses a significant emotional, functional and health-economic burden[261]. Aside from CNVs and methylation abnormalities, a multitude of single gene defects cause IDD with more to be discovered[262, 263]. Diagnosis is essential for accurate genetic counseling, ending the diagnostic odyssey, informed decision-making by families and the health care team, and accessing medical support and services in the community, but does not easily translate into disease-modifying treatments. The exception is inborn errors of metabolism (IEMs), the largest group of genetic IDDs amenable to causal therapy, i.e. interventions directly targeting pathogenesis at the cellular and molecular level such as medical diets, vitamin supplements, medications, hematopoietic stem cell transplantations and gene therapy[264]. Since 2012, the number of treatable IEMs causing IDD has increased from 81 to 89[252, 264].

Many metabolic pathways are yet to be associated with human disease, and thus additional treatable IDDs await discovery. Identification of a genetic basis of an IEM allows for insights into the affected pathway, which sometimes reveal treatment targets, as illustrated by *ALDH7A1* (Lysine catabolism enzyme) causing pyridoxine-dependent epilepsy which allowed implementation of the lysine restricted diet and arginine supplementation to improve suboptimal neurodevelopmental outcomes on vitamin B6 alone[265].

To accelerate treatable IDD identification, we selected carefully characterized IDD patients with unexplained metabolic phenotypes for WES analysis, applying an unbiased semi-automated bioinformatics pipeline and a multi-disciplinary approach to causal variant identification and validation with a focus on translation of diagnosis into precision medicine. Here we report our diagnostic yield as well as novel gene discoveries, and highlight the overall impact on clinical management.

7.3 Methods

7.3.1 Participants

The study was approved by the Ethics Board of the Faculty of Medicine of the University of British Columbia (UBC IRB approval H12-00067); each family provided informed consent for participation in the study and publication. Eligibility criteria included: confirmed IDD or potential for IDD (presence of toxic metabolites in the neonatal period known to cause brain damage) *and* 'metabolic phenotype' of unknown cause; any age; comprehensive clinical phenotyping with extensive previous metabolic / genetic testing. A metabolic phenotype was defined as one or more of (1) (pattern of) abnormal metabolites in urine, blood, CSF; (2) abnormal functional studies at a biochemical / cellular level (e.g. mitochondrial respiratory chain complex deficiency); (3) abnormalities on clinical history (e.g. regression), physical exam (e.g. organomegaly), neuro-imaging/physiology (e.g. leukodystrophy), pathology (e.g. storage vacuoles) suggestive of neuro-metabolic disease. During the informed consent process, the risks and benefits of research-based WES analysis were explained to patient and family, and an option for disclosure of medically actionable incidental findings (IFs) provided.

7.3.2 Sequencing and bioinformatics analysis

We isolated genomic DNA using standard techniques from either peripheral blood or saliva for the proband, both parents, and all affected and unaffected siblings (if available). WES analysis was performed on the index as well as any affected siblings, and in the majority of families, parents as well, either using the Agilent SureSelect targeted capture kit on the Illumina HiSeq 2000 sequencer (Perkin-Elmer, Santa Clara, California, USA) or using the Ion AmpliSeq™ Exome Kit and Ion Proton™ System from Life Technologies (Next Generation Sequencing Services, UBC, Vancouver, Canada).

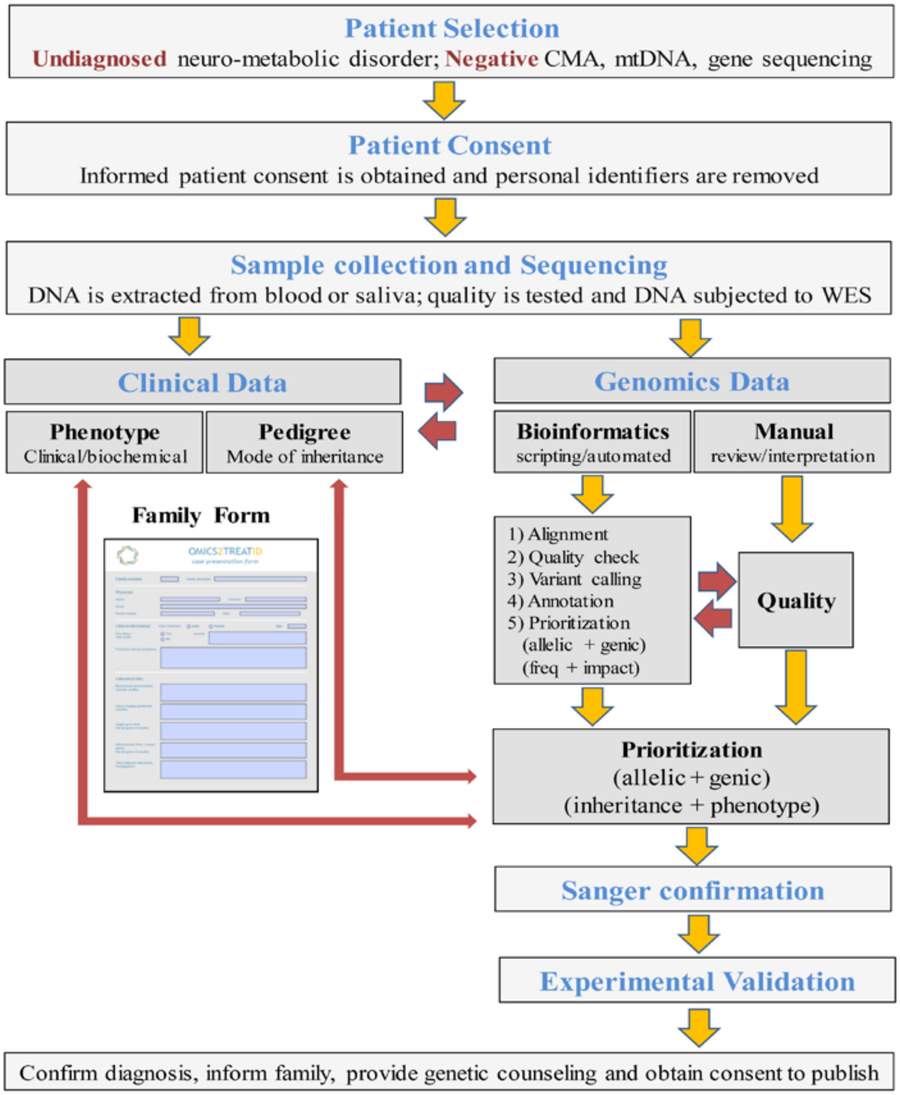


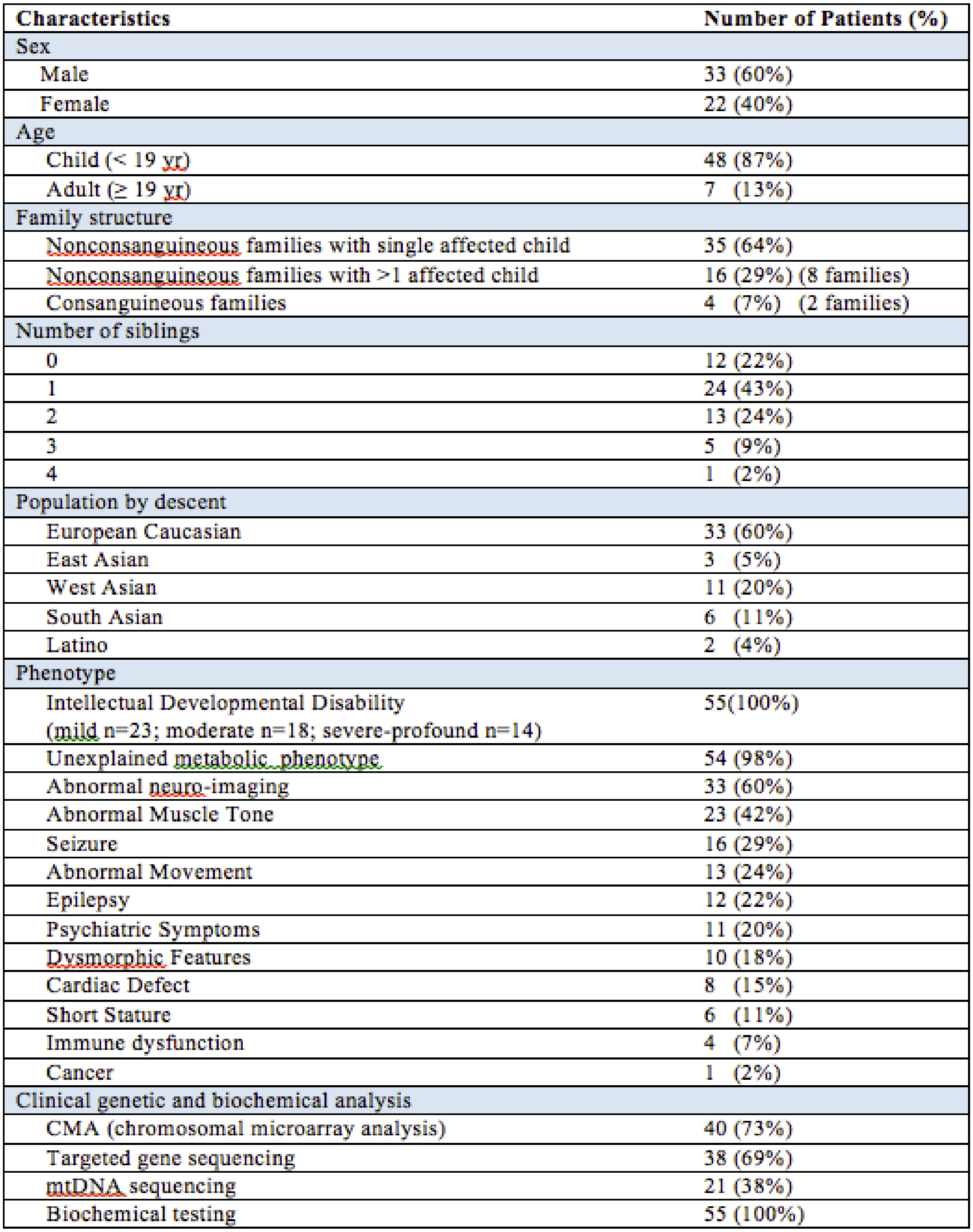
Figure 7-1 Workflow of semi-automated gene-discovery WES approach. CMA (chromosomal microarray analysis), mtDNA (mitochondrial DNA sequencing)

We developed and applied a semi-automated gene-discovery pipeline (Figure 7-1), which takes advantage of minimal, but critical, manual quality inspection of the data as well as essential collaborative interactions between clinicians and bioinformaticians. The referring clinician provided a form populated with data on phenotype, family history with pedigree, ethnicity, prior diagnostic testing results, which was used for WES data interpretation by bioinformaticians. Validation of pathogenicity (classified according recent ACMG Standards and Guidelines)[266] and causality of variants in novel genes (previously unreported in human disease) were pursued according to recent guidelines (MacArthur et al 2014[267]; CCMG 25951830).

7.4 Results

7.4.1 Study group characteristics

Between October 2012 and January 2015, we recruited and completed WES in 59 IDD patients (47 families), meeting selection criteria. Subsequently, 8 patients with negative WES results were excluded for the following reasons: the etiology was confirmed as: teratogen exposure (n=1), congenital infection (n=2), auto-immune disorder (n=3) or pathogenic chromosomal CNV (n=2). The 51 remaining patients (in 42 families) comprised predominantly children (n=45[88%]) with age at enrollment ranging from 0.7 to 31 years (median 5.4 years); 21 females [41%] and 30 males [59%]); 22 with mild IDD; 17 with moderate IDD; and 12 with profound IDD) with a spectrum of additional clinical and biochemical manifestations (Table 7-1). The majority of patients are of European-Caucasian descent (n=32[63%]) born to non- consanguineous parents without family history (n=33[65%]). In all patients, biochemical testing according to a published diagnostic algorithm for treatable IDDs[268] had been performed along with a combination of clinical genetics tests without revealing a diagnosis, prior to recruitment for WES analysis (Table 7-1).



Characteristics	Number of Patients (%)
Sex	
Male	33 (60%)
Female	22 (40%)
Age	
Child (< 19 yr)	48 (87%)
Adult (≥ 19 yr)	7 (13%)
Family structure	
Nonconsanguineous families with single affected child	35 (64%)
Nonconsanguineous families with >1 affected child	16 (29%) (8 families)
Consanguineous families	4 (7%) (2 families)
Number of siblings	
0	12 (22%)
1	24 (43%)
2	13 (24%)
3	5 (9%)
4	1 (2%)
Population by descent	
European Caucasian	33 (60%)
East Asian	3 (5%)
West Asian	11 (20%)
South Asian	6 (11%)
Latino	2 (4%)
Phenotype	
Intellectual Developmental Disability (mild n=23; moderate n=18; severe-profound n=14)	55(100%)
Unexplained metabolic phenotype	54 (98%)
Abnormal neuro-imaging	33 (60%)
Abnormal Muscle Tone	23 (42%)
Seizure	16 (29%)
Abnormal Movement	13 (24%)
Epilepsy	12 (22%)
Psychiatric Symptoms	11 (20%)
Dysmorphic Features	10 (18%)
Cardiac Defect	8 (15%)
Short Stature	6 (11%)
Immune dysfunction	4 (7%)
Cancer	1 (2%)
Clinical genetic and biochemical analysis	
CMA (chromosomal microarray analysis)	40 (73%)
Targeted gene sequencing	38 (69%)
mtDNA sequencing	21 (38%)
Biochemical testing	55 (100%)

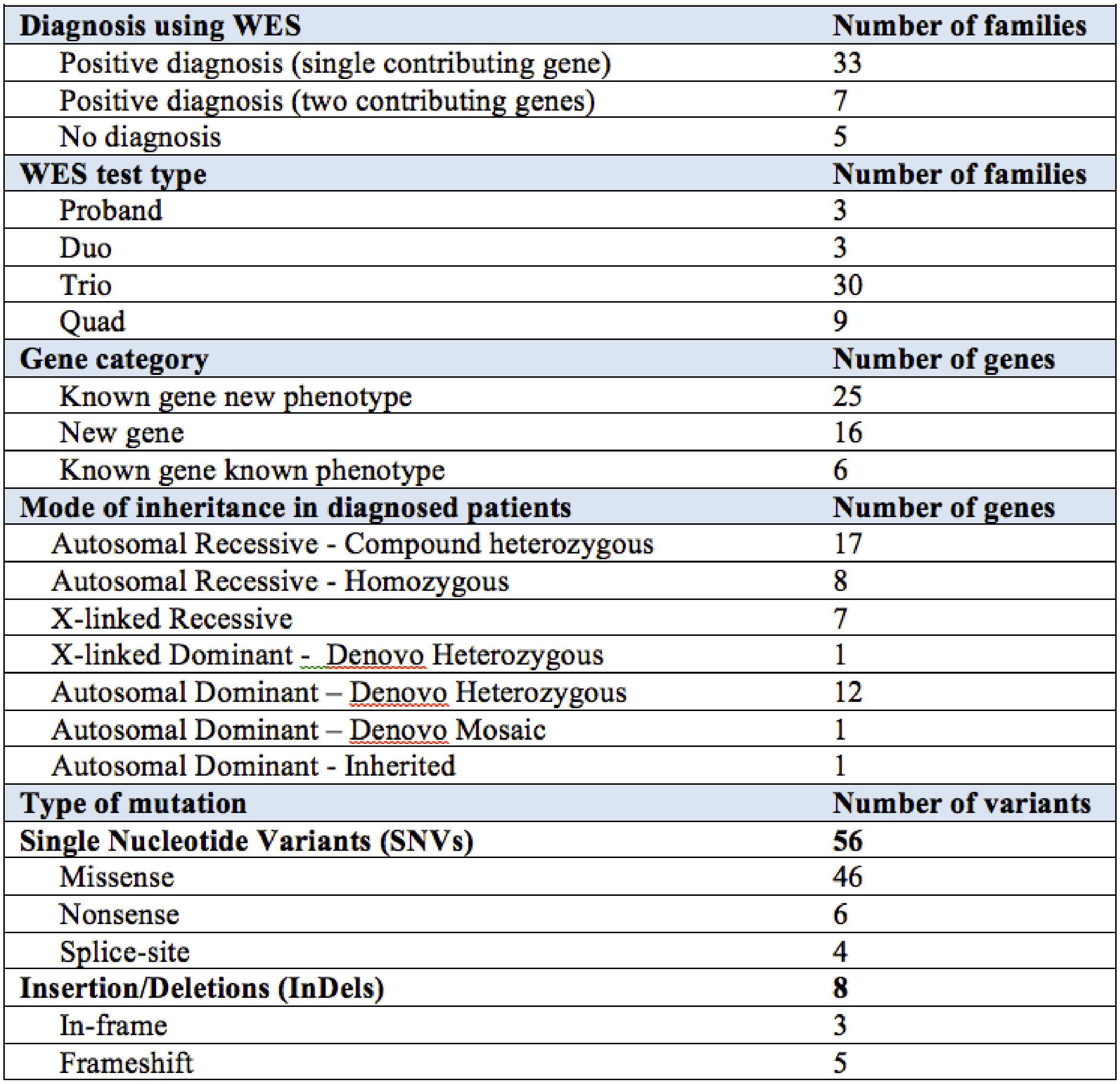
Table 7-1 Clinical characteristics of the 55 patients in 45 families.

7.4.2 Diagnostic yield

WES identified the likely genetic diagnosis in 38 of 47 families enrolled, translating into a diagnostic yield of 90% after exclusion of 5 families with other etiology (Table 7-2). For the majority (n=29[69%]) of the 42 included families WES had been performed on the family trio (mother-father-index).

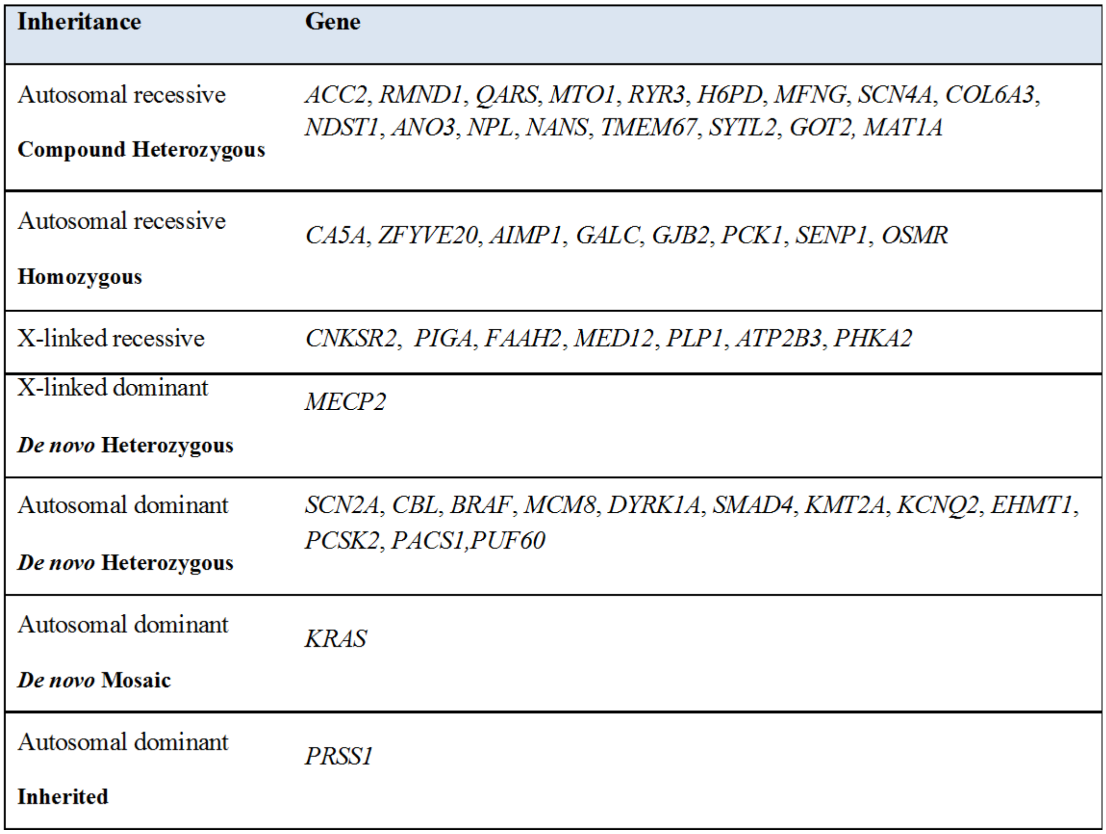
In total, 59 diagnostic mutations were identified in 43 genes; all except the previously reported somatic *KRAS* mutations were germline. The majority of the mutations were classified as pathogenic (n=36[61%]) or likely pathogenic (n=21[36%]) according to recently published ACMG Standards and Guidelines (Appendix F Table 1). Most mutations (n=46[78%]) were not present in either our database of more than 350 individual exomes / genomes or dbSNP (version 138), while 13 variants with dbSNP records were rare (average allele frequency 0.006). Eight diagnostic mutations (14%) were previously identified as pathogenic, which is comparable to previous reports[269]. When compared against Exome Aggregation Consortium (ExAC), a database of 61,486 unrelated individuals, 32 [54%] mutations were novel, while 27 were rare (average allele frequency 0.004). The ExAC dataset includes patients with mental illnesses and thus variants potentially relevant to our cohort. For instance, two previously reported *de novo* pathogenic mutations, in *CBL* and *PACS1* genes (Appendix F Table 1), have been observed in the ExAC population.

The 59 diagnostic mutations consisted predominantly (n=52 [88%]) of single nucleotide variants (SNVs) (Table 7-2). These were further classified as missense (n=44[74%]), nonsense (n=4[7%]) or splice-site mutations (n=4[7%]). We identified InDels (Insertions and Deletions less than 20bp in length) that resulted in either a frameshift (n=3[5%]) or an in-frame deletion of conserved amino acids (n=4[7%]). For the 43 identified genes, the main mode of variant inheritance is recessive (70%), including compound heterozygous (n=16[37%]), homozygous (n=8[19%]), and X-linked recessive (n=6[14%]) (Tables 6-2 and 6-3). Dominant mutations were identified in 30% of patients, including 12 *de novo* mutations (11 heterozygous and 1 mosaic) and a single familial autosomal dominant mutation with variable penetrance.



Diagnosis using WES	**Number of families**
Positive diagnosis (single contributing gene)	33
Positive diagnosis (two contributing genes)	7
No diagnosis	5
WES test type	**Number of families**
Proband	3
Duo	3
Trio	30
Quad	9
Gene category	**Number of genes**
Known gene new phenotype	25
New gene	16
Known gene known phenotype	6
Mode of inheritance in diagnosed patients	**Number of genes**
Autosomal Recessive - Compound heterozygous	17
Autosomal Recessive - Homozygous	8
X-linked Recessive	7
X-linked Dominant - Denovo Heterozygous	1
Autosomal Dominant – Denovo Heterozygous	12
Autosomal Dominant – Denovo Mosaic	1
Autosomal Dominant - Inherited	1
Type of mutation	**Number of variants**
Single Nucleotide Variants (SNVs)	**56**
Missense	46
Nonsense	6
Splice-site	4
Insertion/Deletions (InDels)	**8**
In-frame	3
Frameshift	5

Table 7-2 Diagnostic yield and summary of WES analysis.



Inheritance	**Gene**
Autosomal recessive **Compound Heterozygous**	*ACC2, RMND1, QARS, MTO1, RYR3, H6PD, MFNG, SCN4A, COL6A3, NDST1, ANO3, NPL, NANS, TMEM67, SYTL2, GOT2, MAT1A*
Autosomal recessive **Homozygous**	*CA5A, ZFYVE20, AIMP1, GALC, GJB2, PCK1, SENP1, OSMR*
X-linked recessive	*CNKSR2, PIGA, FAAH2, MED12, PLP1, ATP2B3, PHKA2*
X-linked dominant ***De novo* Heterozygous**	*MECP2*
Autosomal dominant ***De novo* Heterozygous**	*SCN2A, CBL, BRAF, MCM8, DYRK1A, SMAD4, KMT2A, KCNQ2, EHMT1, PCSK2, PACS1,PUF60*
Autosomal dominant ***De novo* Mosaic**	*KRAS*
Autosomal dominant **Inherited**	*PRSS1*

Table 7-3 Inheritance patterns of the 46 genes identified in the study.

7.4.3 Impact on clinical management

Of all families in whom WES identified the (likely) causal gene(s), the novel diagnosis significantly impacted clinical management in 18 (47%) families: preventive measures such as regular malignancy screening and avoidance of disease triggers in 4 (*CBL[270], SMAD4, MTO1, PRSS1*), more precise symptomatic management such as neurotransmitter, serine, folinic acid supplementation in 6 (*CKNSR2, SCN2A, ANO3, BRAF, ATP2B3, MeCP2*), immune- modulating therapies such as chemotherapy or stem cell transplantation in 3 (*SENP1, SYTL2, KRAS*), and causal treatments targeting the pathophysiology at a cellular/molecular level in 5 (*CA5A, ACC2, GOT2, PCK1, NANS*) as further described below.

7.4.4 Illustrative cases

7.4.4.1 Novel diseases amenable to causal therapy

We identified 5 novel IEMs potentially amenable to dietary restriction, supplementation, and/or pharmacological interventions.

The first discovery was carbonic anhydrase VA deficiency as already described in the previous chapter. The second discovery was a homozygous 12-bp deletion in *PCK1* (OMIM 614168), phosphoenolpyruvate carboxykinase, in a 3-year old boy with liver steatosis, and mild hypoglycemia, hyperammonemia, lactic acidosis, elevated tricyclic acid metabolites responding to a metabolic diet and emergency regimen (rich in complex carbohydrates); *in vitro* mutant enzymatic activity was significantly reduced. Third, in a 6-year old boy with acquired microcephaly, severe seizure disorder, spasticity, sleep disturbances, abdominal spasms, and low serine in body fluids, we identified mutations in *GOT2* (OMIM 138150) encoding glutamate oxaloacetate transaminase[271]. The patient responded to oral serine and pyridoxine supplements with improved head growth, psychomotor development and seizure control. Fourth, compound heterozygous mutations and corresponding deficiency of NANS (n-acetylneuraminic acid phosphate synthase; OMIM 605202) in a 3-year old presenting with epileptic encephalopathy and dysmorphic features; targeted metabolomics techniques confirmed the increased concentration of the direct substrate of the enzyme in fibroblasts. We have identified 8 probands in 6 unrelated families with similar phenotype but different alleles. Potential treatment strategies to restore the product of the defective enzyme, 5-neuraminic acid, are being investigated in patient fibroblasts and model organisms. Finally, validation of acetyl-coA carboxylase-beta deficiency as potentially novel treatable IEM is still underway.

7.4.4.2 Expansion of the phenotypic spectrum

Aside from 8 additional novel human disease genes including Rabenosyn-5 deficiency *(ZFYVE20; O*MIM 609511)[272], we identified mutations in 21 genes previously reported to cause monogenic conditions, for which we observed novel / additional clinical symptoms.

Hitherto unappreciated treatment targets can be revealed as part of the phenotypic delineation. An example in this study is an 8-year old boy with IDD, autism, movement disorder, intractable epileptic encephalopathy and persistently abnormal neurotransmitter profiles (low CSF homovanillic acid, 5HIAA, and neopterin) in whom WES identified a pathogenic splice site variant (resulting in a validated exon 14 deletion) in a voltage-sensitive sodium channel, *SCN2A* (OMIM 182390). We hypothesized that this channelopathy causes abnormal synaptic mono-amine metabolite secretion / uptake via impaired vesicular release and imbalance in electrochemical ion gradients, which in turn aggravate the seizures. Treatment with oral 5-hydroxytryptophan, L- Dopa / Carbidopa, and a dopa agonist normalized the CSF profile and correlated with significant improvement in attention and mild improvement of seizure control, the latter most likely via dopamine and serotonin receptor activated signal transduction and modulation of glutamatergic, GABA-ergic and glycinergic neurotransmission.

7.4.4.3 Combined phenotypes due to two monogenic defects

Multiple genetic events leading to complex phenotypes may be mistaken for new disorders or novel phenotypes of a known disorder, and thus remind us that a layer of unbiased and systematic interpretation of NGS data is necessary in any clinical pipeline. In fact, recent NGS reports support the notion that blended phenotypes is an appreciable cause of disease[273]. This is demonstrated in our study group with 7 (18%) of 38 diagnosed families harboring mutations at 2 distinct disease loci related to the phenotype (Appendix F Table 2). For instance, in a 19-year old, cognitively normal male born to non-consanguineous Filipino parents with progressive dilated cardiomyopathy and sensorineural hearing loss, WES revealed compound heterozygous rare, damaging mutations in *NPL* (OMIM 611412) encoding N-acetylneuroaminate pyruvate lyase that controls the final step of sialic acid metabolism. The deafness is attributed to a known homozygous mutation in *GJB2* (connexin 26) (OMIM 121011).

7.4.5 Medically actionable incidental findings (IF)

In these 42 families, we identified only one medically actionable IF in *CFTR* (OMIM 602421). Both alleles were previously reported as pathogenic: rs78655421 and rs121908745; however the family (whose clinical phenotype did not suggest cystic fibrosis) chose not to be informed in case of IF and thus the result was not disclosed.

7.4.6 Discussion

Our study reports an integrated deep phenotyping and customized WES bio-informatics approach to the discovery of novel neuro-metabolic conditions and phenotypes in 51 patients from 42 IDD families, with a focus on therapeutic tractability. Overall, our approach achieved a molecular diagnostic yield of 90% in a highly selected group 42 IDD families with unexplained metabolic phenotypes, including 13 potential gene discoveries. Studies to validate causality in a subset are ongoing; we acknowledge this limitation and have thus provided extensive information on pathogenicity of variants using most current recommendations[266] as well as available experimental data to motivate our findings. Although our diagnostic rate exceeds that of most published studies applying NGS in rare diseases (16% - 73%)[274-

277], the most important outcome of our genomics study is the significant impact of the WES findings on clinical management in half of our patients. One-third of the newly discovered human diseases are potentially amenable to pharmacological and/or dietary treatments.

In IEMs, knowledge of the precise defect in a metabolic pathway provides the opportunity to modify disease using nutritional manipulation, which although under regulatory control and worthy of careful study before implementation, do not require the expensive and time-consuming trials inherent to orphan drugs. The discovery of GOT2 deficiency with severe neurologic symptoms amenable to oral serine and pyridoxine supplements (resp. its end product and cofactor), which are both affordable and previously used and deemed safe for other IEMs[278], nicely illustrates this advantage. We report a single patient however, further highlighting the (ultra-) rare disease predicament – small patient numbers requiring global collaborations and use of matchmaking approaches[279] combined with novel trial methodologies to generate sufficient evidence for treatment effects. For more challenging interventions, such as replenishing the intracellular 5-neuraminic acid in NANS deficiency, testing of existing or novel treatments on cellular and model organisms is a crucial first step. Preventive measures such as metabolic diets and emergency regimens to respectively support proper somatic and psychomotor development and avoid metabolic crisis, further illustrate precision medicine made possible by a genomic diagnosis (e.g. CA-VA and PCK1 deficiency[280]). Notably, discovery of the first recessive germline mutation in the *PCK1* gene in our study is a good example of the power of NGS advances to confirm a four-decade old hypothesis. Namely, in 1975, Sovik *et al.* described a patient with persistent neonatal hypoglycemia as a result of a defective gluconeogenesis due to abnormal subcellular distribution of PCK1[281]. Here, we provide evidence at the molecular and biochemical level that indeed

PCK1 deficiency results in this phenotype and suggest that the metabolic dysregulation is amenable to treatment. Finally, difficult decisions for invasive and costly procedures such as hematopoietic stem cell transplant or chemotherapy (e.g. SENP1 deficiency[282]) are enabled by a precise genetic diagnosis and understanding of pathophysiology; outcome reports of such cases in the literature is essential to help other clinicians faced with a similar challenge.

Potential contributors to the high diagnostic success in this study include: restriction to patients with observed metabolic phenotype, a bioinformatics pipeline tied to close consultation with clinical specialists, the prevalence of recessive conditions in metabolic disorders. We also observed a higher portion (13%) with mutations at two distinct disease loci leading to blended phenotypes, compared to past studies reporting (6%)[273] and (4.6%)[283]. This may be related to inclusion of two phenotypes in the patient selection criteria (metabolic and IDD). As a result, we strongly advocate unbiased analysis of NGS data for multiple "hits" in all patients.

In 4 families, repeated semi-annual re-analysis of exome data failed to identify a genetic diagnosis (Table 7-2). Three of these 4 families were studied using proband-only WES, indicating a possibility that a pathogenic *de novo* variant was missed. In one trio-WES analysis family, the proband presented with neonatal hyperammonemia, hyperlactatemia, methylmalonic aciduria which resolved completely, showing normal development and metabolic profiles at age 2 years; a large 600 gene panel and our WES did not yield disease-causing variants and possibly this child does not suffer from a rare monogenic disease but resolved immaturity of enzymes. In another family, WES quad analysis failed to identify a diagnosis due to lack of coverage, in 2 siblings presenting with neurodegenerative phenotype and neurotransmitter abnormalities, whose seizures responded to Levocarbidopa and 5OH-tryptophan. Subsequent WGS analysis revealed a previously described pathogenic mutation (c.10G>C [p.Gly4Arg]) in the *CSTB* (OMIM 601145) resulting in Unverricht-Lundborg syndrome (OMIM 254800). WGS analysis of the remaining 2 families is underway.

Finally, translational genomics requires collaborations between patients & families, a variety of subspecialist clinicians for careful phenotyping, expert bioinformaticians for accurate data-analysis, and basic scientists engaged in specific gene or pathway research. Data-sharing and open communications are key to maximize NGS' diagnostic potential and its clinical benefit to health outcomes in rare diseases.

Chapter 8: Conclusion

8.1 Introduction

The affordability of next-generation DNA sequencing (NGS) allows exploitation of the technology in specialized genetics clinics to uncover the genetic etiologies of diverse disorders. The new discipline that forms around it, genomic medicine, provides patients with personalized strategies for disease prevention, etiology identification, and therapeutic selection[284]. While clinical utility and revolutionary impacts have been demonstrated for classical Mendelian disorders and diverse cancers, ultimately genomic medicine will impact complex chronic disorders and thereby most individuals. Such a vision is embraced by both biomedical researchers and pioneering clinicians alike, as evident in emergent scientific funding programs such as Horizon 2020 (http://ec.europa.eu/programmes/horizon2020/) and Precision Medicine Initiative (https://www.nih.gov/research-training/precision-medicine-initiative).

Clinical practice will be transformed by DNA sequencing. Initial signs of this transformation include the identification of therapeutic targets in individual cancer cases, specific risks for inherited cancers and common diseases, and personalization of drug choice and dosage[285, 286]. In this chapter, I conclude the thesis by exploring future challenges of genomic medicine, focusing on the roadblocks in the path to achieving the full potential. My exploration traverses 5 sections: section 1 explores ways to improve the diagnostic rate of exomes and whole-genomes; section 2 highlights the difficulties of incorporating multiple types of “omics” data; section 3 illustrates the ongoing evolution of hardware and software for efficient data interpretation and storage; section 4 highlights a need for better training programs for clinicians and patients; and section 5 summarizes ongoing ethical issues. Due to the multi-“omics” direction that this field is heading (examined in section 2), the phrase “genomic medicine" is used interchangeably with "personalized medicine" to better capture the types of data and knowledge involved.

8.2 Improving next-generation sequencing diagnostic rate

The current reported diagnostic rates from major consortiums using exomes and/or whole-genomes vary between 16% - 73%. Studies that focus on specialized patient populations (i.e. consanguineous populations) and/or stringent in-take criteria tend to have higher diagnostic rates. While these diagnoses represent a major advance for patients, there remain a large fraction of patients with unresolved diagnoses. In order to elevate the yield and to further promote high-throughput sequencing to a routine clinical screening, greater attention will need to be placed upon the undiagnosed subset of patients. If such cases were originally selected due to strong evidence for genetic etiology, there may be opportunities to resolve additional subsets by improved approaches.

8.2.1 Resolution of detection

The subset of the undiagnosed cases may be attributable to limitations of short-read NGS technology to detect pathogenic structural variations (SVs) and triplet repeat expansions[287]. Even with pair-end reads, the reliability of variant callers for such alterations appear limited to a maximum spacing change of 35-75bp[288, 289]. Complementary technological platforms such as arrayCGH lack adequate sensitivity and specificity to analyze SVs under 1kb[290]. Therefore, pathogenic SVs too large for NGS but too small for array probes are problematic. Longer read lengths promised by 3rd generation sequencing technologies will provide insights into how frequent such alterations are in the undiagnosed set of patients. For instance, companies such as

Ion Torrent and Pacific Biosciences can produce average read lengths up to 400bp and 15kb respectively[291, 292], which can theoretically overcome the limitations of the short-read data and provide insights into genomic regions that previously could not be investigated. Presently, relatively poor read qualities, low throughput and high cost prohibit such technologies for clinical adoption[293]. Yet a retrospective look at the historical development of preceding technologies suggest that the aforementioned technological challenges will eventually be overcome, and access to long reads for clinical genetics should be resolved within a few years.

8.2.2 Variants of unknown significance (VUS)

Clinical diagnosis is limited by our biological understanding of the molecular mechanisms within cells and tissues. Variants of unknown significance, VUS, constitute the largest category of rare variations in any WGS/WES output[294-296]. This holds true even when looking at well-annotated genes with known biological functions (e.g. BRCA family[297]). The latest guidelines from the American College of Medical Genetics indicate that VUS should be excluded from clinical decision-making[298]. To simplify interpretation, many clinical genetics interpretation procedures restrict focus to variants with obvious functional impacts (e.g. nonsense mutations) in genes for which different alterations are known to contribute causally to related phenotypes[294-296]. While the exclusion of VUS for clinical diagnosis is presently justified, an expanding population of sequenced individuals will allow for statistical correlation of VUS with related patient phenotypes for novel genes. Such advances will ultimately require connection of data from patients around the world for extremely rare conditions.

Our inability to assess VUSs is especially problematic when concerning the analysis of regulatory and intronic variants. Such potential regulatory alterations may be frequent in the undiagnosed patients, as current interpretive analyses have largely been restricted to protein coding alterations. The HGMD (as of 2014) contained reports of more than 3000 such mutations as being pathogenic[299]. At present there is a robust global bioinformatics effort to develop approaches to study the impact of regulatory variants, but few tools are sufficiently mature to provide reliable predictions for clinical use. In part the limitation is due to the fact that WGS was very cost-prohibitive until recently, and WES technology does not have the capacity to systematically capture the regions that contain regulatory variants. Furthermore, the underlying data used to build the predictive models are drawn from limited selections of cell lines strongly biased to cancer cells (e.g. Encyclopedia of DNA Elements, ENCODE[300]). Both the compilation of large number of patient genomes and single cell studies of diverse tissues and cell-types will contribute to the capacity to detect causal regulatory alterations in the future.

8.2.3 Promote collaborative exchange of data and knowledge

The sharing of diagnosed genetic cases has been largely accomplished by initiatives such as ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) from ClinGen. For the undiagnosed cases, sharing data within a collaborative environment may be essential. GeneYenta[233] and Phenomizer[301] are matchmaking software tools that were developed to connect clinicians working on similar patients with similar disease phenotypes. GeneTalk (https://www.gene-talk.de) is a web-based platform that allows clinical researchers to share potentially disease-relevant sequence variants in a crowd-sourced database, and connect with experts working on the same variant(s)/gene(s). LOVD (http://www.lovd.nl/3.0/home) is an open-source gene-centric database within which users can submit both DNA variations and patient information, and allows browsing of submitted variants for a given gene. GenomeConnect (https://www.clinicalgenome.org/genomeconnect/) from ClinGen is a portal that engages patients to participate in the sharing of their own de-identified genetic and health information to form connections between patients and the healthcare providers and researchers who wish to study it. At the moment, these tools allow for distinct component tasks that ultimately need to be united within an encompassing system. Having one centralized repository or a federation of data cross repositories would ensure better coverage on all the reported variants. The Global Alliance for Genetics Health (http://genomicsandhealth.org) has been working toward this aim.

At the moment patient phenotype descriptions are manually entered, however the information overlaps with information within electronic healthcare records and therefore there may be opportunities to create appropriately controlled mechanisms (i.e. protective of patient privacy and permission) to unite such systems. Automated inclusion of digital health data would promote better adherence to standards and streamline clinical analysis. As a proof-of-concept, EHR4CR is a European consortium that utilizes the i2b2 infrastructure to incorporate electronic healthcare records for public health and research[302]. The organization specifically assesses the promises of re-using health records for identifying eligible patients for recruitments to clinical research.

As the software and resources are changing rapidly, it is a challenge for analysts to keep up. Open forums such as SEQanswers (http://seqanswers.com) and BioStar (https://www.biostars.org) promote bioinformaticians to exchange experience and troubleshoot common errors, but they are done so in an unorganized manner. OMICStools (http://omictools.com) provides an easy portal for researchers to find the appropriate tool for their specialized needs, but its layout is not designed for users of a given tool to come together and discuss. Ultimately, a well-structured virtual toolshed is required to allow bioinformaticians to discuss the strengths and weaknesses of tools and workflows for tackling specific problems, and the sharing of feedback to facilitate software development.

8.2.4 Gene networks analysis

Most variants from WES/WGS data are evaluated either at the individual variant/gene level, or based on inheritance pattern. Interactions between two genes may be contributing causally to a portion of unexplained cases, as the current focus on simple genetic models may not detect such cases. Network-based algorithms that capture biological pathways, systems or complexes could be helpful. For analogy, Ingenuity's Pathway Analysis is a commercial software and database package that improves upon gene expression enrichment analysis tools by assessing expression data across pathways. Progress in pathway-based analysis has been made for cancer studies[303]. Standardization may help with broadening network analysis adoption, particularly for the visualization of networks.

8.3 Integrating diverse "omics" data

To understand complex biological interactions, the success of genomic medicine will be enabled by access to molecular profiling beyond the genome sequence. The varied "omics" profiles emerging across the life sciences include transcriptomics, proteomics, epigenomics, metagenomics, metabolomics, nutriomics, etc. Metabolomics, for instance, is expected to bring increased diagnostic potential because metabolic markers represent the functional end point of physiological mechanisms[304]. While metabolic profiles have been produced in laboratory medicine for many years, the parallelization enabled by new technologies allows thousands of molecules to be assessed simultaneously. While the name of the detailed molecular analysis on a per patient basis remains in flux (genomic medicine, personalized medicine, precision medicine, and more), it is clear that the revolution is not restricted to DNA sequence data.

A key challenge is therefore to integrate the signals at these disparate levels. For instance, while current genetic testing may suggest that a drug is unsafe for an individual patient, such classification may be refined by better understanding of epigenetic influences or metabolic activities[305]. Additionally, there is the trend to move from static data towards dynamic profiling. Sequencing the whole genome of a newborn reveals some risk markers, but profiles of other "omics" data over time may provide enhanced resolution of disease risk. Virtual Liver[306] and the Physiome project[307] seek to integrate bioinformatic signals across distinct biological levels and temporal timeframes. To provide clinically useful information it will be necessary to combine a variety of distinct heterogeneous data sources (e.g. DNA, RNA, protein, metabolites, environment) while still addressing the limitations of sample size, cost and incorporating better phenomic profiles.

8.4 Hardware and software revolution

As high-throughput technologies mature, data generation concerns recede and data interpretation challenges remain. For many diseases (e.g. acute neurodevelopmental disorders) the ability to diagnose and provide treatment before the onset of symptoms is critical[308]. This is especially true for neurological disorders in which the damage resulting from inappropriate formation over key developmental periods or the death of essential cells is irreversible[308]. Thus a key long-term concern is minimization of the time to diagnosis. While the identification of early biomarkers and the capacity to collect data earlier are both contributors, increasingly the complexity of the analysis over diverse classes of data and prediction of complex effects of changes are likely to result in computational time being amongst the greatest limitations, especially as technologies move into standard clinical practice and the number of deeply profiled patients grows by orders of magnitude. Specialized computational hardware may be required. Kingsmore *et al.* described a workflow requiring only 26 hours for providing diagnosis of genetic disorders using whole genome sequencing[309]. The reduced time is largely driven by the DRAGEN Bio-IT Processor, which adapts a graphical processing unit (GPU) via a PCIe form-factor card with accompanying analytical software. The graphics card can be integrated to next-generation sequencing servers, allowing the analysis of over 50 whole genomes from raw FASTQ data to VCF in a single day (http://www.edicogenome.com/dragen/). A similar study by Hall *et al*. published an open-source genome analysis platform that accomplishes alignment, variant detection and functional annotation of a 50X human genome in 13h on a low-cost server architecture[310]. In this case, the efficiency is achieved at the programming level via a superior processing of the alignment file and maximizing the simultaneous execution of nondependent pipeline components. As the technologies stabilize and the specific computational algorithms become standard, such hardware and software-based solutions become more practical.

In the near-term, as most hospitals do not posses high-performance computational infrastructure, cloud computing is being adopted to close the gap between data generation and data analysis. The EasyGenomics (http://www.easygenomics.com) internet service provided by the Beijing Genomics Institute exemplifies such an approach. Bioinformatics methods developers will need to allow for such infrastructure for methods that will be broadly used in applied genetics.

8.4.1 The role of industry

Throughout the thesis, much emphasis has been placed upon open-source software due to most commercial systems being unpublished and proprietary. Nonetheless, the role of industry in analyzing and interpreting variants is critical, with example applications such as Alamut (http://www.interactive-biosoftware.com/alamut-visual/) and Ingenuity Variant Analysis (http://www.ingenuity.com/products/variant-analysis). For a compilation of popular commercial applications, refer to table 4 by Klee *et al[311]*. Since the field of genomic medicine remains a highly dynamic area of research, multiple open source and commercial software solutions invariably exist for any single analysis step. The breadth of the field further means bioinformatics solutions are customized to perform under specific clinical contexts, the performance of pipelines becoming exquisitely sensitive to highly-tuned parameters. Presently, aside from the issue of cost and affordability, there is still a balance between innovation and stability when choosing between open-source bioinformatics software and commercial solutions. My research methodologies impact these efforts by placing emphasis upon software usability in the context of clinical users and their interactions with genomic data and related patient health information. In addition, my research results reveal that it is critical for software design to account for ever-growing multidisciplinary care team as healthcare industry shifts from single providers to integrative interpersonal networks. Finally, my research into variant prioritization model informs the NGS developers at large a need for better-automated clinical decision support in selecting candidate variant(s). Overall, the research presented in this thesis can inform industrial software development, informing developers that the design and implementation of systems need to take into account of human factors, with appropriate clinician involvement and attention to workflow and training.

8.5 Training clinicians to use genome sequence data

While new technologies are embedded within research programs, highly specialized analysts work with the data. During the process of translating the methods into standard practice, it becomes imperative to provide streamlined analysis results to clinicians and therefore to develop sufficient capacities in the clinic. A hybrid education is necessary to train new generation of scientists and physicians that are capable to understand the underlying biological problem, the methods of data analysis, interpretation of the data, and the advantages and disadvantages of the new technologies and analytical approaches. While much of this thesis emphasized the need for user-friendly software, it is equally critical for users to have the necessary education. For example, the eMERGE network found that the education of providers about the tools for delivering results about genetic risk is important to ensure the tools are used effectively[312-314].

8.5.1 Training programs

It is reported that 90% of scientists are self-taught in programming, but they often lack basic practices such as task automation, code review, unit testing, version control, and issue tracking[315]. In a survey of 68 specialist healthcare providers, 42% did not believe they had adequate preparation to implement personalized medicine in the clinical setting[315]. This highlights a need for educational programs to bridge the disciplines. Improved understanding of how data is used in informatics can indirectly help transform clinical research and practice. For instance, when clinicians better understand the potential of electronic health records, they develop better appreciation to take the time to fill out the electronic forms[316, 317]. Developments of online webinars and podcasts can be useful for directing busy clinicians to appropriate professional conferences and suitable primer literature. Due to the diverse multidisciplinary domains involved in the healthcare process, it is likely the training methods themselves would have to be tailored to each specialized domain. For instance, while online education methods may be preferred for meeting credential standards, behavioral strategies such as training in genetic counseling would not be conducive under distance-based methods. The training would also need to be tailored to particular user groups; for instance, the clinical geneticists and genetic counselors should receive training on the return of genomic results and secondary findings, while the effective use of clinical decisions support system for appropriate therapies should be more targeted towards nurses and primary physicians.

8.5.2 Patient engagement

The earlier chapters of the thesis highlight the need for usability analysis on healthcare providers interacting with genomics software to make treatment decisions based on the delivered outputs. A currently unexplored future direction is to include patients in the evaluation, instead of solely restricting the evaluation to healthcare providers. Engaging the patients in their healthcare can improve patient outcomes by reducing the imbalance of information that typically exists between healthcare providers and their patients. There are informatics opportunities to design systems to facilitate open communications between patients and clinicians. Outpatient portals such as Open Notes and BMT Roadmap demonstrate how software can facilitate engagement by delivering patient health records in a condensed, user-friendly format, improving the coordination of care and quality of patient-doctor communication[318].

8.6 Ethics

Thus far, this thesis has not touched upon the ethical considerations regarding genomic medicine. While the concept of personalized medicine is not new, the arrival of genomics and other high-throughput patient-specific data impacts social contexts in which the approach may be disadvantageous to some patients. Due to the large scope of the topic, I will point to key papers addressing major topics of exploration[319-325]. The responsibility for payment needs to be resolved, as there needs to be clear health benefit if insurers (or society) are to be obliged to provide for it. Should insurers be allowed to include genetic information in the determination of rates or eligibility for certain policies? How should the privacy of personalized health data be protected? As genetic information can be relevant to multiple members of a family, should there be protections for untested individuals? Which incidental findings should be reported and to whom? How should information be regulated in light of concerns about eugenics? If testing is performed in the earliest days of life, what information should be reserved for when a subject becomes an adult and indicates a desire to obtain it? What is the appropriate interaction with patients who refuse certain testing based on religious or political grounds? While there are few simple answers and some of the questions have been explored robustly outside of the specific context of genome medicine, it should be clear that in the coming years there will be broad discussions around the globe to explore these topics and develop standards. As DNA sequencing technology has arrived to the clinic earlier than expected, much of the exploration will be concurrent with implementation.

8.7 Conclusion

In the context of the Precision Medicine Initiative, United States President Barack Obama comprehensively observed:

> “delivering the right treatments, at the right time, every time to the right person. And for a small but growing number of patients, that future is already here…So if we combine all these emerging technologies, if we focus them and make sure that the connections are made, then the possibility of discovering new cures, the possibility of applying medicines more efficiently and more effectively so that the success rates are higher, so that there’s less waste in the system, which then means more resources to help more people – the possibilities are boundless”

To achieve these ambitious goals, computational improvements are necessary to integrate complex data and allow successful adoption of technologies within clinical settings. My thesis addresses this matter by tackling the usability challenges in genome analysis software targeted towards clinicians. Through a series of interface evaluation methodologies, my doctoral work presents various software interface prototypes desired by the clinical specialists to address domain-specific scenarios in the analysis of exomes and whole-genomes. Informed by the findings from design evaluations, I introduced a novel computational approach to better prioritize exome variants based on automated appraisal of patient phenotypes. The clinical impacts of my works can be seen in my clinical collaborations (achieving diagnosis rate of 38/47 families; 81%), which show how considerations of the designs and visualization of genomic information and algorithmic improvements in variant prioritization come together for improving patient care.

The coming future will require new generations of multidisciplinary research teams to design, operate and evaluate user-tailored tools for clinical decisions supports between informed patients and their physicians. As with any historical major human landmark achievement, the overcoming of obstacles to personalized medicine will require a concerted community effort. My thesis represents an important piece of work in facilitating collaborations between fields and can be continued to expand upon the future as the fields evolve.

Bibliography

1. Portin P: **The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA**. *Journal of genetics* 2014, **93**(1):293-302.
2. Orkin S: **Molecular Genetics of Chronic Granulomatous Disease**. *Annual Review of Immunology* 1989, **7**:277-307.
3. Siva N: **1000 Genomes project**. *Nature biotechnology* 2008, **26**(3):256.
4. Mueller J, Gazzoli I, Bandipalliam P, Garber JE, Syngal S, Kolodner RD: **Comprehensive molecular analysis of mismatch repair gene defects in suspected Lynch syndrome (hereditary nonpolyposis colorectal cancer) cases**. *Cancer research* 2009, **69**(17):7053-7061.
5. Davies JC, Alton EW, Bush A: **Cystic fibrosis**. *Bmj* 2007, **335**(7632):1255-1259.
6. Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, Baehner RL, Cole FS, Curnutte JT, Orkin SH: **Cloning the gene for an inherited human disorder--chronic granulomatous disease--on the basis of its chromosomal location**. *Nature* 1986, **322**(6074):32-38.
7. Ohlendieck K, Matsumura K, Ionasescu VV, Towbin JA, Bosch EP, Weinstein SL, Sernett SW, Campbell KP: **Duchenne muscular dystrophy: deficiency of dystrophin-associated proteins in the sarcolemma**. *Neurology* 1993, **43**(4):795-800.
8. Wissinger B, Besch D, Baumann B, Fauser S, Christ-Adler M, Jurklies B, Zrenner E, Leo-Kottler B: **Mutation analysis of the ND6 gene in patients with Lebers hereditary optic neuropathy**. *Biochemical and biophysical research communications* 1997, **234**(2):511-515.
9. Nathans D, Smith HO: **Restriction endonucleases in the analysis and restructuring of dna molecules**. *Annual review of biochemistry* 1975, **44**:273-293.
10. Manolio TA: **Genomewide association studies and assessment of the risk of disease**. *N Engl J Med* 2010, **363**(2):166-176.
11. Shaikh TH: **Oligonucleotide arrays for high-resolution analysis of copy number alteration in mental retardation/multiple congenital anomalies**. *Genet Med* 2007, **9**(9):617-625.
12. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J *et al*: **Strong association of de novo copy number mutations with autism**. *Science* 2007, **316**(5823):445-449.
13. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *Journal of molecular biology* 1975, **94**(3):441-448.
14. de Magalhaes JP, Finch CE, Janssens G: **Next-generation sequencing in aging research: emerging applications, problems, pitfalls and possible solutions**. *Ageing research reviews* 2010, **9**(3):315-323.
15. Wu GS, Zhang Y, Si W, Sha JJ, Liu L, Chen YF: **Integrated solid-state nanopore devices for third generation DNA sequencing**. *Sci China Technol Sc* 2014, **57**(10):1925-1935.

16. Kuehn BM: **1000 Genomes Project promises closer look at variation in human genome**. *Jama* 2008, **300**(23):2715.

17. Thorisson GA, Smith AV, Krishnan L, Stein LD: **The International HapMap Project Web site**. *Genome research* 2005, **15**(11):1592-1593.

18. Oetting WS: **Exome and genome analysis as a tool for disease identification and treatment: the 2011 Human Genome Variation Society scientific meeting**. *Human mutation* 2012, **33**(3):586-590.

19. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.

20. Sorte H, Morkrid L, Rodningen O, Kulseth MA, Stray-Pedersen A, Matthijs G, Race V, Houge G, Fiskerstrand T, Bjurulf B *et al*: **Severe ALG8-CDG (CDG-Ih) associated with homozygosity for two novel missense mutations detected by exome sequencing of candidate genes**. *Eur J Med Genet* 2012, **55**(3):196-202.

21. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC *et al*: **Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome**. *Nat Genet* 2010, **42**(9):790-793.

22. Bras JM, Singleton AB: **Exome sequencing in Parkinson's disease**. *Clin Genet* 2011, **80**(2):104-109.

23. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery**. *Nat Rev Genet* 2011, **12**(11):745-755.

24. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR: **A comparative analysis of exome capture**. *Genome Biol* 2011, **12**(9):R97.

25. Shendure J: **Next-generation human genetics**. *Genome biology* 2011, **12**(9):408.

26. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE *et al*: **Targeted capture and massively parallel sequencing of 12 human exomes**. *Nature* 2009, **461**(7261):272-276.

27. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M *et al*: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing**. *Science* 2010, **328**(5978):636-639.

28. Scacheri CA, Scacheri PC: **Mutations in the noncoding genome**. *Current opinion in pediatrics* 2015, **27**(6):659-664.

29. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

30. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

33. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3**. *Fly* 2012, **6**(2):80-92.

34. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research* 2010, **38**(16):e164.

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

36. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.

37. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nature genetics* 2014, **46**(3):310-315.

38. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nat Biotechnol* 2011, **29**(1):24-26.

39. Genomes Project C: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.

40. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation**. *Nucleic acids research* 2001, **29**(1):308-311.

41. Ng SB, Nickerson DA, Bamshad MJ, Shendure J: **Massively parallel sequencing and rare disease**. *Hum Mol Genet* 2010, **19**(R2):R119-124.

42. Cheung WA, Ouellette BF, Wasserman WW: **Compensating for literature annotation bias when predicting novel drug-disease relationships through Medical Subject Heading Over-representation Profile (MeSHOP) similarity**. *BMC medical genomics* 2013, **6 Suppl 2**:S3.

43. Ricevuto E, Sobol H, Stoppa-Lyonnet D, Gulino A, Marchetti P, Ficorella C, Martinotti S, Meo T, Tosi M: **Diagnostic strategy for analytical scanning of BRCA1 gene by fluorescence-assisted mismatch analysis using large, bifluorescently labeled amplicons**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2001, **7**(6):1638-1646.

44. Lander ES: **Initial impact of the sequencing of the human genome**. *Nature* 2011, **470**(7333):187-197.

45. Graham EA: **DNA reviews: predicting phenotype**. *Forensic Sci Med Pathol* 2008, **4**(3):196-199.

46. Sirrs S, van Karnebeek CD, Peng X, Shyr C, Tarailo-Graovac M, Mandal R, Testa D, Dubin D, Carbonetti G, Glynn SE *et al*: **Defects in fatty acid amide hydrolase 2 in a male with neurologic and psychiatric symptoms**. *Orphanet journal of rare diseases* 2015, **10**:38.

47. Janer A, van Karnebeek CD, Sasarman F, Antonicka H, Al Ghamdi M, Shyr C, Dunbar M, Stockler-Ispiroglu S, Ross CJ, Vallance H *et al*: **RMND1 deficiency associated with neonatal lactic acidosis, infantile onset renal failure, deafness, and multiorgan involvement**. *European journal of human genetics : EJHG* 2015, **23**(10):1301-1307.

48. Stockler S, Corvera S, Lambright D, Fogarty K, Nosova E, Leonard D, Steinfeld R, Ackerley C, Shyr C, Au N *et al*: **Single point mutation in Rabenosyn-5 in a female with intractable seizures and evidence of defective endocytotic trafficking**. *Orphanet journal of rare diseases* 2014, **9**:141.

49. Armstrong L, Biancheri R, Shyr C, Rossi A, Sinclair G, Ross CJ, Tarailo-Graovac M, Wasserman WW, van Karnebeek CD: **AIMP1 deficiency presents as a cortical neurodegenerative disease with infantile onset**. *Neurogenetics* 2014, **15**(3):157-159.

50. Green ED, Guyer MS, National Human Genome Research I: **Charting a course for genomic medicine from base pairs to bedside**. *Nature* 2011, **470**(7333):204-213.

51. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM *et al*: **The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies**. *BMC medical genomics* 2011, **4**:13.

52. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, Spurr C, Khorasani R, Tanasijevic M, Middleton B: **Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality**. *J Am Med Inform Assoc* 2003, **10**(6):523-530.

53. Hunt DL, Haynes RB, Hanna SE, Smith K: **Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review**. *JAMA* 1998, **280**(15):1339-1346.

54. David Lobach AB, Caitlin Houlihan, Kensaku Kawamoto: **Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success**. *BMJ* 2005.

55. J. Emery JF, D. W. Glasspool, A.S. Coulson: **RAGs: A novel approach to computerized genetic risk assessment and decision support from pedigrees**. *Methods of Information in Medicine* 2001.

56. Hou H, Zhao F, Zhou L, Zhu E, Teng H, Li X, Bao Q, Wu J, Sun Z: **MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation**. *Nucleic acids research* 2010, **38**(Web Server issue):W732-736.

57. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy T: **Manipulation of FASTQ data with Galaxy**. *Bioinformatics* 2010, **26**(14):1783-1785.

58. Hawkins RD, Hon GC, Ren B: **Next-generation genomics: an integrative approach**. *Nat Rev Genet* 2010, **11**(7):476-486.

59. Ge D, Ruzzo EK, Shianna KV, He M, Pelak K, Heinzen EL, Need AC, Cirulli ET, Maia JM, Dickson SP *et al*: **SVA: software for annotating and visualizing sequenced human genomes**. *Bioinformatics* 2011, **27**(14):1998-2000.

60. Teer JK, Green ED, Mullikin JC, Biesecker LG: **VarSifter: visualizing and analyzing exome-scale sequence variation data on a desktop computer**. *Bioinformatics* 2012, **28**(4):599-600.

61. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG: **A probabilistic disease-gene finder for personal genomes**. *Genome Res* 2011, **21**(9):1529-1542.

62. Graham TA, Kushniruk AW, Bullard MJ, Holroyd BR, Meurer DP, Rowe BH: **How usability of a web-based clinical decision support system has the potential to contribute to adverse medical events**. *AMIA Annu Symp Proc* 2008:257-261.

63. Judith Olson GO: **Human-Computer Interaction: Psychological aspects of the human use of computing**. *Annu Rev Psychol* 2003.

64. Kawamoto K, Lobach DF, Willard HF, Ginsburg GS: **A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine**. *BMC Med Inform Decis Mak* 2009, **9**:17.

65. Kushniruk A: **Evaluation in the design of health information systems: application of approaches emerging from usability engineering**. *Computers in biology and medicine* 2002, **32**(3):141-149.

66. Kushniruk AW, Patel VL: **Cognitive and usability engineering methods for the evaluation of clinical information systems**. *Journal of biomedical informatics* 2004, **37**(1):56-76.

67. Nielsen J: **Usability engineering**. Boston: Academic Press; 1993.

68. Juristo N: **Guidelines for Eliciting Usability Functionalities**. *Software Engineering, IEEE Transactions on* 2007, **33**(11):744 - 758.

69. Robert Hoffman BC, Nigel Shadbolt: **Use of the Critical Decision Method to Elicit Expert Knowledge: A Case Study in the Methodology of Cognitive Task Analysis**. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 1998, **40**(2):254-276.

70. Leanne Currie PS, David Kaufman, Barbara Sheehan: **Cognitive Analysis of Decision Support for Antibiotic Prescribing at the Point of Ordering in a Neonatal Intensive Care Unit**. *AMIA Annu Symp Proc* 2009.

71. Kushniruk AW, Santos SL, Pourakis G, Nebeker JR, Boockvar KS: **Cognitive analysis of a medication reconciliation tool: applying laboratory and naturalistic approaches to system evaluation**. *Stud Health Technol Inform* 2011, **164**:203-207.

72. Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, Mann DM: **Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support**. *Int J Med Inform* 2012.

73. van Karnebeek CD, Stockler S: **Treatable inborn errors of metabolism causing intellectual disability: a systematic literature review**. *Mol Genet Metab* 2012, **105**(3):368-381.

74. Kim J, Lee YG, Kim N: **Bioinformatics interpretation of exome sequencing: blood cancer**. *Genomics Inform* 2013, **11**(1):24-33.

75. Hinchcliffe M, Webster P: **In silico analysis of the exome for gene discovery**. *Methods Mol Biol* 2011, **760**:109-128.

76. Ionita-Laza I, Makarov V, Yoon S, Raby B, Buxbaum J, Nicolae DL, Lin X: **Finding disease variants in Mendelian disorders by using sequence data: methods and applications**. *Am J Hum Genet* 2011, **89**(6):701-712.

77. Minshall S: **A review of healthcare information system usability & safety**. *Stud Health Technol Inform* 2013, **183**:151-156.

78. Bronnert J, Masarie C, Naeymi-Rad F, Rose E, Aldin G: **Problem-centered care delivery: how interface terminology makes standardized health information possible**. *J AHIMA* 2012, **83**(7):30-35; quiz 36.

79. Theobald S, Nhlema-Simwaka B: **The research, policy and practice interface: reflections on using applied social research to promote equity in health in Malawi**. *Soc Sci Med* 2008, **67**(5):760-770.
80. Marcilly R, Bernonville S, Riccioli C, Beuscart-Zephir MC: **Patient safety-oriented usability testing: a pilot study**. *Stud Health Technol Inform* 2012, **180**:368-372.
81. Yen PY, Bakken S: **Review of health information technology usability study methodologies**. *J Am Med Inform Assoc* 2012, **19**(3):413-422.
82. Coiera E, Westbrook J, Wyatt J: **The safety and quality of decision support systems**. *Yearb Med Inform* 2006:20-25.
83. Zhang Z, Wang B, Ahmed F, Ramakrishnan I, Zhao R, Viccellio A, Mueller K: **The Five W's for Information Visualization with Application to Healthcare Informatics**. *IEEE Trans Vis Comput Graph* 2013.
84. Elkin PL: **Human Factors Engineering in HI: So What? Who Cares? and What's in It for You?** *Healthc Inform Res* 2012, **18**(4):237-241.
85. Saleem JJ, Flanagan ME, Wilck NR, Demetriades J, Doebbeling BN: **The next-generation electronic health record: perspectives of key leaders from the US Department of Veterans Affairs**. *J Am Med Inform Assoc* 2013, **20**(e1):e175-177.
86. Jaderlund Hagstedt L, Rudebeck CE, Petersson G: **Usability of computerised physician order entry in primary care: assessing ePrescribing with a new evaluation model**. *Inform Prim Care* 2011, **19**(3):161-168.
87. Bolchini D, Finkelstein A, Perrone V, Nagl S: **Better bioinformatics through usability analysis**. *Bioinformatics* 2009, **25**(3):406-412.
88. Neri PM, Pollard SE, Volk LA, Newmark LP, Varugheese M, Baxter S, Aronson SJ, Rehm HL, Bates DW: **Usability of a novel clinician interface for genetic results**. *Journal of biomedical informatics* 2012, **45**(5):950-957.
89. Nygren E, Johnson M, Henriksson P: **Reading the medical record. II. Design of a human-computer interface for basic reading of computerized medical records**. *Computer methods and programs in biomedicine* 1992, **39**(1-2):13-25.
90. Nygren E, Henriksson P: **Reading the medical record. I. Analysis of physicians' ways of reading the medical record**. *Computer methods and programs in biomedicine* 1992, **39**(1-2):1-12.
91. Hripcsak G, Albers DJ: **Next-generation phenotyping of electronic health records**. *Journal of the American Medical Informatics Association : JAMIA* 2013, **20**(1):117-121.
92. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, Middleton B, Bates DW: **Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records**. *Journal of the American Medical Informatics Association : JAMIA* 2013, **20**(3):489-493.
93. Thyvalikakath TP, Dziabiak MP, Johnson R, Torres-Urquidy MH, Acharya A, Yabes J, Schleyer TK: **Advancing cognitive engineering methods to support user interface design for electronic health records**. *International journal of medical informatics* 2014, **83**(4):292-302.
94. Fonteyn M, Fisher A: **Use of think aloud method to study nurses' reasoning and decision making in clinical practice settings**. *The Journal of neuroscience nursing : journal of the American Association of Neuroscience Nurses* 1995, **27**(2):124-128.

95. Jaspers MW, Steen T, van den Bos C, Geenen M: **The think aloud method: a guide to user interface design**. *International journal of medical informatics* 2004, **73**(11-12):781-795.

96. Falk MJ, Pierce EA, Consugar M, Xie MH, Guadalupe M, Hardy O, Rappaport EF, Wallace DC, LeProust E, Gai X: **Mitochondrial disease genetic diagnostics: optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome**. *Discov Med* 2012, **14**(79):389-399.

97. Kirakowski J, Corbett M: **Sumi - the Software Usability Measurement Inventory**. *Brit J Educ Technol* 1993, **24**(3):210-212.

98. Kirakowski J: **Is ergonomics empirical?** *Ergonomics* 2002, **45**(14):995-997; discussion 1042-1046.

99. Kirakowski J: **'The Software Usability Measurement Inventory: Background and Usage**. Taylor and Frances, London, UK.: In: P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), Usability Evaluation in Industry. ; 1995.

100. Kushniruk A, Turner P: **A framework for user involvement and context in the design and development of safe e-Health systems**. *Stud Health Technol Inform* 2012, **180**:353-357.

101. Kushniruk AW, Borycki EM, Kuwata S, Kannry J: **Emerging approaches to usability evaluation of health information systems: towards in-situ analysis of complex healthcare systems and environments**. *Studies in health technology and informatics* 2011, **169**:915-919.

102. Amini A, Shrimpton PJ, Muggleton SH, Sternberg MJ: **A general approach for developing system-specific functions to score protein-ligand docked complexes using support vector inductive logic programming**. *Proteins* 2007, **69**(4):823-831.

103. Warr WA: **Scientific workflow systems: Pipeline Pilot and KNIME**. *Journal of computer-aided molecular design* 2012, **26**(7):801-804.

104. Assimakopoulos NA: **Workflow management with systems approach: anticipated and ad-hoc workflow for scientific applications**. *ISA transactions* 2000, **39**(2):153-167.

105. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine**. *Human genetics* 2014, **133**(1):1-9.

106. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR: **ClinVar: public archive of relationships among sequence variation and human phenotype**. *Nucleic acids research* 2014, **42**(Database issue):D980-985.

107. Coonrod EM, Margraf RL, Voelkerding KV: **Translating exome sequencing from research to clinical diagnostics**. *Clin Chem Lab Med* 2012, **50**(7):1161-1168.

108. Yu Y, Wu BL, Wu J, Shen Y: **Exome and whole-genome sequencing as clinical tests: a transformative practice in molecular diagnostics**. *Clin Chem* 2012, **58**(11):1507-1509.

109. Dimmock D: **Whole genome sequencing: a considered approach to clinical implementation**. *Curr Protoc Hum Genet* 2013, **Chapter 9**:Unit9 22.

110. Bao R, Huang L, Andrade J, Tan W, Kibbe WA, Jiang H, Feng G: **Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing**. *Cancer informatics* 2014, **13**(Suppl 2):67-82.

111. Welch BM, Kawamoto K: **Clinical decision support for genetically guided personalized medicine: a systematic review**. *Journal of the American Medical Informatics Association : JAMIA* 2013, **20**(2):388-400.

112. Biesecker LG, Green RC: **Diagnostic clinical genome and exome sequencing**. *The New England journal of medicine* 2014, **371**(12):1170.

113. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: **A survey of tools for variant analysis of next-generation genome sequencing data**. *Briefings in bioinformatics* 2014, **15**(2):256-278.

114. Shyr C, Kushniruk A, Wasserman WW: **Usability study of clinical exome analysis software: top lessons learned and recommendations**. *Journal of biomedical informatics* 2014, **51**:129-136.

115. Gray SW, Martins Y, Feuerman LZ, Bernhardt BA, Biesecker BB, Christensen KD, Joffe S, Rini C, Veenstra D, McGuire AL *et al*: **Social and behavioral research in genomic sequencing: approaches from the Clinical Sequencing Exploratory Research Consortium Outcomes and Measures Working Group**. *Genetics in medicine : official journal of the American College of Medical Genetics* 2014, **16**(10):727-735.

116. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, Merker JD, Goldfeder RL, Enns GM, David SP *et al*: **Clinical interpretation and implications of whole-genome sequencing**. *Jama* 2014, **311**(10):1035-1045.

117. Huser V, Sincan M, Cimino JJ: **Developing genomic knowledge bases and databases to support clinical management: current perspectives**. *Pharmacogenomics and personalized medicine* 2014, **7**:275-283.

118. Denzin NK, Lincoln YS: **The Sage handbook of qualitative research**, 4th edn. Thousand Oaks: Sage; 2011.

119. Crabtree BF, Miller WL: **Doing qualitative research**, 2nd edn. Thousand Oaks, Calif.: Sage Publications; 1999.

120. Kushniruk AW, Patel VL, Cimino JJ: **Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces**. *Proceedings : a conference of the American Medical Informatics Association / AMIA Annual Fall Symposium AMIA Fall Symposium* 1997:218-222.

121. Daniels J, Fels S, Kushniruk A, Lim J, Ansermino JM: **A framework for evaluating usability of clinical monitoring technology**. *Journal of clinical monitoring and computing* 2007, **21**(5):323-330.

122. Morgan DL: **Qualitative content analysis: a guide to paths not taken**. *Qualitative health research* 1993, **3**(1):112-121.

123. Vaismoradi M, Turunen H, Bondas T: **Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study**. *Nursing & health sciences* 2013, **15**(3):398-405.

124. Moretti F, van Vliet L, Bensing J, Deledda G, Mazzi M, Rimondini M, Zimmermann C, Fletcher I: **A standardized approach to qualitative content analysis of focus group discussions from different countries**. *Patient education and counseling* 2011, **82**(3):420-428.

125. Oliver GR, Hart SN, Klee EW: **Bioinformatics for Clinical Next Generation Sequencing**. *Clinical chemistry* 2015, **61**(1):124-135.

126. Glusman G, Cox HC, Roach JC: **Whole-genome haplotyping approaches and genomic medicine**. *Genome medicine* 2014, **6**(9):73.

127. Carson AR, Smith EN, Matsui H, Braekkan SK, Jepsen K, Hansen JB, Frazer KA: **Effective filtering strategies to improve data quality from population-based whole exome sequencing studies**. *BMC bioinformatics* 2014, **15**.

128. Maranhao B, Biswas P, Duncan JL, Branham KE, Silva GA, Naeem MA, Khan SN, Riazuddin S, Hejtmancik JF, Heckenlively JR *et al*: **exomeSuite: Whole exome sequence variant filtering tool for rapid identification of putative disease causing SNVs/indels**. *Genomics* 2014, **103**(2-3):169-176.

129. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M *et al*: **The UCSC Genome Browser database: 2015 update**. *Nucleic acids research* 2014.

130. Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA: **Genie: literature-based gene prioritization at multi genomic scale**. *Nucleic acids research* 2011, **39**(Web Server issue):W455-461.

131. Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, Bouffard GG, Chines PS, Cruz P, Hansen NF *et al*: **The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine**. *Genome research* 2009, **19**(9):1665-1674.

132. Sekiyama K, Takamatsu Y, Waragai M, Hashimoto M: **Role of genomics in translational research for Parkinson's disease**. *Biochemical and biophysical research communications* 2014, **452**(2):226-235.

133. Razzouk S: **Translational genomics and head and neck cancer: toward precision medicine**. *Clinical genetics* 2014, **86**(5):412-421.

134. Jung S, Main D: **Genomics and bioinformatics resources for translational science in Rosaceae**. *Plant biotechnology reports* 2014, **8**:49-64.

135. Henderson GE, Wolf SM, Kuczynski KJ, Joffe S, Sharp RR, Parsons DW, Knoppers BM, Yu JH, Appelbaum PS: **The challenge of informed consent and return of results in translational genomics: empirical analysis and recommendations**. *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics* 2014, **42**(3):344-355.

136. Muenke M: **Individualized genomics and the future of translational medicine**. *Molecular genetics & genomic medicine* 2013, **1**(1):1-3.

137. Pilbrow A, Arora P, Martinez-Fernandez A: **Top advances in functional genomics and translational biology for 2013**. *Circulation Cardiovascular genetics* 2014, **7**(1):89-92.

138. Zimmern RL, Brice PC: **Realizing the potential of genomics: translation is not translational research**. *Genetics in medicine : official journal of the American College of Medical Genetics* 2009, **11**(12):898-899; author reply 899.

139. Yu JH, Crouch J, Jamal SM, Bamshad MJ, Tabor HK: **Attitudes of non-African American focus group participants toward return of results from exome and whole genome sequencing**. *American journal of medical genetics Part A* 2014, **164A**(9):2153-2160.

140. Wright MF, Lewis KL, Fisher TC, Hooker GW, Emanuel TE, Biesecker LG, Biesecker BB: **Preferences for results delivery from exome sequencing/genome sequencing**.

Genetics in medicine : official journal of the American College of Medical Genetics 2014, **16**(6):442-447.

141. Atkinson NL, Massett HA, Mylks C, Hanna B, Deering MJ, Hesse BW: **User-centered research on breast cancer patient needs and preferences of an Internet-based clinical trial matching system**. *Journal of medical Internet research* 2007, **9**(2):e13.

142. Mirkovic J, Kaufman DR, Ruland CM: **Supporting cancer patients in illness management: usability evaluation of a mobile app**. *JMIR mHealth and uHealth* 2014, **2**(3):e33.

143. Amland RC, Hahn-Cover KE: **Clinical Decision Support for Early Recognition of Sepsis**. *American journal of medical quality : the official journal of the American College of Medical Quality* 2014.

144. Gierl L, Steffen D, Ihracky D, Schmidt R: **Methods, architecture, evaluation and usability of a case-based antibiotics advisor**. *Computer methods and programs in biomedicine* 2003, **72**(2):139-154.

145. Xie M, Weinger MB, Gregg WM, Johnson KB: **Presenting multiple drug alerts in an ambulatory electronic prescribing system: a usability study of novel prototypes**. *Applied clinical informatics* 2014, **5**(2):334-348.

146. Kohli M, Dreyer KJ, Geis JR: **Rethinking radiology informatics**. *AJR American journal of roentgenology* 2015, **204**(4):716-720.

147. Rosenkrantz AB, Doshi AM: **Continued evolution of clinical decision support tools for guiding imaging utilization**. *Academic radiology* 2015, **22**(4):542-543.

148. Hovenga EJ, Grain H: **Health information systems**. *Studies in health technology and informatics* 2013, **193**:120-140.

149. Yuan MJ, Finley GM, Long J, Mills C, Johnson RK: **Evaluation of user interface and workflow design of a bedside nursing clinical decision support system**. *Interactive journal of medical research* 2013, **2**(1):e4.

150. Horsky J, Phansalkar S, Desai A, Bell D, Middleton B: **Design of decision support interventions for medication prescribing**. *International journal of medical informatics* 2013, **82**(6):492-503.

151. Sacchi L, Fux A, Napolitano C, Panzarasa S, Peleg M, Quaglini S, Shalom E, Soffer P, Tormene P: **Patient-tailored workflow patterns from clinical practice guidelines recommendations**. *Studies in health technology and informatics* 2013, **192**:392-396.

152. Luo J, Liang S: **Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data**. *Journal of biomedical informatics* 2014.

153. Masino AJ, Dechene ET, Dulik MC, Wilkens A, Spinner NB, Krantz ID, Pennington JW, Robinson PN, White PS: **Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology**. *BMC bioinformatics* 2014, **15**:248.

154. Fatahi N, Krupic F, Hellstrom M: **Quality of radiologists' communication with other clinicians-As experienced by radiologists**. *Patient education and counseling* 2015.

155. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, The Mouse Genome Database G: **The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease**. *Nucleic acids research* 2014.

156. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE: **Linking human diseases to animal models using ontology-based phenotype annotation**. *PLoS biology* 2009, **7**(11):e1000247.

157. Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y: **KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model**. *Molecular bioSystems* 2014, **10**(9):2441-2447.

158. Roncaglia P, Martone ME, Hill DP, Berardini TZ, Foulger RE, Imam FT, Drabkin H, Mungall CJ, Lomax J: **The Gene Ontology (GO) Cellular Component Ontology: integration with SAO (Subcellular Anatomy Ontology) and other recent developments**. *Journal of biomedical semantics* 2013, **4**(1):20.

159. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A: **OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders**. *Nucleic acids research* 2014.

160. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN: **The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution**. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2012, **Chapter 1**:Unit1 13.

161. Dib-Hajj SD, Waxman SG: **Translational pain research: Lessons from genetics and genomics**. *Science translational medicine* 2014, **6**(249):249sr244.

162. Overby CL, Tarczy-Hornoch P: **Personalized medicine: challenges and opportunities for translational bioinformatics**. *Personalized medicine* 2013, **10**(5):453-462.

163. Fitzpatrick G, Ellingsen G: **A Review of 25 Years of CSCW Research in Healthcare: Contributions, Challenges and Future Agendas**. *Comput Supp Coop W J* 2013, **22**(4-6):609-665.

164. Cavusoglu H, Frisch L, Fels S: **Sociotechnical Challenges and Progress in Using Social Media for Health**. *Journal of medical Internet research* 2013, **15**(10):1-1.

165. Monsted T, Reddy MC, Bansler JP: **The Use of Narratives in Medical Work: A Field Study of Physician-Patient Consultations**. *Ecscw 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work* 2011:81-100.

166. Papagelis M, Rousidis I, Plexousakis D, Theoharopoulos E: **Incremental collaborative filtering for highly-scalable recommendation algorithms**. *Foundations of Intelligent Systems, Proceedings* 2005, **3488**:553-561.

167. Cain J: **Online social networking issues within academia and pharmacy education**. *Am J Pharm Educ* 2008, **72**(1).

168. Montserrat Moliner A, Waligora J: **The European union policy in the field of rare diseases**. *Public health genomics* 2013, **16**(6):268-277.

169. Baird PA, Anderson TW, Newcombe HB, Lowry RB: **Genetic disorders in children and young adults: a population study**. *American journal of human genetics* 1988, **42**(5):677-693.

170. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM**. *American journal of human genetics* 2007, **80**(4):588-604.

171. Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC: **[Orphanet: a European database for rare diseases]**. *Nederlands tijdschrift voor geneeskunde* 2008, **152**(9):518-519.

172. Samuels ME: **Saturation of the human phenome**. *Current genomics* 2010, **11**(7):482-499.

173. Cooper DN, Chen JM, Ball EV, Howells K, Mort M, Phillips AD, Chuzhanova N, Krawczak M, Kehrer-Sawatzki H, Stenson PD: **Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics**. *Human mutation* 2010, **31**(6):631-655.

174. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB *et al*: **A systematic survey of loss-of-function variants in human protein-coding genes**. *Science* 2012, **335**(6070):823-828.

175. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nature protocols* 2009, **4**(7):1073-1081.

176. Adzhubei I, Jordan DM, Sunyaev SR: **Predicting functional effect of human missense mutations using PolyPhen-2**. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]* 2013, **Chapter 7**:Unit7 20.

177. Gill N, Singh S, Aseri TC: **Computational disease gene prioritization: an appraisal**. *Journal of computational biology : a journal of computational molecular cell biology* 2014, **21**(6):456-465.

178. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB: **Genic intolerance to functional variation and the interpretation of personal genomes**. *PLoS genetics* 2013, **9**(8):e1003709.

179. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA: **Genenames.org: the HGNC resources in 2013**. *Nucleic acids research* 2013, **41**(Database issue):D545-552.

180. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE: **Rare-disease genetics in the era of next-generation sequencing: discovery to translation**. *Nature reviews Genetics* 2013, **14**(10):681-691.

181. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al*: **Ensembl 2014**. *Nucleic acids research* 2014, **42**(Database issue):D749-755.

182. Piton A, Redin C, Mandel JL: **XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing**. *American journal of human genetics* 2013, **93**(2):368-383.

183. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J *et al*: **The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data**. *Nucleic acids research* 2014, **42**(Database issue):D966-974.

184. van Karnebeek CD, Shevell M, Zschocke J, Moeschler JB, Stockler S: **The metabolic evaluation of the child with an intellectual developmental disorder: diagnostic algorithm for identification of treatable causes and new digital resource**. *Mol Genet Metab* 2014, **111**(4):428-438.

185. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL *et al*: **De novo mutations revealed by whole-exome sequencing are strongly associated with autism**. *Nature* 2012, **485**(7397):237-241.

186. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V *et al*: **Patterns and rates of exonic de novo mutations in autism spectrum disorders**. *Nature* 2012, **485**(7397):242-245.

187. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD *et al*: **Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations**. *Nature* 2012, **485**(7397):246-250.

188. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A *et al*: **De novo gene disruptions in children on the autistic spectrum**. *Neuron* 2012, **74**(2):285-299.

189. Chen WH, Zhao XM, van Noort V, Bork P: **Human monogenic disease genes have frequently functionally redundant paralogs**. *PLoS computational biology* 2013, **9**(5):e1003073.

190. Castellana S, Mazza T: **Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools**. *Briefings in bioinformatics* 2013, **14**(4):448-459.

191. Stearns FW: **One hundred years of pleiotropy: a retrospective**. *Genetics* 2010, **186**(3):767-773.

192. Yamaguchi T, Sharma P, Athanasiou M, Kumar A, Yamada S, Kuehn MR: **Mutation of SENP1/SuPr-2 reveals an essential role for desumoylation in mouse development**. *Molecular and cellular biology* 2005, **25**(12):5171-5182.

193. Yu L, Ji W, Zhang H, Renda MJ, He Y, Lin S, Cheng EC, Chen H, Krause DS, Min W: **SENP1-mediated GATA1 deSUMOylation is critical for definitive erythropoiesis**. *The Journal of experimental medicine* 2010, **207**(6):1183-1195.

194. Van Nguyen T, Angkasekwinai P, Dou H, Lin FM, Lu LS, Cheng J, Chin YE, Dong C, Yeh ET: **SUMO-specific protease 1 is critical for early lymphoid development through regulation of STAT5 activation**. *Molecular cell* 2012, **45**(2):210-221.

195. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER *et al*: **MuSiC: identifying mutational significance in cancer genomes**. *Genome research* 2012, **22**(8):1589-1598.

196. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA *et al*: **Mutational heterogeneity in cancer and the search for new cancer-associated genes**. *Nature* 2013, **499**(7457):214-218.

197. Moreau Y, Tranchevent LC: **Computational tools for prioritizing candidate genes: boosting disease gene discovery**. *Nature reviews Genetics* 2012, **13**(8):523-536.

198. Micale L, Augello B, Maffeo C, Selicorni A, Zucchetti F, Fusco C, De Nittis P, Pellico MT, Mandriani B, Fischetto R *et al*: **Molecular analysis, pathogenic mechanisms, and readthrough therapy on a large cohort of Kabuki syndrome patients**. *Human mutation* 2014, **35**(7):841-850.

199. Schulz Y, Freese L, Manz J, Zoll B, Volter C, Brockmann K, Bogershausen N, Becker J, Wollnik B, Pauli S: **CHARGE and Kabuki syndromes: a phenotypic and molecular link**. *Human molecular genetics* 2014, **23**(16):4396-4405.

200. Cheon CK, Sohn YB, Ko JM, Lee YJ, Song JS, Moon JW, Yang BK, Ha IS, Bae EJ, Jin HS *et al*: **Identification of KMT2D and KDM6A mutations by exome sequencing in Korean patients with Kabuki syndrome**. *Journal of human genetics* 2014, **59**(6):321-325.

201. Puffenberger EG, Jinks RN, Wang H, Xin B, Fiorentini C, Sherman EA, Degrazio D, Shaw C, Sougnez C, Cibulskis K *et al*: **A homozygous missense mutation in HERC2 associated with global developmental delay and autism spectrum disorder**. *Human mutation* 2012, **33**(12):1639-1646.

202. Bohm J, Leshinsky-Silver E, Vassilopoulos S, Le Gras S, Lerman-Sagie T, Ginzberg M, Jost B, Lev D, Laporte J: **Samaritan myopathy, an ultimately benign congenital myopathy, is caused by a RYR1 mutation**. *Acta neuropathologica* 2012, **124**(4):575-581.

203. Harlalka GV, Baple EL, Cross H, Kuhnle S, Cubillos-Rojas M, Matentzoglu K, Patton MA, Wagner K, Coblentz R, Ford DL *et al*: **Mutation of HERC2 causes developmental delay with Angelman-like features**. *Journal of medical genetics* 2013, **50**(2):65-73.

204. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA *et al*: **Guidelines for investigating causality of sequence variants in human disease**. *Nature* 2014, **508**(7497):469-476.

205. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome research* 2005, **15**(8):1034-1050.

206. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies**. *Genome research* 2010, **20**(1):110-121.

207. Jung JY, DeLuca TF, Nelson TH, Wall DP: **A literature search tool for intelligent extraction of disease-associated genes**. *Journal of the American Medical Informatics Association : JAMIA* 2014, **21**(3):399-405.

208. Xu R, Li L, Wang Q: **Towards building a disease-phenotype knowledge base: extracting disease-manifestation relationship from literature**. *Bioinformatics* 2013, **29**(17):2186-2194.

209. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ: **A navigator for human genome epidemiology**. *Nature genetics* 2008, **40**(2):124-125.

210. Kocot KM, Citarella MR, Moroz LL, Halanych KM: **PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics**. *Evolutionary bioinformatics online* 2013, **9**:429-435.

211. Altenhoff AM, Dessimoz C: **Inferring orthology and paralogy**. *Methods in molecular biology* 2012, **855**:259-279.

212. Pryszcz LP, Huerta-Cepas J, Gabaldon T: **MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score**. *Nucleic acids research* 2011, **39**(5):e32.

213. Samuels DC, Han L, Li J, Quanghu S, Clark TA, Shyr Y, Guo Y: **Finding the lost treasures in exome sequencing data**. *Trends in genetics : TIG* 2013, **29**(10):593-599.

214. Shyr C, Kushniruk A, van Karnebeek CD, Wasserman WW: **Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists, and genetic counselors**. *Journal of the American Medical Informatics Association : JAMIA* 2015.

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

216. Robinson PN, Mundlos S: **The human phenotype ontology**. *Clinical genetics* 2010, **77**(6):525-534.

217. Kohler S, Schulz MH, Krawitz P, Bauer S, Dolken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN: **Clinical diagnostics in human genetics with semantic similarity searches in ontologies**. *American journal of human genetics* 2009, **85**(4):457-464.

218. Robinson PN, Kohler S, Oellrich A, Sanger Mouse Genetics P, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D *et al*: **Improved exome prioritization of disease genes through cross-species phenotype comparison**. *Genome research* 2014, **24**(2):340-348.

219. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, Durtschi J, Eilbeck K, Reese MG, Jorde LB *et al*: **Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families**. *American journal of human genetics* 2014, **94**(4):599-610.

220. Smith CL, Eppig JT: **The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data**. *Mammalian genome : official journal of the International Mammalian Genome Society* 2012, **23**(9-10):653-668.

221. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D *et al*: **Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data**. *Nucleic acids research* 2015, **43**(Database issue):D1071-1078.

222. Karakoc E, Alkan C, O'Roak BJ, Dennis MY, Vives L, Mark K, Rieder MJ, Nickerson DA, Eichler EE: **Detection of structural variants and indels within exome data**. *Nature methods* 2012, **9**(2):176-178.

223. Villanueva A, Willer JR, Bryois J, Dermitzakis ET, Katsanis N, Davis EE: **Whole exome sequencing of a dominant retinitis pigmentosa family identifies a novel deletion in PRPF31**. *Investigative ophthalmology & visual science* 2014, **55**(4):2121-2129.

224. Ogata T, Niihori T, Tanaka N, Kawai M, Nagashima T, Funayama R, Nakayama K, Nakashima S, Kato F, Fukami M *et al*: **TBX1 mutation identified by exome sequencing in a Japanese family with 22q11.2 deletion syndrome-like craniofacial features and hypocalcemia**. *PloS one* 2014, **9**(3):e91598.

225. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova JL, Abel L: **Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants**. *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(17):5473-5478.

226. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.

227. Sanli K, Karlsson FH, Nookaew I, Nielsen J: **FANTOM: Functional and taxonomic analysis of metagenomes**. *BMC bioinformatics* 2013, **14**:38.

228. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJ, van Karnebeek C, Wasserman WW: **FLAGS, frequently mutated genes in public exomes**. *BMC medical genomics* 2014, **7**:64.

229. Richesson RL, Horvath MM, Rusincovitch SA: **Clinical research informatics and electronic health record data**. *Yearbook of medical informatics* 2014, **9**(1):215-223.

230. Shemeikka T, Bastholm-Rahmner P, Elinder CG, Veg A, Tornqvist E, Cornelius B, Korkmaz S: **A health record integrated clinical decision support system to support prescriptions of pharmaceutical drugs in patients with reduced renal function: design, development and proof of concept**. *International journal of medical informatics* 2015, **84**(6):387-395.

231. Shaikh U, Berrong J, Nettiksimmons J, Byrd RS: **Impact of electronic health record clinical decision support on the management of pediatric obesity**. *American journal of medical quality : the official journal of the American College of Medical Quality* 2015, **30**(1):72-80.

232. Monsen KA, Finn RS, Fleming TE, Garner EJ, LaValla AJ, Riemer JG: **Rigor in electronic health record knowledge representation: lessons learned from a SNOMED CT clinical content encoding exercise**. *Informatics for health & social care* 2014:1-15.

233. Gottlieb MM, Arenillas DJ, Maithripala S, Maurer ZD, Tarailo Graovac M, Armstrong L, Patel M, van Karnebeek C, Wasserman WW: **GeneYenta: a phenotype-based rare disease case matching tool based on online dating algorithms for the acceleration of exome interpretation**. *Human mutation* 2015, **36**(4):432-438.

234. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E *et al*: **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 : official journal of the American College of Medical Genetics* 2015, **17**(5):405-424.

235. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel**. *American journal of human genetics* 2011, **88**(4):440-449.

236. Liu X, Jian X, Boerwinkle E: **dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations**. *Human mutation* 2013, **34**(9):E2393-2402.

237. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S *et al*: **Ensembl 2015**. *Nucleic acids research* 2015, **43**(Database issue):D662-669.

238. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI: **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. *Database : the journal of biological databases and curation* 2015, **2015**:bav028.

239. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature**. *Nucleic acids research* 2009, **37**(Web Server issue):W141-146.

240. Popovic D, Sifrim A, Davis J, Moreau Y, De Moor B: **Problems with the nested granularity of feature domains in bioinformatics: the eXtasy case**. *BMC bioinformatics* 2015, **16 Suppl 4**:S2.

241. Sun Y, Ruivenkamp CA, Hoffer MJ, Vrijenhoek T, Kriek M, van Asperen CJ, den Dunnen JT, Santen GW: **Next-generation diagnostics: gene panel, exome, or whole genome?** *Human mutation* 2015, **36**(6):648-655.

242. Baillie CA, Epps M, Hanish A, Fishman NO, French B, Umscheid CA: **Usability and impact of a computerized clinical decision support intervention designed to reduce urinary catheter utilization and catheter-associated urinary tract infections**. *Infection control and hospital epidemiology* 2014, **35**(9):1147-1155.

243. Bahcall OG: **Genetic testing. ACMG guides on the interpretation of sequence variants**. *Nature reviews Genetics* 2015, **16**(5):256-257.

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

245. Xie B, Agam G, Balasubramanian S, Xu J, Gilliam TC, Maltsev N, Bornigen D: **Disease gene prioritization using network and feature**. *Journal of computational biology : a journal of computational molecular cell biology* 2015, **22**(4):313-323.

246. Luo Y, Riedlinger G, Szolovits P: **Text mining in cancer gene and pathway prioritization**. *Cancer informatics* 2014, **13**(Suppl 1):69-79.

247. Meher PK, Sahu TK, Rao AR, Wahi SD: **A statistical approach for 5' splice site prediction using short sequence motifs and without encoding sequence data**. *BMC bioinformatics* 2014, **15**:362.

248. Li JL, Wang LF, Wang HY, Bai LY, Yuan ZM: **High-accuracy splice site prediction based on sequence component and position features**. *Genetics and molecular research : GMR* 2012, **11**(3):3432-3451.

249. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M: **FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer**. *Genome biology* 2014, **15**(10):480.

250. Haberle J: **Clinical and biochemical aspects of primary and secondary hyperammonemic disorders**. *Archives of biochemistry and biophysics* 2013, **536**(2):101-108.

251. Daniotti M, la Marca G, Fiorini P, Filippi L: **New developments in the treatment of hyperammonemia: emerging use of carglumic acid**. *International journal of general medicine* 2011, **4**:21-28.

252. van Karnebeek CD, Sly WS, Ross CJ, Salvarinova R, Yaplito-Lee J, Santra S, Shyr C, Horvath GA, Eydoux P, Lehman AM *et al*: **Mitochondrial carbonic anhydrase VA deficiency resulting from CA5A alterations presents with hyperammonemia in early childhood**. *American journal of human genetics* 2014, **94**(3):453-461.

253. Nagao Y, Batanian JR, Clemente MF, Sly WS: **Genomic organization of the human gene (CA5) and pseudogene for mitochondrial carbonic anhydrase V and their localization to chromosomes 16q and 16p**. *Genomics* 1995, **28**(3):477-484.

254. Krebs JF, Fierke CA: **Determinants of catalytic activity and stability of carbonic anhydrase II as revealed by random mutagenesis**. *The Journal of biological chemistry* 1993, **268**(2):948-954.

255. Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN: **Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing**. *Human mutation* 2007, **28**(2):150-158.

256. Nagao Y, Platero JS, Waheed A, Sly WS: **Human mitochondrial carbonic anhydrase: cDNA cloning, expression, subcellular localization, and mapping to chromosome 16**.

Proceedings of the National Academy of Sciences of the United States of America 1993, **90**(16):7623-7627.

257. Shah GN, Hewett-Emmett D, Grubb JH, Migas MC, Fleming RE, Waheed A, Sly WS: **Mitochondrial carbonic anhydrase CA VB: differences in tissue distribution and pattern of evolution from those of CA VA suggest distinct physiological roles**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(4):1677-1682.
258. Sly WS, Hu PY: **Human carbonic anhydrases and carbonic anhydrase deficiencies**. *Annual review of biochemistry* 1995, **64**:375-401.
259. de Ligt J, Willemsen MH, van Bon BWM, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C *et al*: **Diagnostic Exome Sequencing in Persons With Severe Intellectual Disability EDITOR COMMENT**. *Obstet Gynecol Surv* 2013, **68**(3):191-193.
260. Salvador-Carulla L, Reed GM, Vaez-Azizi LM, Cooper SA, Martinez-Leal R, Bertelli M, Adnams C, Cooray S, Deb S, Akoury-Dirani L *et al*: **Intellectual developmental disorders: towards a new name, definition and framework for "mental retardation/intellectual disability" in ICD-11**. *World Psychiatry* 2011, **10**(3):175-180.
261. Meerding WJ, Bonneux L, Polder JJ, Koopmanschap MA, van der Maas PJ: **Demographic and epidemiological determinants of healthcare costs in Netherlands: cost of illness study**. *Brit Med J* 1998, **317**(7151):111-+.
262. van Bokhoven H: **Genetic and Epigenetic Networks in Intellectual Disabilities**. *Annu Rev Genet* 2011, **45**:81-104.
263. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N *et al*: **Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study**. *Lancet* 2012, **380**(9854):1674-1682.
264. van Karnebeek CDM, Stockler S: **Treatable inborn errors of metabolism causing intellectual disability: A systematic literature review**. *Mol Genet Metab* 2012, **105**(3):368-381.
265. Coughlin CR, van Karnebeek CDM, Al-Hertani W, Shuen AY, Jaggumantri S, Jack RM, Gaughan S, Burns C, Mirsky DM, Gallagher RC *et al*: **Triple therapy with pyridoxine, arginine supplementation and dietary lysine restriction in pyridoxine-dependent epilepsy: Neurodevelopmental outcome**. *Mol Genet Metab* 2015, **116**(1-2):35-43.
266. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E *et al*: **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* 2015, **17**(5):405-424.
267. MacArthur DG: **A systematic survey of loss-of-function variants in human protein-coding genes (vol 335, pg 823, 2012)**. *Science* 2012, **336**(6079):296-296.
268. van Karnebeek CDM, Shevell M, Zschocke J, Moeschler JB, Stockler S: **The metabolic evaluation of the child with an intellectual developmental disorder: Diagnostic algorithm for identification of treatable causes and new digital resource**. *Mol Genet Metab* 2014, **111**(4):428-438.

269. Zhu XL, Petrovski S, Xie PX, Ruzzo EK, Lu YF, McSweeney KM, Ben-Zeev B, Nissenkorn A, Anikster Y, Oz-Levi D *et al*: **Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios**. *Genetics in Medicine* 2015, **17**(10):774-781.

270. Perez B, Mechinaud F, Galambrun C, Ben Romdhane N, Isidor B, Philip N, Derain-Court J, Cassinat B, Lachenaud J, Kaltenbach S *et al*: **Germline mutations of the CBL gene define a new genetic syndrome with predisposition to juvenile myelomonocytic leukaemia**. *Journal of medical genetics* 2010, **47**(10):686-691.

271. McKenna MC, Waagepetersen HS, Schousboe A, Sonnewald U: **Neuronal and astrocytic shuttle mechanisms for cytosolic-mitochondrial transfer of reducing equivalents: Current evidence and pharmacological tools**. *Biochemical pharmacology* 2006, **71**(4):399-407.

272. Stockler S, Corvera S, Lambright D, Fogarty K, Nosova E, Leonard D, Steinfeld R, Ackerley C, Shyr C, Au N *et al*: **Single point mutation in Rabenosyn-5 in a female with intractable seizures and evidence of defective endocytotic trafficking**. *Orphanet journal of rare diseases* 2014, **9**.

273. Yang YP, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu ZY *et al*: **Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders**. *New Engl J Med* 2013, **369**(16):1502-1511.

274. Wright CF, Fitzgerald TW, Jones WD, Clayton S, Mcrae JF, van Kogelenberg M, King DA, Ambridge K, Barrett DM, Bayzetinova T *et al*: **Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data**. *Lancet* 2015, **385**(9975):1305-1314.

275. Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Femandez BA, Bernier FP, Brudno M, Knoppers B, Marcadier J, Dyment D *et al*: **FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project**. *American journal of human genetics* 2014, **94**(6):809-817.

276. Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, Hijazi H, Alshammari M, Aldahmesh MA, Salih MA *et al*: **Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families**. *Cell Rep* 2015, **10**(2):148-161.

277. Hu H, Haas SA, Chelly J, Van Esch H, Raynaud M, de Brouwer AP, Weinert S, Froyen G, Frints SG, Laumonnier F *et al*: **X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes**. *Molecular psychiatry* 2015.

278. van der Crabben SN, Verhoeven-Duif NM, Brilstra EH, Van Maldergem L, Coskun T, Rubio-Gozalbo E, Berger R, de Koning TJ: **An update on serine deficiency disorders**. *Journal of inherited metabolic disease* 2013, **36**(4):613-619.

279. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL *et al*: **ClinGen - The Clinical Genome Resource**. *New Engl J Med* 2015, **372**(23):2235-2242.

280. Furukawa F, Tseng YC, Liu ST, Chou YL, Lin CC, Sung PH, Uchida K, Lin LY, Hwang PP: **Induction of Phosphoenolpyruvate Carboxykinase (PEPCK) during Acute Acidosis and Its Role in Acid Secretion by V-ATPase-Expressing Ionocytes**. *Int J Biol Sci* 2015, **11**(6):712-725.

281. O Sv, Vidnes J, Falkmer S: **Persistent neonatal hypoglycaemia. A clinical and histopathological study of three cases treated with diazoxide and subtotal

pancreatectomy. *Acta pathologica et microbiologica Scandinavica Section A, Pathology* 1975, **83**(1):155-166.

282. Fasci D, Anania VG, Lill JR, Salvesen GS: **SUMO deconjugation is required for arsenic-triggered ubiquitylation of PML**. *Science signaling* 2015, **8**(380).
283. **Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing**. *Obstet Gynecol Surv* 2015, **70**(3):164-167.
284. Orlando LA, Henrich VC, Hauser ER, Wilson C, Ginsburg GS, Connection G: **The genomic medicine model: an integrated approach to implementation of family health history in primary care**. *Personalized medicine* 2013, **10**(3):295-306.
285. Bauer DC, Gaff C, Dinger ME, Caramins M, Buske FA, Fenech M, Hansen D, Cobiac L: **Genomics and personalised whole-of-life healthcare**. *Trends in molecular medicine* 2014, **20**(9):479-486.
286. Blix A: **Personalized Medicine, Genomics, and Pharmacogenomics: A Primer for Nurses**. *Clin J Oncol Nurs* 2014, **18**(4):437-441.
287. Chu HT, Hsiao WW, Tsao TT, Hsu DF, Chen CC, Lee SA, Kao CY: **SeqEntropy: genome-wide assessment of repeats for short read sequencing**. *PloS one* 2013, **8**(3):e59484.
288. Yeo ZX, Wong JC, Rozen SG, Lee AS: **Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes**. *BMC genomics* 2014, **15**:516.
289. Daber R, Sukhadia S, Morrissette JJ: **Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets**. *Cancer genetics* 2013, **206**(12):441-448.
290. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z: **Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives**. *BMC bioinformatics* 2013, **14 Suppl 11**:S1.
291. Rhoads A, Au KF: **PacBio Sequencing and Its Applications**. *Genomics, proteomics & bioinformatics* 2015.
292. Wei N, Bemmels JB, Dick CW: **The effects of read length, quality and quantity on microsatellite discovery and primer development: from Illumina to PacBio**. *Molecular ecology resources* 2014, **14**(5):953-965.
293. Sanchez-Flores A, Abreu-Goodger C: **A practical guide to sequencing genomes and transcriptomes**. *Current topics in medicinal chemistry* 2014, **14**(3):398-406.
294. Jez S, Martin M, South S, Vanzo R, Rothwell E: **Variants of unknown significance on chromosomal microarray analysis: parental perspectives**. *Journal of community genetics* 2015, **6**(4):343-349.
295. Guidugli L, Carreira A, Caputo SM, Ehlen A, Galli A, Monteiro AN, Neuhausen SL, Hansen TV, Couch FJ, Vreeswijk MP *et al*: **Functional assays for analysis of variants of uncertain significance in BRCA2**. *Human mutation* 2014, **35**(2):151-164.
296. Sijmons RH, Greenblatt MS, Genuardi M: **Gene variants of unknown clinical significance in Lynch syndrome. An introduction for clinicians**. *Familial cancer* 2013, **12**(2):181-187.
297. Murray ML, Cerrato F, Bennett RL, Jarvik GP: **Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: variant reclassification and surgical**

decisions. *Genetics in medicine : official journal of the American College of Medical Genetics* 2011, **13**(12):998-1005.

298. Landau YE, Lichter-Konecki U, Levy HL: **Genomics in newborn screening**. *The Journal of pediatrics* 2014, **164**(1):14-19.

299. Mathelier A, Shi W, Wasserman WW: **Identification of altered cis-regulatory elements in human disease**. *Trends in genetics : TIG* 2015, **31**(2):67-76.

300. Consortium EP: **The ENCODE (ENCyclopedia Of DNA Elements) Project**. *Science* 2004, **306**(5696):636-640.

301. Ullah MZ, Aono M, Seddiqui MH: **Estimating a ranked list of human hereditary diseases for clinical phenotypes by using weighted bipartite network**. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference* 2013, **2013**:3475-3478.

302. Ouagne D, Hussain S, Sadou E, Jaulent MC, Daniel C: **The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology**. *Studies in health technology and informatics* 2012, **180**:534-538.

303. Li L, Hur M, Lee JY, Zhou W, Song Z, Ransom N, Demirkale CY, Nettleton D, Westgate M, Arendsee Z *et al*: **A systems biology approach toward understanding seed composition in soybean**. *BMC genomics* 2015, **16 Suppl 3**:S9.

304. Botas A, Campbell HM, Han X, Maletic-Savatic M: **Metabolomics of Neurodegenerative Diseases**. *International review of neurobiology* 2015, **122**:53-80.

305. Zhou ZW, Chen XW, Sneed KB, Yang YX, Zhang X, He ZX, Chow K, Yang T, Duan W, Zhou SF: **Clinical association between pharmacogenomics and adverse drug reactions**. *Drugs* 2015, **75**(6):589-631.

306. Holzhutter HG, Drasdo D, Preusser T, Lippert J, Henney AM: **The virtual liver: a multidisciplinary, multilevel challenge for systems biology**. *Wiley interdisciplinary reviews Systems biology and medicine* 2012, **4**(3):221-235.

307. Hunter P, Robbins P, Noble D: **The IUPS human Physiome Project**. *Pflugers Archiv : European journal of physiology* 2002, **445**(1):1-9.

308. Petrikin JE, Willig LK, Smith LD, Kingsmore SF: **Rapid whole genome sequencing and precision neonatology**. *Seminars in perinatology* 2015, **39**(8):623-631.

309. Miller NA, Farrow EG, Gibson M, Willig LK, Twist G, Yoo B, Marrs T, Corder S, Krivohlavek L, Walter A *et al*: **A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases**. *Genome medicine* 2015, **7**(1):100.

310. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM: **SpeedSeq: ultra-fast personal genome analysis and interpretation**. *Nature methods* 2015, **12**(10):966-968.

311. Oliver GR, Hart SN, Klee EW: **Bioinformatics for clinical next generation sequencing**. *Clinical chemistry* 2015, **61**(1):124-135.

312. Ritchie MD, Verma SS, Hall MA, Goodloe RJ, Berg RL, Carrell DS, Carlson CS, Chen L, Crosslin DR, Denny JC *et al*: **Electronic medical records and genomics (eMERGE) network exploration in cataract: several new potential susceptibility loci**. *Molecular vision* 2014, **20**:1281-1295.

313. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA *et al*: **The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future**. *Genetics in medicine : official journal of the American College of Medical Genetics* 2013, **15**(10):761-771.

314. Fullerton SM, Wolf WA, Brothers KB, Clayton EW, Crawford DC, Denny JC, Greenland P, Koenig BA, Leppig KA, Lindor NM *et al*: **Return of individual research results from genome-wide association studies: experience of the Electronic Medical Records and Genomics (eMERGE) Network**. *Genetics in medicine : official journal of the American College of Medical Genetics* 2012, **14**(4):424-431.

315. Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SH, Huff KD, Mitchell IM, Plumbley MD *et al*: **Best practices for scientific computing**. *PLoS biology* 2014, **12**(1):e1001745.

316. Duffy RL, Yiu SS, Molokhia E, Walker R, Perkins RA: **Effects of electronic prescribing on the clinical practice of a family medicine residency**. *Family medicine* 2010, **42**(5):358-363.

317. Lagerin A, Nilsson G, Tornkvist L: **An educational intervention for district nurses: use of electronic records in leg ulcer management**. *J Wound Care* 2007, **16**(1):29-32.

318. Bell SK, Folcarelli PH, Anselmo MK, Crotty BH, Flier LA, Walker J: **Connecting Patients and Clinicians: The Anticipated Effects of Open Notes on Patient Safety and Quality of Care**. *Joint Commission journal on quality and patient safety / Joint Commission Resources* 2015, **41**(8):378-384.

319. Resnik DB: **Are DNA patents bad for medicine?** *Health policy* 2003, **65**(2):181-197.

320. Martindale VE: **Challenges, opportunities, and liabilities in the genomic age**. *Aviation, space, and environmental medicine* 2013, **84**(1):77-79.

321. Holm IA: **Clinical Management of Pediatric Genomic Testing**. *Current genetic medicine reports* 2014, **2**(4):212-215.

322. DeMets DL: **Clinical trials in the new millennium**. *Statistics in medicine* 2002, **21**(19):2779-2787.

323. Conley JM, Doerr AK, Vorhaus DB: **Enabling responsible public genomics**. *Health matrix* 2010, **20**(2):325-385.

324. Slosar JP: **Ethical issues in genetic testing**. *Health care ethics USA : a publication of the Center for Health Care Ethics* 2005, **13**(3):E1.

325. Wilcken B: **Ethical issues in genetics**. *Journal of paediatrics and child health* 2011, **47**(9):668-671.

326. Goecks J, Nekrutenko A, Taylor J, Galaxy T: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome biology* 2010, **11**(8):R86.

327. Vandeweyer G, Van Laer L, Loeys B, Van den Bulcke T, Kooy RF: **VariantDB: a flexible annotation and filtering portal for next generation sequencing data**. *Genome medicine* 2014, **6**(10):74.

328. Aleman A, Garcia-Garcia F, Salavert F, Medina I, Dopazo J: **A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies**. *Nucleic acids research* 2014, **42**(Web Server issue):W88-93.

329. Yao J, Zhang KX, Kramer M, Pellegrini M, McCombie WR: **FamAnn: an automated variant annotation pipeline to facilitate target discovery for family-based sequencing studies**. *Bioinformatics* 2014.

330. Li MX, Gui HS, Kwan JS, Bao SY, Sham PC: **A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases**. *Nucleic acids research* 2012, **40**(7):e53.

331. Nakai K, Tokimori T, Ogiwara A, Uchiyama I, Niiyama T: **Gnome--an Internet-based sequence analysis tool**. *Computer applications in the biosciences : CABIOS* 1994, **10**(5):547-550.

332. Trakadis YJ, Buote C, Therriault JF, Jacques PE, Larochelle H, Levesque S: **PhenoVar: a phenotype-driven approach in clinical genomics for the diagnosis of polymalformative syndromes**. *BMC medical genomics* 2014, **7**:22.

333. The Gene Ontology C: **Gene Ontology Consortium: going forward**. *Nucleic acids research* 2014.

334. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic acids research* 2012, **40**(Database issue):D109-114.

335. Pang CN, Tay AP, Aya C, Twine NA, Harkness L, Hart-Smith G, Chia SZ, Chen Z, Deshpande NP, Kaakoush NO *et al*: **Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing**. *Journal of proteome research* 2014, **13**(1):84-98.

336. Hillman SC, Pretlove S, Coomarasamy A, McMullan DJ, Davison EV, Maher ER, Kilby MD: **Additional information from array comparative genomic hybridization technology over conventional karyotyping in prenatal diagnosis: a systematic review and meta-analysis**. *Ultrasound Obst Gyn* 2011, **37**(1):6-14.

337. Giefing M, Zemke N, Brauze D, Kostrzewska-Poczekaj M, Luczak M, Szaumkessel M, Pelinska K, Kiwerska K, Tonnies H, Grenman R *et al*: **High Resolution ArrayCGH and Expression Profiling Identifies PTPRD and PCDH17/PCH68 as Tumor Suppressor Gene Candidates in Laryngeal Squamous Cell Carcinoma**. *Gene Chromosome Canc* 2011, **50**(3):154-166.

338. Hodgetts J, Boonham N, Mumford R, Dickinson M: **Panel of 23S rRNA Gene-Based Real-Time PCR Assays for Improved Universal and Group-Specific Detection of Phytoplasmas**. *Appl Environ Microb* 2009, **75**(9):2945-2950.

339. Narumi Y, Nishina S, Tokimitsu M, Aoki Y, Kosaki R, Wakui K, Azuma N, Murata T, Takada F, Fukushima Y *et al*: **Identification of a Novel Missense Mutation of MAF in a Japanese Family With Congenital Cataract by Whole Exome Sequencing: A Clinical Report and Review of Literature**. *American Journal of Medical Genetics Part A* 2014, **164**(5):1272-1276.

340. Khoury MJ, McCabe LL, McCabe ERB: **Genomic medicine - Population screening in the age of genomic medicine.** *New Engl J Med* 2003, **348**(1):50-58.

341. Liu PY, Morrison C, Wang L, Xiong DH, Vedell P, Cui P, Hua X, Ding F, Lu Y, James M *et al*: **Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing**. *Carcinogenesis* 2012, **33**(7):1270-1276.

342. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C *et al*: **Comparison of solution-based exome capture methods for next generation sequencing**. *Genome biology* 2011, **12**(9).

343. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD *et al*: **Natural selection on protein-coding genes in the human genome**. *Nature* 2005, **437**(7062):1153-1157.

344. Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BWM, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A *et al*: **Genome sequencing identifies major causes of severe intellectual disability**. *Nature* 2014, **511**(7509):344-+.

345. Rebbeck TR, Spitz M, Wu X: **Assessing the function of genetic variants in candidate gene association studies**. *Nature reviews Genetics* 2004, **5**(8):589-597.

346. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C: **An integrative approach to predicting the functional effects of non-coding and coding sequence variation**. *Bioinformatics* 2015.

347. Dumas L, Dickens CM, Anderson N, Davis J, Bennett B, Radcliffe RA, Sikela JM: **Exome sequencing and arrayCGH detection of gene sequence and copy number variation between ILS and ISS mouse strains**. *Mammalian Genome* 2014, **25**(5-6):235-243.

348. Poultney CS, Goldberg AP, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwara Y, De Rubeis S, Durand S, Stevens C, Rehnstrom K *et al*: **Identification of Small Exonic CNV from Whole-Exome Sequence Data and Application to Autism Spectrum Disorder**. *American journal of human genetics* 2013, **93**(4):607-619.

349. Robinson PN, Krawitz P, Mundlos S: **Strategies for exome and genome sequence data analysis in disease-gene discovery projects**. *Clinical genetics* 2011, **80**(2):127-132.

350. Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg EJ, Mensenkamp AR, Rodenburg RJ, Yntema HG, Spruijt L, Vermeer S *et al*: **A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases**. *Human mutation* 2013, **34**(12):1721-1726.

351. Ammenwerth E, Graber S, Herrmann G, Burkle T, Konig J: **Evaluation of health information systems-problems and challenges**. *International journal of medical informatics* 2003, **71**(2-3):125-135.

352. Timpka T, Sjoberg C, Hallberg N, Eriksson H, Lindblom P, Hedblom P, Svensson B, Marmolin H: **Participatory design of computer-supported organizational learning in health care: methods and experiences**. *Proceedings / the Annual Symposium on Computer Application [sic] in Medical Care Symposium on Computer Applications in Medical Care* 1995:800-804.

353. Berg M: **Patient care information systems and health care work: a sociotechnical approach**. *International journal of medical informatics* 1999, **55**(2):87-101.

354. Weerakkody G, Ray P: **CSCW-based system development methodology for health-care information systems**. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* 2003, **9**(3):273-282.

355. Gottlieb MM, Arenillas DJ, Maithripala S, Maurer ZD, Tarailo Graovac M, Armstrong L, Patel M, van Karnebeek C, Wasserman WW: **GeneYenta: A Phenotype-Based Rare Disease Case Matching Tool Based on Online Dating Algorithms for the Acceleration of Exome Interpretation**. *Human mutation* 2015.

356. Metzger MJ: **Making sense of credibility on the web: Models for evaluating online information and recommendations for future research**. *J Am Soc Inf Sci Tec* 2007, **58**(13):2078-2091.

357. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee J, van Karnebeek C, Wasserman WW: **FLAGS, frequently mutated genes in public exomes**. *BMC medical genomics* 2014, **7**(1):64.

358. Takacs G, Pilaszy I, Nemeth B, Tikk D: **Scalable Collaborative Filtering Approaches for Large Recommender Systems**. *J Mach Learn Res* 2009, **10**:623-656.

359. Vassy JL, McLaughlin HL, MacRae CA, Seidman CE, Lautenbach D, Krier JB, Lane WJ, Kohane IS, Murray MF, McGuire AL *et al*: **A one-page summary report of genome sequencing for the healthy adult**. *Public health genomics* 2015, **18**(2):123-129.

360. Welch BM, Eilbeck K, Del Fiol G, Meyer LJ, Kawamoto K: **Technical desiderata for the integration of genomic data with clinical decision support**. *Journal of biomedical informatics* 2014, **51**:3-7.

361. Masys DR, Jarvik GP, Abernethy NF, Anderson NR, Papanicolaou GJ, Paltoo DN, Hoffman MA, Kohane IS, Levy HP: **Technical desiderata for the integration of genomic data into Electronic Health Records**. *Journal of biomedical informatics* 2012, **45**(3):419-422.

362. Jensen PB, Jensen LJ, Brunak S: **Mining electronic health records: towards better research applications and clinical care**. *Nature reviews Genetics* 2012, **13**(6):395-405.

363. Kohane IS: **Using electronic health records to drive discovery in disease genomics**. *Nature Reviews Genetics* 2011, **12**(6):417-428.

364. Denny JC: **Chapter 13: Mining Electronic Health Records in the Genomics Era**. *PLoS computational biology* 2012, **8**(12).

365. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, Almoguera B, Basford MA, Bielinski SJ, Brautbar A, Brilliant MH, Carrell DS, Connolly JJ *et al*: **Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems**. *Clinical pharmacology and therapeutics* 2014, **96**(4):482-489.

366. Kullo IJ, Haddad R, Prows CA, Holm I, Sanderson SC, Garrison NA, Sharp RR, Smith ME, Kuivaniemi H, Bottinger EP *et al*: **Return of results in the genomic medicine projects of the eMERGE network**. *Frontiers in genetics* 2014, **5**:50.

Appendices

Appendix A Chapter 2

A.1 Preliminary survey

To gather a brief initial perspective on how clinicians viewed next-generation sequencing (NGS), we conducted one-on-one free-style (i.e. unstructured) interview with ten clinical geneticists based within CFRI (Children and Family Research Institute). The interviews focused on three main issues: 1) how are they currently incorporate next-generation sequencing into clinical/research practice; 2) what interface they currently use; and 3) what challenges or difficulties they experience, whether interpreting the data, integrating to workflow, or using the interface. The interviews gathered personal background information such as the kind of computer-based work they perform, how much time they spend each week using computers, and their experience with computer programming. These ten clinical geneticists were the same representatives that were later recruited to the evaluation in the study. The interviews were done in May-June 2012 separately for each individual in their own respective offices, lasted around 20 minutes per individual. Nine subjects lacked any fundamental computational training, while the remaining individual was proficient with programming in Visual Basic. All the interviewees have prior exposure to working with NGS data – 8/10 had worked with exome data, while the other two worked with targeted gene panels. We note that these two individuals later got exposed to working with exome datasets prior to the usability evaluation. Two subjects were only using NGS within genetics research projects, and expressed reservations about the current clinical utility due to poor specificity of results (a perceived high rate of false variants being reported). Another subject with experience using array-based genotyping for CNV analysis was more enthusiastic about incorporating NGS into clinical practice. Six of the participants reported concerns with their current interface (NextGene), noting many key variant annotations such as type of codon change (synonymous versus nonsynonysmous) or the name of the impacted gene were missing without explanation. Five subjects had worked with GoldenHelix's interface – three of which have also used NextGene before. Two subjects used Excel instead of the NextGene viewer due to prior familiarity with Excel interface, although the data they were reviewing was first generated from NextGene. When asked for what they desired from the software interfaces based on their current experience - the interviewees expressed the desire for more feature annotation for each variant in order to facilitate prioritization of candidate causal variants. Besides reporting quality scores, coverage, frequency in populations, presence in known disease database, evolutionary conservation, and predicted effect on proteins, they expressed desires for information about gene expression, characteristics of mouse models if any, and key terms mined from literature about that variant/gene. The latter is especially time-consuming and is currently done manually by the participating clinicians.

A.2 Clinical scenarios

For ethical reasons, the data provided in each scenario were aimed to be as clinically realistic as possible while avoiding the use of actual confidential patient data. As is the nature of any simulation studies, it is impossible to cover every task that may arise in the real world, but the scenarios were setup based upon a literature review of exome studies published as well as under the guidance and consultation of exome/genome analysis experts who were not recruited to the study.

In the first clinical scenario, a 1st-degree consanguineous family with a single case of mitochondrial disorder is presented. The user is presented with the exome variant calls and ROH for the index patient. The clinicians were presented with a basic biochemical patient background that justifies the prediction of a mitochondrial disorder. The report also states that no defects in mitochondrial DNA were observed. A disruptive mutation in gene SUCLA2 was embedded in the exome to represent the intended causal variant. The mutation was introduced in such a way that it would emerge as a top candidate by filtering for rare non-synonymous mutations present in the listed mitochondrial genes that fall within a region of homozygosity.

In the second clinical case, a patient is described as having episodic muscle weakness and paroxysmal dystonia. Clinicians were supplied with a quartet of exomes corresponding to the unaffected parents, unaffected sibling, and affected proband. The subjects were first asked to conduct mutation analysis on two specific genes previously implicated in this type of disease, before searching for variations consistent with classical Mendelian inheritance models. To accelerate the analysis, the subjects were instructed to focus on *de novo* heterozygous model, through which they should identify a novel mutation disrupting the N-terminal domain of gene KCNJ18, a gene previously implicated to cause this type of disorder.

In the sections below we provide details to the hypothetical scenarios that were provided to the clinicians during the usability evaluation.

Hypothetical Scenario 1:

Key disease trait:

Multiple mitochondrial respiratory chain (MRC) enzyme deficiency, a biochemical signature due to diverse gene defects, including mtDNA or nuclear genes.

Data:

- 1 exome sample corresponding to the proband. Parental exome not available.
- Homozygosity mapping data for proband from SNP-array
- A list of human mitochondrial genes from MitoCarta database.

Patient background:

- Male, currently 9 years old
- European descent (France)
- Parents are consanguineous, second-cousin
- Normal pregnancy and delivery
- During first months of life, see muscle hypotonia, failure to thrive, poor weight gain, frequent vomit
- At 1 year of age, see good visual contact but marked axial hypotonia with absence of head control, poor active movements, brisk tendon reflexes. Put on nasogastric tube feeding because of severe dysphagia.
- Neurosensory hearing loss abnormal.
- Liver and kidney functions normal, but biochemical exams revealed increased levels of lactate and pyruvate in both plasma (3000 uM vs <2000 uM normal, 200 uM versus <140 uM normal) and CSF (2300 uM vs <1800 uM normal, 155 uM vs <120 uM normal)
- By age 2, see persistent lesions of bilateral abnormal signals in caudate and putamina nucleic, from brain MRI

- Clinical features progressively worsened, by 3 years old patient shows dystonic tetraparesis associated with bilateral ptosis and opthalmoparesis, and severe cognitive impairment with no verbal development
- Despite severe clinical features, patient never presented with metabolic crisis and EEG was always normal
- Last diagnosis is at age 6, showing severe bilateral ptosis, incomplete ophthalmoparesis, spastic-dystonic tetraparesis with absence of head control, scoliosis, and marked irritability.
- Biochemical analysis revealed multi-enzymatic defect of mtDNA-dependent MRC activities in muscles and fibroblasts, pointing to a defect of mtDNA maintenance or expression. Sequence of entire mtDNA from skeletal muscle fail to show pathogenic mutations.

Tasks:

1) Upload exome dataset (Proband_exome.txt) to the Varsifter software
2) Retrieve the total number of mutations in this exome dataset
3) Give the number of missense mutations, nonsense mutations, and InDels
4) Filter the dataset against polymorphisms based on your desired choice of frequency threshold.
5) From step 4, filter the remaining variations against homozygosity mapping data (provided as ROH_bed.bed)
6) From step 5, give the list of variants predicted in silico to be damaging after homozygosity filtering. Save this list as a separate file.

7) From step 6, give the list of variants associated to mitochondrial genes (gene list provided as mitochondrial_gene_list.txt).

8) Give a final list of variants that you deem worthy for validation and further follow-ups.

Hypothetical Scenario 2:

Key disease trait:

Episodic muscle weakness and paroxysmal dystonia: an autosomal dominant disorder typically characterized by acute, episodic and usually flaccid loss of skeletal muscle tone in the context of low serum potassium (1-3 mmol/l). Age of onset varies between 5 to 20 years. Episodes can last hours to days, often precipitated by carbohydrate-rich meals and rest after prolonged exercise. Previous studies have shown 80% of such disease is caused by mutations in CACNA1S and SCN4A.

Data:

- 4 exomes in total: 2 exomes from healthy parents (non-consanguineous), 1 exome of unaffected sibling (brother, 2 years older), and 1 exome of female index.

Patient background:

- 5 year old female from Canada with EU ancestry with a 2-year history of episodic lower limb weakness, manifested as difficulty with weight bearing, stumbling and clumsiness and subjective descriptions of pain. Episodes occurred 2 to 3 times per week, lasting from 30 min to 4h.
- Also has early-onset scoliosis, high arched feet, lower limb hypertonia, clumsy gait, and frequent toe-walking

- Physically, non-dysmorphic with tight hell cords, pes cavus, increased plantar reflexes, lordosis, positive Gower sign, and stiff toe-walking and in-toeing of right foot
- Clinical diagnose included testing for HypoKPP (primary hypokalemic periodic paralysis) based on evaluation of exon 5 and 13 for CACNA1S and exons 10, 15 and 21 for SCN4A. Results were unrevealing.

Tasks:

1) Upload exome dataset (Family_exome.txt) into the software
2) Get the total # of non-reference variations for each family member
3) Identify mutations, if any that fall within CACNA1S and SCN4A. Decide if these are worthy candidates for additional follow-up.
4) From the original list, perform intersection to obtain homozygous mutations in index, not homozygous in the other sibling, and heterozygous in both parents (i.e. recessive model).
5) From the original list, perform intersection to mutations that are only present in the index and not in any other family members (i.e. *de novo* model).
6) From the *de novo* mutation list (step 5), filter for heterozygous, nonsynonymous mutations.
7) From step 6, filter against polymorphisms based a frequency threshold of your choice.
8) Give the list of variant-gene pairs that you think are worthy for follow-up from the *de novo* hypothesis.

A.3 Additional details to methodology

Study setting and participants (additional details)

Each evaluation took place within the subject’s office, reflecting the typical environment in which the subjects analyze DNA sequencing data in real life. These selected experts are representative of the hospital’s medical specialists who most closely interact with patient exomes for clinical diagnosis. There was no standard software established within the clinic at the time of testing, with the staff using a variety of self-selected packages for their own work.

Experimental procedure (additional details)

The breakdown of the 45-minute introductory period is as follow: a 10-minute tutorial video for both software was presented to each subject prior to the evaluation. This video showed how each of the tasks that would arise in the simulated study can be accomplished using the software. Subsequently each of the subjects was initially given 30 minutes to work with the two software packages– including interacting with the software and/or reading the software manual. Afterwards, subjects were given approximately 5 minutes to familiarize themselves with the input data provided for each hypothetical clinical scenario, and to go over the tasks that were assigned within each scenario. Throughout the introduction session, participants were given opportunities to raise any questions they may have, and the experimenter (CS) provided answers as required for the subjects to progress through the simulations.

Clinicians were instructed to “think-aloud” while they worked through the scenarios. If the clinicians remained silent for more than five seconds, they were reminded to “keep talking”. If the participant appeared confused or expressed frustration with the provided tool for more than ten seconds, the experimenter (CS) provided tips on overcoming the specific challenges that they were facing.

Interviews and surveys (additional comments)

The pre-evaluation interviews occurred before the usability evaluation. They addressed the subject’s prior experience working with genomic data (e.g. “Please describe the type of analyses you have personally done with patient DNA data”, “what challenges have you faced when using next-generation sequencing in clinics”), and the amount of computational expertise that they have (e.g. “What software have you previously tried in exome or whole-genome analysis”, “are you familiar with any scripting or programming languages”). The post-evaluation interviews occurred immediately after the usability evaluation. They addressed specific issues that came up during the evaluation, and included pre-defined questions such as “what is the most useful function that you find about the software”, “what is the biggest flaw you find with this system?” -- questions that are pertinent to the scope of this research but could not be directly extracted from the screen capture data. While SUMI provided open-ended questions in addition to the 50 multiple-choice questions, we instead incorporated the 3 open-ended questions into the post-evaluation interview. Both interviews were meant to be brief, targeted to be less than 5 minutes in duration. For the clinicians who had exposure to commercial platforms, they were asked if the challenges they faced during the evaluation had been encountered when using these other platforms.

SUMI (Software Usability Measurement Inventory (SUMI))

In this section, we provide the survey questionnaires provided to our subjects. For each question there are three responses: agree, undecided, and disagree. Open-ended questions at the end were incorporated to our post-evaluation interviews. More information on the construction of the survey can be found on the SUMI website7. Below are the questionnaires taken from SUMI:

==

This questionnaire has 50 statements. Please answer them all. After each statement there are three boxes.

- Check the first box if you generally AGREE with the statement.
- Check the middle box if you are UNDECIDED, or if the statement has no relevance to your software or to your situation.
- Check the right box if you generally DISAGREE with the statement.

In checking the left or right box you are not necessarily indicating strong agreement or disagreement but just your general feeling most of the time.

[7] Kirakowski, J. (1995), 'The Software Usability Measurement Inventory: Background and Usage.' In: P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), Usability Evaluation in Industry. Taylor and Frances, London, UK.

Statements 1-10 out of 50	**Agree**	**Undecided**	**Disagree**
This software responds too slowly to inputs.			
I would recommend this software to my colleagues.			
The instructions and prompts are helpful.			
This software has at some time stopped unexpectedly.			
Learning to operate this software initially is full of problems.			
I sometimes don't know what to do next with this software.			
I enjoy the time I spend using this software.			
I find that the help information given by this software is not very useful.			
If this software stops it is not easy to restart it.			
It takes too long to learn the software functions.			
Statements 11-20 out of 50	**Agree**	**Undecided**	**Disagree**
I sometimes wonder if I am using the right function.			
Working with this software is satisfying.			
The way that system information is presented is clear and understandable.			
I feel safer if I use only a few familiar functions.			
The software documentation is very informative.			
This software seems to disrupt the way I normally like to arrange my work.			
Working with this software is mentally stimulating.			
There is never enough information on the screen when it's needed.			
I feel in command of this software when I am using it.			
I prefer to stick to the functions that I know best.			
Statements 21-30 out of 50	**Agree**	**Undecided**	**Disagree**
I think this software is inconsistent.			
I would not like to use this software every day.			
I can understand and act on the information provided by this software.			
This software is awkward when I want to do something which is not standard.			
There is too much to read before you can use the software.			
Tasks can be performed in a straightforward manner using this software.			
Using this software is frustrating.			
The software has helped me overcome any problems I have had in using it.			
The speed of this software is fast enough.			
I keep having to go back to look at the guides.			

Statements 31-40 out of 50	**Agree**	**Undecided**	**Disagree**
It is obvious that user needs have been fully taken into consideration.			
There have been times in using this software when I have felt quite tense.			
The organization of the menus seems quite logical.			
The software allows the user to be economic of keystrokes.			
Learning how to use new functions is difficult.			
There are too many steps required to get something to work.			
I think this software has sometimes given me a headache.			
Error messages are not adequate.			
It is easy to make the software do exactly what you want.			
I will never learn to use all that is offered in this software.			
Statements 41-50 out of 50	**Agree**	**Undecided**	**Disagree**
The software hasn't always done what I was expecting.			
The software presents itself in a very attractive way.			
Either the amount or quality of the help information varies across the system.			
It is relatively easy to move from one part of a task to another.			
It is easy to forget how to do things with this software.			
This software occasionally behaves in a way which can't be understood.			
This software is really very awkward.			
It is easy to see at a glance what the options are at each stage.			
Getting data files in and out of the system is not easy.			
I have to look for assistance most times when I use this software.			

#2: How important for you is the kind of software you have just been rating? (check the box on the right)

Extremely important	
Important	
Not very important	
Not important at all	

#3: How would you rate your software skills and knowledge?

Very experienced and technical	
I'm good but not very technical	
I can cope with most software	
I find most software difficult to use	

3 open-ended questions:

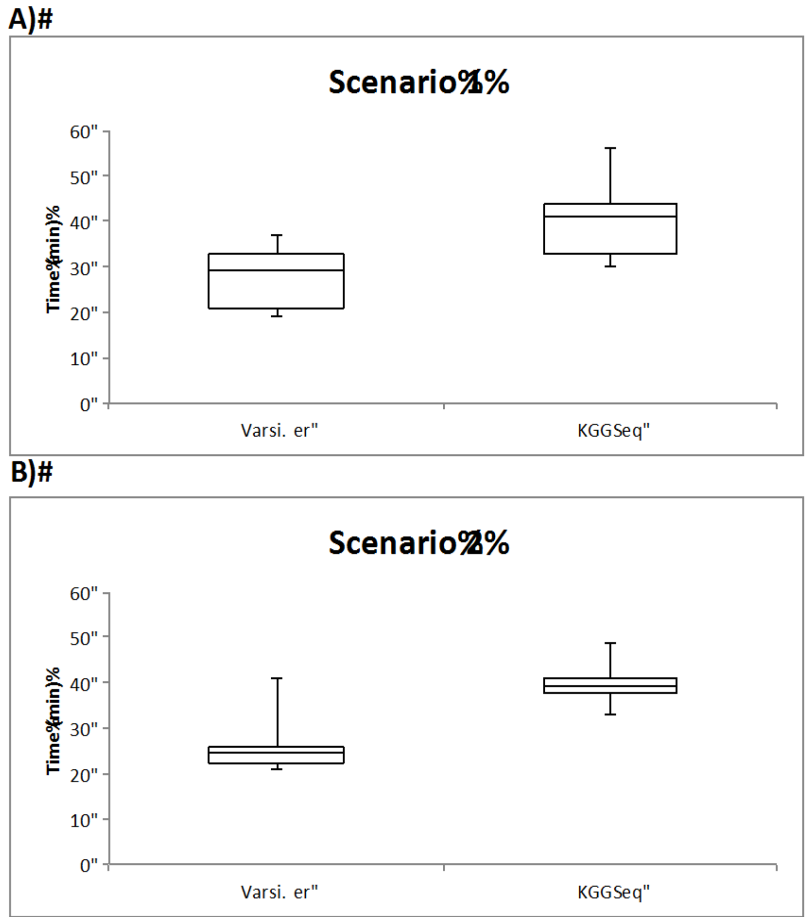
#1: what software have you tried out before?

#2: What do you think is the best aspect of this software, and why?

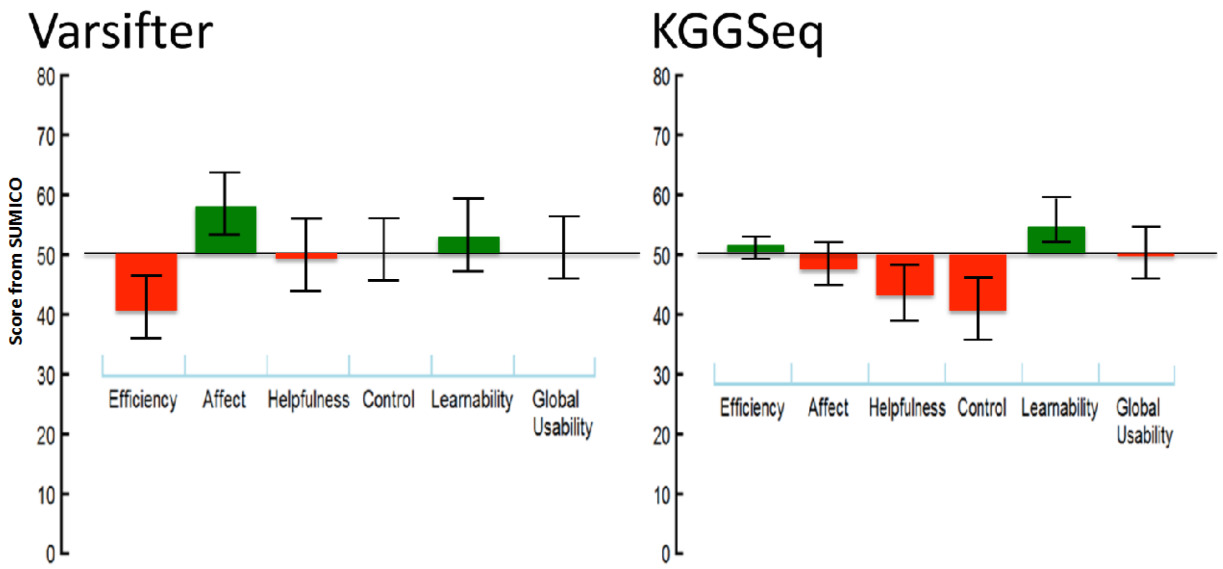
#3: What do you think needs most improvement, and why?

===

The data collected from SUMI survey were assessed using SUMICO software by the Human Factors Research Group (http://www.ucc.ie/hfrg/), which can be found on the SUMI website. As assessed using the SUMICO software, the two tools were perceived differently by the users (see Appendix A Figure 2). As a summary of the output from SUMICO, for each evaluated software, the usability is broken down into six categories: "efficiency", "affect", "helpfulness", "control", "learnability", and "global usability". Details to the descriptions of the meanings of these labels are discussed in the result section. For each category, SUMICO assigns a score that can range from 0 to 80. A score of 50 refers to a response score that is comparable to the expected values from the SUMI database. A value below 50 is colored as red and refers to a negative deviation away from the expected, and a value above 50 refers to a positive deviation away from the expected. The significance of each score is calculated based upon 95% confidence intervals derived from SUMICO using the SUMI database. Interested readers can refer to the SUMI website for more information (http://sumi.ucc.ie/sumipapp.html#sumidev). For the quantified performance measures, Varsifter scored the lowest on efficiency, a measure of how users perceive that the software assists them in completing the given tasks. Regardless, Varsifter scored high on affect, revealing that despite the difficulties in using the software, the clinicians retained a favorable impression of the software. KGGSeq scored low in “Helpfulness” and “Control”, which respectively measure the degree to which the software is self-explanatory and the extent to which the user feels in control of the software, as opposed to being controlled by the software. While users apparently perceive Varsifter as being inefficient, the users preferred its interface and, as noted above, achieved better performance using it in terms of successful mutation discovery and time efficiency.



Appendix A Figure 1 Time required for clinicians to complete each scenario and successfully identify the embedded causal mutation. Panel A shows the time for scenario 1, and panel B shows the time for scenario 2. Y-axis plots the distribution of time (minutes, rounded) from the start on a scenario, and X-axis indicates the performance for each evaluated software.



Appendix A Figure 2. An overall summary of usability across 6 quantified attributes by SUMI. A deviation below or above the score 50 represents a negative or positive attitude on that particular usability category. The precise meaning of these subscales is given in the SUMI manual (http://sumi.ucc.ie/sumipapp.html).

A.4 Performance data

In the sections below, we provide the performance data, the additional relevant comments and feedbacks regarding the two software that were captured during the usability evaluation.

Individual performances on clinical scenarios

	Scenario 1		Scenario 2	
	Varsifter	KGGSeq	Varsifter	KGGSeq
Subject 1	19	41	21	33
Subject 2*	37	56		
Subject 3	21	30	25	49
Subject 4*	33			
Subject 5*	33			
Subject 6	33	44		
Subject 7*	29	33	26	41
Subject 8	29	44	24	

	Scenario 1		Scenario 2	
	Varsifter	KGGSeq	Varsifter	KGGSeq
Subject 9*	21	33	35	39
Subject 10	20	41	22	38

Appendix A Table 1. The time (minutes) of the performance for the ten evaluated clinicians. The time was rounded off to the nearest minute. The table is organized such that the first five subjects (subjects 1-5) used Varsifter first before using KGGSeq for scenario 1, and the remaining subjects used KGGSeq before Varsifter for scenario 1. The subjects with * marked (subjects 2,4,5,7,9) were selected to use Varsifter first before using KGGSeq for scenario 2, while the remaining subjects performed scenario 2 using KGGSeq before Varsifter.

Themes	Categories	Example of comments captured
VISUALIZATION	Navigation	"I don't see why compound heterozygous button is grouped under 'View'. I would have expected it to be under 'Tools'"
	Layout	"I remembered seeing a 'finalize' [Finalize Query] button in the tutorial, but why is it not showing up on my screen?"
	Operation consistency	"This 'search' button is confusing. Most of the time the software refreshes the current screen, but in some cases it opens a new window.”
	Graphics	"I like how you can drag this [referring to the icon showing up during construction of custom query] around"
INFORMATION	Resolution	"There is a lot of information in this file (referring to software output). I feel overwhelmed. I can't make sense out of it"
	Label	"What is SLR_test_statistics? What do all these other columns mean? I don't understand and I don't see any help pop-ups"
	System messages	"The error message says I am missing the sample ID, but doesn’t tell me how or where to specify it”
SYSTEM RESPONSE	Response time	"The software is running, I can see that...but how long would it take?"
	System status	"I've clicked this button twice and I am not getting any response. Is it [the system] stuck?"
FUNCTIONALITIES	Compatibility	"Is the software able to take in a list of genes commonly seen in exome studies? I have a list of commonly mutated genes in Excel but I don't know how to overlay that to the variants"
	Scope of functionalities	"No, I cannot do this task [select for homozygous recessive variants], it is too difficult for me to figure out. There needs to be an easier way"
OVERALL USABILITY	Overall usage	"I do not see myself using this tool at this point. Perhaps if there is a hands-on session where someone can walk me through one of my datasets face-to-face..."

Appendix A Table 2: We assigned the detected usability problems into 5 main themes that are subdivided into 12 categories. These categories cover the visual representation of the system, the information presented by the system, the response of the system, and the functionalities offered by the system. Example comments from each category are shown (emphasis on the negative, unless not available).

In-depth analysis of feedback

Visualization of features on the visible screen (additional comments)

For Varsifter, 42/92 (46%) of the negative comments related to GUI design. For example, the default window size of the software failed to display all options and buttons, resulting in a portion of the screen requiring scrolling to access. Every surveyed clinician complained that texts and/or functions were hidden from view. For example, in the window to create a custom query, 8 of the 10 participants sought a button that they had observed in the tutorial video, but did not know that it was necessary to use the scrollbar to access the remainder of the options. The feedback about KGGSeq also indicated concerns with incomplete display in the command-line generator GUI. One clinician commented “the scrolling means I have to remember what is hidden behind this panel and that is a pain”. Even in cases where all buttons were displayed, usability problems were encountered in navigation and execution. For instance, the procedure to setup custom queries in Varsifter requires a series of clicks in the correct boxes, and every user reported difficulty in learning to use it. One clinician stated: “the way of how I imagine I would go about setting up a query is not the same as the way the software is setup”; A clinician who was proficient with databases and the use of structured query language (SQL) commands said “I know how to do this in SQL, but I cannot do it on here”.

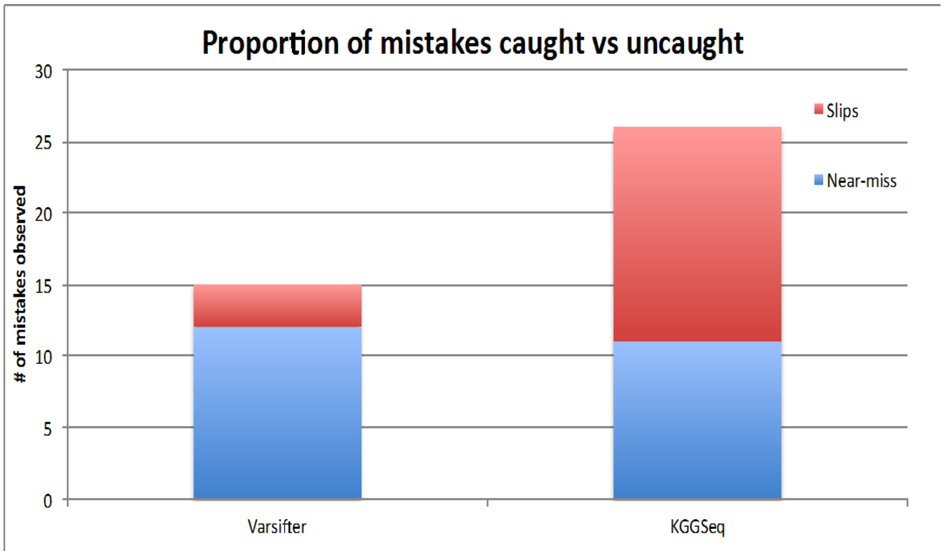
Scope of functionalities (additional comments)

While both software packages have a wide range of functionalities, they suffer both in terms of difficulty in executing the functions and relevance to clinical exome-analysis tasks (26/92 and 22/79 negative comments were captured). For both software, the majority of problems (16/19 for Varsifter, 13/20 for KGGSeq) related to software functions reflect the clinician's inability to perform a task, rather than the absence of a functionality within the system. For instance, 4 out of 10 clinicians were unable to setup Varsifter's custom query to filter variations for a particular Mendelian inheritance model. Although Varsifter does have a pre-set function to filter the data by Mendelian genetic models, those functions were not accessible through the GUI (there were indications on screen, but the functions were not active). For KGGSeq, the clinicians were unfamiliar with the terminal-style interface and had trouble knowing which parameters were required to setup a command, as well as how to provide them. The command-line GUI generator received some initial praise in the early stages of the tests, but negative comments were subsequently expressed as the clinicians realized only a portion of the functions could be access through the GUI. The remaining problems (3 for Varsifter, 7 for KGGSeq) reflected the absence of the necessary function to complete a task. For instance, in using KGGSeq, when trying to filter the variations by a list of genes, there is no easy way to upload a gene list. The program instead accepts a text string consisting of comma separated gene names; a format that is infeasible with a long gene list.

In addition, the clinicians faced tasks that the software seemed incapable to address (3/19 and 7/20 for Varsifter and KGGSeq respectively). For instance, the ability to identify compound heterozygous mutations (i.e. different mutations in the two copies of a gene) but limited to amino-acid changing or splice site mutations that arise infrequently across a population was indicated as a desired trait, but was not achievable by the users. In support of the generalizability of the findings, many (n=6) clinicians commented how they were unable to carry out analyses using other exome analysis packages (recalling that all our recruits have had prior experience analyzing exome data, and there is wide diversity in software used by the specialists). In section 4 and 5 of the main paper, we discuss in more details the nature of these important tasks and our design recommendations.

The usability issues detected for Varsifter and KGGSeq were not similar (Table 2 in the main paper). For Varsifter, a total of 106 comments pertaining to usability were made (92 comments were negative, 12 were positive, and 2 were neutral). For KGGSeq, a total of 90 comments pertaining to usability were provided (79 comments were negative, 10 were positive, and 1 was neutral). From previous studies, it is expected that most comments are likely to be negative in think-aloud protocols involving evaluation of user interfaces[95]. Varsifter received the most negative comments about navigation and system layout, while KGGSeq received the most negative comments about the semantics and the depth of information contained in the output. Both software received criticism about the scope of functionality offered, reflecting those instances when the clinicians were either unable to find the desired function to complete a task, or did not know how to properly execute the function.

In the study, we define a user-induced mistake that is not fixed by the user or the system as a slip, and a user-induced mistake that is caught and fixed by the user or the system as a near-miss. More mistakes were resolved by Varsifter (12/15) compared to KGGSeq (11/26) because clinicians were able to view the results at each filter step. For instance, a comment about KGGSeq quoted "…I am looking at the output file and there is nothing here. Did I do something wrong? Or is this expected from the simulated data?").



Appendix A Figure 3. The distribution of mistakes being caught by the user/system (near-miss) or uncaught (slips). Y-axis is the total number of user-induced mistakes encountered by each software. The iterative design of Varsifter enables fewer mistakes made by the users, and also a statistically significant greater proportion of mistakes to be caught (one-tailed chi-square test yields chi-square 4.0365, p-value < 0.05).

A.5 Data formats and databases

Common data formats for clinical genomic data analysis
BAM (http://samtools.sourceforge.net/)
BED (https://genome.ucsc.edu/FAQ/FAQformat.html)
Fasta, Fastq (http://maq.sourceforge.net/fastq.shtml)
GTF, GFF, GFV (https://genome.ucsc.edu/FAQ/FAQformat.html)
TSV (http://www.cs.tut.fi/~jkorpela/TSV.html)
VCF (and its many variations) (http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41)
Relevant databases for clinical genomic data analysis
1000 Genomes (http://www.1000genomes.org/)
dbNSFP (http://varianttools.sourceforge.net/Annotation/DbNSFP)
dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/)
Ensembl/UCSC (http://uswest.ensembl.org/index.html, http://genome.ucsc.edu/)
Exome variant server (http://evs.gs.washington.edu/EVS/)
Expression Atlas (http://www.ebi.ac.uk/gxa/)

Relevant databases for clinical genomic data analysis
Gen2Phen (http://www.gen2phen.org/)
GeneCards (http://www.genecards.org/)
HapMap (http://hapmap.ncbi.nlm.nih.gov/)
HGMD (http://www.hgmd.cf.ac.uk/)
Human Variome Project (http://www.humanvariomeproject.org/)
LOVD (http://www.lovd.nl/3.0/home)
OMIM (http://www.ncbi.nlm.nih.gov/omim)
Phenotips (http://phenotips.cs.toronto.edu/)

Appendix A Table 3. A list of the commonly encountered data formats in patient genomic analysis, and the databases often explored by clinicians.

A.6 Usability issues

The raw lists of usability issues specific to the evaluated tools are available online[114].

Appendix B Chapter 3

B.1 Example software for WES/WGS analysis

Existing software programs address differing portions of the analysis process, with emphasis tending to fall either on categories 1-2, 3-4 or 5, discussed in the Introduction section of the main text. For example, Galaxy[326] and NextGENe software from Softgenetics (http://www.softgenetics.com/NextGENe.html) emphasize the read processing stages, allowing users to upload the individual DNA sequences (in the FastQ format) and perform both genomic alignment and variant calling. VariantDB[327], BiERapp[328], and FamANN[329] focus on variant interpretation, taking as input a set of called variants (in a VCF or BAM format) and providing users with annotations and filtering functions. KGGSeq[330], gNOME[331], and PhenoVar[332] address the annotation and prioritization stages through the integration of gene annotation from resources, such as OMIM[159], Gene Ontology[333] and KEGG[334] databases. MagicViewer[56], and IGV[335] address quality control changes using graphical visualization of reads aligned to the reference genome, while allowing for prioritization based on read qualities and mapping thresholds. It is possible to incorporate additional types of data, exemplified by the commercial SNP&Variation suite (GoldenHelix) that, in addition to the range of functionalities described above, allows users to filter variants based on regions of heterozygosity determined using arrayCGH methods.

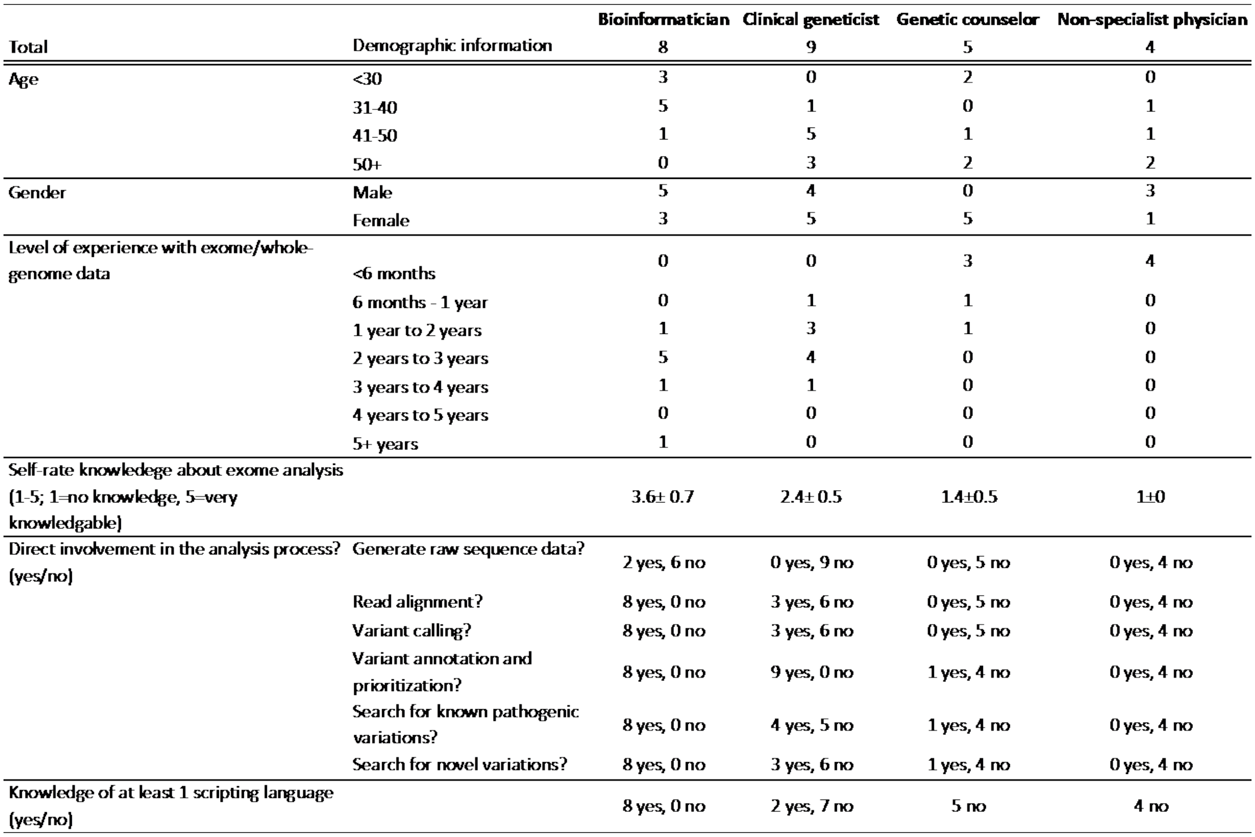
During this study's focus group sessions, we asked the subjects to share perspectives on these existing tools, or similar tools designed for to accomplish the same analytical tasks. Out of our recruits, 8/8 bioinformaticians, 7/9 clinical geneticists, and 2/5 genetic counselors had prior experience. Based on user feedback, two re-occurring issues that inhibit the adoption of these tools emerged: 1) incompatibility with an existing workflow, and 2) lack of focus on individual domain needs. Below we present examples from each user group for illustration.

Bioinformaticians did not find graphical user interface (GUI)-based commercial systems useful due to lack of freedom for constructing custom pipelines, and for the incapacity to modify specific components as new methods emerge. GUI packages were challenging to link with computer networks and be fitted to existing informatics pipelines. One bioinformatician remarked "It takes me three times longer to process an exome using the commercial system installed on a single machine when I can process it much faster on my local network". Clinical geneticists' chief complaint was the steep learning curve required to operate the software, even among the users who attended tool-specific workshops. The extensive learning curve problem is compounded when considering their overloaded schedules, as one clinician stated, "If I have the time, I am sure I can learn it, but I simply don't have that luxury". Genetic counselors criticized the over-abundance of graphical buttons and functionalities that have no relevance to their line of work, and whose presence were distracting from the clinically relevant information ("It is not only visually distracting, but also makes the software a lot more difficult to master than it has to be"). Finally, all evaluated users felt the existing systems are restricted to single workplace setting, and there is a growing need for a network-based system structured around shared project data to better foster collaboration. One bioinformatician expressed frustration that "I have to first convert my data into Excel Spreadsheet for the clinicians, and that always take up a lot of my time", and a geneticist recalled "when you have multiple versions of the same data floating around passing between different people, you can quickly loose track on what is the most current up-to-date analysis".

B.2 Participants

The study was approved by UBC research ethics board. Recruitments were conducted through emails and direct solicitations (crude response rate estimated to range between 20%-45%, but this was a rough estimate due to the fact many emails were sent out to specialized mailing lists, rather than directly to individual accounts). The resulting number of recruits for this study is the number of sought participants who agreed to the study. Each participant signed informed consent prior to the study. All participants identified as bioinformaticians, clinical geneticists or genetic counselors had prior experience working with WES/WGS data. The assignment to each of the four categories was based upon their professional and job title. Appendix Table B-1 shows the demographic distributions of the participants, as well as experience with exome/whole-genome analysis and computer programming.

Clinical geneticists were slightly skewed towards a higher age group, compared to bioinformaticians and genetic counselors. All user groups were composed of males and females except genetic counselors, which had only females as participants. Bioinformaticians had the highest self-rated competency in computer programming and greater experience with the various analytical steps within an exome pipeline, followed by clinical geneticists and genetic counselors respectively. As expected, non-specialist physician had the lowest amount of clinical experience with genomic data.



		Bioinformatician	Clinical geneticist	Genetic counselor	Non-specialist physician
Total	Demographic information	8	9	5	4
Age	<30	3	0	2	0
	31-40	5	1	0	1
	41-50	1	5	1	1
	50+	0	3	2	2
Gender	Male	5	4	0	3
	Female	3	5	5	1
Level of experience with exome/whole-genome data	<6 months	0	0	3	4
	6 months - 1 year	0	1	1	0
	1 year to 2 years	1	3	1	0
	2 years to 3 years	5	4	0	0
	3 years to 4 years	1	1	0	0
	4 years to 5 years	0	0	0	0
	5+ years	1	0	0	0
Self-rate knowledege about exome analysis (1-5; 1=no knowledge, 5=very knowledgable)		3.6± 0.7	2.4± 0.5	1.4±0.5	1±0
Direct involvement in the analysis process? (yes/no)	Generate raw sequence data?	2 yes, 6 no	0 yes, 9 no	0 yes, 5 no	0 yes, 4 no
	Read alignment?	8 yes, 0 no	3 yes, 6 no	0 yes, 5 no	0 yes, 4 no
	Variant calling?	8 yes, 0 no	3 yes, 6 no	0 yes, 5 no	0 yes, 4 no
	Variant annotation and prioritization?	8 yes, 0 no	9 yes, 0 no	1 yes, 4 no	0 yes, 4 no
	Search for known pathogenic variations?	8 yes, 0 no	4 yes, 5 no	1 yes, 4 no	0 yes, 4 no
	Search for novel variations?	8 yes, 0 no	3 yes, 6 no	1 yes, 4 no	0 yes, 4 no
Knowledge of at least 1 scripting language (yes/no)		8 yes, 0 no	2 yes, 7 no	5 no	4 no

Appendix B Table 1. Demographic information and self-rated computational competencies on the recruited participants (n=26) segregated into four different professional domains.

B.3 Additional details on focus group structure

For each specific issue, participants were asked about the type of information they preferred to see, and the properties of user interface design desired. Participants were encouraged to collectively draw out their ideal envisioned user interface design(s) on the whiteboard. If the participants appeared stuck or confused, mock-ups prepared by the moderator (CS) were presented for inspiration to elicit further responses (see Supplementary PowerPoint online). Participants were further instructed to let the moderator know if a particular question/scenario presented was not relevant to their line of work. A second round of interviews were held with the same participants in the same group composition in order to ensure that the digital images (translated from the drawings on the whiteboard) reflected the designs envisaged on the whiteboard. Each second round lasted less than 20 minutes.

B.4 Perspectives from genetic counselors and general physicians on raw sequence data

Genetic counselors and general physicians expressed no desire to access raw sequences, indicating that they did not consider it as part of their professional role. They described technological limitations that would preclude the data processing role in their perspective.

"I don't view myself qualified to process raw sequence data. Isn't that supposed to be what bioinformaticians do?" [Genetic counselor 03]

"Even if I know what to do with raw data, I doubt my computer can handle the processing of such large data!" [Genetic counselor 04]

B.5 Preferred file formats

There were contrasts between geneticists and genetic counselors versus the bioinformaticians regarding the preferred file formats. Bioinformaticians generally accepted diverse file formats (aside from technical complaints not directly related to WES/WGS, and lack of standardization of vocabularies across different institutions). Clinical geneticists and genetic counselors did not perceive canonical data formats (e.g. VCF, BAM) as being user friendly.

"I absolutely hate working with VCF files. I always have difficulties trying to load them into Excel and getting them to display properly." [Genetic counselor 01]

“I find it difficult to manipulate BAM files. I only use it to visualize the quality of alignment, but anything else is beyond my capabilities. And even after attending multiple workshops, I find there is too much of a learning curve for me to dive in.” [Clinical geneticist 04]

B.6 Re-occurring desired quality measurements that are not commonly available in current toolkits

There was a strong overlap between bioinformaticians and clinical geneticists when commenting on the quality measures desired, including such properties as average coverage, percentage of mapped reads, transition to transversion ratio, average read score. They desired a clear indication of the parameters used in alignment and variant-calling. When uploading multiple exomes from a common pedigree, both user groups (bioinformaticians and clinical geneticists) indicated a desire for the software to verify if the family assignment was correct based on the input exomes (e.g. ‘if an exome is assigned as “father”, does the underlying genetic data actually reflect this relationship?’). The ability to calculate degree of consanguinity from exome data was another attribute desired, as this information can be unreliable in patient testimony.

B.7 Bioinformaticians desired diverse variant information

When looking at a single nucleotide variation, while both clinicians and bioinformaticians would like to know the assembly version, genomic position, reference allele, alternative allele, genotype, and alignment quality, bioinformaticians further conveyed a desire for other information such as whether the variant overlaps a specific annotated feature type, such as a promoter sequence.

B.8 Automated literature mining or pathway analysis

Text mining algorithms exist which operate based on user-supplied keywords reflecting patient phenotype or a biological process of interest. They return a list of potentially relevant gene candidates, and are being more broadly incorporated into WES/WGS interpretation. This feature was desired by clinical geneticists and bioinformaticians. Since such algorithms are often based on mining a pre-compiled database, users indicated the importance of recording version numbers and dates of data updates, in order to assess the program's adequacy in quarrying ever-expanding clinical literature.

B.9 Comparative analysis of genetic tests

In this section, we summarize the variety of genetic tests employed in clinical diagnosis, and contrast them against WES/WGS. Traditional karyotyping had been the standard cytogenetic approach to detect large abnormal genomic deletions and duplications[336], but is gradually being superseded by array-based molecular techniques, which detect small genomic copy-number variants (CNVs) that are not routinely detected with karyotyping[337]. Gene-panel and PCR-targeted studies provide enhanced resolution by enabling the detection of single-base substitutions, or small insertions and deletions within a small subset of selected clinically relevant and disease-focused genes[338].

As the cost of DNA sequencing decreases, exome sequencing has become a clinical reality with potential as routine practice, with demonstrable successes in providing genetic diagnosis to rare, clinically unrecognizable, or puzzling disorders suspected to be genetic in origin[339]. Exomes are being considered for preventative medicine screening of healthy persons[340], as well as for individualized cancer therapy[341]. No exome capture kit reliably captures every exon in the human protein-coding genes[342]. The breadth of coverage is nonetheless significantly larger than gene panel approaches (~100 genes in a typical panel versus 20,000 captured protein-coding genes in an exome). Exome sequencing places less weight on the clinical assumptions of the patient's genetic makeup, as clinicians who utilize panel sequencing have to assume the gene(s) selected are of clinical relevance. If no pathogenic mutations are observed from a panel, no reliable conclusion can be made for rest of the genes outside the panel. Exomes, however, still come with certain genetic presuppositions, the most critical being that pathogenic variant(s) of interest rest within the protein coding region (making up ~2-3% of the human genome[343]). Whole-genome sequencing is currently the least biased approach to genetic testing, making no assumption about the location of a causal alteration. Veltman *et al.* demonstrated the power of whole-genome sequencing through a cohort of patients with intellectual disability, where previous genetic tests (including exome sequencing) provided a diagnosis for 42 percent of the patients, versus a potential diagnostic yield of 62 percent from whole-genome sequencing[344]. However, overall the clinical utility of whole genomes is limited by our capacity to assign biological significances to most non-coding variants[345, 346].

Exome sequencing and whole-genome sequencing do not yet replace array-based technologies, or traditional Sanger sequencing. The reliability and capacity of software to detect large genomic variations from sequencing data remains to be determined, while array comparative genomic hybridization (arrayCGH) is clinically confirmed for detection of large aberrations[347, 348]. Array-based methods also can reveal regions of homozygosity – information highly useful with patients of consanguineous background[276]. Therefore, current clinical workflows often include both arrayCGH and genome sequencing components to maximize the chance for successful genetic diagnosis[349]. The variant calls from exomes and whole-genomes have been unreliable compared to variants called by targeted Sanger sequencing[350], thus it is a common practice in both research and clinical pipelines to rely on Sanger confirmation.

From karyotyping to panels to arrays to whole-genomes, we see an exponential increase in computational complexity to generate, analyze, annotate, and store genomic data. An array panel of fifty genes may reveal ten polymorphisms, but a single exome can return over two hundred thousands variants, and a whole-genome returns variants reports millions[349]. These sociotechnical challenges when incorporating to clinical practice are discussed further in Appendix B-10.

B.10 Insights from Computer-Supported Cooperative Work

In this section, we highlight themes important that have emerged relative to past healthcare technology adoption in studies related to Computer-Supported Cooperative Work (CSCW). While not comprehensive, we seek to convey that certain workflow concepts across heterogeneous medical practices are applicable to clinical genomics. The perspectives from CSCW also serve as reminder that adoption of technical advances is hindered by inadequate consideration of the multidisciplinary team's needs and interactions (with each other and with systems).

Certain CSCW concepts may inform the process of assigning clinical importance to variants. One example is the need of methods for processing narrative and numeric data[351]. Our findings indicated the integration of patient phenotype that often comes in free-text narrative format is crucial for variant prioritization. Keeping in context of patient history is especially beneficial to assign clinical importance to a variant (e.g. a variant of unknown significance in MYH7 gene is unlikely to be important in a healthy individual, but highly significant in patient with familial cardiomyopathy). Another CSCW theme is the importance of working with ‘lay’ concepts and language[352]. This draws parallel in the collaborative genomic environment, where certain vocabularies and jargon surrounding genomic data would need to be conveyed to clinicians to inform their decision if a particular variant is worthy for further clinical pursuit (e.g. a variant in MYH7 may not be selected as the first candidate for additional testing if the variant has low read coverage).

CSCW cites themes concerning the multidisciplinary environment in collaborative healthcare practices. One example is the extension from single workplace to multi-workplaces setting. Various CSCW research studies emphasize the role of social networking to resolve individual problems, with examples primarily drawn from patients seeking other patients sharing similar problems[353, 354]. We perceive this to be relevant to genome interpretation for clinical professionals. Clinicians studying rare diseases often have to reach out to hospitals across boundaries. Through case matchmaking services like GeneYenta[355], clinical networks help clinicians connect and find patients with similar rare phenotypes and foster the compilation of deeper insight into disorders. Information credibility and interpretation of automated results has been addressed in CSCW studies[356]. In the context of WES/WGS, this aspect is apparent in the prediction of variants likely to disrupt a gene. The prioritization of variants requires a careful consideration of the aggregated meta-analysis of multiple outputs from different prediction programs or across an ensemble of biological features[357]. Therefore, it is critical that the limitation of each software and performance measurements such as sensitivity and specificity conveyed to the clinical user.

On the analytical side, CSCW studies have addressed technical issues surrounding high-throughput biological data. An example cited in Takacs et al is the need for scalable methods for handling increasingly large data sets[358]. In the context of genomic medicine, the transition from WES to WGS put an exponential increase of computational burden on both the time it takes to process the data, as well as the hardware required to store and load data for clinical interpretation. CSCW further illustrated the need to consider patient privacy. Most current clinical pipelines are in-house, but interactions related to rare genetic disorders will need to be established.

B.11 Standards for data incorporation into electronic health records (EHRs)

In this section, we discuss standards and technical challenges that arise when integrating genomic information into electronic health records for clinical practice. Incorporating the new information and transitioning from older genetic analysis methods will require adjustment of the data infrastructure of diverse organizations. The speed and ease of adoption can be improved by the establishment of standards for workflows and data formats. The workflow for processing genome sequence data is becoming more consistent across groups, but will have continuing volatility for several years. Establishing standards now for the output of the workflows, will allow for the internal mechanisms to continue development while not hindering the clinical adoption process.

The workflow for determining causal alterations and reporting the information is becoming standardized as well. Vassy *et al.* described a process where the ordering physicians would receive concise summaries of the key variants prepared by bioinformaticians, reviewed by geneticists and genetic counselors, and relayed to patients[359]. In figure 5, we highlighted an example of a concise report, similarly supporting the importance of down-weighting clinically non-relevant information from busy physicians and only report what are the most clinically significant for the patient. This maturation will allow for the key required advance of connecting the analysis results to health records.

The challenges of integrating genomic data into EHRs remain ongoing. Various laboratories and organizations (e.g. HL7) have laid out roadmaps and technical desiderata that need to be achieved for successful integration. Levy *et al.* offered a technical approach to compactly and efficiently represent genome information in operational systems, citing seven different considerations including the support of lossless data compression from primary molecular observations to clinically manageable subsets, simultaneous support of human-viewable formats and machine-readable formats for implementation of decision support rules, maintaining linkage of molecular observations to the laboratory methods used to generate them, and the anticipation of the continuously evolving understanding of human molecular variation[360, 361].

To date, no commercial EHR system has been described that systematically integrates genomic data. Electronic Medical Records and Genomics (eMERGE) is a consortium of nine institutions that has set out to provide pioneer experience using commercial prototypes and home-grown systems. Their experience, unexpectedly, overlapped with feedback from HL7[362, 363], and revealed additional core EHR functions that are needed in order to incorporate genomic information. These include storing genetic information as structured data conforming to standards that allow information to be moved freely between EHR systems, phenotypic information must also be stored as structured data and be associated with relevant genetic information, and EHR system must be able to obtain and display the information needed by clinician to interpret genotypic and phenotypic data[364-366].

B.12 Limitations

All of our subjects are employed within the same region, and therefore work within a single socialized medical system. Additional research with more participants within and beyond the current evaluated hospital/academic research centers would likely reveal additional insights. Secondly, our participants were self-motivated to enroll in the study. Even the non-specialist physicians had been exposed to WES/WGS and its utility. Finally, due to time constraints, we could not cover the entire scope of analysis that arises during exome analysis, rather the study was limited to key issues. For instance, our study did not address pharmacogenomics and the types of interaction pharmacologists might have with genetics professionals. Our focus groups were further guided around the genetic diagnosis of rare diseases, or puzzling diseases with suspected genetic etiology. Other clinical utilities such as preventive healthcare in screening of health individuals and sequencing of cancer tumors to find somatic mutations for individualized cancer therapy were not explored. These limitations serve as opportunities for future research.

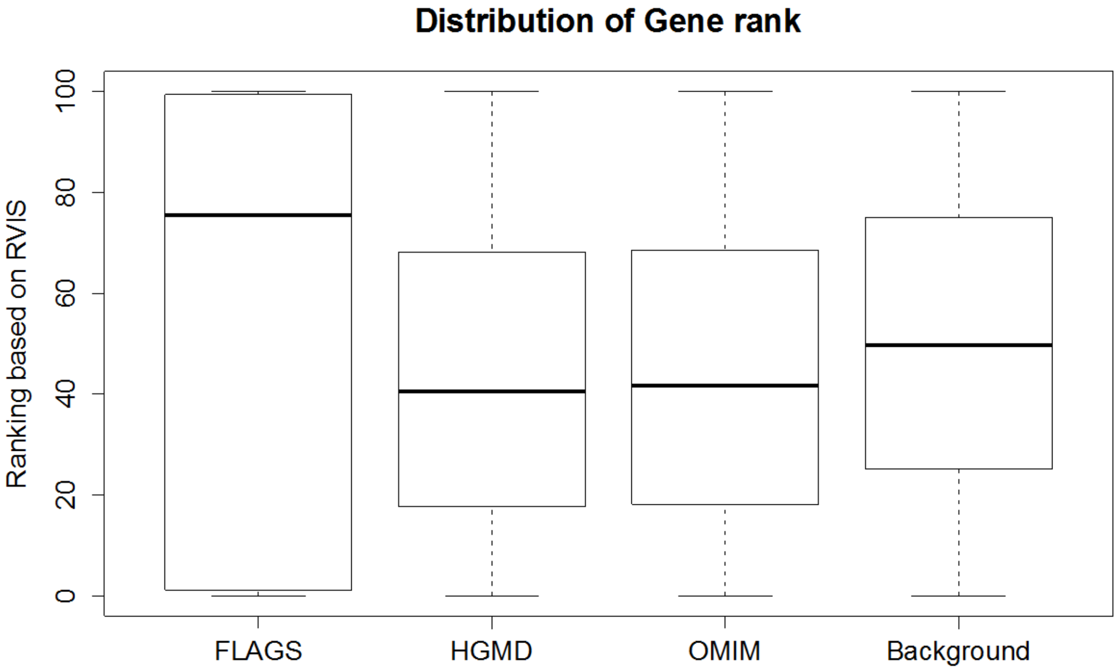
Appendix C Chapter 4

C.1 Distribution of variants across genes

For each gene, only rare coding variants derived from dbSNPv138 and EVS (including short InDels) within the longest coding transcript that results in amino acid change were considered. Polymorphic alleles were excluded based on the same allelic frequency criteria as described above. The location of the affected amino acid was derived from annotation by SnpEff[24] software (as described above). For InDels, only the first affected amino acid location was considered, such that if an InDel affected multiple amino acids, we only considered the location of the first one. To achieve meaningful statistical evaluations, any gene with < 20 remaining variants was not included in this part of the analysis. For each gene with ≥ 20 remaining variants, the same number of variants was randomly selected uniformly across the gene using Python version 2.7 random.randrange function. Mann-Whitney two-sided test was conducted between the locations of the observed mutations versus the locations from randomly selected ones using SciPy's[27] mannwhitneyu function. The p-value from this test was recorded, and the procedure repeated 20,000 times. Treating the p-value as a score, the p-value from this list corresponding to 99% statistical confidence was determined, reflecting how likely is the distribution of the observed variants to deviate from the uniform distribution. A Bonferroni multiple testing correction was applied when interpreting the significance of each p-value.

C.2 RVIS

Next, we analyzed the genic tolerance of the FLAGS gene set to variants. We expected FLAGS to be predicted to be more tolerant to variations and thus less likely to be impacted by pathogenic variants resulting in rare human diseases. To investigate this, we used a method published by Petrovski *et al.* (2013)[40] to assess the residual variation intolerance score (RVIS) for each gene based on their published supplementary dataset. This intolerance scoring system was developed by surveying whether a gene has relatively more or less functional genetic variation compared to the expected value based on neutral variations found in the same gene within the exomes from EVS. We chose this measurement because to our knowledge this is the only reliable published scoring system that is gene-centric rather than variant-centric. For each FLAGS gene, we extracted the relative rank based on the published intolerance score (the lower rank, the more intolerant the gene to variations), and we find that these FLAGS genes have a higher median score of 76 compared to OMIM, HGMD and Background which have medians of 42, 41 and 50 respectively (Appendix C Figure 1). However, Mann-Whitney U one-tailed tests revealed no significant differences (p-value between 0.05 and 0.1), likely attributable to the bimodal distribution of the ranks within the FLAGS, as there are genes within the FLAGS that have low RVIS ranks (n=32 with rank < 20). While this supports our findings that majority of the genes in FLAGS are ranked as more tolerant to variations, there are FLAGS that are predicted not to tolerate variation well. We found that these genes tend to have greater proportion of rare functional mutations over polymorphic functional mutations (Online Supplemental Data), which may explain why they receive RVIS ranks of <20. Namely, RVIS methodology does not consider rare functional variations, it ranks those genes as intolerant to genetic variation, despite the presence of numerous rare functional variants. We believe this may be a limitation on RVIS, because if a gene is observed to be frequently mutated with rare functional mutations yet is highly ranked as pathogenic in RVIS system, then by expectation that gene should not be highly ranked.



Appendix C Figure 1. Distribution of gene ranking across gene sets. The Y-axis plots the boxplot distribution of gene rank based on RVIS score.

C.3 In-house bioinformatics pipeline

In this section we discussed briefly the bioinformatics pipeline that we have setup in-house to process clinical exome data from TIDE-BC project. Because the project spans across multiple years, the software and genome versions have undergone various updates, so we will only provide the name of the software used but not the actual version.

The pipeline starts with pair-end 100bp Illumina reads in FASTQ format. The coverage of each exome or whole-genome ranges from as low as 30X to as high as 150X. Reference genome is hg19. Reads are aligned with Bowtie2 aligner under default parameter settings in a cluster server maintained in-house with 13 compute nodes, each with 16 CPUs and 32Gb RAM available per node. Aligned reads are sorted and merged into BAM using Samtools. Reads with

< 20 mapping quality score are discarded. Picard adds the read group and library information to the BAM file. GATK performs local re-alignment on the BAM file. BCF file is called from the re-aligned BAM using Samtools. VCF is generated using vcfutils.pl varFilter with mapping quality score 20 and a minimum of 2 alternative bases. Variants from VCF with less than 20 SNP quality score are further filtered out. Variant annotation is done by SnpEff with parameter –SpliceSiteSize 7 using always the latest available genomic annotation available at the time. Custom perl scripts are used to filter variants by Mendelian inheritance models (*de novo* dominant, homozygous recessive from either one or both parents, compound heterozygous), and filtering against dbSNP database (downloaded from UCSC Genome Browser) and ESP6500 downloaded from Exome variant server, and against the in-house already processed VCFs. Genomic coverage is analyzed using GATK on all the known exons downloaded from Ensembl Biomart. Candidate variants selected for further follow-ups are first manually screened on IGV for quality inspection before Sanger confirmation.

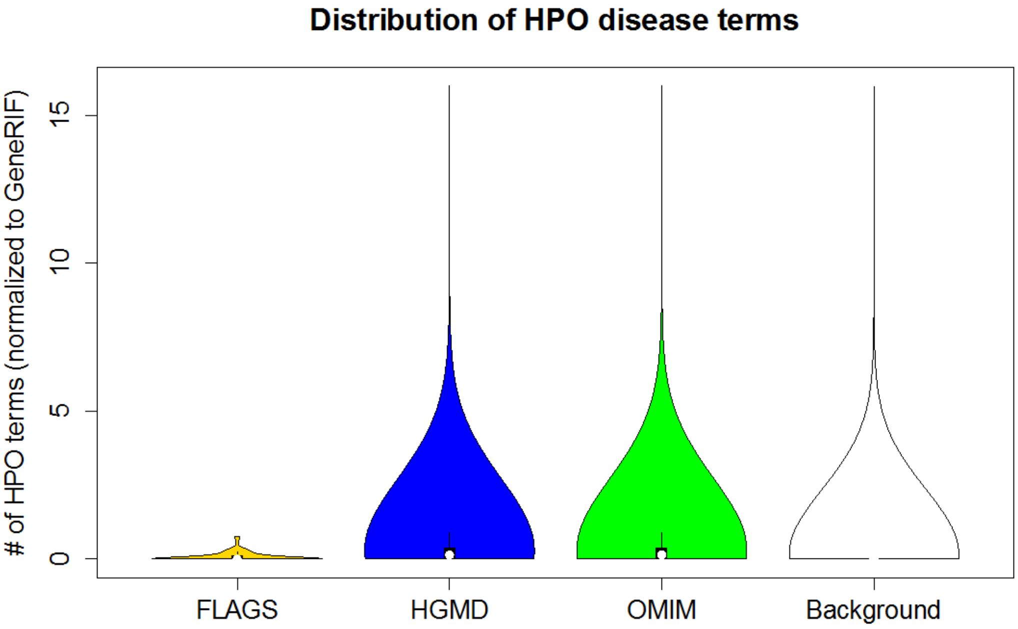
C.4 TIDE-BC

Tide BC (http://www.tidebc.org) is a new collaborative care & research initiative with a focus on prevention and treatment of Intellectual disability (ID). We have shown that the ID seen in some children is due to treatable genetic conditions known as inborn errors of metabolism (IEM). Many of these IEM’s can be treated with diet or drugs. Presently, health care policy and institutional culture is still operating under the old premise that all ID is incurable and thus, many children born with treatable ID are at risk of not being treated. At BC Children’s Hospital (BCCH) in Vancouver, Canada, 1500 patients with ID are seen for diagnostic assessment per year by various services, such as neurology, medical genetics, biochemical diseases, developmental pediatrics and child psychiatry. With the local expertise of all these specialists, existing diagnostic laboratory methods, and the major advances in diagnostic and therapeutic technologies, BCCH is the ideal academic location to implement our evidence-based protocol to identify treatable causes of ID. TIDEX was designed by TIDE-BC investigators to take advantage of new technologies to help crack the code for those families who have undergone the million dollar workup and are still unable to receive a diagnosis for their child's debilitating condition. These technological advances, coupled with TIDE-BCs already proven approach, has every promise in providing much needed answers to help those families.

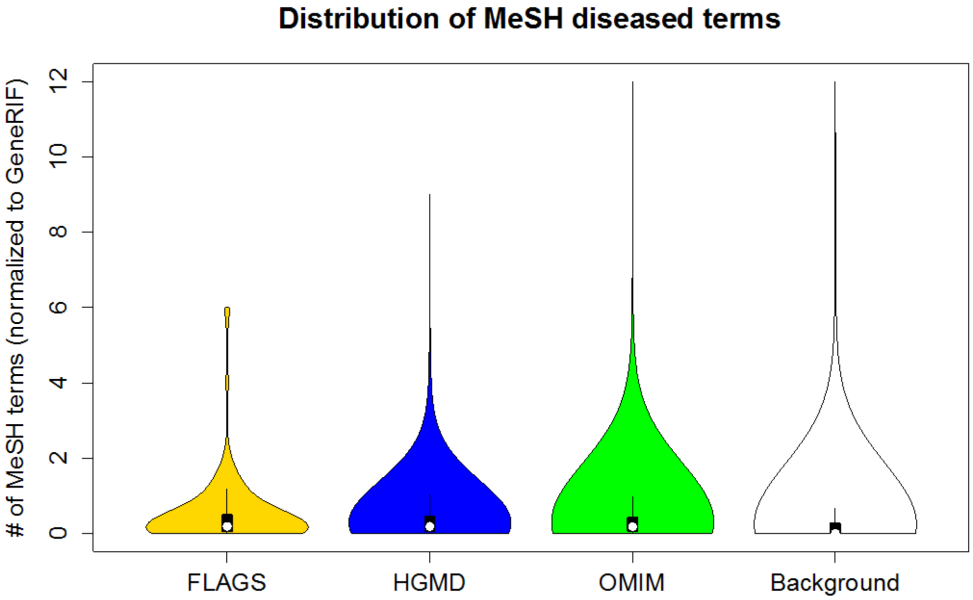
In order to provide those answers, TIDE-BC investigators are presently looking for those undiagnosed patients who have some evidence of an interrupted metabolic pathway or enzyme deficiency. This may be abnormal chemicals in body fluids such as blood or urine or test results that provide a clue that a biochemical pathway may be altered. Then by comparing the protein coding regions or "whole exome" of DNA they hope to find the cause. As sequencing cost continues to decrease, the project now shifts more and more towards whole-genome sequencing, rather than only restricted to exomes. The additional sequencing of one or more healthy family members helps them to eliminate sequence variations that do not contribute to the disorder. The informatics team, based within Dr. Wyeth Wasserman lab, uses a new, CFI-funded computational system. It features high-capacity storage (~0.3 petabytes), a set of high-performance servers supporting virtualized computing, a computing cluster with ~100 computing cores, and a tape system for long-term genome data archiving. The system is interconnected with 10 gigabyte channels for efficiency. Once the genetic cause is found, this group of metabolic disorders are often amenable to simple and successful treatments, sometimes only involving dietary changes or dietary supplementation.

C.5 HPO and MeSH terms normalized to GeneRIF

To adjust for the potential bias that genes with more articles are likely to have more MeSH and HPO terms attached, we repeated the analysis by normalizing the MeSH and HPO terms to the number of publications in GeneRIF. Appendix C Figure 2A and 2B show the violin distribution of HPO and MeSH terms per gene after normalization.



Appendix C Figure 2A. The Y-axis plots the number of HPO disease terms per gene after normalizing to the number of entries from GeneRIF for the same given gene. FLAGS have significantly fewer terms than OMIM, HGMD and significantly more terms than Background (each p-value << 0.00001; Mann-Whitney 1-tailed test).



Appendix C Figure 2B. The Y-axis plots the number of MeSH disease terms per gene from MeSHOP after normalizing to the number of entries from GeneRIF for the same given gene. There are no significant differences observed between FLAGS and OMIM and HGMD, but FLAGS have significantly more terms than Background (p-value << 0.00001; Mann-Whitney 1-tailed test)

C.6 Application in in-house WES /WGS database

To further demonstrate the utility of this study, we evaluated how frequently FLAGS appear as gene candidates in an in-house collection of 150 exomes and 13 whole genomes – comprising of 53 independent families suffering from distinct rare inborn errors of metabolism (IEM) (http://www.tidebc.org). These cases represent a collection of exome and whole genomes collected over a period of 3 years to study rare intellectual disorders exhibiting metabolic defects. Each family displayed a unique undiagnosed IEM, and the family structures range from singleton case (i.e. proband only) to paired (mother-proband; proband-affected sibling) to trio (father-mother-proband) to quartet (father-mother-proband-sibling) [for more details on exact breakdown of family structure, see Appendix D Table 4]. In each family, rare functional variants falling into Mendelian inheritance patterns were extracted by Wasserman laboratory in-house pipeline, which we then overlapped against FLAGS. When focusing only on the top 100 frequently mutated genes from FLAGS, on average across all 53 families, we see ~3 genes from the recessive models overlapping with the FLAGS per family, which is around ~8% of the recessive candidates per family. From the *de novo* dominant model, on average ~4 genes overlapped with FLAGS, which is around ~3% of the *de novo* candidates per family. This demonstrates that many top genes in FLAGS do indeed show up at a relatively frequent rate across exome families despite after applying rigorous canonical filtering at the variant level. While these results are drawn from data processed by an in-house pipeline based on a specific class of disorder, our processing methodology is built on popular tools setup in a workflow as recommended by Broad Institute (http://www.broadinstitute.org/gatk/guide/best-practices) using standard parameters and common filtering strategies such that they should be reproducible in other labs using a similar approach in studying other classes of rare Mendelian disorders.

C.7 Supplementary tables

The Supplementary tables referred to in this chapter are available online[228].

Appendix D Chapter 5

D.1 Classifying variants in the training set

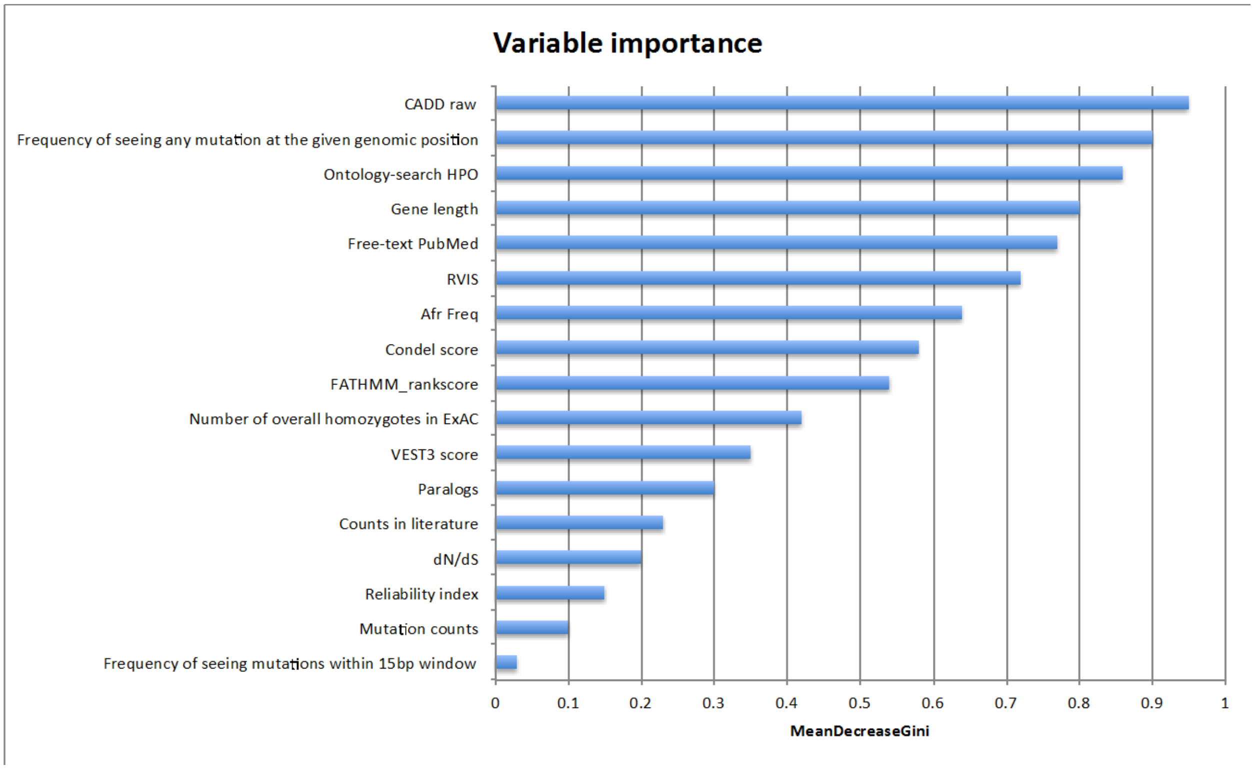
Each ClinVar variant was classified according to the following rules: ‘pathogenic’ if it had multiple independent reviewers reporting its pathogenicity (CLNSIG=5 in the CLNSIG attribute); ‘likely pathogenic’ if it had only one reviewer annotating the variant as pathogenic or likely pathogenic (CLNSIG=4), or if it had multiple reviewers but not all reviewers marked it as pathogenic; ‘VUS’ or ‘benign’ if the variation only had a score of 0 or 2 in CLNSIG attribute respectively.

D.2 Additional details on feature selection and model building

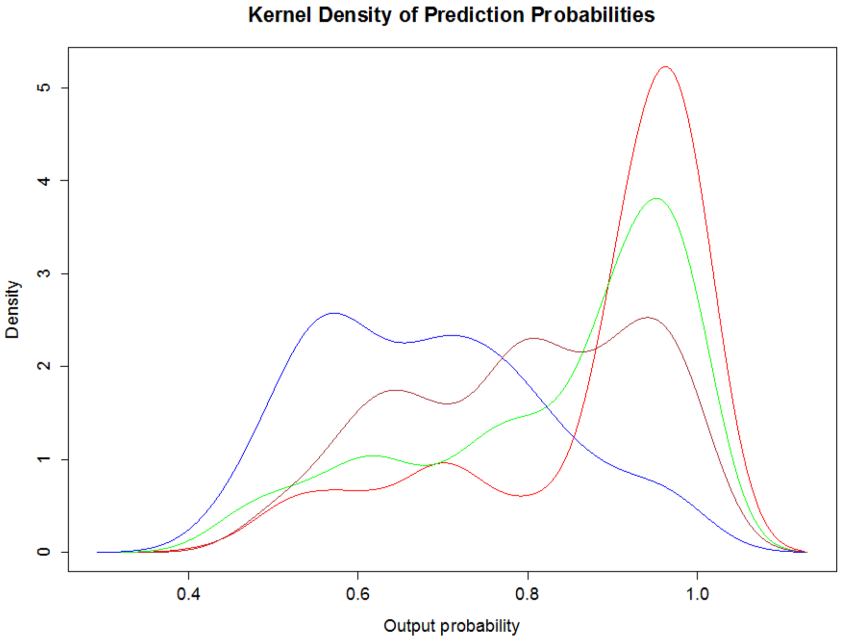
For each feature considered in this study, we assigned it to a hierarchy level within a hierarchical tree (Appendix D Table 5). This assignment was done based on manual assessment of the primary literature behind each feature. A feature that was not dependent upon other features was placed at the bottom level, while a feature that built upon previous features was placed at a progressively higher level. Using the features ranked at the top of the hierarchy (e.g. coarsest domain), we divided the training set into 3548 distinct partitions with each partition corresponding to inputs sharing the same values at the coarsest domain. A total of 1000 trees were used to train the random forest. In each tree-building iteration, an input row from the data was randomly selected under uniform distribution until the number of selected rows equal to the number of partitions, and this training subset was used to train for that one particular tree. Each tree was trained using hie-ran-forest package in R 3.2.0 with number of clusters set to 6 using the hierarchy tree discussed earlier. This taking-one-example-per-stratum approach assured that no gene-level information would be over-represented in a single tree, and the hierarchical selection during each tree construction ensured the compositions of scores were accounted for. Table S2.1 summarizes the statistical performance on the training set after a 10-fold cross-validation. Overall, VPA achieved an 81.2% true positive rate, a 3.1% false positive, with a receiver operating characteristic area-under-the-curve (AUROC) of 0.912. Classification performance was best for "Pathogenic" variants, followed by "Likely pathogenic" variants, and then "Benign" variants. VUS got the lowest performance, which is not surprising given that VUSs were expected to display the greatest amount of heterogeneity in their biological characteristics. Appendix D Figure 1 shows the importance of each selected feature based on the MeanDecreaseGini value. The values were scaled to be between 0 and 1. Appendix D Figure 2 shows the kernel density distribution of the probability assigned to each prediction. Appendix D Table 2 shows the overall statistical performance obtained from other explored machine-learning techniques.

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.941	0.027	0.894	0.941	0.896	0.908	0.986	0.93	Pathogenic
	0.888	0.029	0.879	0.888	0.873	0.851	0.926	0.891	Likely pathogenic
	0.713	0.034	0.765	0.713	0.777	0.737	0.869	0.698	VUS
	0.872	0.031	0.794	0.872	0.827	0.796	0.929	0.808	Benign
Overall	0.812	0.031	0.810	0.812	0.822	0.797	0.912	0.789	

Appendix D Table 1 Statistical performance on training set.



Appendix D Figure 1 Importance weight for the selected features.



Appendix D Figure 2 Density plot of the probabilities assigned to each predicted mutation class. In this figure, the probabilities are plotted according to the variant class in which the variants are annotated in the training set. Red corresponds to variants from the pathogenic class, green corresponds to the likely pathogenic class, blue corresponds to the VUS, and brown corresponds to the benign class.

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.699	0.057	0.684	0.699	0.712	0.683	0.789	0.672	Decision tree
0.753	0.039	0.779	0.753	0.743	0.729	0.858	0.737	Random forest (without hierarchical feature selection)
0.401	0.068	0.583	0.401	0.422	0.401	0.599	0.374	Naïve Bayes
0.721	0.047	0.741	0.721	0.719	0.699	0.832	0.698	Multinomial logistic
0.739	0.042	0.744	0.739	0.726	0.711	0.847	0.712	Multi-class SVM

Appendix D Table 2 Statistical performances on other evaluated machine-learning methodologies. The algorithms were executed from the e1071 package in R. The numbers here represent the overall performance across all the 4 mutation classes under 10-fold cross validation. Hierarchical feature selection was not applied to any of these methods.

D.3 Constructing patient-level features

In the simulated cases, the phenotypic terms were drawn directly from ClinVar if available, or were drawn from OMIM by first converting the ClinVar disease label to an OMIM entry, and mapped to the closest HPO vocabularies using WordNet word-to-word mapping implemented in SEMILAR (http://semanticsimilarity.org), taking the best hit using default threshold. Variants without clinical annotations in ClinVar (e.g. benign mutations) were randomly assigned clinical annotations drawn from a pool of phenotypic terms that were observed in the “Pathogenic” and “Likely Pathogenic” classes with a probability corresponding to each term’s observed frequency. In actual clinical sets, the terms were derived from clinicians’ input and mapped to HPO using the same approach as described above. In both simulated and real test sets, annotations that could not be mapped to HPO vocabularies were disregarded for ontology-based scoring, but kept for calculation of free-text-based scoring. The HPO OBO file was incorporated using OntoCAT API (http://sourceforge.net/projects/ontocat/) by extracting the ancestral and descendent terms for each HPO term and writing to custom MySQL tables.

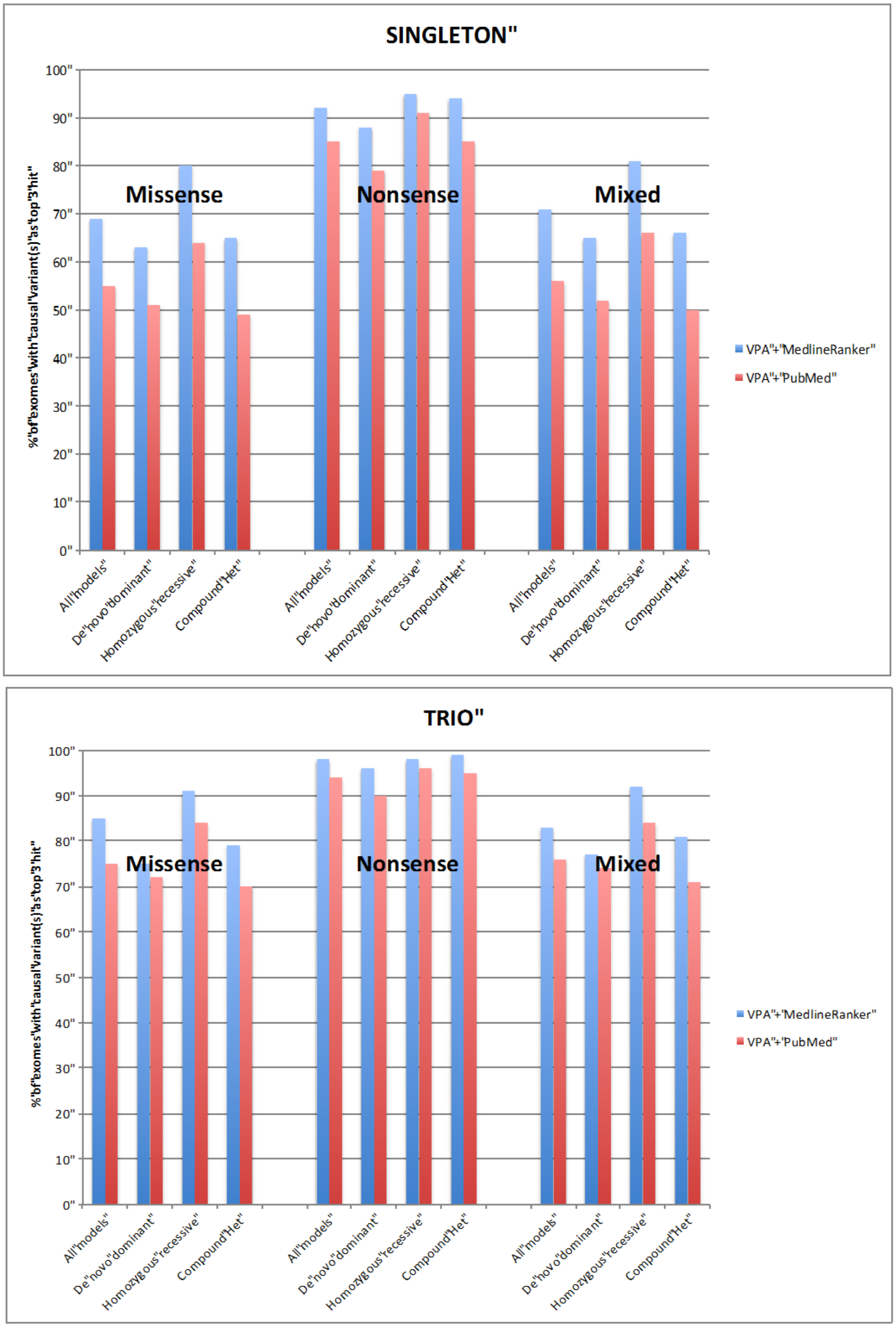
D.4 Constructing Sim(t,t’) and R_t

Part of the procedure was previously described in the GeneYenta publication (cited in the Main text) under the context of patient-to-patient matching, and is partially repeated here under the context of patient-to-gene matching for reader convenience. Sim(t,t’) is equal to the information content for t if t is equal to t’ or if t is a member of the set of all ancestors of t’. Otherwise, sim(t,t’) returns the highest value score among the sets of ancestors for term t and t’. The matching algorithm essentially sums up the weighted similarities between each term in the patient set and the most similar term in the gene set, and then divides it by the highest possible score that could be associated with the terms in the patient set. In the patient phenotype set (pat), R_t is an importance-ranking integer specified by the user (e.g. clinicians) at the beginning as part of data input, ranging from 1 to 5, 1 = least important and 5 = most important. By default, R_t is assigned to value of 3 unless specified otherwise. R_t for the gene set (dis) was derived from HPO consortium’s annotation of frequency for observing a phenotype in a given disease. If a phenotype is annotated as “very rare”, “rare” or “occasional”, R_t was assigned a weight of 1, if “frequent” then 2, if “typical” or “variable” then 3, if “common” then 4, and if “hallmark” or “obligate” then 5. If frequencies were given instead of categorical words, then the assignment of R_t is as follow: <10% = 1, 10%-35% = 2, 35%-60% = 3, 60%-85% = 4, and 85%-100% = 5. If multiple phenotypes were assigned to the same gene with different R_t score, the higher one was taken. If no annotation was available for frequency of phenotype, then a default 3 was assigned.

These thresholds were arbitrarily assigned to reflect phenotype frequency, and in theory could have been allocated in multiple similar methodologies.

D.5 Free-text matching using PubMed

Custom Python scripts were written to extract genes returned by PubMed when searching for clinical keywords in the NCBI Gene database. The extraction was done through PubMed API, restricting to Homo sapiens as the species and sorting by relevance. The returned genes were ranked according to total number of genes returned minus the rank at which the gene appeared in the returned output plus 1, divided by the square of the total number in the returned list, and multiplied by the weight of clinical importance (default=3). For example, if the returned list has 10 genes, the gene at the top of the list would be assigned a rank of (10-1+1)/10^^{2}*3. If multiple clinical terms were provided per case, then each term was searched individually and multiplied by the clinical weight (which can be different for each term, if user specified), and if a gene had multiple scores from multiple results, then the max of that would be taken. Performance was compared between VPA + PubMed and VPA + MedlineRanker in the same way as described in the Results section. VPA + MedlineRanker displayed superior performance over VPA + PubMed query (Appendix Figure 3). This was expected given PubMed's naïve implementation for gene query based on clinical descriptors, compared against MedlineRanker, which is a specialized tool developed for the purpose.



Appendix D Figure 3 Performance for singletons and trios between using MedlineRanker versus PubMed for free-text matching.

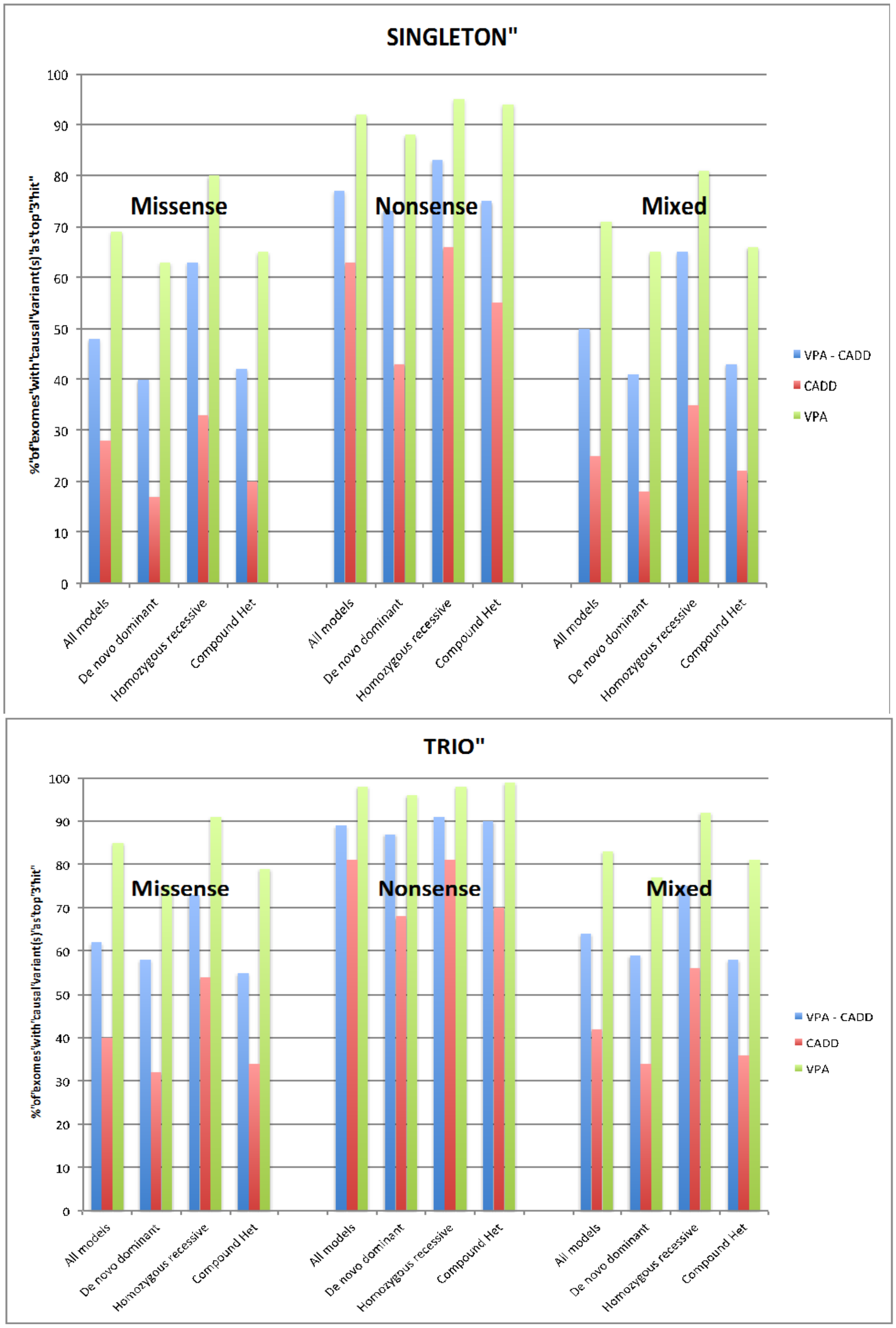
D.6 Extracting variant-gene rank

For Exomiser and eXtasy, the ranks of the pathogenic mutations were extracted directly from the software output under default thresholds. For VPA, each variant was assigned a score reflecting likelihood for pathogenicity. Each variant was classified individually without taking other variants into context, except for the variants in compound heterozygous model where the probabilities returned by VPA for variants affecting the same gene were summed and normalized before being assigned a final classification label. The compound heterozygous variants were ranked with respect to rest of the variants in other Mendelian models. For variants that were homozygous for the mutation, VPA took the score for the individual variant and multiplied it by 2 to reflect the genotype. Ultimately, VPA ranked all the mutations across the genetic models collectively and mutations with the highest likelihood score were ranked at the top. To derive a rank from CADD, for compound heterozygous variants, CADD scores were summed up per gene and then ranked with all the other variants across the considered genetic models. CADD scores for variants of homozygous genotype were multiplied by 2. Variants (or gene-variant pairs for the compound heterozygous list) with the highest CADD score were returned at the top. The CADD scores for InDels were computed by first submitting to the CADD website, and storing the returned result in a custom MySQL table.

D.7 Performance without CADD

We first retrained our model with all the CADD-related scores removed from the list of features considered. In their absence, additional features were recruited after the hierarchical sampling + feature selection as described in the Methods: GERP++ NR, VEST3_score, and Fold degenerate. Below we compared the performance of VPA with CADD versus VPA without

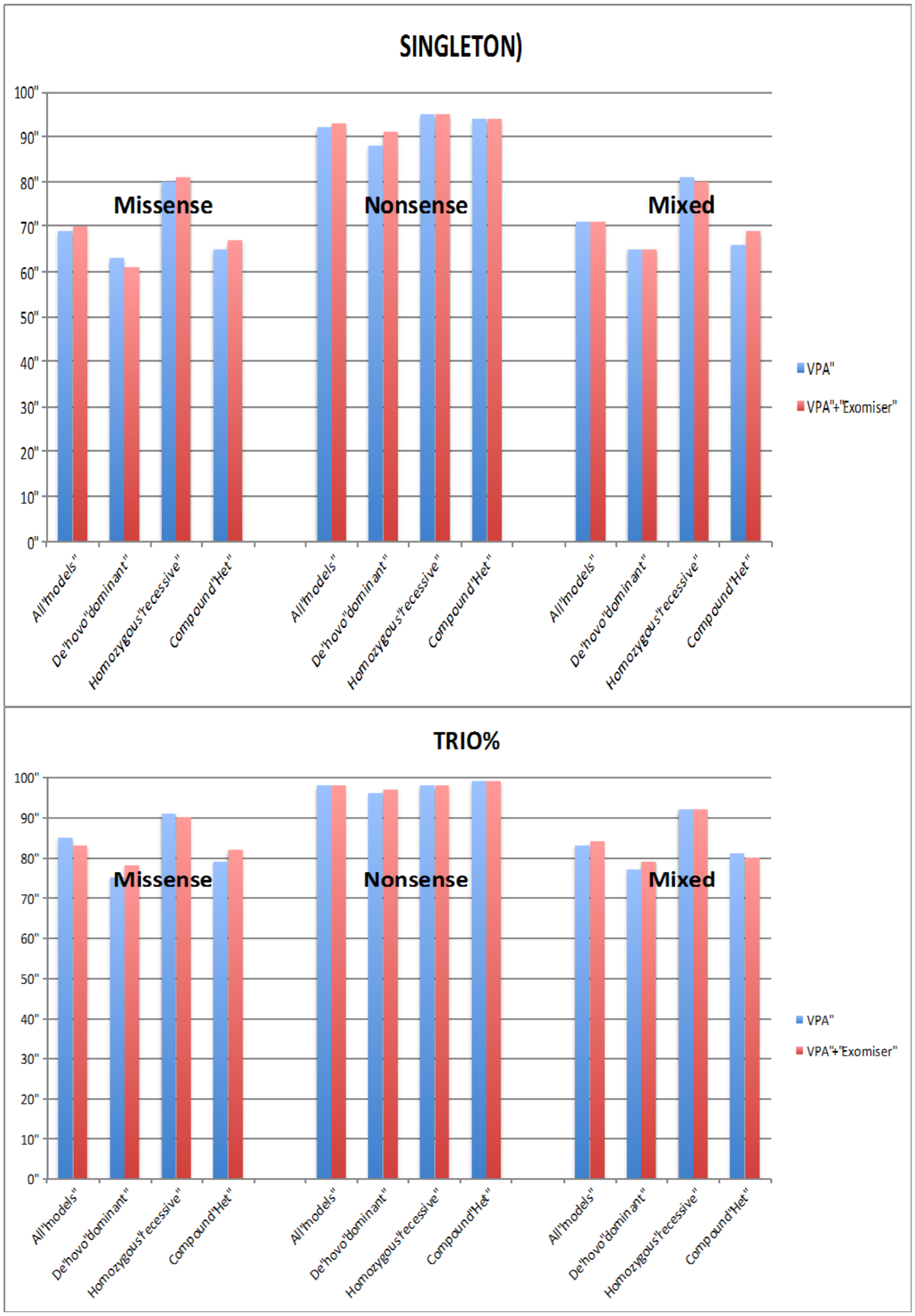
CADD to see how much information is CADD bringing. The type of evaluation was the same as described in Methods. From Appendix D Figures 4, it became apparent that CADD was an important contributor to the overall performance, which is unsurprising given its demonstrated superiority over previous variant-level prediction methodologies, and its high feature weight assigned during cross-validation (Appendix D-2).



Appendix D Figure 4 Performance in singletons and trios for VPA with versus without CADD as a feature.

D.8 Including Exomiser and eXtasy as features

Exomiser score was incorporated into VPA to evaluate how much additional information Exomiser supplies. The model was retrained and we found Exomiser was not selected from our feature list. Nevertheless, we still tried to see what are the performance differences by incorporating Exomiser via looking at the top 3 predictions that appeared in both software and testing the performance against VPA without Exomiser as a feature. Appendix D Figure 5 below shows the performance on the simulated dataset, evaluating the ability to predict embedded pathogenic mutations as among the top 3 predicted candidates based on the same methodology as previously described. We observed fluctuations in performance, where the incorporation of Exomiser sometimes improved predictions but in other cases it caused a decrease in the performance. Overall, for singleton cases we do not see any predictive improvement with Exomiser incorporated (both models achieved 71% averaging across all genetic models in the mixed mutations category). For trio cases, VPA with Exomiser incorporated saw a very slight increase in the overall performance (84% versus 83%). These observations agreed with the exclusion of the Exomiser score in the model. A similar exercise was repeated for eXtasy score but as this feature did not make it past feature selection, we did not pursue it further.



Appendix D Figure 5 Performance in singletons and trios for VPA with versus without Exomiser as a feature.

D.9 Clinical example of mutation that is not missense/nonsense

An illustrative example: a boy with a 13bp deletion in PLP1 (OMIM 300401) presenting with global developmental delay, spasticity, nystagmus, ataxia, and most notably severe hypomyelination of early myelinating structures (HEMS). Result is published in PMID 26125040. VPA predicted the pathogenicity of the deletion with a rank of 4. Exomiser and eXtasy were not able to assign prediction to this illustrative mutation.

D.10 Constructing simulations with novel phenotype associations

In this section, we describe how we constructed simulations to represent cases with novel phenotype associations. The goal at this stage is to introduce phenotypic terms into each simulated case where the terms realistically represent novel disease phenotypes. The number of terms introduced is derived by taking the ceiling of the number of HPO terms that we started with per simulated case multiplied by 0.3. The newly introduced terms were selected based on three criteria: 1) they could not be one of the descendants from the already-included HPO terms in the current simulated case, 2) they could not be one of the ancestral nodes located on the shortest path from each of the included HPO terms to the root, and 3) they had to be within ± 2 hierarchical levels with respect to the HPO hierarchical tree with one of the already-included HPO nodes. These restrictions were imposed to ensure the randomly inserted phenotypes are still clinically realistic to represent novel phenotypes.

D.11 Novel genes

From the 53 clinical families, we derived a subset of clinical patients harboring pathogenic mutations in genes that had not been previously directly cited in clinical literature for human diseases. Among 10 such families with novel gene associations, VPA achieved better performance in 9 families (the last family was a draw between VPA and Exomiser) based on the comparison on the predicted ranks for pathogenic mutations. At the time of writing, these families are described under a separate manuscript that is currently under review. We note that 5 of 10 families had one or more primary keywords that could not be mapped to HPO, so the performance differences reported here could not be solely attributed to novel gene associations.

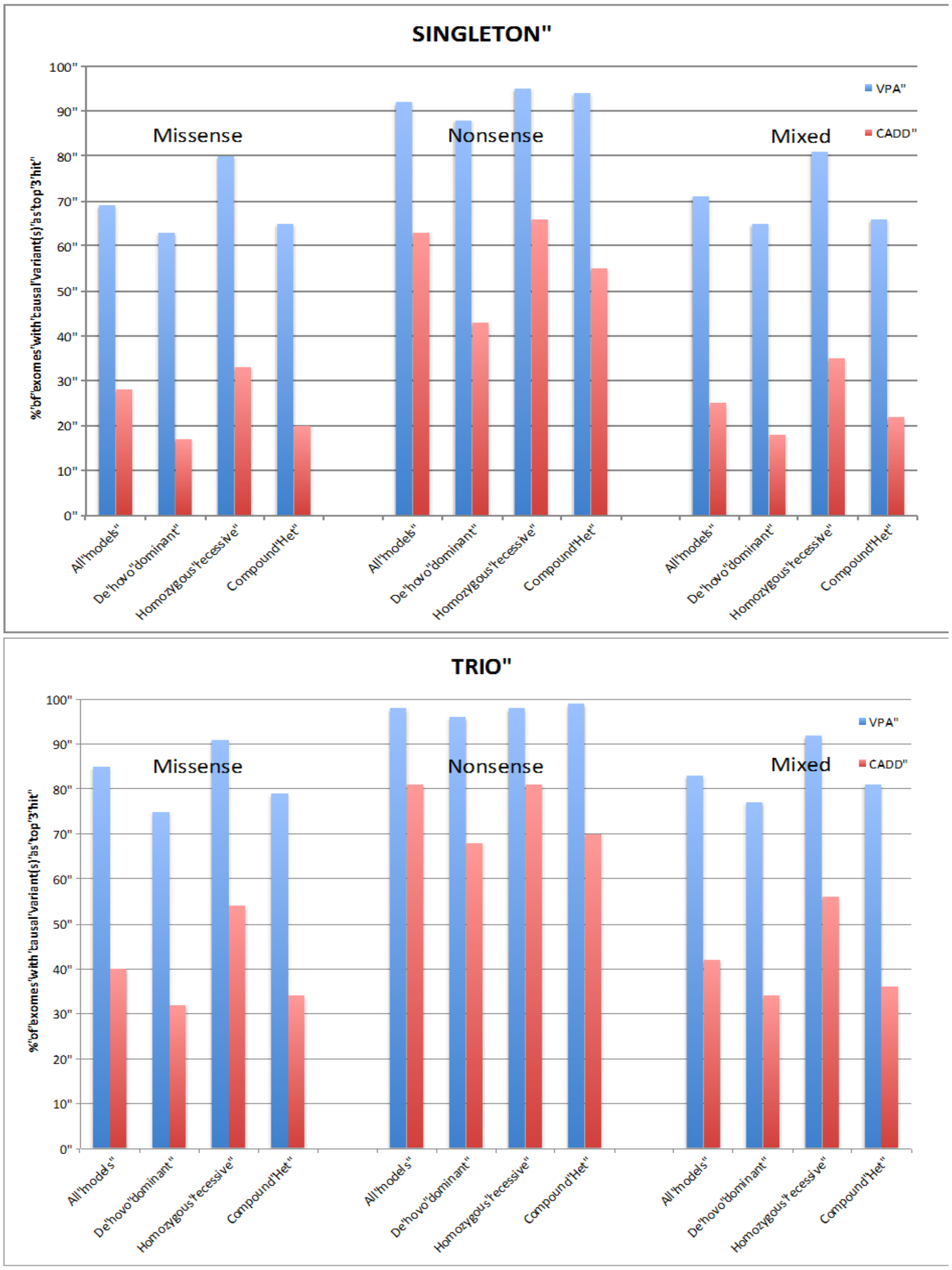
D.12 Constructing simulations with polygenic/oligogenic phenotypes

To construct simulations representing polygenic/oligeogenic phenotypes, we embedded two causal mutations in distinct genes into each simulation. Each causal mutation was randomly drawn from the pool of HGMD mutations as described in Methods, and each mutation has an attached set of HPO terms (derived by the protocols described in Methods). The final set of HPO terms that was fed into the predictive models was derived by first performing a union of HPO terms from the two causal mutations, and then randomly choosing 50% of these terms. Each model received the same final set of input terms. Performance was evaluated in two ways: 1) percentage of success to identify one of the two embedded pathogenic mutations as among the top 3 candidates, and 2) percentage of success to identify the second pathogenic mutation as among the top 5 candidates.

D.13 Assigning clinical weights to each phenotypic term

Keywords which could not be clearly distinguished between primary versus secondary were left with a score of “3” (the same default as used to produce the earlier results). These assignments were provided by clinicians who oversaw the patients. They were the same healthcare providers who provided the original clinical descriptors for each family.

D.14 VPA versus CADD on simulated test sets



Appendix D Figure 6 Performance of VPA and CADD on 750 simulated exomes for singletons and trios, broken down by category of embedded pathogenic mutations that were either missense or nonsense or mixed, e.g.

missense/nonsense pathogenic mutations from HGMD were selected without discrimination of its type, and family structure (trios versus singletons). The Y-axis shows the percentage of cases where the embedded pathogenic mutation(s) were predicted to be among the top 3 candidates. The "All models" category shows the average performance across *de novo*, compound heterozygous and homozygous recessive for the given mutation type and family structure. We consider the "All models" under the "Mixed" category to be representative of the overall performance within each respective family structure.

D.15 Size of simulated dataset

Genetic models	**Numbers of mutations after standard filtering**			
	Singleton - average	**Singleton - SD**	**Trio - average**	**Trio - SD**
Homozygous recessive (including hemizygous)	16	5	10	4
Compound heterozygous	54	8	21	5
De novo heterozygous	89	18	33	11

Appendix D Table 3 The number of rare, protein-coding single nucleotide variants that remained on average across the evaluated simulated cases, broken down by family structure and genetic models, after filtering against dbSNPv142 for allelic frequency ≤ 1% in the general population, and focusing on protein-coding variants that result in a change in the sequence of the protein transcript. 750 cases were considered for each family structure. SD = standard deviation.

D.16 Structure of clinical test set

Family Number	**Genetic model that contains the pathogenic mutation(s)**	**Family structure**
1	Compound	Quart
2	Homo Rec	Trio
3	Compound het	Trio
4	Compound het	Trio
5	Homo Rec	Quart
6	Homo Rec	Trio
7	Compound het	Trio
8	Compound het	Trio
9	*De novo*	Trio

Family Number	Genetic model that contains the pathogenic mutation(s)	Family structure
10	Homo Rec	Singleton
11	Homo Rec	Trio
12	Homo Rec	Trio
13	*De novo*	Trio
14	Compound het	Quart
15	Compound het	Duo
16	Compound het	Trio
17	Compound het	Trio
18	Compound het	Singleton
19	*De novo*	Trio
20	Compound het	Quart
21	Compound het	Trio
22	Homo Rec	Duo
23	*De novo*	Trio
24	*De novo*	Duo
25	*De novo*	Quart
26	*De novo*	Trio
27	Compound het	Trio
28	*De novo*	Trio
29	Homo Rec	Duo
30	Compound het	Singleton
31	Homo Rec	Duo
32	*De novo*	Trio
33	Compound het	Trio
34	Compound het	Quart
35	Compound het	Trio
36	*De novo*	Trio
37	Homo Rec	Duo
38	Homo Rec	Quart
39	Homo Rec	Trio
40	Compound het	Trio
41	*De novo*	Trio
42	Homo Rec	Trio
43	Compound het	Singleton
44	Compound het	Duo
45	Compound het	Trio
46	Compound het	Trio
47	*De novo*	Singleton

Family Number	Genetic model that contains the pathogenic mutation(s)	Family structure
48	Compound het	Trio
49	Homo Rec	Trio
50	Homo Rec	Trio
51	*De novo*	Trio
52	Homo Rec	Duo
53	Homo Rec	Trio

Appendix D Table 4 The first column contains the ID assigned to each clinical case. The second column contains the genetic model in which the pathogenic variant(s) was identified from. Homo Rec =homozygous recessive, *De novo* = *de novo* heterozygous, and Compound het = compound heterozygous. The last column contains the information on the availability of exomes in the family at the time of analysis that led to the identification of the pathogenic variant(s). The "Duo" category can refer to either exomes for index + one parent (typically the mom), or can refer to exomes on multiple affected individuals (e.g. siblings).

D.17 Complete list of features considered

Type	Granularity	Features	Source
V	1	In-house allelic frequency	In-house database of 370+ exomes and whole-genomes
V	1	Frequency of seeing any mutation at the given genomic position	In-house database of 370+ exomes and whole-genomes
V	1	Frequency of seeing mutations within 15bp window	In-house database of 370+ exomes and whole-genomes
V	1	Frequency of seeing mutations within 25bp window	In-house database of 370+ exomes and whole-genomes
V	1	Frequency of seeing mutations within 35bp window	In-house database of 370+ exomes and whole-genomes
V	1	Frequency in ExAC	ExAC
V	1	Number of overall homozygotes in ExAC	ExAC
V	1	Afr Freq	ExAC
V	1	Amr Freq	ExAC
V	1	Eas freq	ExAC
V	1	Fin freq	ExAC
V	1	Nfe freq	ExAC
V	1	Oth freq	ExAC
V	1	Sas freq	ExAC
V	1	Average # of reads covering the genomic position	ExAC

Type	Granularity	Features	Source
V	1	Site quality	ExAC
V	2	SLR_test_statistic	dbNSFP
V	1	codonpos	dbNSFP
V	1	fold-degenerate	dbNSFP
V	2	SIFT_score	dbNSFP
V	2	SIFT_converted_rankscore	dbNSFP
V	2	Polyphen2_HDIV_rankscore	dbNSFP
V	2	Polyphen2_HVAR_rankscore	dbNSFP
V	2	PROVEAN score	PROVEAN
V	2	Condel score	CONDEL
V	2	LRT_score	dbNSFP
V	2	LRT_converted_rankscore	dbNSFP
V	2	MutationTaster_score	dbNSFP
V	2	MutationTaster_converted_rankscore	dbNSFP
V	2	MutationAssessor_score	dbNSFP
V	2	MutationAssessor_rankscore	dbNSFP
V	2	FATHMM_rankscore	dbNSFP
V	2	MetaSVM_score	dbNSFP
V	2	MetaSVM_rankscore	dbNSFP
V	2	MetaLR_score	dbNSFP
V	2	MetaLR_rankscore	dbNSFP
V	2	Reliability_index	dbNSFP
V	2	VEST3_score	dbNSFP
V	2	VEST3_rankscore	dbNSFP
V	2	CADD_raw	CADD
V	2	CADD_raw_rankscore	CADD
V	2	CADD_phred	CADD
V	2	GERP++_NR	UCSC
V	2	GERP++_RS	UCSC
V	2	GERP++_RS_rankscore	UCSC
V	2	phyloP46way_primate	UCSC
V	2	phyloP46way_primate_rankscore	UCSC
V	2	phyloP46way_placental	UCSC
V	2	phyloP46way_placental_rankscore	UCSC
V	2	phyloP100way_vertebrate	UCSC
V	2	phyloP100way_vertebrate_rankscore	UCSC
V	2	phastCons46way_primate	UCSC
V	2	phastCons46way_primate_rankscore	UCSC
V	2	phastCons46way_placental	UCSC
V	2	phastCons46way_placental_rankscore	UCSC
V	2	phastCons100way_vertebrate	UCSC

Type	Granularity	Features	Source
V	2	phastCons100way_vertebrate_rankscore	UCSC
V	2	Presence in known regulatory regions	FANTOM
V	2	SiPhy_29way_logOdds	dbNSFP
V	2	SiPhy_29way_logOdds_rankscore	dbNSFP
V	2	LRT_Omega	dbNSFP
V	1	1000Gp1_AC	ESP6500
V	1	1000Gp1_AF	ESP6500
V	1	1000Gp1_AFR_AC	ESP6500
V	1	1000Gp1_AFR_AF	ESP6500
V	1	1000Gp1_EUR_AC	ESP6500
V	1	1000Gp1_EUR_AF	ESP6500
V	1	1000Gp1_AMR_AC	ESP6500
V	1	1000Gp1_AMR_AF	ESP6500
V	1	1000Gp1_ASN_AC	ESP6500
V	1	1000Gp1_ASN_AF	ESP6500
V	1	ESP6500_AA_AF	ESP6500
V	1	ESP6500_EA_AF	ESP6500
V	1	ARIC5606_AA_AC	ESP6500
V	1	ARIC5606_AA_AF	ESP6500
V	1	ARIC5606_EA_AC	ESP6500
V	1	ARIC5606_EA_AF	ESP6500
V	1	Average coverage at genomic position	ESP6500
V	1	COSMIC_CNT	dbNSFP
G	3	Mutation counts	FLAGS
G	3	dN/dS	FLAGS
G	3	dN/dS version 2	Ensembl
G	3	Gene length	Ensembl
G	4	# of MeSH terms	FLAGS
G	4	# of HPO terms	FLAGS
G	3	Paralogs	FLAGS
G	3	Paralogs version 2	HOGENOM
G	3	Counts in literature	FLAGS
G	3	# of mutations in HGMD	FLAGS
G	4	RVIS	RVIS
G	3	Interactions(IntAct)	dbNSFP
G	3	Interactions(BioGRID)	dbNSFP
G	3	Interactions(ConsensusPathDB)	dbNSFP
G	4	P(HI)	dbNSFP
G	4	P(rec)	dbNSFP
G	4	Known_rec_info	dbNSFP
G	4	Essential_gene	dbNSFP

Type	Granularity	Features	Source
G	3	ZFIN_zebrafish_phenotype_tag	dbNSFP
G	3	Mouse phenotype entry	MGI
P	3	Free-text PubMed	Presented in paper
P	3	Ontology-search HPO	Presented in paper
P	4	eXtasy score	eXtasy
P	4	Exomiser score	Exomiser

Appendix D Table 5 A total of 103 features were considered. In the first column, V= variant-level, G = gene-level, P = patient-level. In the second column, “1” correspond to the bottom of the hierarchy, “4” is at the top of the hierarchy. The third column describes the property of the feature, and the last column shows the source in which the feature was derived from. Whenever possible, we used the latest command-line accessible version for each tool/database at the time of the analysis. ExAC version 0.3 was derived from http://exac.broadinstitute.org version 0.3. dbNSFP was downloaded from https://sites.google.com/site/jpopgen/dbNSFP, version v3.0 beta2. PROVEAN version 1.1 was downloaded from http://provean.jcvi.org/downloads.php. CONDEL was downloaded from FannsDB version 2.0. CADD version 1.2 was downloaded from http://cadd.gs.washington.edu. UCSC refers to the data tables attached to reference genome GRCh37 on the UCSC Genome Browser. The data was downloaded using the Table Browser function. RVIS was downloaded from http://chgv.org/GenicIntolerance/ based on the unpublished version on ExAC sequencing datasets. eXtasy and Exomiser refer to the scores output from the respective software (software version cited in the Main text). ESP6500 was downloaded from http://evs.gs.washington.edu/EVS/ version V2. FLAGS was derived from our publication in 2014 BMC Medical Genomics, doi:10.1186/s12920-014-0064-y. Ensembl referred to the Ensembl databases version 80 accessed via Ensembl API. FANTOM data was accessed via collaboration with FANTOM5 consortium. MGI phenotype data was downloaded from Mouse Genome Informatics in April 2015. HOGENOM version 06 was downloaded from its home website.

D.18 Diagnosis breakdown per family

ID	Nature of diagnosis	Multiple genes	Non-mappable words
1	New phenotype		
2	New gene		Yes
3	New gene	2 genes	
4	New phenotype		

ID	Nature of diagnosis	Multiple genes	Non-mappable words
5	New gene		
6	New gene		Yes
7	New phenotype		
8	New gene		
9	New phenotype		Yes
10	New phenotype		
11	New phenotype		
12	New phenotype		
13	New phenotype		
14	New phenotype		Yes
15			Yes
16	New gene		
17	New gene		
18	New gene		Yes
19	New phenotype		Yes
20	New phenotype	2 genes	
21	New phenotype		
22			Yes
23	New phenotype		
24			
25	New phenotype		
26	New phenotype		
27			
28	New gene		Yes
29			Yes
30			Yes
31			Yes
32			
33			
34			Yes
35			Yes
36	New phenotype		Yes
37		2 genes	
38			Yes
39		2 genes	Yes
40			
41			Yes
42			
43		2 genes	

ID	Nature of diagnosis	Multiple genes	Non-mappable words
44			Yes
45	New gene		Yes
46			
47			
48		2 genes	Yes
49		3 genes	Yes
50			
51			Yes
52			Yes
53			

Appendix D Table 6 The first column corresponds to the ID for each family. The second column indicates if the family harbored a novel/rare pathogenic variant(s) in a gene not previously reported in human diseases (e.g. novel gene-disease associations, referred in the column as “New gene”), or if the family harbored novel/rare pathogenic variant(s) in known disease genes but the patient(s) displayed symptoms not previously reported (e.g. novel disease-phenotype associations, referred to as “New phenotype”). The assignments of novel associations were provided by a systematic review of literature by a team of clinical geneticists, molecular biochemists, bioinformaticians and genetic counselors involved in each clinical case. The third column corresponds to if the family contained multiple pathogenic variants in distinct genes. Six families had 2 impacted genes; one family had 3 impacted genes. The last column specifies if the clinical descriptors supplied by the clinicians prior to the exome analysis could be mapped to controlled terminologies in Human Phenotype Ontology (HPO). Only descriptors corresponding to the primary phenotypes were considered.

D.19 Polygenic/oligogenic families

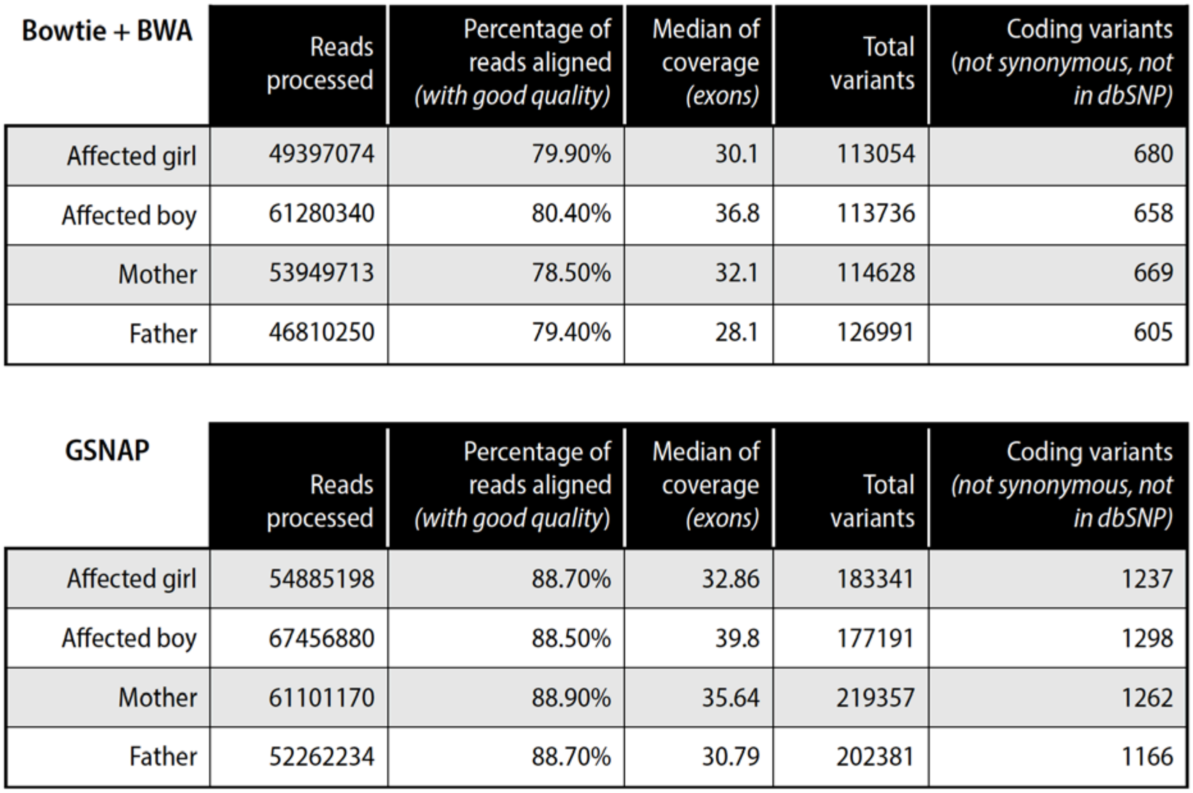
Family	Gene 1				Gene 2				Gene 3			
	VPA	Exomiser	eXtasy	CADD	VPA	Exomiser	eXtasy	CADD	VPA	Exomiser	eXtasy	CADD
3	1	6	4	7	3	9	6	8				
20	1	5	3	4	6	10	5	10				
37	4	6	7	10	4	8	9	16				
39	1	3	7	9	3	7	9	9				
43	8	10	14	13	9	14	17	15				

Family	Gene 1				Gene 2				Gene 3			
	VPA	Exomiser	eXtasy	CADD	VPA	Exomiser	eXtasy	CADD	VPA	Exomiser	eXtasy	CADD
48	5	7	13	9	8	NA	NA	11				
49	2	5	8	11	4	6	NA	13	7	15	13	16

Appendix D Table 7 The table displays the 7 families that possessed multiple pathogenic mutations hitting in distinct genes, contributing to blended phenotypes. We showed the performance capability to predict these mutations per family from VPA, Exomiser, eXtasy and CADD by displaying the predicted rank at which the diseased gene-variant appeared. Exomiser and eXtasy were unable to predict for family #48's second pathogenic variant because the second variant is an insertion. eXtasy was unable to predict the variant in family #49 due to it being a nonsense mutation. The leftmost column corresponds to the family ID. The numbers in each subsequent column correspond to the rank assigned to the pathogenic variant(s) for each model. Only the last family (ID 49) had three disease genes impacted.

Appendix E Chapter 6

E.1 Coverage and variants of the exome dataset



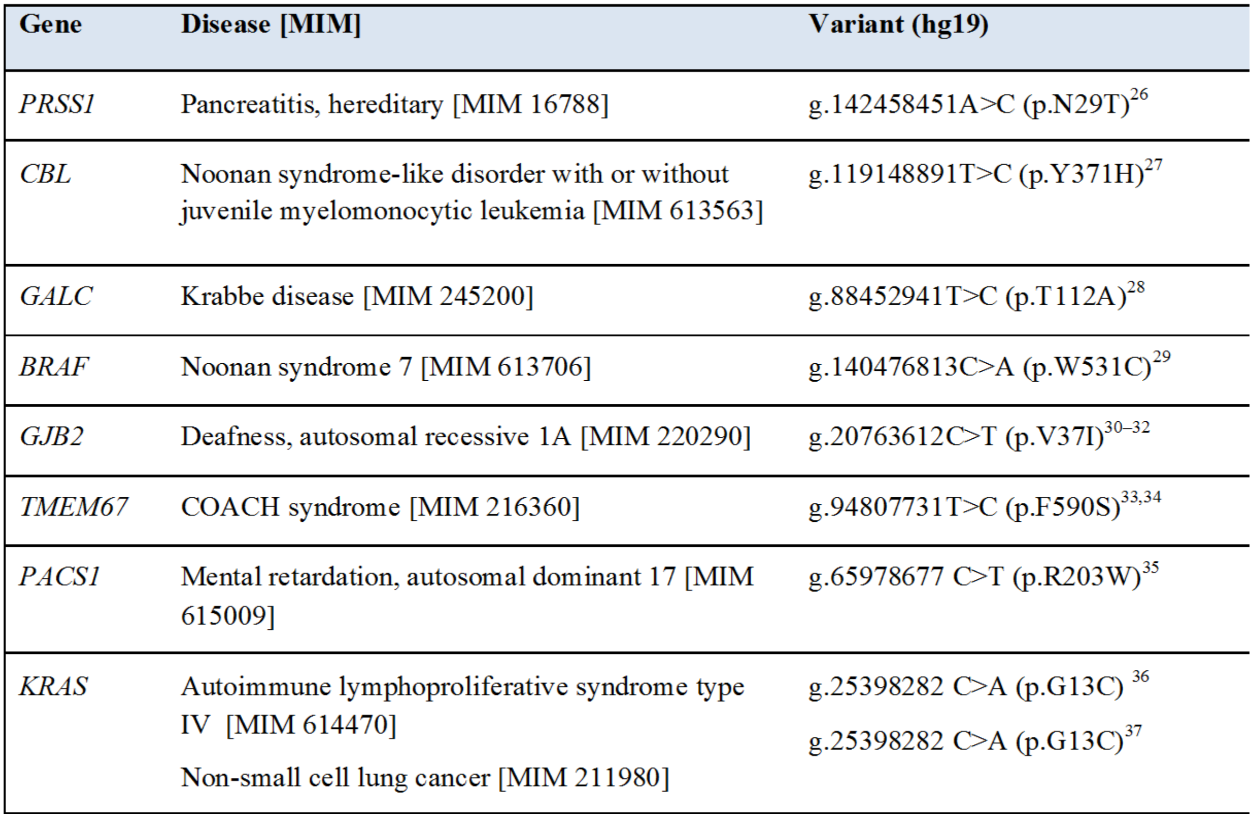
Bowtie + BWA	Reads processed	Percentage of reads aligned *(with good quality)*	Median of coverage *(exons)*	Total variants	Coding variants *(not synonymous, not in dbSNP)*
Affected girl	49397074	79.90%	30.1	113054	680
Affected boy	61280340	80.40%	36.8	113736	658
Mother	53949713	78.50%	32.1	114628	669
Father	46810250	79.40%	28.1	126991	605

GSNAP	Reads processed	Percentage of reads aligned *(with good quality)*	Median of coverage *(exons)*	Total variants	Coding variants *(not synonymous, not in dbSNP)*
Affected girl	54885198	88.70%	32.86	183341	1237
Affected boy	67456880	88.50%	39.8	177191	1298
Mother	61101170	88.90%	35.64	219357	1262
Father	52262234	88.70%	30.79	202381	1166

Appendix E Figure 1 A combination of Bowtie, BWA, and GSNAP were used to map the reads to the hg19 reference genome, and Samtools was used to identify variants. SnpEff was used to assign annotations to the variations, with respect to the hg19 database. Allele frequency was assessed in dbSNP (version 137; downloaded from UCSC Table Browser "All SNPs(137") on Feb 19, 2013. ESP data was downloaded from the NHLBI ESP server on June 2, 2013. The total number of starting reads for each individual is listed in the second column, and only reads with mapping quality of #20 are kept (percentage shown in third column). Coverage is shown in the fourth column, and is based on all known human exons compiled from Ensembl Biomart. The total number of variations, including InDels, is listed in the fifth column. The sixth column lists the number of variations that remain after filtering against intergenic or intronic variations, polymorphisms, and synonymous mutations.

Appendix F Chapter 7

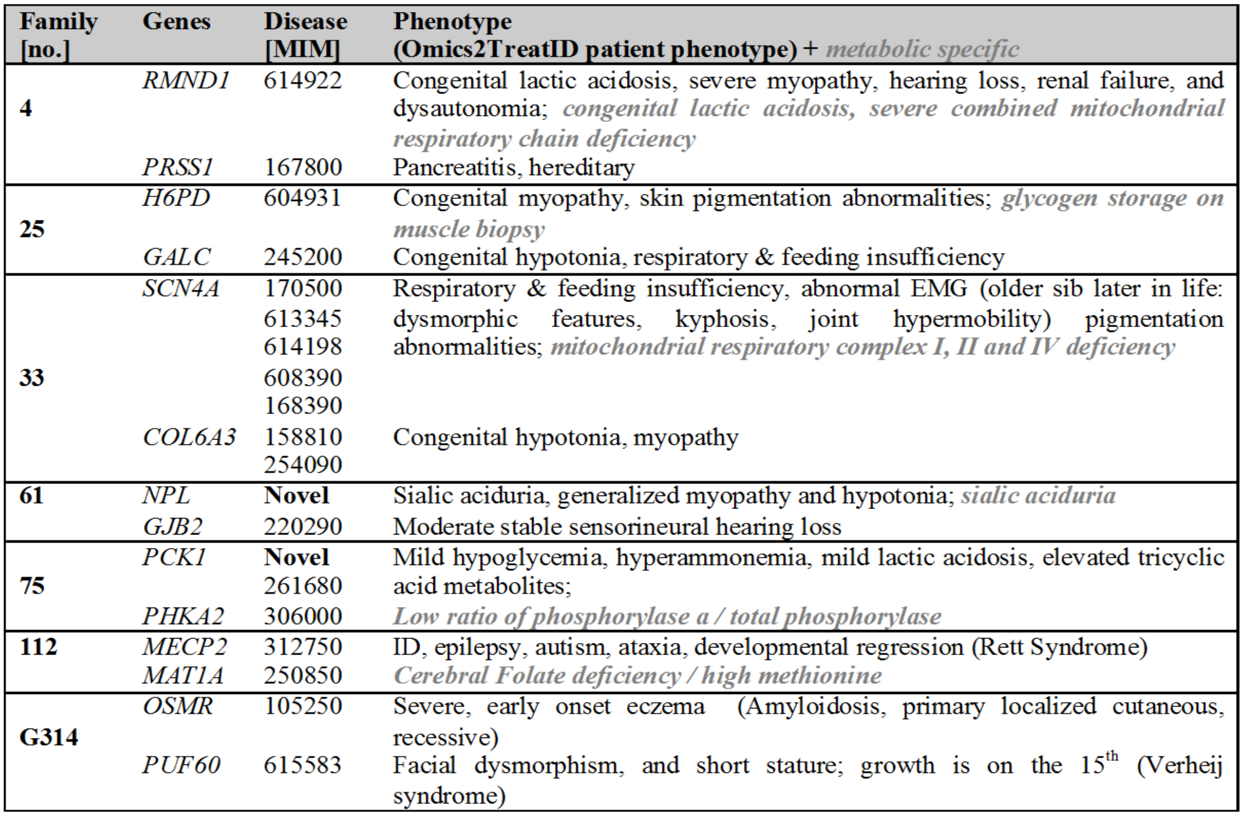
F.1 Known pathogenic variants



Gene	Disease [MIM]	Variant (hg19)
PRSS1	Pancreatitis, hereditary [MIM 16788]	g.142458451A>C (p.N29T)[26]
CBL	Noonan syndrome-like disorder with or without juvenile myelomonocytic leukemia [MIM 613563]	g.119148891T>C (p.Y371H)[27]
GALC	Krabbe disease [MIM 245200]	g.88452941T>C (p.T112A)[28]
BRAF	Noonan syndrome 7 [MIM 613706]	g.140476813C>A (p.W531C)[29]
GJB2	Deafness, autosomal recessive 1A [MIM 220290]	g.20763612C>T (p.V37I)[30–32]
TMEM67	COACH syndrome [MIM 216360]	g.94807731T>C (p.F590S)[33,34]
PACS1	Mental retardation, autosomal dominant 17 [MIM 615009]	g.65978677 C>T (p.R203W)[35]
KRAS	Autoimmune lymphoproliferative syndrome type IV [MIM 614470] Non-small cell lung cancer [MIM 211980]	g.25398282 C>A (p.G13C) [36] g.25398282 C>A (p.G13C)[37]

Appendix F Table 1. Known pathogenic variants.

F.2 Blended phenotypes resulting from two single gene defects



Family [no.]	**Genes**	**Disease [MIM]**	**Phenotype (Omics2TreatID patient phenotype) +** *metabolic specific*
4	*RMND1*	614922	Congenital lactic acidosis, severe myopathy, hearing loss, renal failure, and dysautonomia; ***congenital lactic acidosis, severe combined mitochondrial respiratory chain deficiency***
	PRSS1	167800	Pancreatitis, hereditary
25	*H6PD*	604931	Congenital myopathy, skin pigmentation abnormalities; ***glycogen storage on muscle biopsy***
	GALC	245200	Congenital hypotonia, respiratory & feeding insufficiency
33	*SCN4A*	170500 613345 614198 608390 168390	Respiratory & feeding insufficiency, abnormal EMG (older sib later in life: dysmorphic features, kyphosis, joint hypermobility) pigmentation abnormalities; ***mitochondrial respiratory complex I, II and IV deficiency***
	COL6A3	158810 254090	Congenital hypotonia, myopathy
61	*NPL*	**Novel**	Sialic aciduria, generalized myopathy and hypotonia; ***sialic aciduria***
	GJB2	220290	Moderate stable sensorineural hearing loss
75	*PCK1*	**Novel** 261680	Mild hypoglycemia, hyperammonemia, mild lactic acidosis, elevated tricyclic acid metabolites;
	PHKA2	306000	***Low ratio of phosphorylase a / total phosphorylase***
112	*MECP2*	312750	ID, epilepsy, autism, ataxia, developmental regression (Rett Syndrome)
	MAT1A	250850	***Cerebral Folate deficiency / high methionine***
G314	*OSMR*	105250	Severe, early onset eczema (Amyloidosis, primary localized cutaneous, recessive)
	PUF60	615583	Facial dysmorphism, and short stature; growth is on the 15th (Verheij syndrome)

Appendix F Table 2 Blended phenotypes resulting from two single gene defects