DEVELOPMENT OF HUMAN-COMPUTER INTERACTIVE APPROACHES FOR RARE DISEASE GENOMICS

by

Jessica J. Y. Lee

B.Sc., The University of British Columbia, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

November 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Development of Human-Computer Interactive Approaches for Rare Disease Genomics

submitted by Jessica J. Y. Lee in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Genome Science and Technology

Examining Committee:

Clara van Karnebeek, Pediatrics; Genome Science and Technology
Co-supervisor

Wyeth Wasserman, Medical Genetics
Co-supervisor

William Hsiao, Pathology and Laboratory Medicine
Supervisory Committee Member

Martin Dawes, Family Practice
University Examiner

Sabrina Wong, Nursing
University Examiner

Additional Supervisory Committee Members:

Sara Mostafavi, Statistics
Supervisory Committee Member

Raymond Ng, Computer Science
Supervisory Committee Member

Abstract

Clinical genome sequencing is becoming a tool for standard clinical practice. Many studies have presented sequencing as effective for both diagnosing and informing the management of genetic diseases. However, the task of finding the causal variant(s) of a rare genetic disease within an individual is often difficult due to the large number of identified variants and lack of direct evidence of causality. Current computational solutions harness existing genetic knowledge in order to infer the pathogenicity of the variant(s), as well as filter those unlikely to be pathogenic. Such methods can bring focus to a compact set (less than hundreds) of variants. However, they are not sufficient to interpret causality of variants for patient phenotypes; interpretation involves expert examination and synthesis of complex evidence, clinical knowledge, and experience. To accelerate interpretation and avoid diagnostic delay, computational methods are emerging for automated prioritization that capture, translate, and exploit clinical knowledge. While automation provides efficiency, it does not replace the expert-driven interpretation process. Moreover, knowledge and experience of human experts can be challenging to fully encode computationally.

This thesis, therefore, explores an alternative space between expert-driven and computer-driven solutions, where human expertise is deeply embedded within computer-assisted analytic and diagnostic processes via facilitated human-computer interactions. First, clinical experts and their work environment were observed via collaborations in an interdisciplinary exome analysis project as well as in a clinical resource development project. From these observations, we identified two elements of human-computer interaction: characteristic cognitive processes underlying the diagnostic process and information visualization. Exploiting these findings, we

designed and evaluated an interactive variant interpretation strategy that augments cognitive processes of clinical experts. We found that this strategy could expedite variant interpretation. We then qualitatively assessed current information visualization practices during clinical exome and genome analyses. Based on the findings of this assessment, we formulated design requirements that can enhance visual interpretation of complex genetic evidence. In summary, this research highlights the synergistic utility of human-computer interaction in clinical exome and genome analyses for rare genetic diagnoses. Furthermore, it exemplifies the importance of empowering the skills of human experts in digital medicine.

Lay Summary

Diagnosing rare genetic diseases is a race against time. For conditions that are amenable to treatment, early diagnosis and treatment prevent irreparable damages to the health of affected children. Recent advances in DNA sequencing technology are allowing healthcare experts an unprecedented opportunity to examine genomic mutations en masse in efforts to rapidly diagnose rare genetic diseases. Unfortunately, with the ability to examine all, comes the challenge of identifying causal mutations within a haystack of millions of DNA variations in any individual. Collaborative global efforts are being made to encode available knowledge into computers and to create computational methods that expedite DNA analyses. However, human expertise and knowledge are difficult to fully encode into computers. Thus, this thesis explores a hybrid approach, where experts and computers collaboratively analyze genomic data through facilitated human-computer interactions. The research findings will contribute to future genome analysis methods that empower experts to expedite critical diagnoses.

Preface

Part of the research described in this thesis is based on collaborations with the TIDEX gene discovery project, guided by my supervisors, Dr. Clara D. M. van Karnebeek and Dr. Wyeth W. Wasserman. As part of the TIDEX bioinformatics team, I regularly performed applied genome analyses for patients and their families enrolled in the study. I would like to acknowledge that my experience within this project has motivated the work described in chapter 3 and 4. The bioinformatics contribution resulted in one first-author and two co-authored publications, but they will not be explicitly discussed in this thesis.

A version of chapter 2 has been published: Lee, J.J.Y., Wasserman, W.W., Hoffmann, G.F., van Karnebeek, C.D.M., and Blau, N. (2018) Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. Genet Med. 20(1):151-158. This collaborative work was initiated by NB as part of the RD-CONNECT initiative. The nascent clinical database which this work is based on was originally compiled by NB. With WWW, CDMvK, and NB, I contributed to the study design and conception of the research. I implemented the knowledge base and mini-expert platform using the nascent database provided by NB. I conducted retrospective validation and user testing with CDMvK, GFH, and NB. I performed all performance analyses and interpreted the results with WWW, CDMvK, and NB. I wrote the manuscript, which WWW, CDMvK, GFH, and NB revised. I would like to make additional acknowledgement towards G. Frauendienst-Egger for review and revision of clinical terminology mappings to HPO; D. Pak for research management support; M. Hatas for system support, as well as all validators who contributed retrospective cases (Carlos Ferreira, USA;

Saumya Jamuar, Singapore; Stefan Kölker, Germany; Ylmaz Yildiz, Turkey; Luis Umana, USA; Saikat Santra, UK; Consuelo Duran, Argentina; Bryce Medelsohn, USA; Somesh Kumar, India; G.F.H., Germany, N.B., Germany; Himanshu Goel, Australia; Dorothea Haas, Germany; Susan Brooks, USA; Burcu Öztürk-Hismi, Turkey; C.D.M.v.K., Amsterdam & Canada; Christian Staufer, Germany; Stanley Korman, Israel; Friedrich K. Trefz, Germany). This work was supported by the FP7-HEALTH-2012-INNOVATION-1 EU grant 305444 (to NB), funding from the Canadian Institutes of Health Research (CIHR) (to CDMvK and WWW), funding from Genome Canada/Genome BC/CIHR Large Scale Applied Research Grant ABC4DE project Ref 174CDE (to WWW), and funding from the Dietmar Hopp Foundation (to GFH and NB).

A version of chapter 3 has been accepted for publication by the Journal of the American Medical Informatics Association: Lee, J.J.Y., van Karnebeek, C.D.M., and Wasserman, W.W. (2018) Development and user evaluation of a rare disease gene prioritization workflow based on cognitive ergonomics. With the guidance of my supervisors, WWW and CDMvK, I contributed to the study design and conception of the research. I conducted the user study and data analyses. I interpreted the results with WWW and CDMvK. I wrote the manuscript, which CDMvK and WWW edited. CDMvK facilitated recruitment of clinicians within Care4Rare initiative (with an approval from Dr. K. Boycott) as well as clinicians in Netherlands. I thank all participants of this study who generously contributed their time and knowledge to complete the survey. In addition, I would like to acknowledge A. Kushniruk for discussions in the planning phase of the project, X. C. Ye and M. Voulgaris for comments and discussion regarding the early version of the manuscript, D. Pak for research management support, as well as M. Hatas and D. Arenillas for system support. Special thanks go towards Dr. K. Boycott of Care4Rare Canada and Dr. N. Blau

of Dietmar-Hopp-Metabolic Center for assistance in recruiting study participants and to Dr. D. Vilet for assistance in selecting and reviewing simulated clinical scenarios. This work was approved by the UBC Children's and Women's Research Ethics Board (H17-00872). This work was supported by the funding from BC Children's Hospital Foundation (Treatable Intellectual Disability Endeavour in British Columbia: 1st Collaborative Area of Innovation http://www.tidebc.org), the Canadian Institutes of Health Research (CIHR) (to CDMvK and WWW), National Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program (RGPIN-2017-06824) (to WWW), and funding from Genome Canada/Genome BC/CIHR Large Scale Applied Research Grant ABC4DE project (Ref 174CDE) (to WWW).

Chapter 4 describes original work, and a version of this chapter is under review by a leading biomedical informatics journal: Lee, J.J.Y., van Karnebeek, C.D.M., and Wasserman, W.W. (2018) The role and design of information visualization in genome sequence analysis for clinical genetics. With the guidance of WWW and CDMvK, I contributed to the study design and conception of the research. I conducted the online survey and local interviews. I performed data analyses and interpreted the results with WWW and CDMvK. WWW and CDMvK assisted with participant recruitment. I wrote the manuscript, which CDMvK and WWW revised. I am especially grateful for all survey and interview participants for contributing their time and knowledge to this study. In addition, I would like to acknowledge A. Kushniruk for discussions in the planning phase of the project, X. C. Ye and M. Voulgaris for comments and discussion regarding the early version of the manuscript, D. Pak for research management support, as well as M. Hatas and D. Arenillas for system support. Special thanks go towards R. Docking (Canada's Michael Smith Genome Sciences Centre), J. Majewski (McGill University), M.

Tarailo-Graovac (University of Calgary), C. Marshall (The Hospital for Sick Children), and D. Azzariti (Matchmaker Exchange) for facilitating participant recruitment. This work was approved by the UBC Children's and Women's Research Ethics Board (H17-02809). This work was supported by the funding from BC Children's Hospital Foundation (Treatable Intellectual Disability Endeavour in British Columbia: 1st Collaborative Area of Innovation http://www.tidebc.org), the Canadian Institutes of Health Research (CIHR) (to CDMvK and WWW), funding from Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2017-06824) (to WWW), and Genome Canada/Genome BC/CIHR Large Scale Applied Research Grant ABC4DE project (Ref 174CDE) (to WWW).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

EHR Electronic health record

HCI Human-computer interaction

IEM Inborn errors of metabolism

WES Whole exome sequencing

WGS Whole genome sequencing

Acknowledgements

With sincere gratitude, I would like to acknowledge the amazing guidance of my supervisors, Dr. Clara van Karnebeek and Dr. Wyeth Wasserman. Thank you for your patience, encouragement, and constant support in this endeavour. Your mentorship truly went beyond the work described in this thesis and your exceptional leadership that I experienced in collaborative projects has allowed me to grow professionally.

This thesis research could not have progressed without the guidance and insights of my committee members, Dr. William Hsiao, Dr. Sara Mostafavi, and Dr. Raymond Ng. Thank you for providing your time, expertise, and invaluable feedback throughout this journey.

I am grateful for the support from all members of the Wasserman lab. I would also like to thank Dora Pak, Miroslav Hatas, and Dave Arenillas for their administrative and technological support. Special thanks go towards all members and alumni of the genome analysis group - Dr. Maja Tarailo-Graovac, Dr. Allison Matthews, Dr. Magda Price, Dr. Jill Mwenifumbo, Dr. Robin van der Lee, Phil Richmond, and Cynthia Ye - for critical advice on all aspects of my research and this thesis.

For funding, I thank BC Children's Hospital Foundation for Jan M. Friedman Studentship, RD-CONNECT, Dr. Nenad Blau and other collaborators of IEMbase, as well as my supervisors.

I owe most special thanks to my partner and my family for their patience, love, and many, many forms of support that sometimes do not have exact words for. Above all, I thank them for putting up with me since the very beginning of - or perhaps even before - this adventure.

Chapter 1: Introduction

Recent advances in DNA sequencing technology have revolutionized understanding of the connection between genes and human diseases. The ability to examine DNA variants *en masse* with a single test is offering researchers and clinicians an unprecedented opportunity to identify molecular causes of rare genetic diseases. Unfortunately, with the ability to examine all comes a challenge of identifying causal variants (or mutations) within a haystack of millions of DNA variations in any individual. Accurate and efficient analysis of DNA sequencing data is crucial for timely diagnosis. Current analytic approaches employ computational methods that help bring focus on a compact set of variants, which are subsequently examined by human experts for interpretation and diagnosis [1-3]. Recognizing an opportunity to bridge computational tools and human experts, this thesis explores approaches that embed human expertise within the computational process of analyzing genome sequence data for rare genetic disease diagnoses.

1.1 Motivation

Inspiration for this thesis originated from my contributions to the TIDEX gene discovery project at BC Children's Hospital [1, 4, 5]. TIDEX is a study within the Treatable Intellectual Disability Endeavor in British Columbia (TIDE-BC) [6], which aims to (a) raise awareness as well as (b) deliver early diagnoses and effective treatment of inborn errors of metabolism (IEMs). IEMs are rare genetic conditions that cause defective metabolism, resulting in clinical symptoms such as intellectual disability. TIDEX focuses on the discovery of novel genetic defects that underlie IEM patients by applying whole exome, and more recently whole genome, sequencing technology. Whole genome sequencing (WGS) is a process of mapping out the entire sequence

of a person's DNA, while whole exome sequencing (WES) applies the same approach to isolated DNA segments (usually targeting protein coding regions which comprises less than two percent of the genome).

At the outset of my study, I joined the TIDEX bioinformatics team and have since been contributing to applied WES and WGS analyses. This engagement has encouraged me to immerse myself in the wave of clinical WES/WGS analyses during a time when tools of the trade have been transitioning from computer command lines to more user-friendly software [7-12]. Despite this transition, the analyses are still largely bottlenecked by expert interpretation as they require complex evidence that affects clinical decisions [13]. Through the TIDEX project, I have been allowed a unique opportunity to experience both sides of the human-computer spectrum. This experience sparked my interest in pursuing an efficient and effective liaison between clinical experts and computers. How do experts use computational tools and why do they use them the way they do? What tasks can computers support to expedite expert work and how should they support the tasks? For what tasks do experts choose not to use computational tools and why? These questions led to development of this thesis, which explores a space in the human-computer spectrum where both parties interactively analyze genome sequence data.

Upcoming sections of the introduction will lay out the background on clinical WES/WGS analyses in terms of (a) their application in rare genetic disease diagnoses and current challenges of data interpretation, as well as (b) elements of human-computer interaction which can potentially accelerate WES/WGS analyses. The final section will outline specific thesis objectives.

1.2 Clinical WES/WGS analyses for rare genetic disease diagnoses

DNA sequencing technology has progressed rapidly in recent years, allowing WES and WGS to become more efficient and affordable [14, 15], and enabling both to be applied within clinical research studies more frequently [16]. Many studies have demonstrated WES and WGS as effective for timely diagnosis and informed management of rare genetic diseases[1] [1, 17-19]. The following subsections will discuss why such sequencing technologies are useful for rare genetic disease investigations, how their data is used, and what makes the data challenging to interpret.

1.2.1 Rare genetic diseases

Rare genetic diseases are conditions caused by DNA variants that affect the function of a single gene or sometimes multiple genes. The word "rare" generally refers to its extremely low prevalence in a population. The European Commission Regulation on Orphan Medicinal Products specifically defines rare diseases as conditions that affect fewer than 1 in 2,000 people [21]. Although the prevalence is low for individual diseases [22], it is estimated that there are between 6,000 to 7,000 rare genetic diseases [23]. This set of disorders may collectively affect 30 million people in Europe [24]. While these diseases have lifelong impacts, timely diagnoses lead to better clinical management and can improve patient conditions [25, 26].

A classic example of treatable rare diseases is phenylkenonuria (PKU). PKU is a genetic disorder that occurs in approximately 1 in 10,000 births [27, 28]. If untreated, it causes defective

[1] WES and WGS are also widely applied in cancer studies [20], but such work will not be addressed in this thesis.

metabolism of phenylalanine, resulting in intellectual disability, seizure, and/or behavioural/psychiatric problems [28, 29]. Since the identification of PKU by Asbjørn Følling in 1934 [30], decades of PKU research has elucidated molecular mechanisms of the disease, developed treatment, and implemented screening methods [31]. PKU has been incorporated into population newborn screening programs since the 1960s [31], which has enabled early detection, timely treatment, and therefore healthy development of the affected children.

However, most rare genetic diseases have not been as well handled as PKU. Diagnoses of many rare diseases are challenging as they remain poorly characterized due to the limited number of studied patients, genetic/phenotypic heterogeneity, and/or difficulty with distinguishing a novel disease from existing diseases [23, 32]. As more cases are reported, the aforementioned challenges can be resolved by progressively broadened knowledge. This is where WES and WGS are invaluable for accelerating rare disease investigations because they allow for the interrogation of almost all genes, including ones yet to be deeply characterized. For instance, an application of WES identified mutations in the gene *DHODH* as the cause of Miller syndrome (a rare disease that has been reported in only 30 cases to date [22] and presents with undersized jaws, cleft lip/palate, and limb deformities [33]) [34]. The gene discovery and molecular basis was uncovered more than 30 years after the disease was first described by Geneé [35].

1.2.2 A typical clinical WES/WGS analysis pipeline

WES or WGS is performed on genomic DNA that is isolated from a biological sample collected from a patient. For a rare disease investigation, sequencing may also be performed for members of the patient's family when their DNA is available and sequencing is accessible, as this

information can be extremely helpful for interpretation of the patient's genomic data [36]. The output of WES/WGS is then processed through a data analysis pipeline. Figure 1.1 summarizes common WES/WGS analysis pipeline steps.

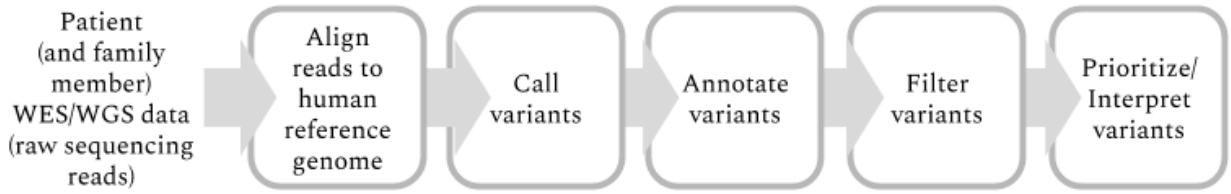


Figure 1.1 A typical clinical WES/WGS analysis pipeline.

A patient's (and family members') DNA is revealed using whole exome or whole genome sequencing. The sequence data is processed by a computational analysis pipeline, which aligns raw sequencing reads to the human reference genome, identifies variants in the patient's (and family members') DNA, annotates variants with multiple information (e.g. minor allele frequency, predicted impact on protein function), and filters out the variants that are common and those that are not expected to affect protein function. The filtered variants are then interpreted with regards to their connection to patient phenotypes.

A typical analysis pipeline begins by aligning WES/WGS sequencing reads to the human reference genome[2] and subsequently detecting DNA variants [19, 38-40]. For WES data, the pipeline commonly focuses on calling small nucleotide variants (SNVs) and small insertions or deletions (indels) [39, 41, 42]. For WGS data, improved variant calling methods have continued

[2] While the human reference genome is commonly used in WES/WGS analyses, it is known in the field that the single representative genome sequence does not fully reflect common human DNA variation [37]. Solutions to this problem are emerging in the form of multiple reference sequences based on diverse cohorts of human genomes [37].

to be developed to identify more types of variants such as copy number variants and structural variants (e.g. duplications and inversions) [43-45], with these methods included in the pipeline in addition to those that call SNVs and small indels [46-48]. Via this process, roughly 60 000[3] variants are identified from the WES data of an individual [49] while roughly 5 000 000[3] are identified from the WGS data of an individual [51]. In order to successfully target rare, potentially deleterious variants, the variants are annotated and filtered by minor allele frequency (rare disease studies will commonly focus on variants that occur in less than 1% of a population), by occurrences within protein-coding regions[4], and/or by predicted impact on protein function [3, 41, 53]. After applying such filters, the number of rare, protein-affecting variants[4] in an individual can be reduced to hundreds [46, 53]. In the case of a family-based analysis, further filtering is applied using sequence data of the family members based on different inheritance models [36, 54]. Additional filtering is also possible by comparing across sequence data of multiple unrelated individuals with similar phenotypes or by assessing sequence data of related individuals of known phenotype [55, 56].

Once a list of rare, potentially pathogenic variants is identified in a patient, the next step is to interpret those variants in the context of patient phenotypes in order to identify connections between a variant and a patient's disease. For this task, experts examine any available information on each variant, the gene that harbours the variant, and the phenotypes caused by

---

[3] The exact number of detected variants differ by the technology used (e.g. capture kits, alignment and variant calling software) [49]. Also, variant calling software can be tuned to include more or less noise [50].

[4] As WGS is capable of capturing variants in non-coding (or non-protein coding) regions, its analyses may include non-coding variants [47, 52]. However, annotation of coding regions has historically outpaced annotation of non-coding regions, and most of the variants with established pathogenicity have been within protein-coding regions [32]. For these reasons, current WGS data analyses tend to prioritize coding variants over non-coding variants. With improved annotation of non-coding regions of the genome in the near future, such priority will likely be eliminated from common data analysis practices.

disruptions of the gene. Table 1.1 summarizes the types of information commonly considered during variant interpretation. When candidate variants are identified from interpretation, their presence and segregation with the disease is confirmed by Sanger sequencing [40, 42, 57]. Afterwards, diagnoses are made if pathogenicity and causality of the variants have been established by experimental validation and/or by published reports/identification of unrelated individuals with similar phenotypes caused by the same or other damaging variants in the same gene, with similar modes of inheritance [1, 54, 57, 58].

Types of information
Sequencing/variant quality
Coverage of targeted regions/consensus coding sequence
Functional annotation (e.g. synonymous/nonsynonymous, nonsense/missense, frameshift)
Location of variant (e.g. overlap with disease-associated region, within exon/intron, exon-intron boundaries)
Variant frequency in population databases (e.g. Genome Aggregation Database (gnomAD) [59])
Variant frequency in in-house databases
in-silico functional prediction (e.g. Polymorphism Phenotyping (PolyPhen) 2 [60], Combined Annotation Dependent Depletion (CADD) [61])
Nucleotide conservation (e.g. PhyloP [62])
Splice-site prediction (e.g. Human Splicing Finder [63])
Inheritance model
Known gene-disease association in disease databases (e.g. Online Mendelian Inheritance in Man (OMIM) [64])
Presence and designation in disease-focused variation databases (e.g. ClinVar [65])
Interaction with known disease-associated genes

Table 1.1 Commonly considered information during variant prioritization/interpretation.

1.2.3 Current data interpretation challenges

As described in the previous section, variant interpretation requires an analysis of multi-dimensional, heterogeneous data including the variant, the variant-harbouring gene, and the phenotypes associated with the gene disruption [66, 67]. An array of computational methods has been developed to facilitate the knowledge-heavy process known as "variant prioritization". Variant prioritization refers to a systematic process of focusing on variants that are more likely to

disrupt genes relevant to patient phenotypes [12]. A distinction is made from variant interpretation in that interpretation determines direct connections between a variant and a disease [12], whereas prioritization is part of the informatics process.

Computational variant prioritization tools commonly rank a list of variants by predicted pathogenicity [60, 61, 68], conservation [62, 69], population allele frequency [2], and/or connections to patient phenotypes based on variant/gene/disease annotations [70-73]. While these computational tools have improved variant prioritization and interpretation, both processes still involve significant engagement of experts, with the most intense involvement in the final steps [41, 46, 74]. This expert-dependence in the end steps is because the processes require understanding of complex evidence and, most importantly, affect diagnoses and treatment of patients [12, 74]. As such, the reliance on experts creates a bottleneck within WES/WGS analyses [13] as time, speed, and energy of human experts are limited. To alleviate this problem, large-scale efforts have been made to encode clinical genetics knowledge into computers [64, 75-78]. These efforts have enabled computational exploitation of expert knowledge and have improved expert variant interpretation workflow especially in the aspects of managing and leveraging phenotypic information. For example, the Human Phenotype Ontology (HPO) provides a standardized medical vocabulary for describing phenotypic abnormalities that are observed in known genetic diseases [75]. Its compilation has spawned development of computational tools that collect and analyze patient phenotypes [7, 79], as well as curate phenotypic information to assist diagnoses and accelerate collaboration for disease gene discovery [80, 81].

However, the above approaches have yet to fully preempt expert involvement because identifying and quantifying human knowledge is a complex task, and diagnostic expertise involves cognitive processes that cannot be explicitly materialized [82, 83]. For instance, recent studies have demonstrated that incorporating clinical geneticists' guidance and knowledge into the variant interpretation process can increase the diagnostic rate of clinical WES analyses, compared to those without explicit engagement of clinicians (e.g. laboratory interpretation or computational prioritization) [84, 85]. These studies suggested that clinicians' experience and abilities to incorporate additional diagnostic modalities as well as context (e.g. family/patient history, negative findings) likely contributed to the enhanced WES data interpretation [84, 85]. As such, this thesis sought an alternative approach, which aimed to augment the expert diagnostic workflow during computer-assisted variant interpretation by facilitating interaction between experts and computers. The next section will discuss the technical background for assisting expert-computer interactions in the context of WES/WGS analyses.

1.3 Elements of human-computer interaction within WES/WGS data interpretation

In variant interpretation, clinical experts and computational tools represent opposite extremes of the human-computer spectrum (Figure 1.2). Each side has its own advantage: experts promise satisfactory answers to complex clinical problems, while computers scale easily and are much faster than individual human efforts. Human-computer interaction (HCI) occupies the middle of the human-computer spectrum (Figure 1.2), offering methods that can harness the aforementioned advantages of both parties by facilitating a collaboration between the two [86, 87]. In this thesis, I sought to devise collaborative solutions for variant interpretation without reinventing existing interpretation methods. To achieve this, I identified and explored two

potential elements of HCI, cognitive characteristics and information visualization, within current variant interpretation processes. The following subsections will discuss these two elements in terms of (a) how they can inform HCI designs, (b) how they relate to variant interpretation, and (c) how they can be exploited for interactive interpretation strategies.

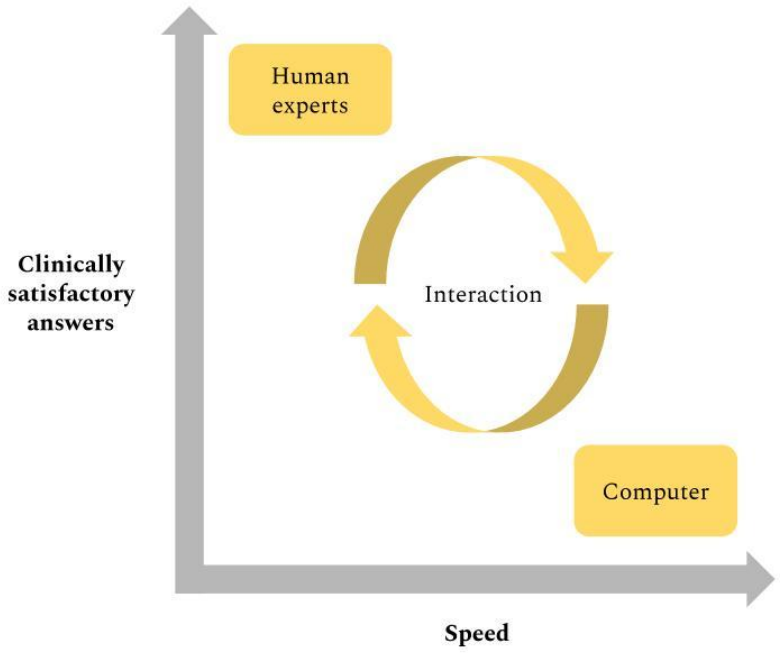


Figure 1.2 Illustration of human-computer spectrum in variant interpretation.

The horizontal axis represents the speed of interpretation performed by human experts or computers. The vertical axis represents the degree of which interpretation performed by human experts or computers is clinically satisfactory. Interpretation by computers are fast, but less clinically satisfactory than human experts. Interpretation by human experts are slow, but more clinically satisfactory than computers. Interaction between the two parties occupies the middle space where interpretation is faster than individual experts and more satisfactory than computers.

1.3.1 Cognitive processes underlying variant interpretation

Cognition is the mental process of "acquiring knowledge and understanding through thought, experience, and senses" [88]. It determines action and reaction; thus, it shapes the way we interact with external entities. For this reason, cognition has been one of core research topics in the field of HCI [89], with a specific focus on understanding and modeling the process to

enhance the efficiency and efficacy of HCI. Historically, cognition-based HCI research began in the 1980s when a traditional view of cognition from psychology and cognitive science was introduced to the field [89-91]. This view described cognition as being part of a linear information processing system (Figure 1.3A), acting as a module that manipulates information based on perception (input) before resulting in an action (output) [90]. This view enabled HCI to model cognitive processes that underlie user behaviours [92, 93]. A well-known modeling method from the era is GOMS (Goals, Operators, Methods, and Selection rules), which analyzes a routine HCI in terms of (a) goals that a user wants to achieve, (b) operations that the user performs on a computer (e.g. mouse-click), (c) methods (or a series of operations) to achieve a goal, and (d) user's selection among multiple methods that achieve the same goal [92, 93]. In the 1990s, the view of cognition in cognitive science began to shift from an isolated process within a linear system to an interrelated process that is shaped by human perception and action in social context (Figure 1.3B) [91, 94, 95]. Following this evolution, the field of HCI adopted alternative views, such as embodied cognition (cognition as embodiment of sensorimotor capabilities of human body) [96] and distributed cognition (cognition as a holistic process that includes interaction with people and environment) [97]. This shift has since inspired a new school of HCI approaches that focus on holistic interactions in a specific context or culture [98-100]. Such approaches emphasize the analysis of interactions that occur within users' environments, such as ethnography, where researchers observe interactions from a user's perspective in naturalistic settings [98, 101].

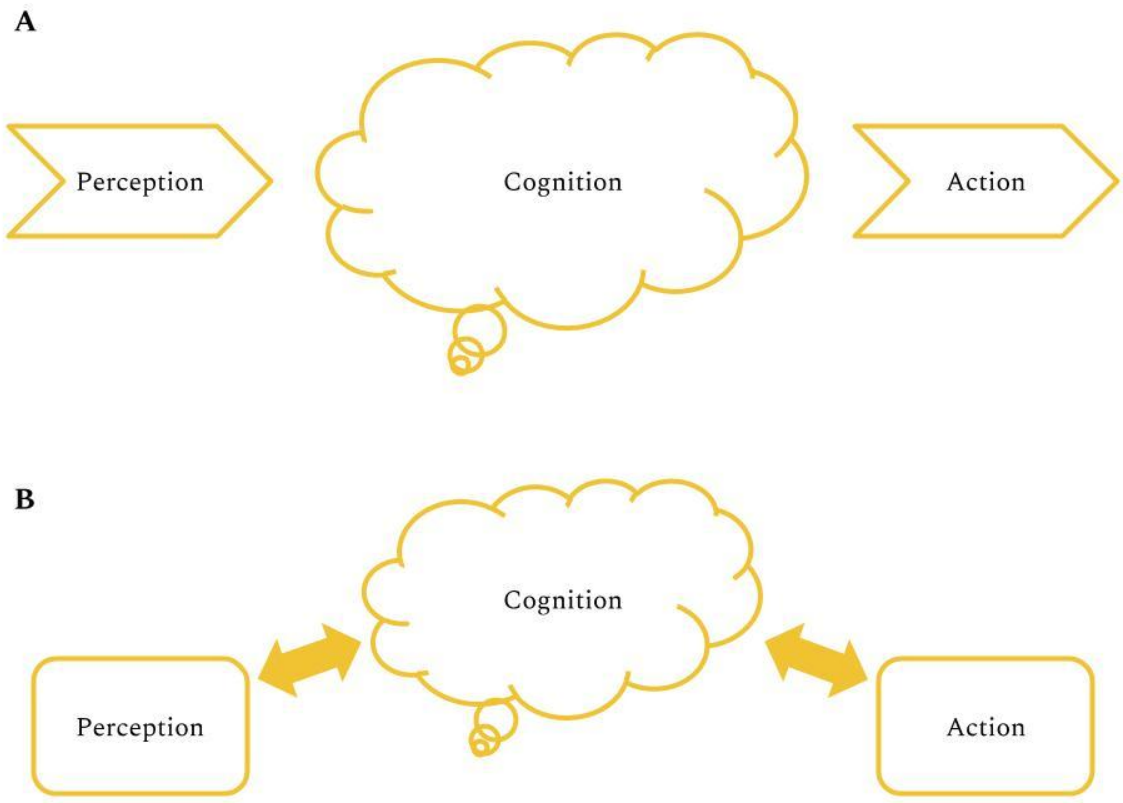


Figure 1.3 Different views of cognition.

A illustrates cognition as a linear system where information is accepted by perception, processed by cognition, and outputted as an action. B illustrates cognition as an interrelated system where perception and action shape cognition and vice versa.

Variant interpretation during clinical WES/WGS analyses involves cognitive processes that can be identified and exploited to improve efficiency of the interpretation process. Aspects of these processes have already been modelled in various ways: they have been translated into variant-filtering logic within WES/WGS analysis pipelines, reported as a decision-making algorithm for reporting variants [3, 54, 102, 103], codified into a clinical guideline [67], or incorporated into features that are exploited by variant prioritization tools (e.g. phenotypic similarity) [70, 72,

104]. However, these aspects tend to focus on specific types of information and how they are processed during variant interpretation, leaving holistic and contextual aspects underexplored. For example, in which context do experts consider certain information a priority when interpreting variants? How do experts interact with each other when discussing WES/WGS results? When do experts seek additional computational tools to further dissect WES/WGS results and how do they use those tools? Such questions belong to the realm of HCI and their answers can inspire effective interactive strategies, which closely reflect cognitive processes that experts engage in during WES/WGS analyses within their natural work environment. Therefore, this thesis attempted to exploit contextual aspects of WES/WGS variant interpretation by (a) ethnographically examining clinical WES/WGS analyses through the TIDEX project, (b) modelling diagnostic processes of biochemical geneticists, and then (c) using the findings from (a) and (b) to design and evaluate interactive variant prioritization strategies.

1.3.2 Information visualization during variant interpretation

Visualization refers to a process of "transforming data, information, and knowledge into visual form" [105]. While visualization is a field of its own (referred to as information visualization), it is of high relevance to HCI as it serves as an interface between human minds and computers [105, 106]. From an HCI standpoint, information visualization helps reduce the cognitive burden of understanding complex data, thereby allowing users to efficiently interact with computers to extract the information they seek from the data in view [107].

In clinical WES/WGS analyses, information visualization has been actively employed for variant interpretation, either as dedicated visualization software or as a feature in analysis software [59,

70, 108-117]. These types of visualization support specialized interpretation tasks, such as (a) visualizing sequencing read alignments to inspect read/variant quality and read coverage, or identify structural variants [59, 108], (b) visualizing 3D protein structures to assess the impact of a mutation [110], (c) visualizing annotated genomic/protein features to identify features overlapping with candidate variants [109, 114-117], (d) visualizing metabolic pathways or protein-protein interaction networks to assess the impact of gene disruption [111, 112], (e) visualizing phenotypic patterns to evaluate phenotypic similarity between patients or against a known phenotypic profile of a disease [70, 113]. Table 1.2 provides a list of representative tools for the above tasks.

Types of information visualization (dedicated tools or features of analysis software)
Integrative Genomics Viewer [108]
UCSC Genome Browser [109]
Protein structure visualization (e.g. Chimera [110])
Network visualization (e.g. Cytoscape [111], GeneMANIA [112])
Phenotype-driven visual prioritization tools (e.g. OMIM Explorer [70])
Phenotype comparison visualization (e.g. PhenoBlocks [113])
Custom R visualization
Visualization features within population databases (e.g. read data browser in gnomAD [59], graphical sequence viewer in dbSNP [114])
Visualization features within sequence databases (e.g. graphical sequence viewer in NCBI Gene [115], Feature viewer in UniProt [116])
Visualization features within disease databases (e.g. protein browser or phenotype browser in Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) [117])
Visualization features within commercial variant analysis tools (e.g. Alamut Visual [118])

Table 1.2 An overview of commonly used information visualization during WES/WGS analyses.

Use of the above visualization tools occurs regularly in WES/WGS analyses [119-123]. However, the specific context of their use is infrequently documented [119-121]. Similar to cognitive processes, understanding how users perform tasks and use data within their specialized domain is the first step towards designing new information visualization [124]. User requirements and domain problems that are identified during this step guide subsequent design processes, which (1) translate users' work into computational data types and operations, (2) design visual representation of the data and HCI, and (3) implement an efficient algorithm that

enables visualization on computer [124]. Furthermore, assessment of users and their work practices can also enhance the utility of existing visualization tools [125].

The above type of contextual evaluation can be conducted on a diverse scale in information visualization research. For example, it could be performed as part of a study involving a full design and validation process to create a novel visualization [126-128], or as a study investigating individual aspects of visualization (e.g. users, computational algorithm, visual encoding theory) [125, 129, 130] to inform future visualization development. This thesis pursued the latter type of study, qualitatively evaluating bioinformatics and healthcare experts as well as their routine WES/WGS analysis practices that involved the use of information visualization. The findings were analyzed to extract formal requirements for visually supported analyses of WES/WGS data.

1.4 Thesis overview and objectives

Clinical sequencing studies have presented WES/WGS as effective for diagnosing and informing the management of rare genetic diseases [17-19]. Common analytic methods for WES and WGS data hone in on a refined set of potentially pathogenic variants that are computationally prioritized among millions of DNA variants found in a patient, before the variant set is examined by experts for evidence of pathogenicity [3, 41, 42, 58]. Currently, reliance on expert interpretation causes a bottleneck during WES/WGS analyses [13]. To relieve this issue, automated methods that capture, curate, and exploit clinical genetic knowledge have been introduced into the field to accelerate interpretation [7, 73 75, 77, 78, 104]. However, human expertise and knowledge are difficult to fully encode computationally. Furthermore, recent

studies have suggested the role of clinical experts in enhancing WES/WGS data interpretation by deriving differential diagnoses based on their experience, additional diagnostic modalities, and/or contextual information such as case history [84, 85].

Recognizing an existing gap between expert-driven interpretation and computer-driven prioritization of rare disease-causing variants, the research described in this thesis focuses on how experts and computers effectively interact during the interpretation of WES/WGS data. The overarching idea explored herein is whether efficient facilitation of expert-computer interaction can accelerate WES/WGS analyses by augmenting experts' analytic and diagnostic workflows within a common computer-assisted data interpretation process. The central objective of this thesis is to, therefore, explore and evaluate expert-computer interactive approaches for WES/WGS analyses. As summarized in Figure 1.4, this thesis is structured to reflect a HCI research process, where examination of users and their work environment (Chapter 2 and my contributions to applied WES/WGS analyses within the TIDEX project) inspire design and evaluation of an HCI solution that supports expert cognitive processes (Chapter 3) and qualitative assessment of information visualizing HCI solutions for synthesis of novel or improved visual solutions (Chapter 4).

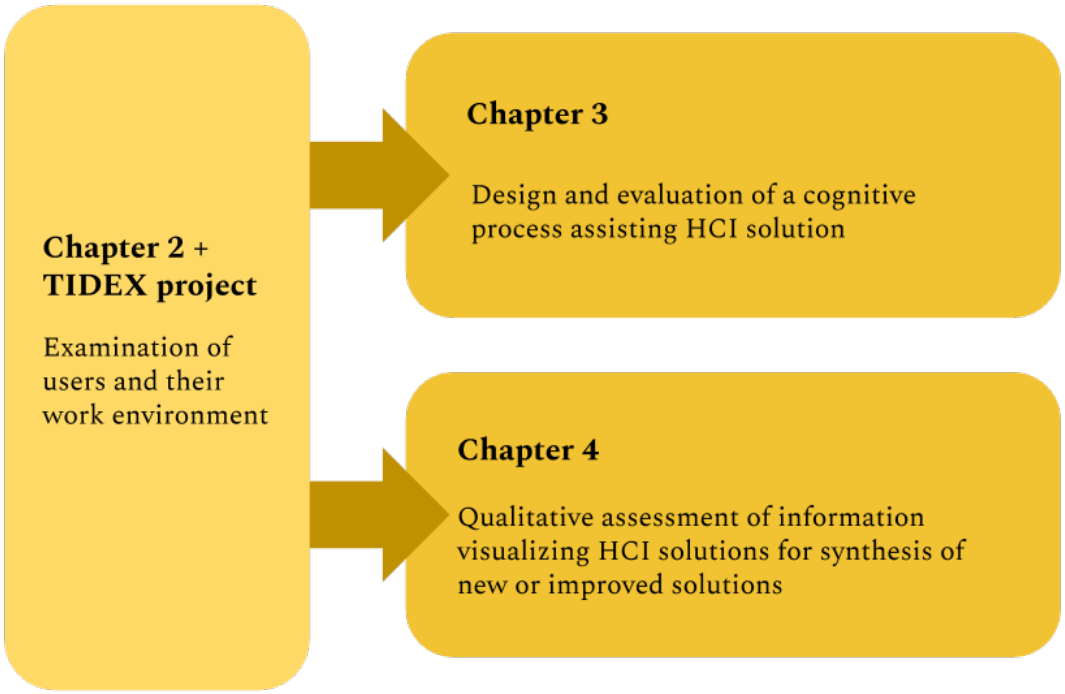


Figure 1.4 A visual summary of thesis structure.

This thesis is structured to reflect a HCI research process where examinations of users and their work environment (Chapter 2 and contributions to applied WES/WGS analyses within the TIDEX project) inspire the work described in Chapter 3 and Chapter 4. In Chapter 3, an HCI approach that assists expert cognitive processes was designed and evaluated. In Chapter 4, current information visualization practices were qualitatively assessed to formulate design requirements for novel or improved visualization for WES/WGS data interpretation.

Inception of this research is owed to my collaboration within the realm of IEMs. As explained in Section 1.1, this journey began with the TIDEX project, where I performed applied WES/WGS analyses for IEM patients and their families [1, 4, 5]. This collaboration helped me to learn the ins-and-outs of the analysis process, and to observe how WES/WGS data was processed and

analyzed by bioinformaticians, as well as how the analysis results were interpreted jointly by a multi-disciplinary team. While each case was unique and its interpretation required complex knowledge, I realized that there was potential to accelerate the process of seeking, sharing, and using this expert knowledge. This discovery led to consideration of the utility in augmenting experts' analytic workflow within computer-assisted diagnostic processes as a means of expediting the WES/WGS analyses. To explore this idea further, I collaborated on a project (as described in Chapter 2) that produced an online diagnostic aide for clinicians seeking to classify IEMs. While data resource development in bioinformatics is not a hypothesis driven activity, the central intention informing the research within Chapter 2 was how we could enable clinicians to efficiently and accurately diagnose inborn errors of metabolism based on the clinical and biochemical phenotypes of their patients. While implementing and validating this resource, it became more apparent that acceleration of WES/WGS analyses could be achieved by an expert-assistive system that was capable of complementing experts' analytic and diagnostic workflows through an effective HCI.

As such, two HCI elements, cognitive processes and information visualization, were identified to examine their potential to expedite WES/WGS analyses. Selection of these elements was based on the following observations of key stakeholders of WES/WGS analyses within the above collaborations. Cognitive processes were based on a pattern that was observed in verbal and written communications of clinicians, where they had a tendency to describe patients in reference to classic presentations of genetic diseases that were hypothesized as compatible diagnoses. Information visualization was based on common uses of visualization tools among bioinformatics experts and healthcare professionals (e.g. cytogeneticists) during WES/WGS data

interpretation. Each of these HCI elements were then explored in Chapter 3 and 4, respectively, as follows.

In Chapter 3, I designed a gene prioritization workflow informed by reports in cognitive science and medical literature, which suggested that clinicians frequently employed prototypical thinking during WES/WGS investigations - a cognitive process where they recalled classic presentations of genetic diseases to inform their assessment of patients and interpretation of WES/WGS results. Such a "prototype"-based approach could be modelled computationally by querying clinicians to specify the disease they felt was most similar to the phenotype of the patient, instead of specifying individual phenotype terms (i.e. "symptom"-based approach) as commonly found in computational variant/gene prioritization tools. The central questions addressed by this work were whether and when there were advantages in a “prototype”-oriented software workflow. As such, through a web-based user study, the designed prototype-based workflow was evaluated against a symptom-based workflow. Clinicians interpreted genetic diagnoses faster using prototype-based workflows. Meanwhile, neither workflow was more accurate, more effective in collecting detailed phenotypic information, or showed higher user satisfaction. These findings suggested that clinicians employed prototypical thinking within gene prioritization and demonstrated potential utility of facilitating such processes within WES/WGS analyses.

In Chapter 4, I qualitatively assessed information visualization practices during routine WES/WGS analyses for diagnoses of rare genetic diseases. As the preceding research studies made it anecdotally apparent that human-computer interactions for genome interpretation were dependent on a diverse range of visualization-based tools, the driving goal of this work was to

conceptualize the design of an ideal visual interface for clinical genome interpretation. For this goal, I needed to determine both common user practices and specific aspects of WES/WGS analyses which were missing or had insufficient information visualization. Therefore, contextual interviews and an online survey were conducted with bioinformatics and healthcare experts who regularly analyzed WES/WGS data. These evaluations produced a comprehensive overview of common WES/WGS analysis and visualization practices. The overview summarized (a) which types of data and visualization tools were frequently used, (b) in what context experts commonly employed visualizations, and (c) experts' suggestions for new visualization. Based on the above findings, design recommendations were formulated to inform for novel or improved information visualization in this domain, which could enhance experts' understanding of complex data during WES/WGS analyses.

In sum, this thesis investigated human-computer interactive approaches for analyzing WES/WGS data for rare disease investigations. Through multi-disciplinary collaborations with IEM clinicians and researchers, key stakeholders in applied WES/WGS analyses were observed and domain knowledge was acquired. This experience led to the identification of HCI elements, cognitive processes and information visualization, which were (a) explored as a novel interactive strategy that could potentially accelerate expert variant interpretation, and (b) evaluated as an existing interactive strategy that could inform the development of enhanced information visualization for WES/WGS data interpretation. The findings of this thesis demonstrate the utility of these interactive strategies, as well as the power of synergy between healthcare experts and computers. For emerging innovators in digital medicine, the research presented herein may

provide a starting point for effectively integrating their technologies powered by big data and advanced artificial intelligence into clinical practice.

Chapter 2: Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism

2.1 Prelude for Thesis Readers

This chapter describes a collaborative project with Dr. Nenad Blau, a renowned investigator in IEMs. The primary objective of this collaboration was to create an online, public resource that curates the clinical knowledge of IEM experts. At the outset of this project, we also identified a secondary objective, which was to explore efficient ways to assist an expert's diagnostic process. The project allowed us to pursue those both objectives throughout its development, from designing of a resource database to implementing a diagnosis supporting algorithm and testing with real clinical users/cases. This experience resulted in deeper understanding of experts' criteria for assistive technology and motivated exploration of two HCI elements, cognitive processes and information visualization, in Chapter 3 and 4, respectively. The remaining sections of this chapter detail the outcome and development of this project with respect to its primary objective.

2.2 Synopsis

Purpose: Recognizing individuals with inherited diseases can be difficult because signs and symptoms often overlap those of common medical conditions. Focusing on IEMs, we present a method that brings the knowledge of highly specialized experts to professionals involved in early diagnoses. We introduce IEMbase, an online expert-curated IEM knowledge base combined with a prototype diagnosis support (mini-expert) system.

Methods: Disease-characterizing profiles of specific biochemical markers and clinical symptoms were extracted from an expert-compiled IEM database. A mini-expert system algorithm was developed using cosine similarity and semantic similarity. The system was evaluated using 190 retrospective cases with established diagnoses, collected from 15 different metabolic centers.

Results: IEMbase provides 530 well-defined IEM profiles and matches a user-provided phenotypic profile to a list of candidate diagnoses/genes. The mini-expert system matched 62% of the retrospective cases to the exact diagnosis and 86% of the cases to a correct diagnosis within the top five candidates. The use of biochemical features in IEM annotations resulted in 41% more exact phenotype matches than clinical features alone.

Conclusion: IEMbase offers a central IEM knowledge repository for many genetic diagnostic centers and clinical communities seeking support in the diagnosis of IEMs.

2.3 Introduction

Identification of rare genetic disorders has been greatly improved by the advent of genome-wide sequencing. The new technology has expanded our knowledge of rare disease genetics and enhanced our ability to diagnose new patients [1, 32]. However, the diagnosis of rare genetic disorders remains a challenge. Misdiagnoses and delayed diagnoses are often [131] due to nonspecificity and heterogeneity of signs and symptoms, rarity of conditions, and also limited access to the knowledge of highly specialized experts [24, 32, 132]. IEMs exemplify these challenges: early signs and symptoms are nonspecific [133] and insufficiently recognized [134]. For example, in a survey of 34 junior pediatric doctors regarding their confidence and knowledge

in the acute management of three IEMs — glutaric aciduria type I (MIM 231670), medium-chain acyl Co-A dehydrogenase deficiency (MIM 201450), and maple syrup urine disease (MIM 248600) — only five respondents were able to identify the correct treatment steps for the former two, while only two respondents identified the correct steps for the latter [134]. Moreover, more than 22 respondents indicated a low level of confidence in their knowledge [134].

The knowledge gap between IEM specialists and other clinicians involved in IEM diagnoses is concerning, given the amenability to targeted treatments for an increasing number of IEMs; a delayed diagnosis can lead to irreversible organ damage or even death. Moreover, this disparity is widening with the explosive amount of knowledge generated by multi-omics technology [32]. Such a divide stands in contrast to the historic efforts by the IEM clinical and research community toward early recognition through the creation and use of diagnostic tests, such as population newborn screening. Thus, a potential solution may be found in the rich disease knowledge base established by the IEM community, dating back to Archibald Garrod's study on alkaptonuria in 1902 [135]. This compiled knowledge base has, however, lagged behind other fields in the transition to digital form, as much of the work occurred before modern data systems came into existence and therefore the information was stuck on paper. Aspects have been incorporated into large-scale rare-disease databases [136, 137]. However, these databases aim to provide an overview of many kinds of individual disorders, and are not designed to guide clinicians in the diagnostic process. Therefore, digital translation and standardization of the IEM community knowledge base are urgently needed to bridge the knowledge gap.

Thus, we created IEMbase, an online application that combines the IEM community knowledge base with a prototype mini-expert system. The expert-compiled knowledge base provides clinical, biochemical, and genetic profiles of 530 known IEMs. The mini-expert system accepts a list of biochemical and clinical phenotypes from users, compares the input phenotypic profile against IEMs in the knowledge base using cosine similarity and semantic similarity, and returns a list of matching IEM diagnoses. With the resulting list, users can generate differential diagnosis charts, biochemical test panels, and targeted gene panels in order to pursue concurrent biochemical and genetic/genomic investigations for a rapid diagnosis. IEMbase aims to renew the existing IEM community knowledge base for the modern age, creating a global resource to facilitate the collection and dissemination of high-quality clinical knowledge for advanced recognition of IEMs.

2.4 Materials and methods

2.4.1 Knowledge base compilation

IEMbase was compiled by extracting 530 disease-characterizing profiles from a nascent disease database, which was previously compiled by more than 100 IEM experts to produce a textbook guide on IEM classification [138]. Table 2.1 shows an example of an extracted IEM profile. Each IEM profile consisted of known disorder names, disorder abbreviations, causal gene information, a MIM number, and a list of associated biochemical markers and clinical symptoms. Additionally, the list of biomarkers/symptoms was annotated with information regarding onset, severity/ pathological level, and whether the biomarker/symptom is characteristic of the associated IEM. The onset information was organized in five categories (neonatal: birth to 1 month, infant: 1–18 months, childhood: 1.5–11 years, adolescence: 11–16 years, and adulthood:

>16 years). The pathological levels of biochemical markers were denoted by up/down/no arrows and the severities of clinical symptoms were denoted by plus/minus signs. The presence or absence of phenotypic characteristics was indicated by yes/no.

Disorder name	Sepiapterin reductase deficiency					
Disorder abbreviation	SRD					
Associated gene	*SPR*					
Chromosomal localization	2p14-p12					
Affected protein	Sepiapterin reductase					
MIM number	182125					
Affected biochemical markers/clinical symptoms[a]	**Neonatal (birth - 1month)**	**Infancy (1-18 months)**	**Childhood (1.5-11 years)**	**Adolescence (11-16 years)**	**Adulthood (>16 years)**	**Is characteristic of disease?**
Axial hypotonia	++	++	++	+	?	No
Cerebral palsy	?	?	±	±	±	Yes
Eye movements, abnormal	±	±	±	?	?	No
Hypokinesia	+	++	±	±	±	Yes
Muscle weakness	+	±	±	±	?	No
5-Hyroxyindoleacetic acid, 5HIAA (cerebrospinal fluid)	↓↓↓	↓↓↓	↓↓↓	↓↓↓	↓↓↓	Yes
Biopterin (cerebrospinal fluid)	↑	↑	↑	↑	↑	Yes
Biopterin (urine)	n	n	n	n	n	No
Dihydrobiopterin (cerebrospinal fluid)	↑↑	↑↑	↑↑	↑↑	↑↑	Yes
Homovanillic acid, HVA (cerebrospinal fluid)	↓↓↓	↓↓↓	↓↓↓	↓↓↓	↓↓↓	Yes
Neopterin (cerebrospinal fluid)	n	n	n	n	n	No

Neopterin (urine)	n	n	n	n	n	No
Phenylalanine (plasma)	n	n	n	n	n	Yes
Prolactin (plasma)	↑	↑	↑	↑	↑	Yes
Sepiapterin (cerebrospinal fluid)	↑↑	↑↑	↑↑	↑↑	↑↑	Yes
Sepiapterin (urine)	?	↑↑	↑↑	↑↑	?	Yes

Table 2.1 An example disorder profile extracted from the nascent database.

For clinical symptoms, + denotes their presence and ± denotes occasional absence/presence. For biochemical markers, ↑ denotes elevated values, ↓ decreased values, and n denotes normal values. ? denotes uncertain/unreported presence of biomarkers/symptoms.

[a] The affected biochemical markers and clinical symptoms are selected for brevity.

The extracted profiles were manually reviewed for consistency and then were imported into IEMbase as three PostgreSQL tables, each representing the type of annotation used in the profiles: disorders, biochemical/clinical phenotypes, or disorder-phenotype associations (Figure A10, Appendix A). In total, the tables contained 530 disorders, 2,323 biochemical/clinical phenotypes, and 8,465 disorder-phenotype associations.

Additional annotations were created within each IEM profile. One was the amenability of individual IEMs to treatment, which was manually annotated based on previous literature [139, 140] and denoted by yes/no/unknown categories.

Another was the prevalence of IEMs as reported in literature or clinical resources [136–138, 140]. The last was a list of links to relevant entries in external databases, such as UniProt [116], NCBI Gene [141], GeneCards [142], Kyoto Encyclopedia of Genes and Genomes [77], National

Institutes of Health Genetic Testing Registry [143], and GeneReviews [76]. The links were created for interoperability with existing systems and were created using a BioMart ID conversion tool [144], as well as URL rules specified on the resource websites [76, 143].

The compiled knowledge base was assigned a version number of 1.0.0. This initial version was used for both the methods and the results described herein. Since the initial compilation, IEMbase has been regularly updated with new information. Thus, the version number has been incremented to indicate such updates.

2.4.2 Mapping to structured vocabulary

A known strategy for matching user-provided phenotypic profiles to diseases is to exploit semantic relationships between phenotypic features, which are defined by a structured vocabulary [145]. The phenotype vocabulary in IEMbase was not structured, but a structure could be imposed based on a compatible external vocabulary. Therefore, the following four standard medical vocabularies were assessed for their compatibility with IEMbase: Human Phenotype Ontology (HPO) [146], Medical Subject Headings (MeSH) [147], Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) [148], and International Classification of Diseases, 10th revision (ICD10) [149].

During compatibility assessment, HPO OBO file (2016-04- 01 release), MeSH ASCII file (2016 version), SNOMED CT RF2 files (2016 versions), and ICD10 XML file (2016 version) were used. The assessment proceeded in three steps. First, unique IDs and medical terms were extracted from IEMbase (version 1.0.0) and the four vocabularies. For SNOMED CT, extraction

was restricted to only the terms categorized under "Clinical finding" and "Substance", to minimize false mapping. The OntoCAT R package [150] was used to parse HPO terms. The standard string library and Nokogiri gem in Ruby programming language were used to parse all others. Second, all extracted terms were normalized using the Norm program included in the SPECIALIST Lexical Tools [151] String normalization removed the differences in alphabetic case, singular or plural variants, punctuations, stop words, and word order. Finally, all IEMbase terms were compared against all terms in each vocabulary. Only the exact matches were recorded as compatible mappings.

The initial compatibility assessment revealed that no single vocabulary could completely cover both the biochemical and the clinical phenotypes in IEMbase (Table 2.2). It also revealed that the most compatible vocabulary was different for biochemical (SNOMED CT) and clinical phenotypes (HPO) (Table 2.2). Therefore, the assessment was adjusted to consider the two phenotype categories separately. Once adjusted, two additional biochemical vocabularies were added: Chemical Entities of Biological Interest (ChEBI; OBO file; 2016-04-01 release) [152] and Logical Observation Identifiers Names and Codes (LOINC; CSV file; version 2.56) [153].

	Biochemical (# phenotypes mapped)	**Clinical** (# phenotypes mapped)	**Total** (# phenotypes mapped)
HPO	0	450	450
ICD 10	6	92	98
SNOMED CT	371	389	760
MeSH	324	283	607
ChEBI	301	3	304
LOINC	367	61	428

Table 2.2 Vocabulary compatibility assessment results.

Total number of biochemical phenotypes in IEMbase is 1123. Total number of clinical phenotypes in IEMbase is 1200. Total number of phenotypes in IEMbase is 2323.

Based on the adjusted assessment (Table 2.2), clinical phenotypes were mapped to the most compatible vocabulary, HPO. A medical expert manually reviewed exact matches identified during the compatibility assessment and manually mapped unmatched clinical phenotypes to HPO terms. In total, 1,193 of 1,200 clinical phenotypes were mapped to HPO. The mapped HPO terms and their ancestor/descendant HPO terms were extracted using the OntoCAT R package and were then written into IEMbase as PostgreSQL tables. For biochemical phenotypes, we allowed matches to terms in any of four vocabularies: SNOMED CT, MeSH, LOINC, and ChEBI. However, manual review of unmatched phenotypes revealed that these terms were highly specialized and thus not present in the vocabularies. Therefore, we implemented an alternative strategy for assessing user-supplied biochemical phenotypes and abandoned the established biochemical vocabularies.

2.4.3 Phenotype-matching algorithm for mini-expert system

The mini-expert system of IEMbase accepts a list of biochemical and clinical phenotypes as input. The system then employs a two-step algorithm that compares the input phenotypic profile against every IEM profile in IEMbase (Figure 2.1).

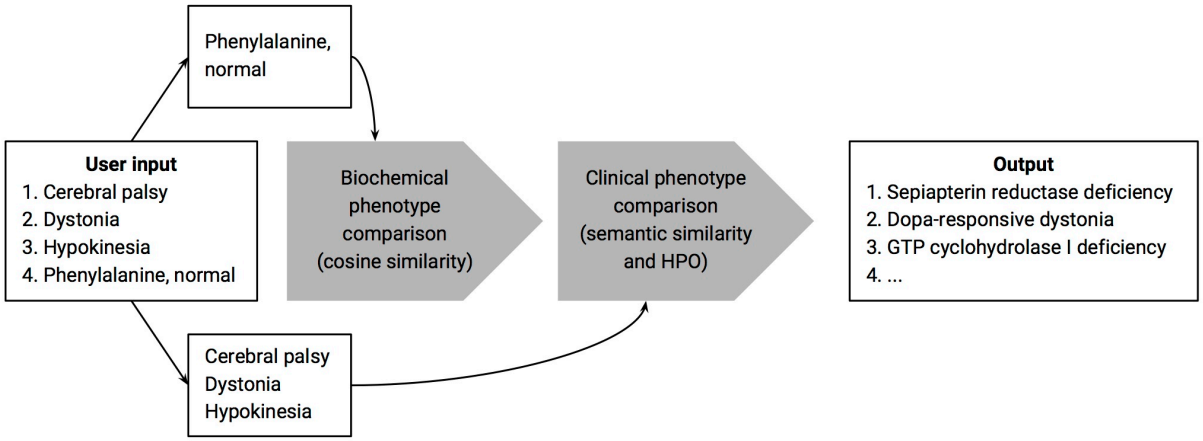


Figure 2.1 Mini-expert algorithm flowchart.

Users enter a list of biochemical/clinical phenotypes into IEMbase's mini-expert system. The system's phenotype-matching algorithm first divides the input list into biochemical and clinical categories. The algorithm then ranks the disorders in IEMbase by comparing the biochemical profile of each disorder against the input biochemical profile, using cosine similarity. Subsequently, the algorithm breaks ties in the ranked list by comparing the clinical profiles, using semantic similarity.

First, the algorithm ranks IEMs by assessing only biochemical phenotypes, using cosine similarity. Cosine similarity [154] is defined as the cosine of two vectors, $TFIDF_Q$ and $TFIDF_D$, which represent the input profile Q and an IEM profile D from IEMbase. The vectors consist of term frequency–inverse document frequency (*tfidf*) scores [154] defined as follows:

$$tfidf(d, D) = tf(d, D) \times idf(d, D)$$

tf(d, D) represents the occurrence of biochemical phenotype *d* in *D*, expressed as 0 or 1. *idf(d, D)* represents the specificity of *d* to *D*, defined as:

$$idf(d, D) = log \frac{Total\ number\ of\ IEMs\ in\ IEMbase}{Number\ of\ IEMs\ associated\ with\ d}$$

Using the above definitions, the algorithm computes *tfidf* scores for all *d* in *D* and all biochemical phenotypes *q* in *Q*. Individual *tfidf* scores are subsequently multiplied by a score for matching the pathological level (i.e., elevated/normal/decreased), which is -1 if the levels of *d* and *q* do not match or 1 if they match. The algorithm then computes the cosine of vector $TFIDF_Q$ and vector $TFIDF_D$:

$$cosSim\left(TFIDF_Q, TFIDF_D\right) = \frac{TFIDF_Q \cdot TFIDF_D}{\left\|TFIDF_Q\right\| \|TFIDF_D\|}$$

The cosine similarity scores are further multiplied by decay factors defined based on severity/characteristics (*sc*) scores for disorder *D*:

$$scDecay(D) = e^{\lambda * dist}$$

λ is a decay constant defined between 0.0 and 1.0. *dist* is a Euclidean distance between a vector of *sc* scores for disorder *D* and a vector of maximum possible *sc* scores. The vector of *sc* scores for *D* consists of *sc* scores for individual phenotypes *d* in *D* that match an input phenotype *q* in *Q*. The *sc* score for individual *d* is defined as follows:

$$sc(d, D) = s(d, D) \times c(d, D)$$

s(d, D) is the severity score of d ranging from 1 to 3, based on the severity annotation of *d*. *c(d, D)* is the characteristic score of *d* assigned either 1 or 2, based on whether *d* is characteristic of *D*.

After the initial ranking of IEMs by biochemical phenotypes, the algorithm breaks ties in this ranking by assessing clinical phenotypes using semantic similarity that is computed based on the work of Kohler et al [145]. The similarity between two clinical phenotypes, p and p', is computed as the information content (IC) of their most informative common ancestor ($MICA_{p,p'}$) in the HPO. IC is a measure of concreteness of a phenotype p in the HPO. It is defined as:

$$IC(p) = -log\left(\frac{Number\ of\ IEMs\ mapped\ to\ p\ and\ its\ descendants}{Total\ number\ of\ IEMs\ in\ IEMbase}\right)$$

The similarity between input profile Q and an IEM profile D is computed by averaging the best match scores for clinical phenotypes q in Q:

$$semSim(Q, D) = \frac{\sum_{i=1}^{n_q} best\ match\ score\ for\ q_i}{n_q}$$

n_q is the number of q in Q. The best match score for each q is defined as $IC(MICA_{q,d_{best}})$, where d_{best} is a clinical phenotype in D whose common ancestor with q has the highest IC and the highest severity score. The similarity score is then multiplied by a decay factor as in biochemical similarity.

2.4.4 Software framework details

IEMbase data is stored in a PostgreSQL database. The front-end user interface was developed using an Angular.js framework. The back-end system was developed in a Ruby on Rails framework.

2.4.5 Code availability

IEMbase is freely available online (http://iembase.org/app) and upon request through an application programming interface. Computer code used for performance evaluation is available upon request.

2.4.6 Mini-expert system case study

To demonstrate a potential use case scenario of the mini-expert system, we used a case of a delayed diagnosis of hyperornithinemia–hyperammonemia–homocitrullinuria syndrome. Case details are described in the Results section.

2.4.7 Performance evaluation of mini-expert system

To evaluate the performance of IEMbase's mini-expert system, 190 retrospective cases were collected from 15 different metabolic centers. For each case, the contributors provided the final diagnosis and biochemical/clinical information. These cases were collected using an online form, which restricted the contributors to providing the case information using only the disorder and phenotype vocabularies in IEMbase.

Each evaluation case was matched to potential diagnoses using the mini-expert system. The system's performance was compared against three phenotype-matching algorithms, each of which uses cosine similarity, with or without semantic similarity, and also with or without severity and characteristic scores.

In addition, the system performance was compared using only biochemical phenotypes, and only clinical phenotypes of retrospective cases. For each retrospective case, the phenotypes were separated into biochemical and clinical categories before each category was evaluated with the mini-expert system. Eighteen cases with phenotypes only in either category, were excluded from this paired comparison (n = 172).

We also tested whether the number of phenotypes specified for each case correlated with the rank of correct diagnoses, in order to assess if some cases ranked better than others because more phenotypes were provided for them.

The above evaluations were conducted using version 1.0.0 of IEMbase. Difference in performance was statistically tested using the Mann-Whitney-U test implemented by wilcox.test in R (version 3.3.1). The correlation test was performed using Spearman's rank correlation test, implemented by cor.test in R. All plots were generated using the ggplot2 R package.

2.5 Results

2.5.1 Overview and walkthrough of IEMbase

We developed IEMbase as an online application which combines a comprehensive IEM knowledge base with a diagnosis support (mini-expert) system. IEMbase curates expert-provided information on 530 IEMs, their treatability and genetics, as well as associated biochemical/clinical phenotypes with detailed annotations on the onset/severity/pathological level of the phenotypes. The application is freely available and can be accessed at http://www.iembase.org/app, or from a link on the project overview website

(http://www.iembase.org). IEMbase is also available through an application-programming interface for integration into other computational systems. Application-programming interface access is available upon request. A detailed walkthrough of the application is presented in Appendix A.1.

2.5.2 Applying mini-expert system in clinical settings

We demonstrate the utility of IEMbase's mini-expert system using a case of a delayed hyperornithinemia–hyperammonemia–homocitrullinuria (HHH) syndrome diagnosis. A girl 2 years and 8 months of age had shown inconspicuous psychomotor development. Following an upper respiratory tract infection, she developed recurrent vomiting, while refusing feeding but drinking occasionally. She was slightly lethargic. Over the following weeks she never fully recovered and continued to undergo episodes of postprandial vomiting, lethargy, and apparent seizures reminiscent of absences. Laboratory tests revealed hyperammonemia (260 μmol/L) together with the constellation of acute liver failure (ASAT 130 U/l, ALAT 233 U/l, ALP 267 U/l, Quick 10%, INR 4.87, aPTT 52sec.). Plasma amino acids demonstrated high to normal glutamine, elevated ornithine, and low citrulline and arginine, all as abnormalities. Orotic acid was highly elevated in urine. Homocitrulline was specifically tested for but could not be identified in plasma or urine. With a presumptive diagnosis of ornithine transcarbamylase deficiency, the patient was referred to a metabolic center and treated, accordingly, with protein restriction and ammonia scavengers. Over the following months, there were several similar episodes, usually triggered by minor intercurrent infections. Molecular analysis of ornithine transcarbamylase was negative.

When the constellation of symptoms was entered into the IEMbase's mini-expert system (Table A1, Appendix A), hyperornithinemia–hyperammonemia–homocitrullinuria syndrome was suggested as the most likely disease candidate, while ornithine transcarbamylase deficiency was listed as the second probable disease candidate. Indeed, molecular analysis of SLC25A15 identified biallelic variants in the gene, confirming the diagnosis of hyperornithinemia–hyperammonemia–homocitrullinuria syndrome and enabling targeted treatment.

2.5.3 Mini-expert system performance evaluation

IEMbase's mini-expert system matched 62% of cases to exact diagnoses, 86% of cases within the top five candidate disorders, and 90% of cases within the top ten. The performance comparison between the mini-expert system algorithm (combined + weighted) and three other phenotype-matching algorithms (combined + unweighted, cosine + weighted, cosine + unweighted) is shown in Table 2.3 and Figure A11 (Appendix A). There was no significant difference in performance between the mini-expert algorithm and the alternative phenotype-matching algorithms. Cases that were ranked out of the top 20 tended to have entries of unspecific biochemical markers, such as "Acylcarnitines, all" or "Amino acids, all." Refer to Table A2 (Appendix A) for an overview of the cases and their ranks. Refer to Table A3 (Appendix A) for more information about the cases that were ranked out of the top 20.

	Combined + Weighted (Mini-expert system)	Combined + Unweighted	Cosine + Weighted	Cosine + Unweighted
MRR	0.72	0.70	0.72	0.68
% success at 1	62	59	63	57
% success at 5	86	85	85	83
% success at 10	90	91	90	89
% success at 20	93	92	92	91

Table 2.3 Mini-expert system performance evaluation results.

Mean reciprocal rank (MRR) measures how close the correct match is to the top rank on average. It ranges from 0 to 1 and values close to 1 indicate that correct matches appear closer to the top on average. % success at N = % of cases with correct diagnoses within top N ranks. Combined = combined cosine and semantic similarity. Cosine = cosine similarity only.

The system performance using only biochemical queries was significantly better than using only clinical queries ($P < 0.001$; Figure 2.2 and Table A4 (Appendix A)). Using only biochemical phenotypes, 60% of cases were matched to exact diagnoses, 83% of cases within the top five candidate disorders, and 89% of cases within the top ten. The success rate of biochemical phenotypes plateaued after 90%, as the number of assessed candidates increased, reflecting 13 cases which failed to produce candidates owing to insufficient/unspecific biochemical information and/or the system's inability to recognize similar biochemical phenotypes. As an example of the latter, the current implementation fails to recognize "Acylcarnitines, all" and "Long-chain acylcarnitine" as related phenotypes. Using only clinical phenotypes, only 19% of cases were matched to exact diagnoses, 38% of cases within the top five candidate disorders, and 49% of cases within the top ten.

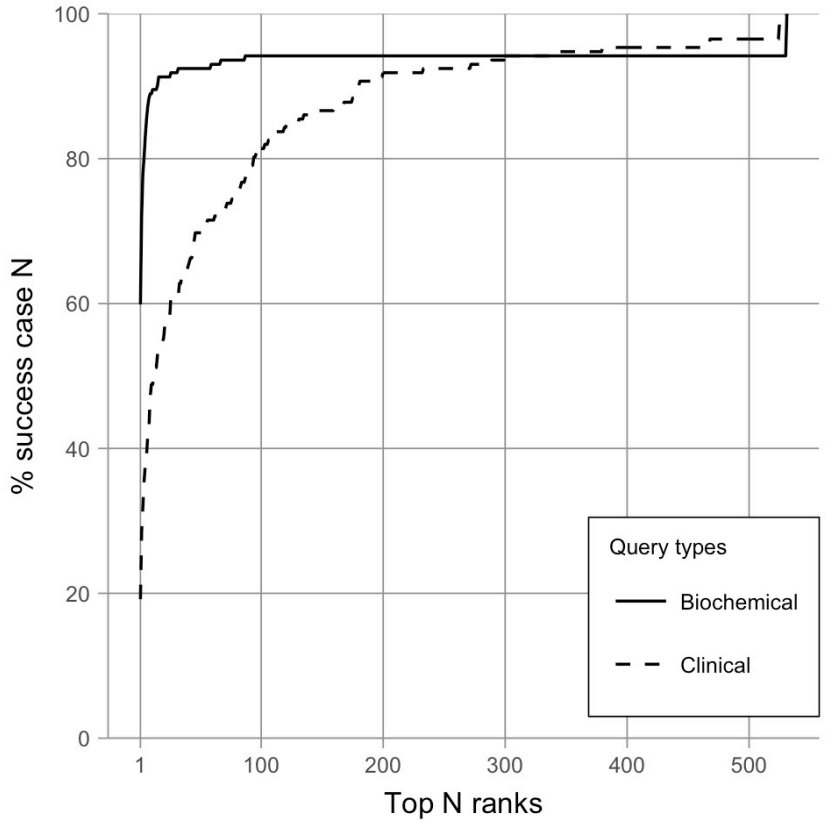


Figure 2.2 Mini-expert system performance using only biochemical/clinical information.

The system performance when using only biochemical phenotypes was compared with that when using only clinical phenotypes of 172 retrospective cases. Percentage success N measures % of cases whose actual diagnoses ranked within the top N ranks. The system performance when using only biochemical phenotypes was significantly better than that when using only clinical phenotypes ($P < 0.001$; Mann-Whitney-U).

There was no significant correlation between the rank of correct diagnoses and the number of provided phenotypes ($P = 0.69$; Figure A12, Appendix A).

2.6 Discussion

Although disease databases for IEMs have been developed in the past [155–157], they were either based on individual case reports [155, 156] or more focused on specific symptoms, such as intellectual disability [157]. Large-scale rare disease databases currently available for a general clinical audience [136, 137] do cover a wide range of rare diseases, but by their nature do not provide the depth of information found in specialized expert knowledge bases. IEMbase is designed to fill this gap, by combining a central knowledge repository with a basic diagnostic support system. This design allows simultaneous collection of the current expert knowledge and its dissemination to the broader clinical community. In addition, it leads to further improvement of the mini-expert system as the depth of knowledge is compiled. Curated knowledge bases are intended to surpass the capacity of any single expert. IEMbase is therefore of utility for all those involved in IEM diagnoses: pediatricians, internists, neurologists, geneticists, and metabolic specialists. As our case study demonstrates, the utility of IEMbase can also be extended to established metabolic centers and biochemical genetics laboratories to help broaden the array of potential differential diagnoses—specifically to include lesser-known diseases when their constellations overlap with typical presentations of better-known diseases.

The evaluation of the mini-expert system revealed that phenotype-matching performance is significantly higher with the use of biochemical phenotypes than that of clinical phenotypes. This probably reflects two influences: (i) many clinical features of IEMs are not specific, while biochemical alterations are frequently so [133, 158] and (ii) the IEM community has made intense efforts toward both disease-specific biomarker discovery and the annotation of biochemical phenotypes [138, 139, 158, 159]. The second point draws upon a hundred years of

IEM community efforts, leading to a depth and breadth of biochemical annotations that constitute a phenome space well suited to research of assisted diagnostic methods. Furthermore, the uniting of biochemical annotations with genetic and clinical annotations aligns with the imminent shift in investigative paradigm, where multi-omics technology allows holistic investigation into an individual's genome, epigenome, transcriptome, proteome, metabolome, and phenome [158]. Extrapolating from our experience, the knowledge bases of other clinical communities may hold untapped high-quality offline information which could be renewed in a similar way to that held in IEMbase.

Owing to a lack of compatible structured vocabulary for biochemical phenotypes in IEMbase, the current mini-expert system algorithm uses a nonsemantic information retrieval metric (tfidf-cosine similarity) to compare biochemical phenotypes. We recognize that this approach is not robust when matching imprecise terms. For example, the use of tfidf-cosine similarity will not take into account the fact that neopterin and biopterin belong to the same group of pterins. The use of structured vocabulary and semantic similarity can mitigate this shortcoming. Therefore, we plan to contribute our biochemical vocabulary to existing ontologies as we make updates to our system.

Biochemical test/gene panel suggestions that are provided with the output of the mini-expert system are currently restricted to basic information (e.g., gene names or chemical test panels), as detailed specification will require future contributions from the expert community. We anticipate that such improvements will be introduced over time as a result of community outreach efforts such as those described below.

For the long-term viability of IEMbase, continuous contribution from the expert community is crucial, especially with the large number of novel IEMs and phenotypes now being revealed with the use of multi-omics technologies. Therefore, we will periodically reach out to the IEM community for knowledge contribution, in addition to assembling an expert panel, which will regularly review and update the knowledge base. To encourage adoption among the new generation of clinicians, we plan to develop a mobile version of the application and a training module.

In summary, IEMbase is a web application intended to provide the clinical community with a comprehensive IEM knowledge base and a tool to facilitate early and accurate diagnoses of IEMs. Its knowledge base features expert-curated clinical resources on 530 IEMs. Its mini-expert system empowers clinicians and complements their workflow with suggested diagnoses, differential diagnosis charts, biochemical test panels, and gene panels. The multitude of suggestions enables clinicians to initiate concurrent biochemical and genetic evaluations, where the former can help focus the latter for rapid diagnosis, especially in clinical exome/genome interpretations. We believe that the power of IEMbase comes from the community of experts who contribute their knowledge for the greater benefit of the broader clinical community and as such, the value of community science should be recognized as a key component of digital medicine in the 21st century.

Chapter 3: Development and user evaluation of a rare disease gene prioritization workflow based on cognitive ergonomics

3.1 Synopsis

Objective The clinical diagnosis of genetic disorders is undergoing a transformation, driven by whole exome sequencing and whole genome sequencing (WES/WGS). However, such nucleotide-level resolution across 3 billion base pairs creates an interpretive challenge. Prior literature suggests that clinicians may employ characteristic cognitive processes during WES/WGS investigations to identify disruptions in genes causal for the observed disease. Based on cognitive ergonomics, we designed and evaluated a gene prioritization workflow that supported these cognitive processes.

Materials and Methods We designed a novel workflow, in which clinicians recalled known genetic diseases with similarity to patient phenotypes to inform the WES/WGS data interpretation. This prototype-based workflow was evaluated against the commonly used computational approach based on physician-specified sets of individual patient phenotypes. The evaluation was conducted as a web-based user study, where 18 clinicians analyzed two simulated patient scenarios using a randomly assigned workflow. Data analysis compared the two workflows with respect to accuracy and efficiency in diagnostic interpretation, efficacy in collecting detailed phenotypic information, and user satisfaction.

Results Participants interpreted genetic diagnoses faster using prototype-based workflows. The two workflows did not differ in other evaluated aspects.

Discussion The user study findings indicate that prototype-based approaches, which reflect the cognitive processes of the experts, can expedite gene prioritization. However, further research is required to study the extent of this accelerated diagnosis across diverse genetic diseases.

Conclusion The findings demonstrate potential for prototype-based phenotype description to accelerate computer-assisted variant/gene prioritization through complementation of skills, knowledge, and experience of clinical experts via human-computer interaction.

3.2 Background and significance

Whole exome sequencing (WES) and whole genome sequencing (WGS) are allowing clinicians an unprecedented opportunity to examine human genes *en masse* and to diagnose rare genetic diseases [1, 39, 46]. An accurate and efficient analysis of DNA sequence data has become crucial for a timely diagnosis of patients, many of which might otherwise suffer a long and costly diagnostic odyssey [32, 160, 161]. However, identifying causal variants among millions of DNA variations (within billions of nucleotides) in any individual is challenging [12]. For this reason, collaborative global efforts have focused on expediting WES/WGS analyses, by encoding available clinical genetic knowledge into computers [64, 65, 75, 78] and creating computational methods that exploit encoded information [7, 8, 70, 72, 80, 162-164]. Such solutions have improved efficiency in multiple aspects of WES/WGS analyses, from collecting comprehensive phenotype information [7], to prioritizing potentially pathogenic variants [8, 70, 72, 162-164], and to matching patients for collaborative investigation of rare, novel genetic diseases [80].

While the above aspects have been improved with computational approaches, variant (or gene in a wider context) prioritization and interpretation during WES/WGS analyses have largely remained expert-driven tasks with computer assistance that connect variant-level, gene-level, phenotype-level, and population-level information to patients [12]. Human experts have been a vital part of prioritization and interpretation as these processes require cross-examination of complex evidence that affect treatment decisions [12, 165]. Recent economic analysis of cancer-related genome diagnosis indicates that the overall cost is increasingly dominated by the informatics/interpretative activity [13]. Considering the importance of human experts, an alternative solution for accelerating WES/WGS analyses may lie in the creation of new computational methods that more efficiently collaborate with highly-trained experts (whose skills and knowledge are difficult to fully encode into computers). For instance, a recent study in this direction has demonstrated that variant prioritization based on a clinician-generated gene list outperformed purely computational methods in the analysis of singleton WES data [84]. The findings suggest the utility of harnessing clinical expertise, such as a clinician's experience, skills in recognizing clinical gestalt, and their ability to evaluate multifactorial information such as disease onset, family history, and negative findings [84, 85, 166].

In this study, we report the design and evaluation of a gene prioritization workflow based on cognitive ergonomics, the study of understanding human cognitive capabilities in interactive systems, and applying this understanding to support human cognition via human-system interaction for optimized system performance [167]. The word “workflow”, within the context of this study, refers to a sequence of interactions between clinical experts and computers during computer-assisted variant/gene prioritization. Using this definition, this study focused on

examining two different designs of interactions (workflows) and their effect on expert performance regardless of variant/gene prioritization algorithms. These two workflows are herein referred to as the prototype-based workflow and symptom-based workflow.

First, we created the prototype-based workflow that aimed to complement the following characteristics of diagnostic reasoning and human cognition reported in literature: (a) clinicians form a "gestalt diagnosis" from perceived clinical information [82, 168, 169], (b) people tend to make categorizations using an ideal/core representation called the "prototype" [170], and (c) people tend to focus on deeper structural information when comparing two examples, whereas they focus on superficial information when considering an isolated example [171]. The “prototype” in this study refers to a representation that effectively describes patient characteristics, in the form of a specific genetic disease that closely resembles a patient. Therefore, in this prototype-based workflow, clinicians suggest a specific genetic disease with resemblance to patient phenotypes before initiating variant/gene prioritization. The computer then retrieves a set of characteristics described for the prototype disorder from an underlying database.

Next, the prototype-based workflow was compared against the symptom-based workflow, which simulated a commonly applied workflow in which experts provide a set of individual characteristics observed in the patient before embarking on a computational variant/gene prioritization process. For workflow comparison, a user study was conducted with expert clinical/biochemical geneticists as subjects. The workflows were assessed with respect to accuracy and efficiency in diagnostic interpretation, efficacy in collecting detailed phenotypic

information, and user satisfaction. Finally, we created a proof-of-concept mobile application for phenotyping based on study findings.

This study explores an alternative in computer-assisted variant/gene prioritization and interpretation where computational methods attempt to harness the intellectual power of clinical experts by aligning human-computer interaction with a natural reasoning process. We hope our findings catalyze further interest to explore human-computer interactive methods in this domain.

3.3 Materials and methods

3.3.1 Workflow definitions

As explained in Background and Significance (Section 3.2), the "workflow" refers to a sequence of interactions between clinical experts and computers during computer-assisted variant/gene prioritization. The two workflows (prototype-based and symptom-based) that were designed for the study are illustrated in Figure 3.1. For clarification, this section briefly summarizes each workflow to help understanding of the remaining sections of Materials and Methods. For a detailed explanation of the workflow designs, please refer to the "Workflow designs" in Results (Section 3.4.1).

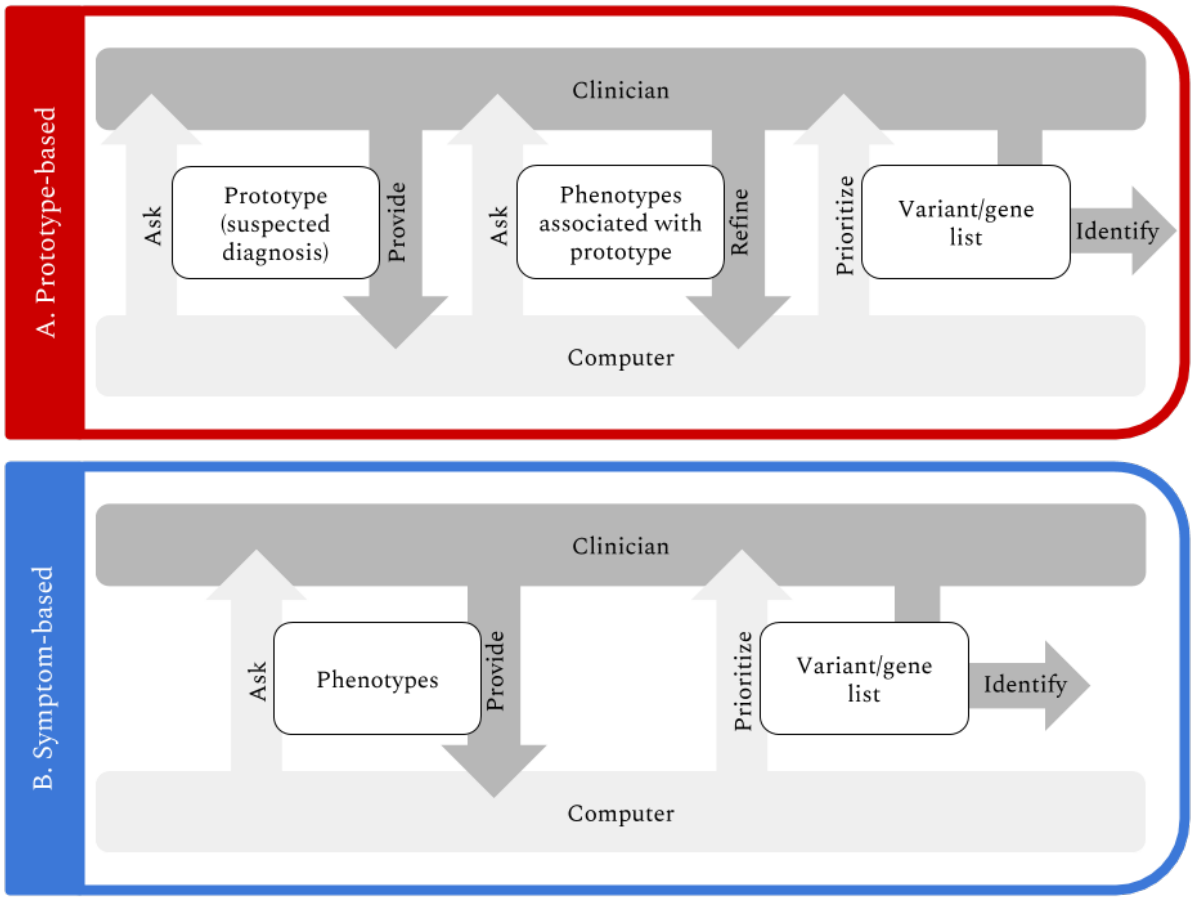


Figure 3.1 Sequence diagram for prototype-based and symptom-based workflows.

(A) Illustration of the prototype-based workflow. (B) Illustration of the symptom-based workflow. In the prototype-based workflow, clinicians provide a prototype in the form of suspected diagnosis, refines a list of phenotypes that are suggested based on the given prototype, and identifies a causal variant/gene from a list of variants/genes that is computationally prioritized by the relevance to given phenotypes. In the symptom-based workflow, clinicians provide a list of phenotypes and identifies a causal variant/gene from a list of variants/genes that is computationally prioritized by the relevance to given phenotypes. The prototype-based workflow is different from the symptom-based workflow in that it explicitly asks clinicians to provide prototypes that they have in mind.

The prototype-based workflow was designed to complement the cognitive properties concerning the use of prototypes - model presentations of genetic diseases - during patient assessment and variant interpretation during WES/WGS investigations. In this prototype-based workflow, prior

to initiating gene prioritization, clinicians are solicited to provide a prototype in the form of a disease with similarity to the observed patient phenotypes. The system then extracts a set of phenotypes described for the known disorder from an underlying database.

To assess the impact of the prototype-based workflow, a symptom-based workflow was implemented for comparison. This symptom-based workflow simulated a common process employed by phenotype-driven variant/gene prioritization tools, where users enter patient phenotypes and variants/genes were assessed based on their relevance to input phenotypes [8, 162-164]. The words "symptom" and "phenotype" refer to characteristics of patients, and will be used in an interchangeable manner in the upcoming sections.

The difference between the two workflows was that the prototype-based workflow explicitly asked for a suspected diagnosis (to populate the set of phenotypes, which the user can refine by eliminating/adding terms) while the symptom-based workflow solicits the user to specify the observed patient phenotypes *de novo*.

3.3.2 User study participants

Between October 2017 and May 2018, 59 clinicians from specialized (tertiary) healthcare institutions within Canada, the Netherlands, Ireland, Germany, and Switzerland were invited to participate in the user study. Participant inclusion criteria were to (a) hold the title of medical geneticist/biochemical geneticist or specialize in rare genetic diseases, and (b) have prior experience working with WES/WGS data as part of their clinical practice. The invitees were identified by consulting hospital staff directories, a rare disease research network, and

collaborators. The invitees were contacted by an email which provided researcher information, explanation on how the contact was obtained, purpose and a brief description of the study, as well as a web link to the user study website. Participation was completely voluntary and consent to participate was implied by submission of responses. Of the 59 invitees, 18 completed their participation in the study.

The user study was reviewed and approved by the University of British Columbia Research Ethics Board (Certificate: H17-00872).

3.3.3 Development of simulated clinical scenarios

Five simulated clinical scenarios were developed for the user study (Table 3.1). One was dedicated to a tutorial exercise and four were for clinical scenario analysis exercises. The latter four scenarios were coupled as two disease-based pairs, with each pair consisting of a scenario that described a typical presentation and a scenario that described an atypical presentation of a genetic disease. The above arrangements were used for scenario assignment in the study so that each participant analyzed one typical scenario from one pair and one atypical scenario from the other pair, while the order of the scenarios was randomized. This ensured (a) elimination of exposure to the same genetic disease diagnosis during analysis exercises, (b) minimizing ordering bias, and (c) examination of the effect of different disease presentations on workflow performance.

		Tutorial scenario	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Diagnosis		CHARGE syndrome (MIM 214800)	Smith-Lemli-Opitz syndrome (MIM 270400)	Smith-Lemli-Opitz syndrome (MIM 270400)	Tuberous sclerosis 1 (MIM 191100)	Tuberous sclerosis 1 (MIM 191100)
Gene		*CHD7*	*DHCR7*	*DHCR7*	*TSC1*	*TSC1*
Typical/Atypical		-	Typical	Atypical	Typical	Atypical
Demographic information		5-month-old girl	18-month-old boy	18-month-old boy	6-year-old girl	6-year-old girl
Family information		Parents were nonconsanguineous and of European ancestry	Parents were nonconsanguineous and of European ancestry	Parents were nonconsanguineous and of European ancestry	Parents were nonconsanguineous and of European ancestry	Parents were nonconsanguineous and of European ancestry
Clinical synopsis*	**Pregnancy and delivery**	Born at term following an uneventful pregnancy and delivery	Born at term following an uneventful pregnancy and delivery	Born at term following an uneventful pregnancy and delivery	Born at term following an uneventful pregnancy and delivery	Born at term following an uneventful pregnancy and delivery
	Phenotypic description	- Asymmetric facial palsy - Bilateral coloboma of the iris - Choanal atresia and ventricular septal defect @ birth - Developmental delay - Missing ear lobes and short, wide ears - Swallowing difficulties	- 2nd-3rd toe syndactyly - Anteverted nares - Broad nasal bridge - Developmental delay - Feeding difficulties and failure to thrive @ 3 months - Hypotonia - Irritable - Low-set ears - Microcephaly - Micrognathia - Postaxial polydactyly - Ptosis	- Brain MRI and MRS: no structural abnormalities - Broad nasal bridge - Developmental delay - Feeding difficulties @ 3 months - Finger clinodactyly - Micrognathia - Mild hypotonia - Mild ptosis - Minimal cutaneous 2nd-3rd toe syndactyly	- Brain MRI: cortical sclerotic tubers - Epileptic seizure - Hypomelanotic macules on the chest - Hypsarrhythmia - Renal cysts - Skin papules on the side of nose	- Brain MRI: normal - Epileptic seizure - Hypsarrhythmia - Intellectual disability - Renal cysts - Skin papules on the side of nose

Table 3.1 Simulated scenarios.

* Clinical synopses are summarized from a paragraph format for brevity

Each scenario consisted of a diagnosis, a patient description, and a gene list, simulating a case involving WES data (equivalent to restricting analysis to exons within WGS data). Normally, WES analyses produce a list of variants at the resolution of nucleotides. In order to limit the time demands on participants, we simplified the results to provide a list of genes impacted by variation. In the user study, participants were explicitly notified that a variant list had been simplified to display only gene level information, and instructed participants to assume that each gene in the list harboured a variant/variants that was/were rare, potentially pathogenic, and aligned with inheritance models (e.g. dominant, recessive).

Each simulated clinical scenario was developed in the following order: diagnosis, patient description, and gene list. Diagnosis selection used the following criteria: (1) the diagnosis was a rare genetic disease that had been described in at least ten peer-reviewed publications; (2) it was widely known so that participants could recognize its associated gene by name/symbol during gene list interpretation, thus minimizing time spent looking up gene information using online tools; and (3) the disease was well-characterized so that participants could formulate a prototype (or a model presentation of the disease) by reading a text description. After reviewing previously published rare genetic disease annotations [145], three diseases, CHARGE syndrome, Smith-Lemli-Opitz syndrome, and tuberous sclerosis, that fulfilled the above criteria were assigned to each scenario as follows: CHARGE syndrome for the tutorial scenario, Smith-Lemli-Opitz syndrome for two analysis scenarios, and tuberous sclerosis for the remaining two analysis scenarios.

Based on the diagnosis assignment, patient descriptions were then generated by extracting typical/atypical characteristics from published case reports (Appendix B) as well as the disease annotations used during the previous step, which contained a list of phenotypes described using the Human Phenotype Ontology (HPO) [75] and their frequency [145].

After patient descriptions were generated, gene lists were compiled. The gene list for each scenario contained 17 genes, one associated with the scenario's diagnosis and the rest associated with diseases that had varying degrees of similarity to the diagnosis. The purpose of such an arrangement was to ensure that investment of thought and time was required before discerning the diagnosis. The following outlines the steps that determined gene lists. For each scenario, the patient description was converted into a list of HPO terms. These terms were then used to compute the scenario's similarity against 6946 diseases in Online Mendelian Inheritance in Man (OMIM) that were annotated by HPO [75] (phenotype_annotation.tab downloaded on June 27, 2017). Similarity was computed using a previously published HPO-based disease similarity score [145] and normalized to a range between 0 and 1. OMIM diseases were then ordered and categorized by their similarity [highly similar (0.6-1.0), similar (0.5-0.6), somewhat similar (0.4-0.5), and irrelevant (0-0.4)]. From each category, four diseases were randomly selected and their associated genes were added to the gene list. All components of the simulated clinical scenarios were reviewed by CDMvK.

3.3.4 User study procedure

The user study was formatted in an online survey. Participants were asked to complete the survey as outlined in Figure 3.2 and were randomly assigned to either prototype-based or symptom-

based workflows. The survey consisted of four sections: introduction, clinical scenario analysis, debriefing, and user satisfaction questionnaire. The introduction section presented three questions regarding participants' demographic information/clinical expertise, an orientation video explaining the study purpose and procedure, and a tutorial exercise which walked through a sample clinical scenario to help participants become acquainted with the survey interface.

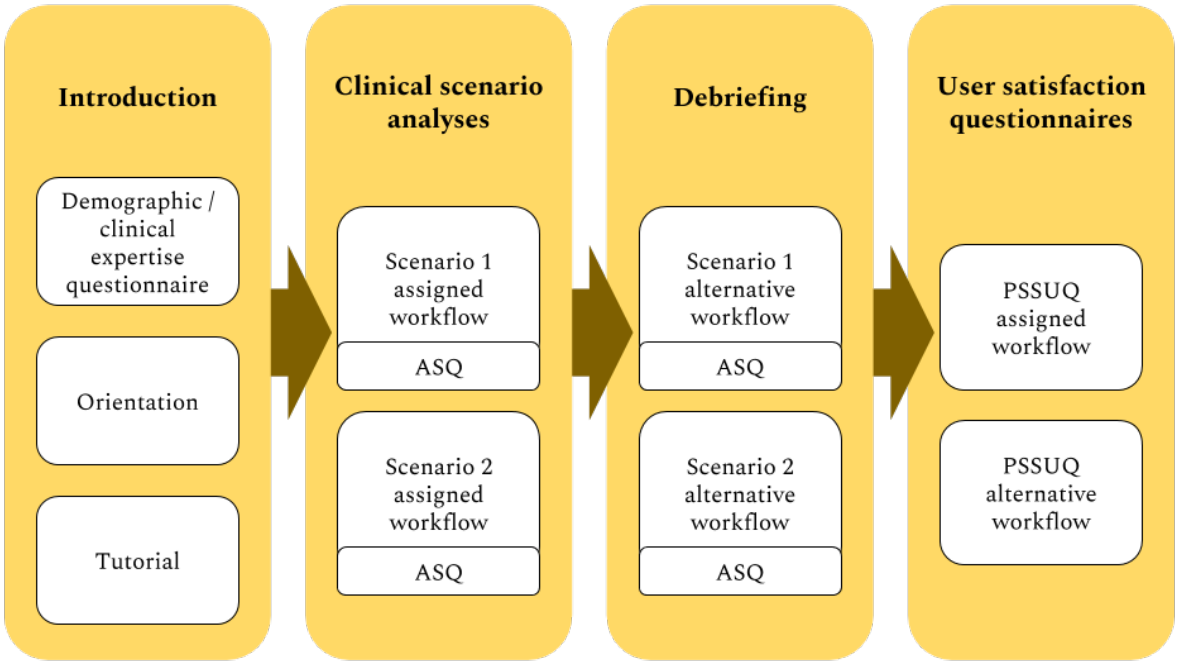


Figure 3.2 User study structure.

The user study consisted of four main sections: introduction, clinical scenario analysis, debriefing, and user satisfaction questionnaire. During the introduction, participants answered questions regarding their demographic information and clinical expertise, watched an orientation video, and walked through a sample clinical scenario. Afterwards, participants analyzed two simulated clinical scenarios using their assigned workflow. At the end of each scenario, participants completed an After-Scenario Questionnaire (ASQ). Upon completion of clinical scenario analyses, participants were debriefed about the workflow that they were not assigned to and tried out the workflow using the same simulated scenarios. Participants also filled out an ASQ at the end of each scenario. Finally, participants filled out Post-Study System Usability Questionnaires, regarding the assigned workflow and the alternative (unassigned) workflow, respectively.

The clinical scenario analysis section invited participants to diagnose two simulated clinical scenarios using their assigned workflow. For each scenario, the analysis exercise proceeded as follows. Participants were presented with a simulated patient description and asked to input prototypes or patient phenotypes according to their assigned workflow (Figure 3.1). The order of the sentences within the description was randomized to minimize ordering bias. For prototype selection, participants were restricted to OMIM disease names (provided by OMIM API) [64]. For phenotype selection (symptom-based workflow) and phenotype refinement (prototype-based workflow), participants were restricted to HPO terms. Such restrictions were imposed to enable accurate comparison of input from different participants. Afterwards, participants were asked to identify a diagnosis within a simulated gene list, which was ordered by the number of phenotypes that overlapped between input and diseases that were associated with each gene. The ordering was performed to mimic the output of common computational variant/gene prioritization tools. Gene-phenotype-disease associations provided by HPO [75] were used to enable this functionality. Participants could freely modify input phenotypes and reorder the gene list until they identified a diagnosis. Following diagnosis selection, the actual diagnosis was revealed to participants, and they were invited to express their satisfaction with the assigned workflow by completing a modified After-Scenario Questionnaire (ASQ) [172].

During each analysis exercise, the following information was collected: prototype/phenotype selections, changes made to prototype/phenotype selections before making diagnoses, final diagnoses, time elapsed between initial display of the gene list and identification of diagnoses, and ASQ responses.

Upon completion of two clinical scenario analyses, participants were debriefed about the alternative (unassigned) workflow. During debriefing, they walked through the alternative workflow using the same scenarios and completed ASQ at the end of each scenario. Only ASQ responses were collected during the walkthrough. Finally, participants were invited to express their overall satisfaction with the workflows by completing two modified Post-Study System Usability Questionnaires (PSSUQ) [172], for the assigned workflow and for the alternative workflow, respectively.

For the survey, a custom online interface was developed using Ruby on Rails and React.js in order to implement functionalities required by clinical scenario analysis exercises.

3.3.5 Data analysis

All data analyses were performed using R version 3.4.4. The two workflows were compared with respect to (a) diagnostic accuracy (measured as the number of correctly diagnosed scenarios), (b) efficiency in gene list interpretation (measured as the time elapsed between when the gene list was presented and when participants selected causal gene from the list), (c) efficacy in phenotype collection (measured as the number of participant-provided phenotypes), and (d) user satisfaction (measured as ASQ and PSSUQ scores). All comparisons except the PSSUQ score comparison were performed using a 2 x 2 analysis of variance (ANOVA) (afex package) with workflow assignments (prototype-based/symptom-based) as a between-subject variable, disease presentations (atypical/typical) as a within-subject variable, and each measurement as a response variable. The primary focus of ANOVA was on the main effect of workflow assignments. PSSUQ scores were compared using the Mann–Whitney U test (wilcox.test). To account for

multiple comparisons within (d), the Bonferroni correction (p.adjust) was applied to ASQ and PSSUQ comparisons. Participant-provided prototypes and phenotypes were qualitatively analyzed for common and workflow-specific information patterns. Optional written comments provided in ASQ and PSSUQ were reviewed to extract common participant opinions.

3.4 Results

3.4.1 Workflow designs

We present the design of the two workflows investigated in this study as follows. The prototype-based workflow (Figure 3.1A) was designed to augment the following properties of clinical reasoning and human cognition during WES/WGS investigations: (a) an ability to form gestalt diagnosis [82, 168, 169], (b) a tendency to categorize using an ideal/core representation called the "prototype" [170], and (c) a tendency to focus on deeper structural information when comparing two examples [171]. The specific steps of this prototype-based workflow follow: (1) the computer solicits the clinician to provide a prototype in the form of suspected diagnosis; (2) the computer presents a list of key phenotypes of the given prototype; (3) the clinician refines the presented list by adding or excluding phenotypes; (4) the computer prioritizes genes based on their overlap with the phenotypes; and (5) the clinician specifies a causal gene (diagnosis) from the prioritized list.

The rationale behind this prototype-based workflow design was that clinicians employ prototypes as proxies to evaluate the characteristics of patients or to gauge the relevance of each candidate variant/gene during WES/WGS data interpretation. In terms of cognitive ergonomics, the prototype-based workflow was anticipated to reduce the cognitive burden of constantly

keeping track of prototypes by (a) explicitly asking for them and (b) guiding the expert-computer interaction based on how experts would employ them during phenotypic assessment and gene interpretation.

The symptom-based workflow (Figure 3.1B) was designed for comparison with the prototype-based workflow. This symptom-based workflow modelled common phenotype-driven variant/gene prioritization tools. Specific steps of the symptom-based workflow follow: (1) the computer solicits clinicians to provide a list of patient phenotypes; (2) computer prioritizes genes based on their overlap with the phenotypes; and (3) clinicians identified a causal gene (diagnosis) from the prioritized list. The difference from the prototype-based workflow was that the symptom-based workflow did not ask for prototypes and instead focused on serially collecting individual patient phenotypes.

3.4.2 User study participant characteristics

Characteristics of 18 participants are summarized in Figure 3.3. 94% of participants have practiced more than 5 years. All participants had experience with cases involving clinical WES/WGS data.

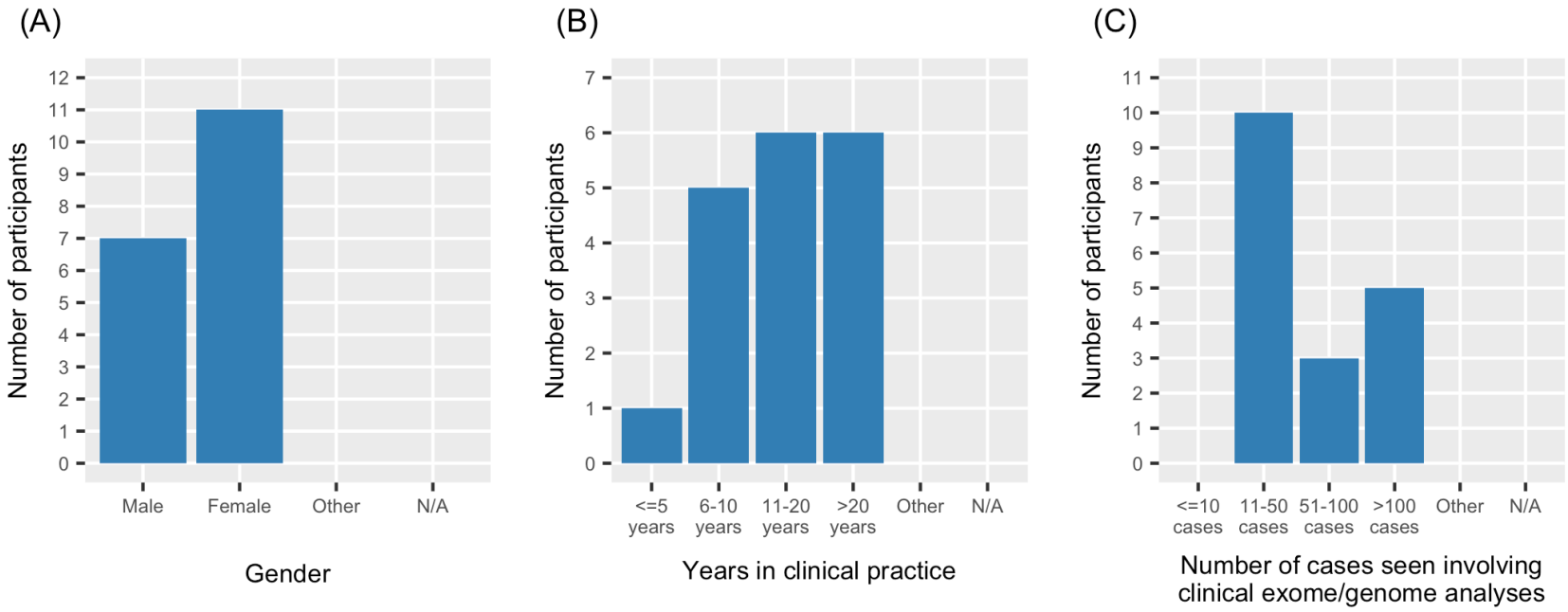


Figure 3.3 Participant characteristics.

(A) Gender of participants; (B) Participants' level of clinical expertise, measured as years in clinical practice; (C) Participants' experience with exome or genome sequencing data, measured as the number of cases involving exome or genome analyses.

3.4.3 Workflow performance evaluation

Figure 3.4 summarizes the performance of the prototype-based workflow and the symptom-based workflow. There was no difference in diagnostic accuracy between the two workflows ($F(1, 16) = 1.0$, $p = .33$, $\eta_p^2 = .059$). Almost all participants, except one, correctly diagnosed assigned scenarios. The participant who incorrectly diagnosed one scenario explained via optional comments that a general diagnosis (tuberous sclerosis) was correctly anticipated and the correct genetic diagnosis (*TSC1*) was considered during gene list interpretation. However, the participant determined that the presented scenario was more compatible with a different genetic diagnosis (*TSC2*) and thus did not select any diagnosis.

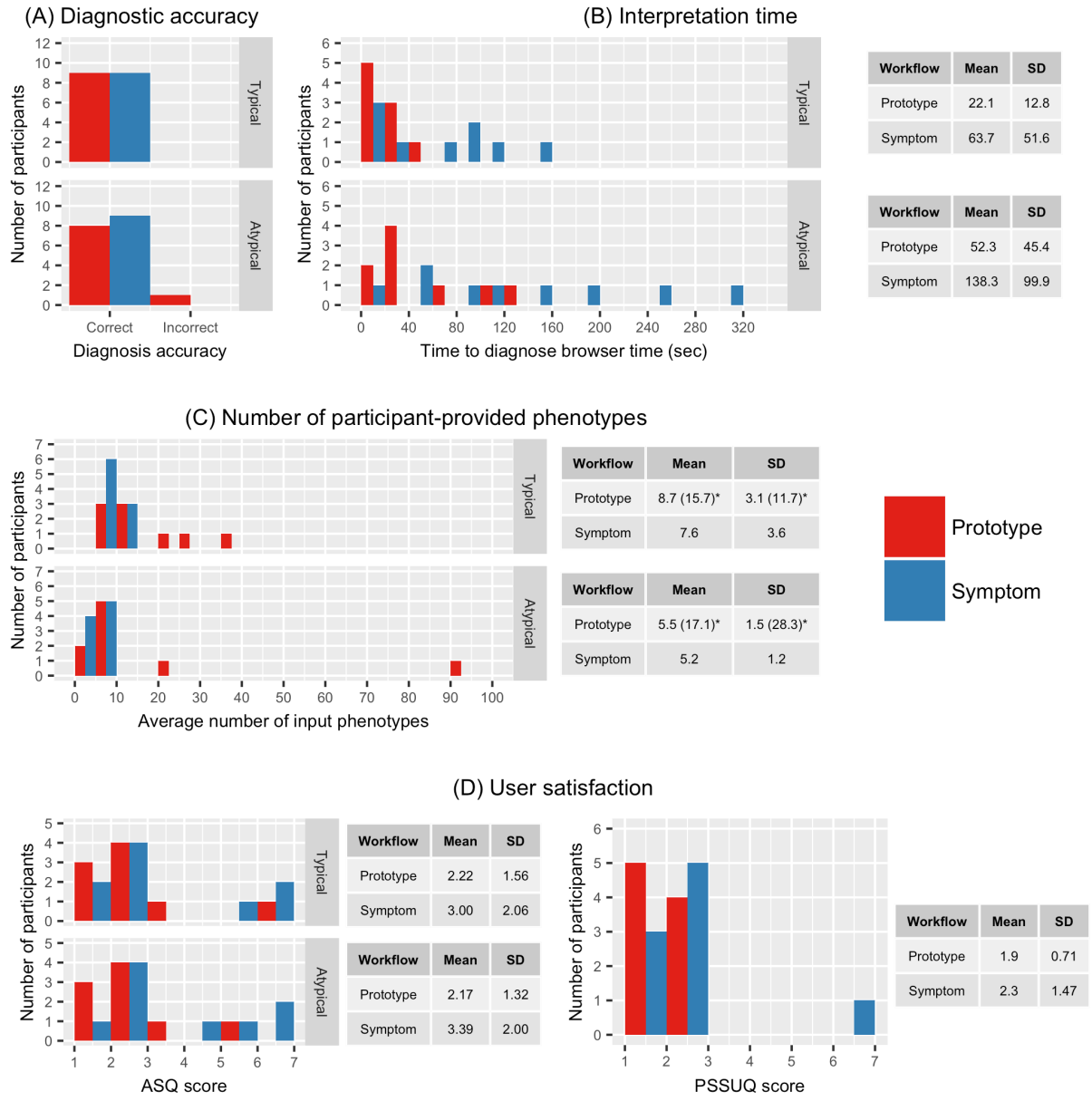


Figure 3.4 Summary of workflow performance evaluation.

Evaluation results are shown in histograms or bar-plots for categorical variables. In (B), (C), and (D), the tables next to histograms summarize descriptive statistics for each corresponding histogram. SD = standard deviation. (A) Diagnostic accuracy, measured as the number of correctly diagnosed scenarios; (B) Interpretation time, measured as the time elapsed between when the gene list was presented and when participants selected causal gene from the list; (C) Number of participant-provided phenotypes. Values denoted by * represent mean or standard deviation including (within brackets) or excluding (without brackets) three outlier individuals assigned to the prototype-based

workflow; (D) User satisfaction, measured as After-Scenario Questionnaire (ASQ) and Post-Study System Usability Questionnaire (PSSUQ) scores.

Participants who were assigned to prototype-based workflows identified diagnoses significantly faster than those assigned to symptom-based workflows ($F(1, 16) = 6.04$, $p = .026$, $\eta_p^2 = .27$). In addition, participants identified diagnoses faster for scenarios with typical presentations than atypical presentations ($F(1, 16) = 18.1$, $p = .0006$, $\eta_p^2 = .53$), while no significant interaction between workflow assignment and disease presentation was observed ($F(1, 16) = 3.26$, $p = .090$, $\eta_p^2 = .17$).

No difference was observed in the number of phenotypes collected by either workflow ($F(1, 16) = 2.71$, $p = .12$, $\eta_p^2 = .14$). Three outliers were observed in the number of input phenotypes collected using prototype-based workflows. Examination of individual responses revealed that at least two participants who were assigned to prototype-based workflows selected almost all of the phenotypes that were suggested based on participant-specified prototypes, regardless of their presence/absence in simulated scenarios (i.e. they chose not to eliminate phenotypes that were not reported in the scenarios). Lastly, there was no difference in user satisfaction between the two workflows (ASQ: $F(1, 16) = 1.50$, $p = .48$ (uncorrected $p = .24$), $\eta_p^2 = .086$; PSSUQ: $W = 37$, $p = 1.0$ (uncorrected $p = .79$), $r = 0.19$).

	Actual scenario diagnosis	Participant-specified prototype	# of participants who selected the prototype**
Scenario 1 (n* = 5)	Smith-Lemli-Opitz syndrome (MIM 270400)	Smith-Lemli-Opitz syndrome (MIM 270400)	5
Scenario 2 (n* = 4)	Atypical Smith-Lemli-Opitz syndrome (MIM 270400)	Smith-Lemli-Opitz syndrome (MIM 270400)	4
Scenario 3 (n* = 4)	Tuberous sclerosis 1 (MIM 191100)	Tuberous sclerosis 1 (MIM 191100)	2
		Tuberous sclerosis 2 (MIM 613254)	2
Scenario 4 (n* = 5)	Atypical tuberous sclerosis 1 (MIM 191100)	Tuberous sclerosis 1 (MIM 191100)	3
		Tuberous sclerosis 2 (MIM 613254)	3

Table 3.2 Prototype selection summary.

* n = number of participants assigned to scenario using the prototype-based workflow

** Counts how many participants selected each prototype as a probable diagnosis. If participants changed prototypes multiple times, they were counted for all prototypes that they had specified.

3.4.4 Qualitative analysis of input prototype and phenotypes

Nine participants who were assigned to prototype-based workflows selected the actual or very close diagnoses as prototypes prior to interpreting gene lists (Table 3.2). Phenotypes that were collected by the two workflows are summarized in Figure 3.5. A detailed list of phenotypes is provided in Appendix B. Phenotypes provided by three individuals assigned to prototype-based workflows were excluded from this comparison, as two likely did not refine phenotype suggestions and one did refine the suggestions for one case but likely did not refine for the other case. Those phenotype lists likely did not involve a conscious assessment of patient phenotypes.

Participants who were assigned to symptom-based workflows had a tendency to input close synonyms of a phenotype. For example, hypotonia in the atypical Smith-Lemli-Opitz scenario was captured in three different terms, generalized hypotonia, central hypotonia, and muscular hypotonia. Meanwhile, synonyms were rarely present in phenotypes captured by prototype-based workflows because participants were offered to select/unselect suggested phenotypes that were associated with the prototype of their choice. Furthermore, the prototype-based suggestions seem to have encouraged participants to enter additional phenotypes that were not collected by symptom-based workflows. For example, terms such as vomiting, gastroesophageal reflux, and poor suck were provided for feeding difficulty in the atypical Smith-Lemli-Opitz scenario.

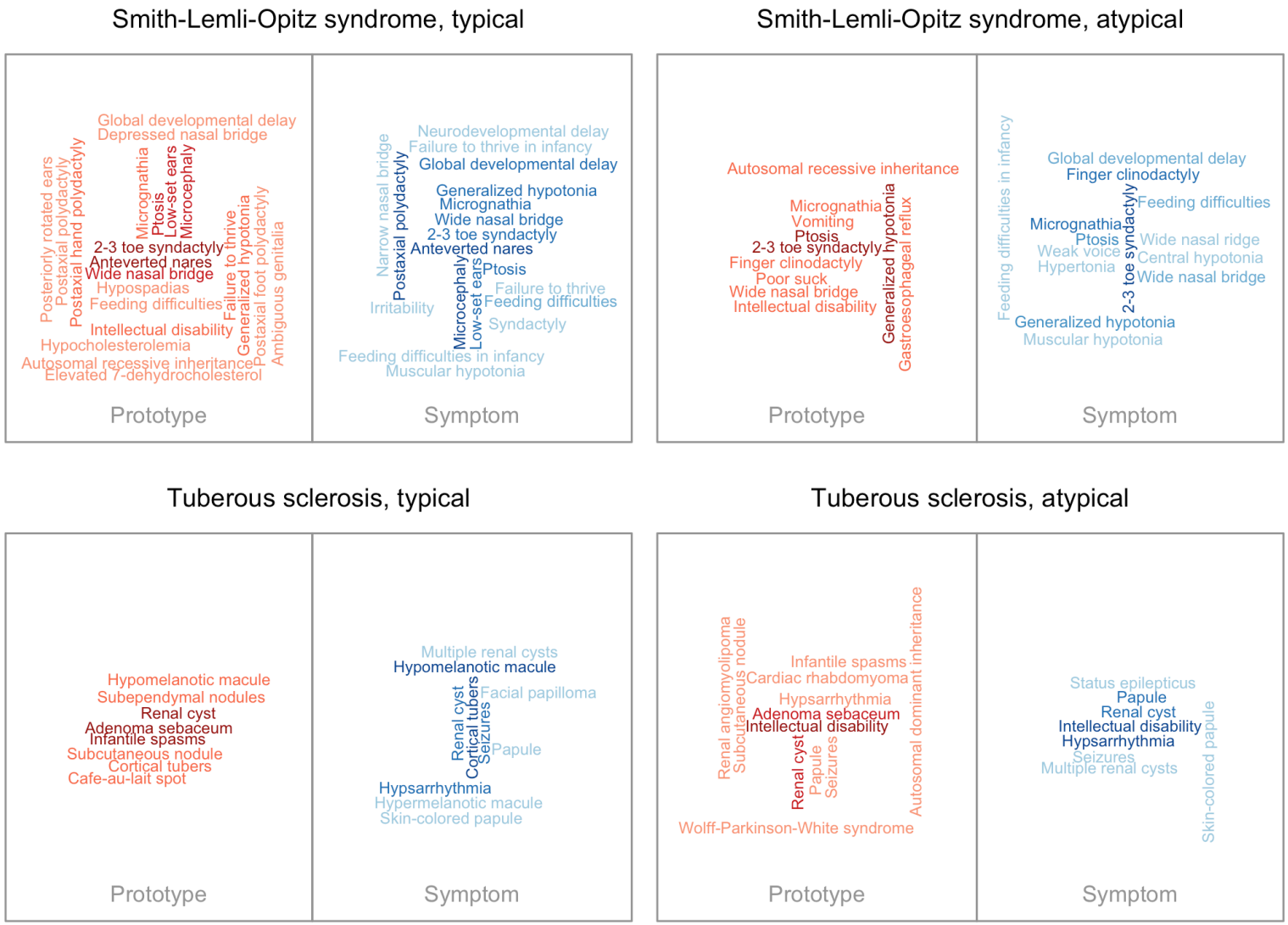


Figure 3.5 Qualitative summary of phenotype selection.

For each scenario, phenotype terms collected by the two workflows are summarized into word clouds. Red clouds represent phenotype terms collected by prototype-based workflows. Blue clouds represent phenotype terms collected by symptom-based workflows. Darker colors represent terms that were collected more frequently. The underlying data is available in Table B2 (Appendix B).

3.4.5 User-requested mobile application for phenotyping

In the survey results, a participant expressed preference for availability of both workflows for use on mobile devices. As such, we created a proof-of-concept, open-source, mobile application, PhenoChat (https://github.com/jes8/phenochat), to demonstrate a potential implementation that combines both workflows. PhenoChat allows users to build phenotypic descriptions by specifying individual phenotypes or by specifying a prototype and subsequently refining phenotypes that are suggested based on known disease-phenotype associations. Users can send the descriptions via email or copy them into a clipboard and share via messaging services. The descriptions are restricted to HPO [75] and OMIM [64] terminologies, and they are generated in a machine-processable format to enable integration into existing framework for phenotyping. PhenoChat was developed using React Native.

3.5 Discussion

Prior literature on diagnostic reasoning and cognitive properties [82, 168-171] suggests that clinicians employ prototypes (paragon disease presentations) to assess patients and identify relevant genetic diagnoses within WES/WGS results. We designed a novel gene prioritization workflow based upon a prototype-based approach and evaluated it against a workflow that simulated a common phenotype-driven variant/gene prioritization process. Finally, we demonstrated that gene interpretation could be accelerated using the prototype-based workflow by facilitating prototypical thinking.

The workflow performance evaluation revealed that time spent on gene list interpretation was significantly shorter for the prototype-based workflow compared to the symptom-based workflow, with no differences observed in other performance measurements. The qualitative analysis of phenotypic information revealed noticeable differences between phenotype terms that were collected by both workflows. However, the above evaluation was limited in scope to focus on only two genetic diseases which were both well-characterized in the literature. More research is needed to generalize the observed workflow performances over different rare genetic diseases, which have not yet reached the same level of characterization or expert awareness. Furthermore, the prototype-based workflow should be assessed in terms of (a) performance with diseases that present with heterogeneous, overlapping, or novel phenotypes and (b) incorporation of information beyond gene-level. In addition to the scope of the evaluation, the study recruitment was also restricted to medical/biochemical geneticists in order to ensure that participants represented a focused group of users, who would exhibit similar areas of attention/interest when approaching WES/WGS data as well as shared desiderata towards WES/WGS analysis software

[173]. Further research should also consider inclusion of other healthcare professionals involved in clinical WES/WGS interpretation (e.g. genetic counsellors and bioinformaticians) to gain further insights into diverse groups of users and their interaction with different workflows.

Within the scope of this study, the observed difference in time spent on gene interpretation suggested that participants likely engaged in prototypical thinking. The main difference in workflow designs was that the prototype-based workflow explicitly kept track of prototypes. Though tracking, the prototype-based workflow likely reminded participants of their diagnostic reasoning process and encouraged prototypical comparison of genetic diagnoses. This notion was also supported by a secondary finding, where time spent on gene interpretation was shorter for both workflows when analyzing typical scenarios compared to atypical scenarios. This difference agreed with reports in prototype theory research regarding faster recall and recognition of typical members of a category compared to atypical members [174-176]. In sum, it was likely that participants employed some level of prototypical thinking in both workflows while the reasoning process was more efficiently facilitated by the prototype-based workflow.

While the two workflows resulted in equivalent phenotypic information amounts, differences in the content of phenotypic information suggested possible involvement of distinct cognitive processes during phenotype assessment. Phenotype terms collected from symptom-based workflows did not deviate greatly from simulated patient descriptions, whereas those collected from prototype-based workflows did. The deviating terms were relevant concepts but not exact synonyms: for example, cafe-au-lait spot was provided in relation to hypomelanotic macule, and renal angiomyolipoma was provided in relation to renal cysts. This observation could be

explained by a known cognitive tendency towards focusing on deeper structural details when comparing two examples as opposed to considering a single example [171]. However, a quantitative experiment is required to conclusively determine involvement of the aforementioned cognitive tendency during phenotype assessment within different workflows.

Upon observing no difference in user satisfaction, optional comments provided in user satisfaction questionnaires were examined. Specific comments suggested that the study findings should be translated by implementing the best of both worlds. Symptom-based workflow participants pointed out that (1) having to enter each phenotype did not enhance productivity and thus opted to enter only those deemed highly discriminatory; and (2) it was occasionally difficult to code phenotypes impromptu. Meanwhile, prototype-based workflow participants highlighted that (1) a typical feature could not be found in phenotype suggestions (likely due to limitations of disease-phenotype annotations); and (2) some thought it was redundant to refine the phenotype list. The above comments suggested that perceived deficiencies of one workflow could be remedied by the other, and flexibility to use either workflow for phenotype specification seemed most desirable. As such, we translated these findings into PhenoChat, by allowing users to build phenotypic descriptions using either workflow.

3.6 Conclusion

In summary, we explored the utility of augmenting clinical reasoning and cognitive characteristics of experts within computer-assisted gene prioritization. We found that clinicians interpreted genes faster and described phenotypes in relevant, but not synonymous, terminologies using a prototype-based gene prioritization workflow. These findings demonstrate

the potential utility of augmenting experts' analytic and diagnostic workflows during gene prioritization. However, further investigation is warranted to confirm the above findings across diverse rare genetic diseases. WES/WGS informatics methods that recognize how human experts approach gene prioritization and use computers as active partners in knowledge discovery offer promise for overcoming the informatics bottleneck in clinical genome analysis.

Chapter 4: Qualitative evaluation of information visualization practices during applied exome and genome sequence data analyses for rare disease diagnoses

4.1 Synopsis

Information visualization facilitates interpretation of complex data during rare disease exome/genome analyses. Its context of use has been infrequently documented, limiting insight into what types of tasks and data are supported by information visualization and for which tasks design of new visualization methods is needed. As such, we qualitatively evaluated contextual aspects of information visualization practices during applied exome/genome investigations.

An online survey and contextual interviews were conducted with 23 bioinformatics/healthcare experts who conducted clinical exome or genome analyses on a regular basis. Data analysis focused on identifying common analysis/information visualization practices, context of using information visualization, participants' experience with current visualization tools, and participant-suggested requirements for new visualization methods.

Information visualization was frequently employed for visual confirmation of data quality and variant interpretation tasks involving multiple layers of evidence. Participants performed prioritized analyses of phenotype and publicly curated variants. These findings and participant suggestions were translated into recommendations for visual support of common exome/genome analysis tasks.

This study provides an overview of and design recommendations for information visualization that assists rare disease exome/genome analyses. Our findings can inform the development of new visualization in this domain.

4.2 Introduction

Clinically applied exome and genome analyses are transforming the diagnosis of genetic disorders, revealing new disorders and accelerating the resolution of diagnostic odysseys [1, 19]. This disruptive technology requires expert users to consider multiple types and scales of data, including DNA sequence, genes, and patient phenotypes [12, 66]. Each layer of information contributes a critical component for revealing pathogenic DNA sequence variants that underlie patient phenotypes [66]. To better interpret each type of data, a multitude of solutions have been developed with respect to computational variant prioritization methods [12, 70], online interpretation resources [59, 64, 65, 75, 116, 117], and information visualization [108, 109, 113]. Use of these solutions have been frequently reported in research literature involving whole exome sequencing (WES) or whole genome sequencing (WGS) data (e.g. [119, 122]). In addition, much literature has documented the performance, utility, and context of use of computational variant prioritization methods and online interpretation resources (e.g. [104, 177]). However, such documentation has been less frequent for information visualization, likely because it has been implemented as a feature of computational tools and online resources that support data interpretation [59, 116, 117], rather than as an independent method.

In this study, we report a qualitative evaluation of current information visualization practices during exome and genome analyses for rare disease diagnoses. The evaluation focused on

investigating contextual aspects of visualization uses, which can generate a deeper understanding of current analysis practices and inform the design of new visualization methods [124, 125]. We conducted contextual interviews and an online survey with bioinformatics experts who routinely performed clinically applied WES or WGS analyses. The following questions were addressed during the evaluation: (1) What types of information visualization are commonly used during routine WES/WGS analyses? (2) In what context is information visualization used during the analyses? (3) How does information visualization facilitate an analysis task? (4) What types of common analysis tasks should be supported by information visualization?

Data collected from interviews and surveys were analyzed to (a) construct a holistic understanding of current visualization practices and common analysis tasks, and (b) extract user requirements for new visualization. Based on these findings, we formulated design recommendations for augmenting applied exome and genome analyses using information visualization. We anticipate that the findings of this study will provide a resource on visually supported WES/WGS data analyses and design considerations for emerging visualization in this domain.

4.3 Materials and methods

4.3.1 Participants

Between March 2018 and June 2018, we recruited (by email) bioinformatics and healthcare experts who routinely performed WES/WGS data analyses for rare disease or cancer investigations. Prospective participants were identified by consulting university/hospital staff directories, rare disease research networks, and online search engines.

For contextual interviews, six participants from biomedical research and healthcare institutions in Vancouver, Canada were recruited. Four of these participants routinely analyzed patient cases involving a diverse array of rare diseases (e.g. neurogenetic/neurodevelopmental disorders, biochemical disorders). Two of these participants conducted WGS analyses primarily for cancer diagnoses. They were included in the study because they occasionally analyzed hereditary cancer cases, which could give insights into analyses involving germline data. Informed consent was obtained in person for all interview participants.

For the online survey, an invitation email was distributed to (a) 11 prospective participants from academic institutions or healthcare institutions in Canada and (b) through a rare disease research consortium mailing list. In addition, the survey was advertised on social media websites, Twitter and Facebook. In total, 17 survey responses were received. Consent to participate in the survey was implied by submission of responses.

The interview participants and survey invitees were mutually exclusive. Participation in either study component was voluntary. This study was reviewed and approved by the University of British Columbia Research Ethics Board (Certificate: H17-02809).

4.3.2 Contextual interview

Prior to conducting contextual interviews, an interview template was developed to guide the interview process. Questions were created based on (a) common WES/WGS analysis practices that were reported in literature (Appendix C) and (b) four further topics: characteristics of routine

WES/WGS analyses, context of using information visualization during routine analyses, perception of current visualization tools, and suggestions for new visualization. Appendix C provides the template and describes its development process in detail.

The contextual interviews were conducted based on the methods described by Raven and Flanders [178]. All interviews were held during participants' regular work hours and within their work environment, which included their office, laboratory, or (if they worked remotely at a location of their convenience) a meeting room.

The interviews consisted of three parts: introduction (10 - 20 minutes), observation (one - two hours), and follow-up discussion (30 minutes - one hour). During the introduction, participants learned about the interview process and answered a brief set of questions regarding their experience with WES/WGS analyses and the characteristics of their routine analyses. After the introduction, participants were observed as they performed routine exome or genome analyses on their computers. Half of the participants (n = 3) demonstrated analyses of new cases. The other half did not have new analyses to demonstrate, and therefore walked through a previous analysis. The observation focused on capturing participants' use of computational analysis/information visualization tools and the context in which such tools were employed. During case demonstration, participants were occasionally interrupted to clarify what task was being performed. The observation ended when (a) the analysis reached a conclusion, (b) participants declared that they had demonstrated all the steps of their routine analyses, or (c) two hours have passed since the beginning of the observation.

A follow-up discussion was conducted upon completion of observation. First, participants were invited to review and correct the interviewer's observations on their analysis and visualization practices. Next, participants were invited to talk about other data types or visualization tools that were regularly used but were not included in their demonstration. To aid the discussion, participants were presented with a catalogue of common data types and visualization used in WES/WGS analyses (Appendix C.3) and asked to check off applicable items in the catalogue. Finally, the interviews concluded once participants reviewed all observations and catalogue items.

Only follow-up discussions were recorded on video or audio as most participants worked in an open-office environment, which risked recording of non-participating individuals. The participants' work space (if the absence of all non-participating individuals was possible) or a vacant meeting room was used as a recording environment for follow-up discussions.

4.3.3 Online survey

The online survey questionnaire consisted of nine items adapted from the interview template, and covered the same four topics as the interview (characteristics of routine WES/WGS analyses, context of using information visualization during routine analyses, perception of current visualization tools, and suggestions for new visualization). Three items in the questionnaire contained follow-up questions that dynamically appeared based on the answers to the preceding question. The online survey interface was developed and hosted using Qualtrics software provided by the University of British Columbia (Vancouver, Canada). The complete questionnaire is provided in Appendix C.4.

4.3.4 Data analysis

Transcripts were generated from audio/video recordings of the contextual interviews. The interview notes and transcripts were analyzed by constructing work flow diagrams [178] which summarized common analysis tasks, their goals, and types of data and visualization required by each task.

The online survey responses were analyzed to identify frequently used common analysis tasks, types of data and information visualization required by each task, data/visualization's context of use, participants' shared experience with current visualization tools, and participants' suggestions for new visualization.

The interview and survey data were compared with respect to the types of data and visualization that were commonly captured by both evaluations. This comparison was used to construct an overview of common analysis and visualization practices.

4.4 Results

4.4.1 Participant and analysis characteristics

Figure 4.1A-4.1C summarize the characteristics of 23 participants (n = 6 for contextual interview; n = 17 for online survey) and their routine WES/WGS analyses. A majority of participants (14 out of 23) had experience with more than 50 exome or genome analysis cases (Figure 4.1A). 17 participants routinely analyzed WES data and 15 analyzed WGS data (Figure 4.1B). WGS analyses performed by three contextual interview participants (excluding

participants who analyzed cancer cases) were restricted to variants affecting coding regions of the genome. It was not known if the same restriction was applied to WGS analyses performed by online survey participants. Most routine analyses (19 out of 23) were performed in a research setting (Figure 4.1C).

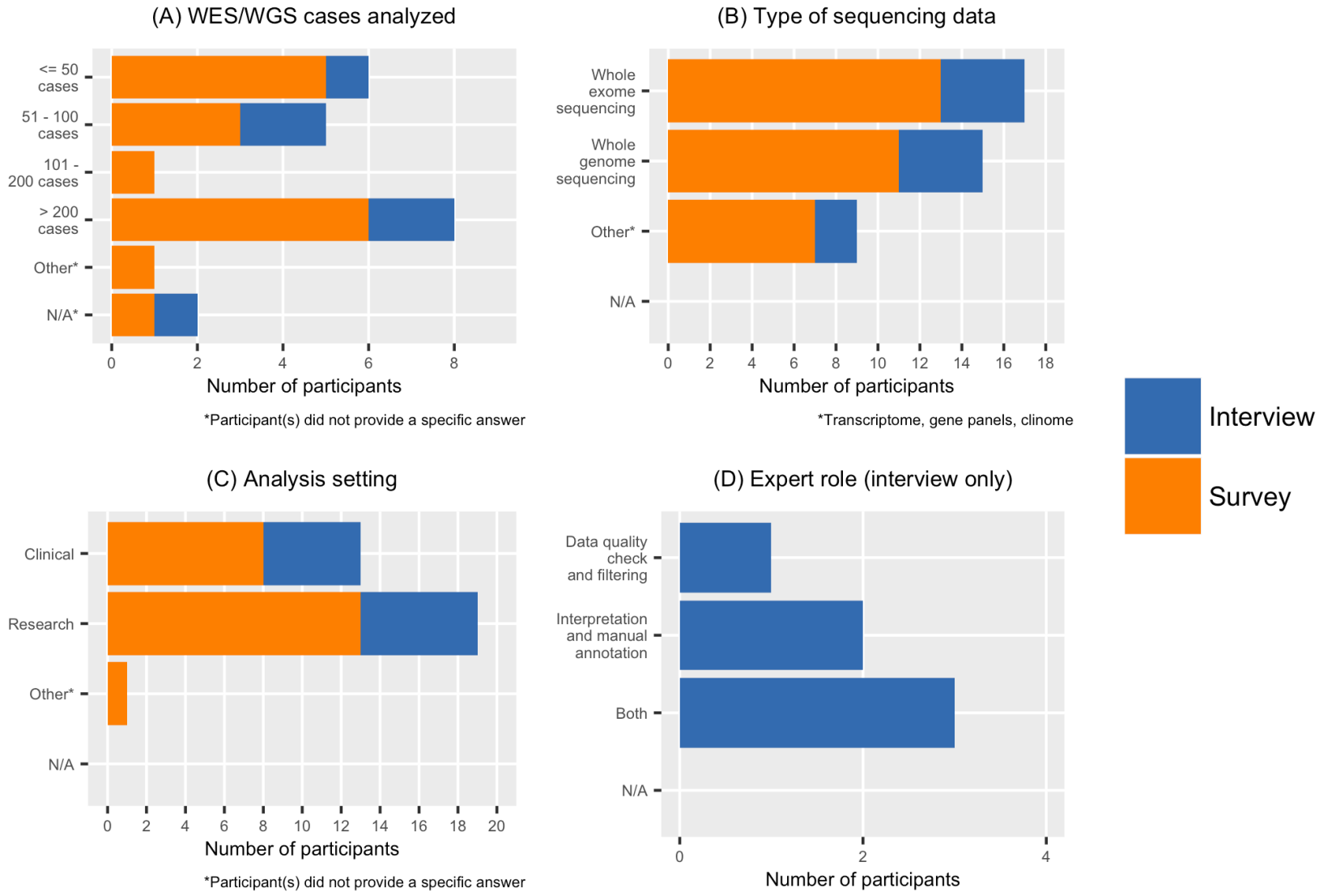


Figure 4.1 Participant and routine WES/WGS analysis characteristics.

1A-C shows characteristics of 23 participants (n = 6 for contextual interview in blue; n = 17 for online survey in orange). D shows characteristics of contextual interview participants only. (A) Number of cases involving WES/WGS data that participants have analyzed to date. (B) Types of sequencing data used in routine analyses. (C) Setting in which routine analyses were performed. (D) Role of interview participants. Three interview participants reported that their analyses were performed by two different experts, each having a separate role: one was focused on checking data quality and filtering variants. The other was focused on interpretation and manual annotation for clinical expert review. N/A = Not Applicable.

Figure 4.1D outlines additional characteristics of interview participants (n = 6). Three interview participants indicated that their routine analyses were divided into two parts, and each part was performed by different experts. These participants were assigned to perform only one of the roles as follows: one role focused on checking data quality and filtering variants, while the other focused on interpretation and manual annotation for clinical expert review.

4.4.2 Common analysis and information visualization practices

Common analysis and information visualization practices were identified and are illustrated in Figure 4.2. Table 4.1 outlines a list of commonly used analysis tools, with information visualization tools highlighted, that were captured in this study. Screenshots of information visualization tools are provided in Appendix C.5. A common goal of participants' analyses was to identify and manually annotate a filtered list of potentially pathogenic variants for clinical expert interpretation. Participants' routine analyses were performed generally in two tiers, with the first-tier focusing on known pathogenic variants or variants in known disease genes, and the second-tier focusing on the remaining variants that were detected exome-wide or genome-wide. For rare disease cases, routine WES/WGS analyses were restricted to coding variants. Regardless of tiers, almost all interview participants (n = 5) visually inspected sequencing read alignment before beginning the clinical interpretation of any variants. Unanimously agreed, the analysts indicated that the read visualization step was intended to detect poor quality variants or data abnormalities that passed through quality control (QC) filters of automated data processing pipelines. While visual inspection was not the only QC mechanism, these participants preferred a visual confirmation in addition to checking automatically derived QC-related annotations. One interview participant indicated that not every variant was visually inspected. This participant

explained that visual inspection was performed only when poor quality was noted by custom QC annotations.

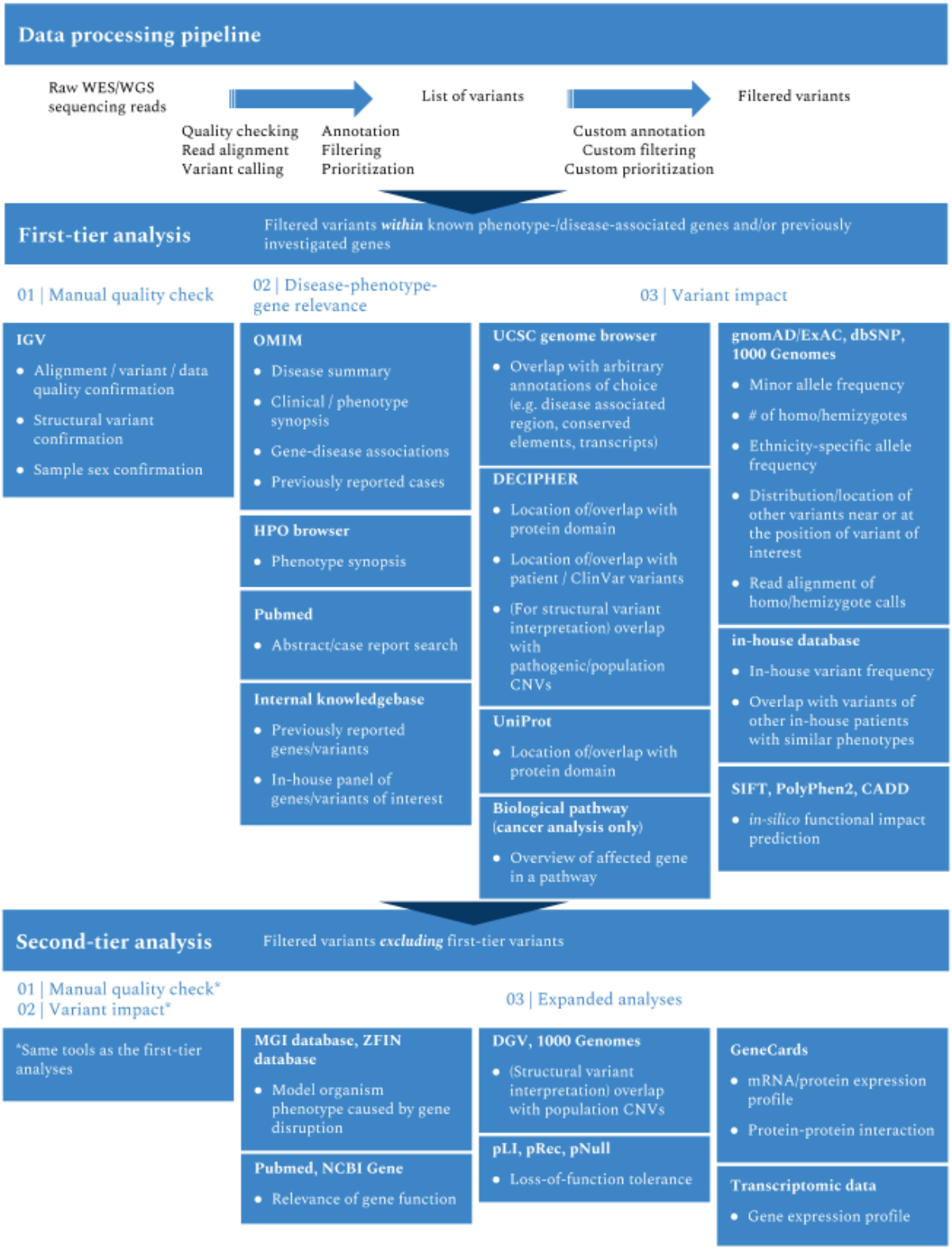


Figure 4.2 A composite workflow diagram of routine WES/WGS analyses.

This figure summarizes common analysis/information visualization practice identified in this study. In addition, this figure outlines commonly used analysis tools and their context of use as captured by contextual interviews and survey responses. First, participants processed raw sequencing reads through automatic data processing pipelines. Next, semi-automated pipelines produced a list of filtered variants and participants analyzed each variant. For rare disease investigations, WES/WGS analyses were restricted to variants predicted to affect coding regions only. The analyses were generally two-tiered. The first-tier focused on variants within known phenotype-associated/disease-associated genes or within previously investigated genes. The second-tier focused on variants that were detected exome-wide or genome-wide.

Types of analysis tools		Specific analysis tools captured in this study	Types of visualization (provided by specific analysis tools) captured in this study
Tools for assessing disease / phenotype / gene / protein level information	Disease databases (e.g. known gene-disease association, overlapping/similar phenotypes)	Online Mendelian Inheritance in Man (OMIM) [64]	-
		Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER) [117]	Genome browser Phenotype browser
	Phenotype resources	Human Phenotype Ontology (HPO) [75]	-
		Mouse Genome Informatics (MGI) database [179]	-
		Zebrafish Information Network (ZFIN) [180]	-
	Gene / protein resources	UniProt [116]	Feature viewer
		National Center for Biotechnology Information (NCBI) Gene [115]	Graphical sequence viewer
		GeneCards [181]	mRNA/protein expression plot Protein-protein interaction diagram
	Literature search	PubMed [182]	-
	Biological pathway/network analysis (e.g. interaction with known disease-associated gene)	GeneMANIA [112]	Pathway visualization Protein-protein interaction visualization

Tools for assessing variant impact	Disease-focused variation databases	ClinVar [65]	-
	Population variation databases	Single Nucleotide Polymorphism database (dbSNP) [114]	Graphical sequence viewer
		1000 Genomes [51]	-
		Genome Aggregation Database (gnomAD) [59] Exome Aggregation Consortium (ExAC) [59]	Illustrated gene summary Read data browser
		Database of Genomic Variants (DGV) [183]	-
	in-silico functional prediction tools	SIFT [68]	-
		PolyPhen2 [60]	-
		Combined Annotation Dependent Depletion (CADD) [61]	-
	Loss-of-function tolerance prediction	pLI/pRec/pNull [59]	-
	Splice-site prediction	Human Splicing Finder [63]	-
	Nucleotide conservation	PhyloP [62]	-
	Commercial variant analysis tools	Alamut [118]	Arbitrary/custom annotation visualization
Tools dedicated for information visualization	Genomic data visualization	Integrative Genomics Viewer (IGV) [108]	Read alignment visualization
	Genome browser	University of California Santa Cruz (UCSC) Genome Browser [109]	Arbitrary/custom annotation visualization (e.g. conserved elements, disease-associated regions, transcripts, ClinVar variants)

	Protein structure visualization	Chimera [110]	3D structure visualization
	Phenotype-driven visual prioritization tools	OMIM Explorer [70]	-
	Phenotype comparison visualization	PhenoBlocks [113]	-
	Custom R visualization	R packages or scripts were not specified by participants	Visualization of arbitrary data (e.g. relatedness, ancestry, data quality metrics)

Table 4.1 A list of analysis/information visualization tools commonly used by participants.

After visual inspection of read alignment, visualization practices changed across the tiers of analyses. The first-tier analyses (Figure 4.2) were phenotype-driven and focused on (a) a panel of genes known to be associated with phenotypes of interest, (b) genes known to be associated with genetic diseases, or (c) an internal knowledgebase of previously reported/investigated genes. A common analysis task within this tier was the examination of minor allele frequencies (MAF) and the number of homozygotes/hemizygotes appearing in population variation databases (Table 4.1). For this task, participants typically checked automatic annotations in customized variant lists. However, when a variant was on the edge of consideration (e.g. MAF was not very low but phenotype matched/one or two homozygotes/hemizygotes detected), participants consulted information visualization that was provided on the websites of population variation databases (Table 4.1). For example, a read browser on the gnomAD [59] or ExAC [59] website was used to check the exome/genome sequencing reads from homozygote/hemizygote individuals and confirm the quality of homozygous/hemizygous variant calls (Figure 4.2). Furthermore, an illustrated gene summary in gnomAD [59] or ExAC [59] was used to assess a region containing a variant of interest with respect to other nearby variants and determine if the variant of interest was located within a mutational hotspot or constrained region (Figure 4.2).

In addition to the above, first-tier analyses involved protein domain visualization, biological pathway visualization (cancer-related analyses only), and visualized distribution of variants curated in disease-focused variation databases (Table 4.1).

The second-tier analyses (Figure 4.2) were variant-driven. Participants first focused on gathering evidence to decide whether to invest time and interpretation on a variant. Such evidence included

extremely low MAF/absence in population databases, agreement of pathogenic prediction by multiple *in-silico* functional prediction tools, and abnormal expression level detected in transcriptomic data. When a variant was selected for deeper investigation, participants used diverse resources and visualization as outlined in Figure 4.2 to gather further evidence for biological relevance to patient cases.

Overall, participants showed two commonalities in their analysis/visualization practices. First, they consulted genome browsers when assessing multiple layers of evidence. For example, participants used the genome browser within DECIPHER [117] when evaluating the same variant in a regional context, such as its overlap with known pathogenic/benign variants curated in ClinVar [65], distribution/location of variants curated in gnomAD [59], and known pathogenic structural variants affecting a single gene.

Second, participants prioritized interpretation of information that had the highest perceived diagnostic value. For example, three participants preferred to begin their first-tier analyses by assessing phenotypic relevance. This task involved examining clinical synopses in OMIM [64], phenotype-disease associations in the HPO browser [75], or literature. To maximize efficiency, they employed specific decision-making rules, such as looking for at least one matching phenotype in disease descriptions or particular keywords in abstracts before deciding to pursue a variant. Of these participants, two noted that if multiple phenotypic features matched patients, they examined patient photographs published in literature. This was to compare clinician-provided patient description with disease-defining phenotypes described in literature, and one

comment illustrated this process as follows: *"if it (patient phenotype) was very obvious, they (clinicians) must have noted it."*

4.4.3 Experience with other information visualization

Online survey participants (n = 17) were asked to indicate (a) if they had actively searched for visualization tools, (b) what tools they had tried but did not use in their routine analyses, and (c) why they did not use those tools. 12 participants reported their experience (Figure 4.3A) with 11 types of visualization as outlined in Figure 4.3B. A common reason for not using the above visualization tools was lower diagnostic value gained compared to the effort spent, preference for alternative tools, or no available funding for subscription to commercial tools (Figure 4.3C). Three participants reported that they had never tested or searched for other visualization tools because they had never felt the need to do so (Figure 4.3A).

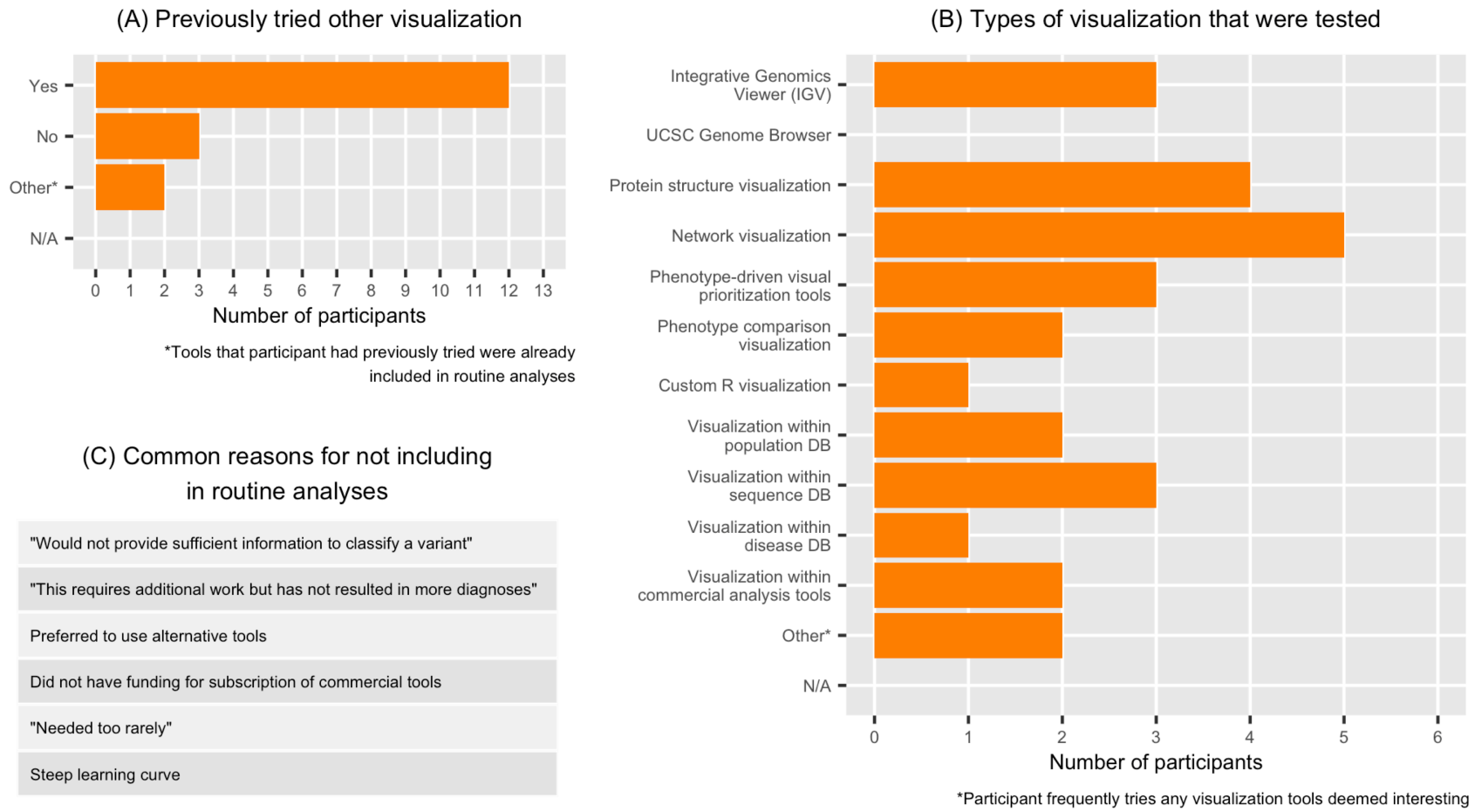


Figure 4.3 Participants' experience with currently available information visualization.

Only the responses from online survey participants (n = 17) were considered for A-C as contextual interview participants were not explicitly asked regarding their experience with currently available information visualization. (A) Number of survey participants who had/had not actively tried or looked for visualization tools other than those used during routine analyses. (B) Types of visualization that survey participants had tried in the past but not used during routine analyses. (C) Common reasons for not using the selected visualization during routine analyses. N/A = Not Applicable.

Interview participants (n = 6) were not explicitly asked the same questions (a-c) as the survey participants, but three participants shared their experience with protein structure visualizations and visualization features within commercial variant analysis tools. These participants indicated that they had tried these tools but no longer used them, expressing the same rationale as the survey participants.

4.4.4 Suggestions for new information visualization

Online survey participants (n = 17) were asked to indicate which commonly used types of data would be helpful or not helpful to visualize for their analyses. A strong consensus was expressed that visualization methods for sequencing quality (n = 12), variant quality (n = 13), and coverage analysis (n = 13) would be helpful (Figure 4.4A) for (a) reducing the effort spent on examining technical details, as well as for (b) interpreting the regional context, such as coverage at the gene level and regional coverage with respect to overall genome-wide coverage (Table C1, Appendix C). Slightly less than the majority of survey participants indicated that it would not be helpful to visualize functional annotation (n = 8), variant frequency in population databases (n = 7), and *in-silico* functional prediction (n = 7) (Figure 4.4A). A common reason expressed for finding visualization unhelpful for the above data types was that textual/numeric information was sufficient for interpretation (Table C1, Appendix C).

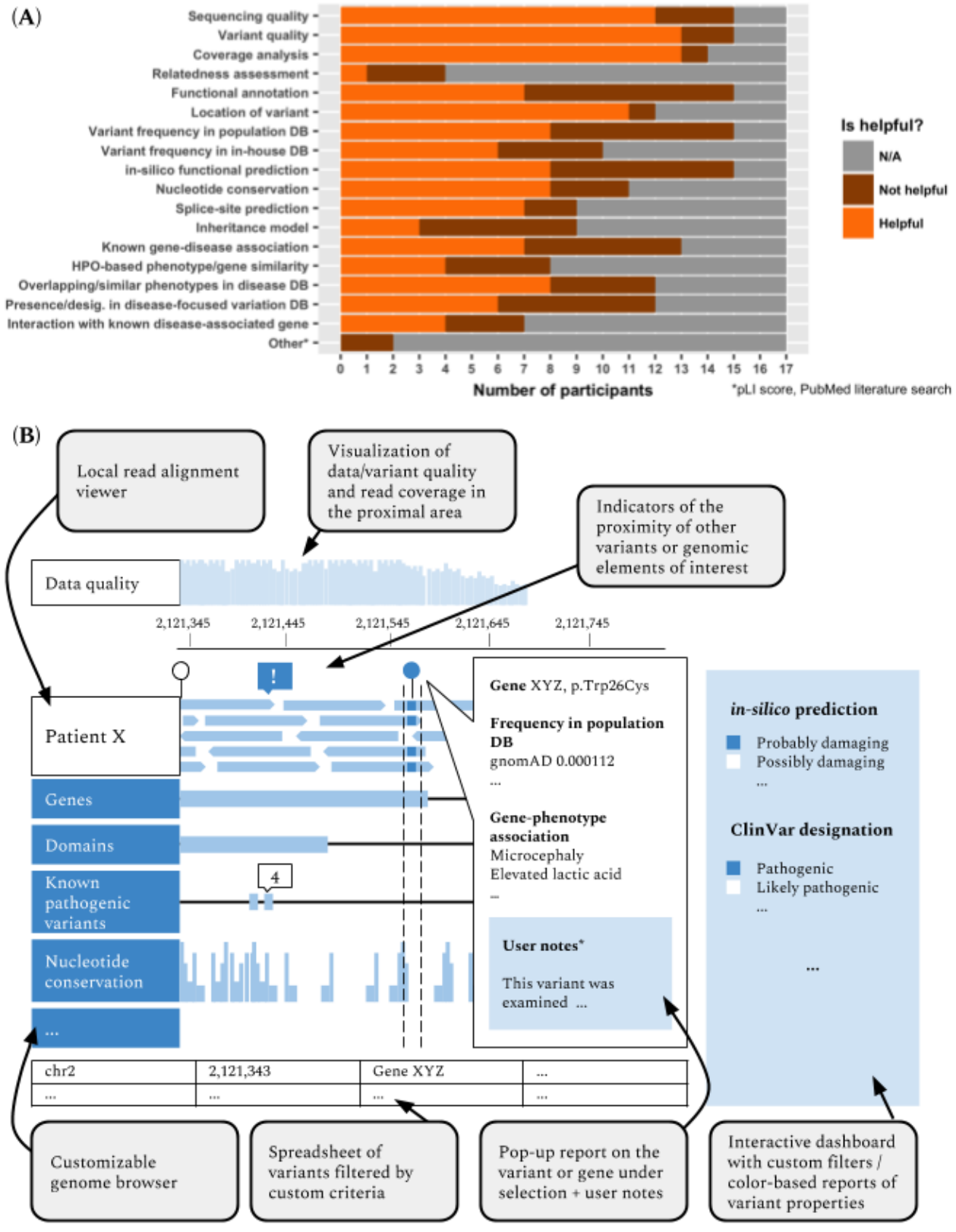


Figure 4.4 Design suggestions for emerging information visualization.

(A) Opinion of online survey participants (n = 17) regarding whether the given data types would be helpful or not helpful to visualize. (B) An illustration of design recommendations for visually supporting WES/WGS analyses based on participant suggestions and observations. Participants desired information visualization which integrated the properties of a customizable genome browser, a read alignment viewer, a spreadsheet of filtered variants, an interactive dashboard with customizable variant filters (which generate color-based reports of variant properties), and a pop-up report about the variant/gene in focus. The pop-up report would include information such as phenotype associations and variant frequency across populations. Suggestions were made regarding visualization of sequencing data quality, visualization of read coverage in the proximal area, indicators of proximity of other variants/genomic elements, and functionality that allows attachment of notes on variant in view.

Interview participants (n = 6) were not explicitly asked to provide suggestions for new information visualization. However, during the observation of common analysis/visualization practices, design recommendations were expressed for visually supporting WES/WGS analyses (Table C2, Appendix C). These recommendations were combined with the feedback from online survey participants to construct a representative model (Figure 4.4B) which outlines key visualization desiderata as follows.

The participants focused primarily on visualization aspects related to candidate variant identification (with less priority given to the interpretive methods for prioritization of novel variants). As illustrated in Figure 4.4B, participants sought an integrated view combining the properties of a customizable genome browser (such as UCSC Genome Browser [109]), a local read alignment viewer (such as provided by IGV [108]), a spreadsheet of filtered variants, and an interactive dashboard with customizable variant filters that generate color-based reports of variant properties. There was a desire for visualization of data/variant quality and read coverage in the proximal area, as well as indicators of the proximity of other variants or genomic elements

of interest. The idealized view (Figure 4.4B) includes a pop-up report about the variant or gene under selection, which would provide information such as associated phenotypes and variant frequency across populations. As all interview participants maintained notes on the variants, a note taking function is also incorporated into the idealized view.

4.5 Discussion

In this report, we identified information visualization practices during routine clinical exome and genome investigations, by conducting contextual interviews and an online survey with bioinformatics and healthcare experts. A comprehensive overview of information visualization practices showed that routine analyses were generally two-tiered: first-tier analyses focused on the assessment of phenotypic similarity and variant frequency/impact in the context of a protein or a population; and second-tier analyses focused on the assessment of diverse evidence (e.g. variant, gene, model organism phenotypes) for biological relevance. Overall, the overview also revealed participants' tendencies to: (a) visually confirm sequencing read alignment, (b) use a genome browser when assessing multiple levels of evidence, and (c) prioritize assessment of information with a high perceived diagnostic value. Next, we evaluated participants' experience with currently available visualization tools and discovered that use of visualization depended on the tradeoff between time/cost and perceived value of evidence that could be added to the analyses. Finally, we extracted participants' suggestions and generated our own idealized representation of their suggestions to inform the creation of new information visualization in this domain.

Currently, variant interpretation is an expert-driven process assisted by computational tools and resources [1, 12, 19, 66], creating a bottleneck within clinical WES/WGS analyses [13], which will likely persist as the process is complex and has a direct impact on patients [12]. Considering this, an acceleration of the interpretation process is beneficial and can be achieved by methods that facilitate experts' understanding of complex data, such as information visualization. Our evaluation demonstrates that information visualization is currently a vital part of WES/WGS analyses, supporting core data interpretation tasks. For example, visual inspections of read alignment data assist in the elimination of poor quality variants as well as identification of structural variants. In addition, genome browsers enable visual integration of heterogeneous evidence (e.g. protein domain location, distribution of known pathogenic variants) allowing for efficient interpretation of variant impact.

Furthermore, our findings also suggest that there are diverse aspects within WES/WGS analyses that can be supported visually. Participant suggestions revealed potential areas of improvement such as visualization of sequencing quality, variant quality, and coverage analysis (Figure 4.4A). The recommendations (Figure 4.4B) highlighted potential visual features that can augment common analysis practices. For instance, initial expert assessment regarding whether or not to pursue further variant interpretation can be expedited by color-coding variant annotations based on custom decision rules (Figure 4.4B). Based on user workflows, we suggest simultaneous display of information in tabular format and in a genome browser format to ease the transition between variant-level and gene-level interpretation (Figure 4.4B).

While this study attempted to capture current visualization practices within rare disease WES/WGS analyses, we acknowledge that this is a quickly evolving domain. New analysis tools emerge constantly, and the transition from WES to WGS has created demand for new types of visualization methods for analysis of non-coding genomic regions [184] as well as for variants beyond the levels of single nucleotide, indels, or splice sites [185]. We believe that a continuous rare disease community-wide effort to evaluate visualization practices is necessary to ensure dissemination of the latest practices and fulfillment of user requirements by new visualization. In addition to the difficulties in capturing current expert practices, this study was also limited in scope due to its small sample size. In order to address the recruitment/logistic limitations of conducting in-person interviews and to improve the rigour of qualitative research by capturing more comprehensive expert opinions [186], two data collection methods (contextual interview and online survey) were performed within this study. However, the online survey participation was limited, and thus, more research with a larger sample is needed to generalize the study findings to common WES/WGS analyses that are performed in clinical genetics. Therefore, the online user-survey will remain open for 1 year following publication (https://ubc.ca1.qualtrics.com/jfe/form/SV_b9kI2jmaAowP2C1), with summary data presented in a report that is available at https://ubc.ca1.qualtrics.com/reports/public/dWJjLTViMzY5ZTllNjhiNzZkMDAwZDAxMjkxNy1VUl9lZmRKYkhPVnJqUWd3UEg=.

In summary, information visualization facilitates understanding of complex data during rare disease exome and genome analyses. The findings presented herein provide an overview of current information visualization practices and recommendations for emerging visualization

tools. More data and annotations are expected to be incorporated into applied WES/WGS analyses, creating an increasing need for information visualization. Efforts to address this need can benefit from not only the type of evaluation demonstrated in this study, but also other types of evaluations in diverse scopes and scales, such as assessment of visualization design and assessment of tasks that visualizations aim to support [124, 125]. We hope our study catalyzes further interest in the evaluation of information visualization that assists WES/WGS investigations, guiding the development of new visual analysis tools that can accelerate expert interpretation.

Chapter 5: Conclusion

"The $1,000 genome, the $100,000 analysis" [187] has been a popular musing about the current state of clinical genome sequence data analyses. The cost of genome sequencing has reduced significantly during recent years [188, 189], but the cost of analysis has not [190, 191]. According to a recent study [13], the analysis cost is predicted to remain over $5,000 during the next ten years. This suggests that clinical exome/genome analyses will remain an expert-driven process coupled with computer assistance for the near future. Recognizing the current status, this thesis explored augmentation of an expert's ability through HCI as a method for expedited exome and genome analyses.

The exploration began with close observations of two main stakeholders in clinical genomics, bioinformatics experts and clinical geneticists, within my collaboration with the TIDEX project. During this collaboration, I performed applied WES/WGS analyses for patients with biochemical diseases [1, 4, 5], acquiring practical knowledge of this domain. This experience led me to discover the potential to accelerate analyses by complementing experts' abilities within a computer-assisted diagnostic process. This idea became more apparent during the collaboration described in Chapter 2, where I developed an online resource that provides an expert knowledgebase and a diagnosis supporting system for IEMs. Implementation and validation of this resource demonstrated how an assistive system could complement an expert's workflow and what experts required of such a system. Motivated by these observations, I decided to explore two HCI elements which could potentially expedite variant interpretation: cognitive process and information visualization.

Chapter 3 presented the design and evaluation of a gene prioritization workflow that aimed to augment clinicians' prototypical thinking process, where they use classic representation of genetic diseases to assess patients and to interpret WES/WGS results. The evaluation demonstrated that clinicians could identify genetic diagnoses faster using the novel workflow compared to a common computer-assisted variant prioritization workflow, suggesting utility in aligning the prioritization workflow with experts' cognitive process.

In Chapter 4, current information visualization practices during WES/WGS analyses for rare genetic disease diagnoses were assessed through contextual interviews and an online survey. Based on this assessment, a comprehensive overview of common WES/WGS analysis and visualization practices was constructed, summarizing frequently used data types, visualization tools, common context in which experts employed visualizations, and user suggestions for new visualization. These findings were then translated into design recommendations to inform subsequent development of visualization in this domain.

Taking these chapters together, this thesis narrates a journey of designing and evaluating novel HCI-based approaches for accelerating WES/WGS analyses. Its findings highlight the utility of empowering healthcare experts through efficient HCI to rapidly diagnose patients using genome sequence data. In the following sections, I will discuss emerging areas within clinical genomics and healthcare in which this thesis work will continue. Figure 5.1 also visually summarizes the future directions of this thesis.

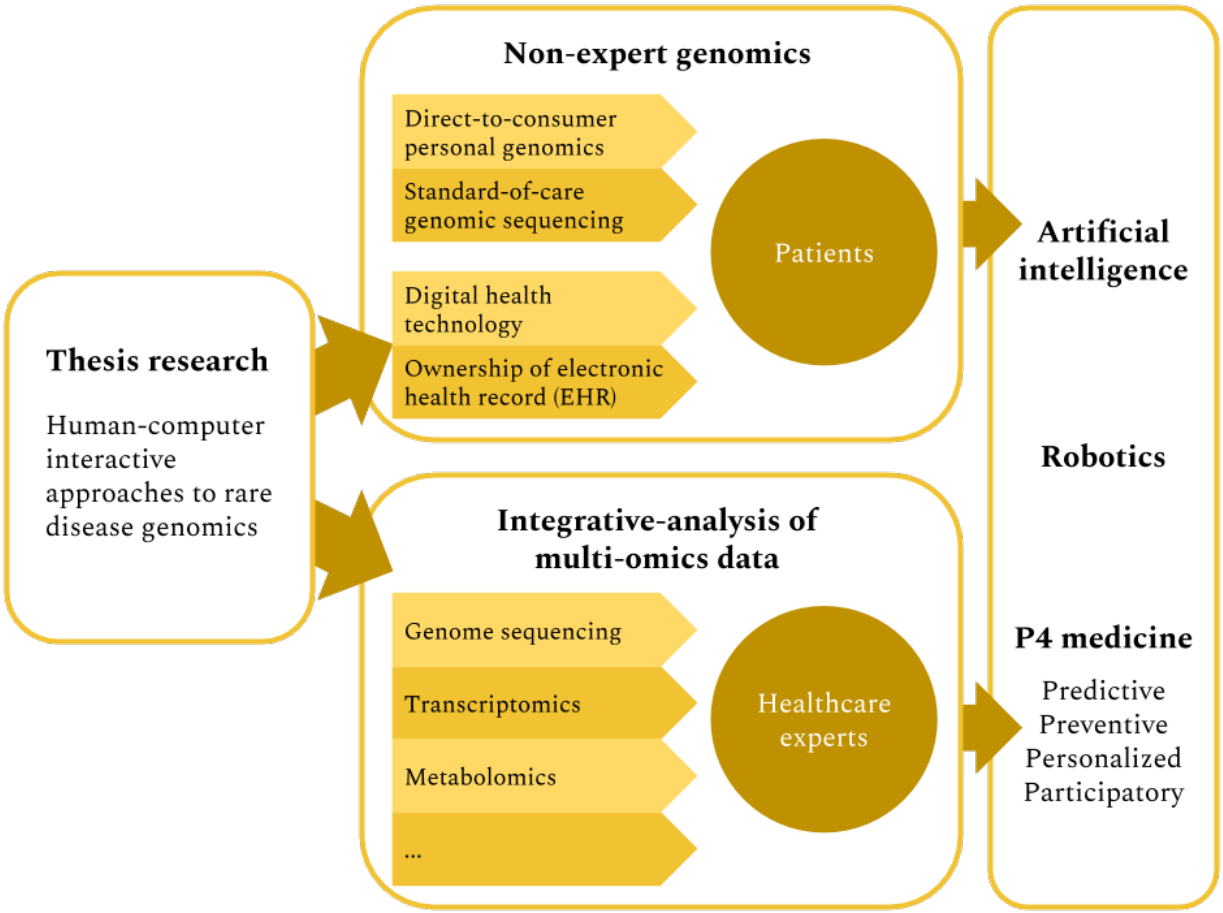


Figure 5.1 A visual summary of future directions.

The figure illustrates how main stakeholders of healthcare will be affected by emerging health innovations, and how this thesis work will continue within these innovations. In the short term, this thesis will likely support (a) exploration of personal genomic data by non-experts as current developments in digital health lead to ownership of electronic health records and personal genome data by patients; as well as (b) collaboration of diverse healthcare experts in an integrative analysis of multiple -omics data. In ten years, the aforementioned developments will transform into a paradigm shift towards incorporation of robotics, artificial intelligence, and P4 (predictive, preventive, personalized, participatory) medicine.

5.1 Future directions

5.1.1 HCI for non-expert stakeholders in clinical genomics

This thesis explored interactions between (a) clinical experts and computers (Chapter 3) and (b) bioinformatics/healthcare experts (e.g. cytogeneticists) (Chapter 4) within the domain of clinical genomics. These interactions represent traditional clinical settings where genomic investigations are initiated by experts. Another driving force in genomics is non-experts who access genome sequencing technology via direct-to-consumer personal genomics services (e.g. 23andMe [192], AncestryDNA [193], MyHeritageDNA [194], Helix [195], Color Genomics [196]). With increasing popularity of these services, HCI research has recently been established in this domain, focusing on effective facilitation of non-expert exploration of personal genomic data through online interactive reports [197, 198]. Currently, the scope of these reports is diverse, spanning from genealogy to general health information [199, 200]. This may change in the near future as non-experts desire more health-focused exploration of their genomic data, motivated by advances in digital medicine that (a) enable access to electronic health records (EHR) for patients and (b) harness patient-provided health data [201, 202]. For example, Apple has recently announced a feature that allows browsing of an EHR within their Health application for their mobile devices [203]. This application collects and exploits user-generated health data from digital health devices such as activity trackers which can connect to mobile devices [204]. Another example towards this direction is a recent initiative to incorporate layperson medical vocabulary into HPO, which will facilitate understanding of genomic investigations and encourage active participation in research by rare disease patient communities [205].

Expected outcomes of the above advances are an active ownership of complete EHR by patients and an ability to share this data with digital health technologies both from the patient and from their healthcare providers [201, 206]. As clinical genomics become the standard of care, personal genomic data will eventually be incorporated into EHR [207, 208]. Following this shift, health technologies that assist with non-expert exploration of their genomic data will be increasingly sought after. This thesis work can inform the design of such technologies and continue within the non-expert domain by trickling down the knowledge of expert-level genomic analysis to the non-expert audience.

5.1.2 HCI in systems medicine

Another imminent shift in clinical genetics is systems medicine, which focuses on holistic investigations involving integration of multi-omics data, such as transcriptomics, metabolomics, lipidomics, and glycomics [158]. The ability to examine multiple levels of biology promises not only improved interpretation of genomic data, but also comprehensive understanding of disease mechanisms, more efficient diagnoses, as well as identification of treatment strategies [209]. A body of literature focusing on approaches to integrate different -omics technologies has been growing in recent years [210-214]. As -omics analyses consider more data, each integrative analysis will require collaboration between multi-disciplinary experts and clinicians to interpret heterogeneous biological information [215]. In this setting, computational assistance will be vital not only for the analysis of big data, but also for translating outcomes of the analysis in a manner understood by all collaborating experts [216], and most importantly, clinicians who oversee patient care [217]. Development of such supportive tools may benefit from adoption of HCI-focused methodologies, whose potential to accelerate tasks that require expert engagement has

been demonstrated by this thesis. The methods presented herein can provide a starting point for collaborative tools that require channeling of guidance from multiple experts into -omics analyses or visually communicate analysis results to diverse experts.

5.1.3 Healthcare in the next ten years

With the aforementioned developments in digital health and integrative -omics, what will future healthcare look like?

Recent media coverages have focused on a possible (or partial) replacement of doctors by machines, powered by artificial intelligence (AI) and robotics [218-220]. Adoption of these technologies is a possibility, as they have the potential to alleviate existing problems (e.g. low doctor-to-patient ratio in developing countries) [221] and impending problems (e.g. increasing demand for healthcare by rapidly aging population) [222, 223]. However, incorporation of AI and robotics into healthcare requires a thorough discussion of ethics and regulating policies [224] due to its impact on trust among all stakeholders [225]. As such, while a radical conversion to these automating technologies may be less likely [226], the next ten years will involve (a) an active public discourse on ethical utilization of these technologies [227], and (b) their steady implementation in a manner that garners the trust of patients and healthcare experts [228].

Another major change in healthcare will likely arise from P4 medicine: a paradigm for predicting the emergence of a disease and preventing it via personalized care and active participation of patients [229]. This paradigm can enable proactive management and timely treatment of genetic conditions such as IEMs [217], as well as common or chronic conditions (e.g. cardiovascular

diseases [230]). Furthermore, it can empower those predisposed to diseases, such as cancer or diabetes, with individualized advice on their lifestyle choices [231, 232]. In recent years, the concept of P4 medicine has been increasingly advocated within healthcare as advances in digital health and -omics technologies have enabled rapid generation of health data for the masses [232]. This trend will likely lead to gradual materialization of P4 medicine during the next ten years and beyond, with its success hinging on (a) continuous efforts to raise awareness among all healthcare stakeholders regarding principles and impact of P4 medicine, (b) establishment of ethical standards, regulating policies, and technical infrastructures for managing and utilizing personal health data, and (c) societal agreement on access to and payment of healthcare services based on P4 medicine [233, 234].

5.2 Final remarks

Life has become digital. As of 2017, 54% of the world population has access to the internet [235]. Recent advances in smart mobile devices, internet of things, and AI have transformed the computer into an essential medium for daily activities. Healthcare is now embracing these digital innovations [236]. Simultaneously, clinical genomics is being incorporated into diagnostic approaches and treatment selection across medical disciplines [234]. Together, digital technologies and genomics are driving a rapid shift towards personalized medicine. Innovators of these technologies, however, should remember that healthcare is one area where one cannot simply "move fast and break things" according to the mantra of the start-up age. Inventions in this area not only impact matters of life and death, but also affect the interpersonal trust that has been established throughout the history of modern medicine [237].

This thesis describes the development of HCI-based methodologies that empower healthcare experts to rapidly diagnose patients using genomic data, demonstrating a case study of technology that accelerates healthcare practice with minimal disruption to the trust that binds patients and healthcare experts. As emerging innovators attempt to actualize their visions for healthcare, this work will provide food-for-thought or a starting point for their trailblazing technology.

Bibliography

1. Tarailo-Graovac M, Shyr C, Ross CJ, et al. Exome Sequencing and the Management of Neurometabolic Disorders. N Engl J Med. 2016;374(23):2246-2255.
2. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nature Protocols. 2015;10(10):1556-1566.
3. Pal LR, Kundu K, Yin Y, Moult J. CAGI4 SickKids clinical genomes challenge: A pipeline for identifying pathogenic variants: PAL et al. Human Mutation. 2017;38(9):1169-1181.
4. Lee JJY, van Karnebeek CDM, Drögemoller B, et al. Further Validation of the SIGMAR1 c.151+1G>T Mutation as Cause of Distal Hereditary Motor Neuropathy. Child Neurol Open. 2016;3:2329048X16669912.
5. Schlingmann KP, Bandulik S, Mammen C, et al. Hypomagnesemia, refractory seizures, intellectual disability and de novo mutations in ATP1A1. (Submitted).
6. TIDE BC: Treatable Intellectual Disability Endeavor in B.C. http://www.tidebc.org/. Accessed June 29, 2018.
7. Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: Patient Phenotyping Software for Clinical and Research Use. Human Mutation. 2013;34(8):1057-1065.
8. Robinson PN, Köhler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014;24(2):340-348.
9. Li M-X, Gui H-S, Kwan JSH, Bao S-Y, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research. 2012;40(7):e53-e53.
10. 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 Med. 2014;6(10):74.
11. Miller CA, Qiao Y, DiSera T, D'Astous B, Marth GT. bam.iobio: a web-based, real-time, sequence alignment file inspector. Nat Methods. 2014;11(12):1189.
12. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nature Reviews Genetics. 2017;18(10):599-612.
13. Weymann D, Laskin J, Roscoe R, et al. The cost and cost trajectory of whole-genome analysis guiding treatment of patients with advanced cancers. Molecular Genetics & Genomic Medicine. 2017;5(3):251-260.
14. Dragojlovic N, Elliott AM, Adam S, et al. The cost and diagnostic yield of exome sequencing for children with suspected genetic disorders: a benchmarking study. GENETICS in MEDICINE. January 2018.
15. Plöthner M, Frank M, von der Schulenburg J-MG. Cost analysis of whole genome sequencing in German clinical practice. The European Journal of Health Economics. 2017;18(5):623-633.
16. Smith HS, Swint JM, Lalani SR, et al. Clinical Application of Genome and Exome Sequencing as a Diagnostic Tool for Pediatric Patients: a Scoping Review of the Literature. Genet Med. May 2018.

17. Tan TY, Dillon OJ, Stark Z, et al. Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions. JAMA Pediatrics. 2017;171(9):855.
18. Bodian DL, Klein E, Iyer RK, et al. Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. Genetics in Medicine. 2016;18(3):221-230.
19. Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genetics in Medicine. 2018;20(4):435-443.
20. Mwenifumbo JC, Marra MA. Cancer genome-sequencing study design. Nature Reviews Genetics. 2013;14(5):321-332.
21. REGULATION (EC) No 141/2000 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 1999 on Orphan Medicinal Products.
22. Orphanet. Prevalence and incidence of rare diseases: Bibliographic data. http://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf. Published June 2017. Accessed June 29, 2018.
23. 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.
24. EURORDIS - Rare Diseases Europe. “Rare Diseases: understanding this Public Health Priority.” https://www.eurordis.org/IMG/pdf/princeps_document-EN.pdf. Published November 2005. Accessed June 29, 2018.
25. van Karnebeek CDM, Stockler S. Treatable inborn errors of metabolism causing intellectual disability: A systematic literature review. Molecular Genetics and Metabolism. 2012;105(3):368-381.
26. Parkman R. The application of bone marrow transplantation to the treatment of genetic diseases. Science. 1986;232(4756):1373-1378.
27. Loeber JG. Neonatal screening in Europe; the situation in 2004. Journal of Inherited Metabolic Disease. 2007;30(4):430-438.
28. Blau N, van Spronsen FJ, Levy HL. Phenylketonuria. Lancet. 2010;376(9750):1417-1427.
29. van Spronsen FJ. Phenylketonuria: a 21st century perspective. Nature Reviews Endocrinology. 2010;6(9):509-514.
30. Følling A. Uber Ausscheidung von Phenylbrenztraubensaure in den Harn als Stoffwechselanomalie in Verbindung mit Imbezillitat. Hoppe Seylers Z Physiol Chem. 1934;277:169-176.
31. Scriver CR. The PAH gene, phenylketonuria, and a paradigm shift. Hum Mutat. 2007;28(9):831-845.
32. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015;97(2):199-215.
33. Vigneron J, Stricker M, Vert P, Rousselot JM, Levy M. Postaxial acrofacial dysostosis (Miller) syndrome: a new case. J Med Genet. 1991;28(9):636-638.
34. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics. 2010;42(1):30-35.

35. Geneé E. Une forme extensive de dysostose mandibulo-faciale. J Genet Hum. 1969;17:45-52.
36. Lee H, Deignan JL, Dorrani N, et al. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. JAMA. 2014;312(18):1880.
37. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Research. 2017;27(5):665-676.
38. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, eds. Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013:11.10.1-11.10.33.
39. Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. European Journal of Human Genetics. 2017;25(2):176-182.
40. Miller KA, Twigg SRF, McGowan SJ, et al. Diagnostic value of exome and whole genome sequencing in craniosynostosis. J Med Genet. 2017;54(4):260-268.
41. Stark Z, Tan TY, Chong B, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet Med. 2016;18(11):1090-1096.
42. Zhu X, Petrovski S, Xie P, et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet Med. 2015;17(10):774-781.
43. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974-984.
44. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15(6):R84.
45. Mohiyuddin M, Mu JC, Li J, et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics. 2015;31(16):2741-2744.
46. Stavropoulos DJ, Merico D, Jobling R, et al. Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. npj Genomic Medicine. 2016;1(1).
47. Carss KJ, Arno G, Erwood M, et al. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. Am J Hum Genet. 2017;100(1):75-90.
48. Geoffroy V, Stoetzel C, Scheidecker S, et al. Whole-genome sequencing in patients with ciliopathies uncovers a novel recurrent tandem duplication in IFT140: GEOFFROY et al. Human Mutation. 2018;39(7):983-992.
49. Clark MJ, Chen R, Lam HYK, et al. Performance comparison of exome DNA sequencing technologies. Nature Biotechnology. 2011;29(10):908-914.
50. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30(20):2843-2851.
51. The 1000 Genomes Project Consortium, Gibbs RA, Boerwinkle E, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
52. Clark L, Odgerel Z, Hernandez N, Ottman R, Louis E. Whole Genome Sequencing and Rare Variant Analysis in Essential Tremor Families (P1.073). Neurology. 2017;88(16 Supplement).

53. Wang N, Zhang Y, Gedvilaite E, et al. Using whole-exome sequencing to investigate the genetic bases of lysosomal storage diseases of unknown etiology. Human Mutation. 2017;38(11):1491-1499.
54. Wright CF, Fitzgerald TW, Jones WD, 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.
55. Rosewich H, Thiele H, Ohlenbusch A, et al. Heterozygous de-novo mutations in ATP1A3 in patients with alternating hemiplegia of childhood: a whole-exome sequencing gene-identification study. Lancet Neurol. 2012;11(9):764-773.
56. van Kuilenburg ABP, Tarailo-Graovac M, Meijer J, et al. Genome sequencing reveals a novel genetic mechanism underlying dihydropyrimidine dehydrogenase deficiency: A novel missense variant c.1700G>A and a large intragenic inversion in DPYD spanning intron 8 to intron 12. Human Mutation. 2018;39(7):947-953.
57. Yang Y, Muzny DM, Reid JG, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. New England Journal of Medicine. 2013;369(16):1502-1511.
58. Bloss CS, Zeeland AAS-V, Topol SE, et al. A genome sequencing program for novel undiagnosed diseases. Genetics in Medicine. 2015;17(12):995-1001.
59. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-291.
60. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-249.
61. 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.
62. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-121.
63. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research. 2009;37(9):e67-e67.
64. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM®. https://omim.org/. Accessed June 29, 2018.
65. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research. 2018;46(D1):D1062-D1067.
66. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-476.
67. Richards S, Aziz N, Bale S, 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):405-423.
68. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003;31(13):3812-3814.
69. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901-913.

70. James RA, Campbell IM, Chen ES, et al. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. Genome Medicine. 2016;8(1).
71. Flygare S, Hernandez EJ, Phan L, et al. The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool. BMC Bioinformatics. 2018;19(1):57.
72. Thuriot F, Buote C, Gravel E, et al. Clinical validity of phenotype-driven analysis software PhenoVar as a diagnostic aid for clinical geneticists in the interpretation of whole-exome sequencing data. Genet Med. February 2018.
73. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Medicine. 2015;7(1).
74. Thiffault I, Lantos J. The Challenge of Analyzing the Results of Next-Generation Sequencing in Children. PEDIATRICS. 2016;137(Supplement):S3-S7.
75. Köhler S, Vasilevsky NA, Engelstad M, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Research. 2017;45(D1):D865-D876.
76. Adam MP, Ardinger HH, Pagon RA, Wallace SE, eds. GeneReviews®. Seattle, WA: University of Washington, Seattle; 1993. https://www.ncbi.nlm.nih.gov/books/NBK1116/. Accessed June 29, 2018.
77. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45(D1):D353-D361.
78. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Research. 2018;46(D1):D608-D617.
79. Hamosh A, Sobreira N, Hoover-Fong J, et al. PhenoDB: A New Web-Based Tool for the Collection, Storage, and Analysis of Phenotypic Features. Human Mutation. February 2013:n/a-n/a.
80. Philippakis AA, Azzariti DR, Beltran S, et al. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. Human Mutation. 2015;36(10):915-921.
81. Pontikos N, Yu J, Moghul I, et al. Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. Bioinformatics. 2017;33(15):2421-2423.
82. Cianci P, Selicorni A. “Gestalt diagnosis” for children with suspected genetic syndromes. Italian Journal of Pediatrics. 2015;41(Suppl 2):A16.
83. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med. 2006;355(21):2217-2225.
84. Stark Z, Dashnow H, Lunke S, et al. A clinically driven variant prioritization framework outperforms purely computational approaches for the diagnostic analysis of singleton WES data. European Journal of Human Genetics. 2017;25(11):1268-1272.
85. Baldridge D, Heeley J, Vineyard M, et al. The Exome Clinic and the role of medical genetics expertise in the interpretation of exome sequencing results. Genetics in Medicine. 2017;19(9):1040-1048.
86. Dubberly H, Pangaro P, Haque U. ON MODELING: What is interaction?: are there different types? interactions. 2009;16(1):69.
87. Fischer G. Identifying and exploring design trade-offs in human-centered design. In: ACM Press; 2018:1-9.
88. cognition. In: OxfordDictionaries.com. Oxford University Press; 2018. https://en.oxforddictionaries.com/definition/cognition. Accessed June 29, 2018.

89. Kaptelinin V. Activity Theory: Implications for Human-computer Interaction. In: Nardi BA, ed. Context and Consciousness: Activity Theory and Human-Computer Interaction. Cambridge, MA, USA: Massachusetts Institute of Technology; 1995:103–116.
90. Card SK, Moran TP, Newell A. The Psychology of Human-Computer Interaction. Hillsdale, N.J: L. Erlbaum Associates; 1983.
91. Hurtienne J. Cognition in HCI: An Ongoing Story. Human Technology: An Interdisciplinary Journal on Humans in ICT Environments. 2009;5(1):12-28.
92. Carroll JM. Human–computer interaction: psychology as a science of design. International Journal of Human-Computer Studies. 1997;46(4):501-522.
93. Newell A, Card SK. The Prospects for Psychological Science in Human-Computer Interaction. Human–Computer Interaction. 1985;1(3):209-242.
94. Kaptelinin V, Nardi B, Bødker S, et al. Post-cognitivist HCI: second-wave theories. In: ACM Press; 2003:692.
95. Kuutti K. Activity Theory As a Potential Framework for Human-computer Interaction Research. In: Nardi BA, ed. Context and Consciousness: Activity Theory and Human-Computer Interaction. Cambridge, MA, USA: Massachusetts Institute of Technology; 1995:17–44.
96. Wilson M. Six views of embodied cognition. Psychonomic Bulletin & Review. 2002;9(4):625-636.
97. Hollan J, Hutchins E, Kirsh D. Distributed cognition: toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction. 2000;7(2):174-196.
98. Harrison S, Tatar D, Sengers P. The three paradigms of HCI. In: Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems San Jose, California, USA. ; 2007:1–18.
99. Bødker S. When second wave HCI meets third wave challenges. In: ACM Press; 2006:1-8.
100. Bødker S. Third-wave HCI, 10 years later---participation and sharing. interactions. 2015;22(5):24-31.
101. Baskerville RL, Myers MD. Design ethnography in information systems: Design ethnography. Information Systems Journal. 2015;25(1):23-46.
102. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. The American Journal of Human Genetics. 2017;100(2):267-280.
103. Whiffin N, Walsh R, Govind R, et al. CardioClassifier: disease- and gene-specific computational decision support for clinical genome interpretation. GENETICS in MEDICINE. January 2018.
104. Harrison SM, Riggs ER, Maglott DR, et al. Using ClinVar as a Resource to Support Variant Interpretation: Using ClinVar as a Resource to Support Variant Interpretation. In: Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, Smith DR, eds. Current Protocols in Human Genetics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2016:8.16.1-8.16.23.
105. Gershon N, Eick SG, Card S. Information visualization. interactions. 1998;5(2):9-15.
106. Nielsen CB. Visualization: A Mind-Machine Interface for Discovery. Trends Genet. 2016;32(2):73-75.
107. Chittaro L. Information visualization and its application to medicine. Artificial Intelligence in Medicine. 2001;22(2):81-88.

108. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-26.
109. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.
110. Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera-A visualization system for exploratory research and analysis. Journal of Computational Chemistry. 2004;25(13):1605-1612.
111. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-2504.
112. Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research. 2010;38(suppl_2):W214-W220.
113. Glueck M, Hamilton P, Chevalier F, et al. PhenoBlocks: Phenotype Comparison Visualizations. IEEE Transactions on Visualization and Computer Graphics. 2016;22(1):101-110.
114. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-311.
115. Maglott D. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research. 2004;33(Database issue):D54-D58.
116. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2017;45(D1):D158-D169.
117. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. The American Journal of Human Genetics. 2009;84(4):524-533.
118. Interactive Biosoftware. Alamut Visual. https://www.interactive-biosoftware.com/alamut-visual/. Accessed June 29, 2018.
119. Tanaka AJ, Cho MT, Millan F, et al. Mutations in SPATA5 Are Associated with Microcephaly, Intellectual Disability, Seizures, and Hearing Loss. The American Journal of Human Genetics. 2015;97(3):457-464.
120. Bayram Y, White JJ, Elcioglu N, et al. REST Final-Exon-Truncating Mutations Cause Hereditary Gingival Fibromatosis. The American Journal of Human Genetics. 2017;101(1):149-156.
121. Holm I, Spildrejorde M, Stadheim B, Eiklid KL, Samarakoon PS. Whole exome sequencing of sporadic patients with Currarino Syndrome: A report of three trios. Gene. 2017;624:50-55.
122. Whitford W, Hawkins I, Glamuzina E, et al. Compound heterozygous SLC19A3 mutations further refine the critical promoter region for biotin-thiamine-responsive basal ganglia disease. Molecular Case Studies. 2017;3(6):a001909.
123. Khateb S, Kowalewski B, Bedoni N, et al. A homozygous founder missense variant in arylsulfatase G abolishes its enzymatic activity causing atypical Usher syndrome in humans. GENETICS in MEDICINE. January 2018.
124. Munzner T. A Nested Model for Visualization Design and Validation. IEEE Transactions on Visualization and Computer Graphics. 2009;15(6):921-928.
125. Lam H, Bertini E, Isenberg P, Plaisant C, Carpendale S. Empirical Studies in Information Visualization: Seven Scenarios. IEEE Transactions on Visualization and Computer Graphics. 2012;18(9):1520-1536.

126. Xu P, Mei H, Ren L, Chen W. ViDX: Visual Diagnostics of Assembly Line Performance in Smart Factories. IEEE Transactions on Visualization and Computer Graphics. 2017;23(1):291-300.
127. Simon S, Mittelstädt S, Kwon BC, et al. VisExpress: Visual exploration of differential gene expression data. Information Visualization. 2017;16(1):48-73.
128. Harrison DG, Efford ND, Fisher QJ, Ruddle RA. PETMiner—A Visual Analysis Tool for Petrophysical Properties of Core Sample Data. IEEE Transactions on Visualization and Computer Graphics. 2018;24(5):1728-1741.
129. von Landesberger T, Fellner DW, Ruddle RA. Visualization System Requirements for Data Processing Pipeline Design and Optimization. IEEE Transactions on Visualization and Computer Graphics. 2017;23(8):2028-2041.
130. Nusrat S, Alam MJ, Kobourov S. Evaluating Cartogram Effectiveness. IEEE Transactions on Visualization and Computer Graphics. 2018;24(2):1077-1090.
131. EURORDIS-Rare Diseases Europe. Survey of the delay in diagnosis for 8 rare diseases in Europe ("EurordisCare2"). http://www.eurordis.org/IMG/pdf/Fact_Sheet_Eurordiscare2.pdf. Accessed June 29, 2018.
132. Shashi V, McConkie-Rosell A, Rosell B, et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genet Med. 2014;16(2):176-182.
133. Leonard JV, Morris AAM. Diagnosis and early management of inborn errors of metabolism presenting around the time of birth. Acta Paediatr. 2006;95(1):6-14.
134. Hawkes CP, Walsh A, O'Sullivan S, Crushell E. Doctors' knowledge of the acute management of Inborn Errors of Metabolism. Acta Paediatr. 2011;100(3):461-463.
135. Garrod A. THE INCIDENCE OF ALKAPTONURIA : A STUDY IN CHEMICAL INDIVIDUALITY. The Lancet. 1902;160(4137):1616-1620.
136. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-517.
137. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum Mutat. 2012;33(5):803-808.
138. Blau N, Duran M, Gibson KM, Dionisi-Vici C, eds. Physician's Guide to the Diagnosis, Treatment, and Follow-up of Inherited Metabolic Diseases. Heidelberg, Germany: Springer; 2014.
139. 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.
140. Zschocke J, Hoffmann GF. Vademecum Metabolicum Diagnosis and Treatment of Inborn Errors of Metabolism. Friedrichsdorf, Germany: Schattauer; 2011.
141. Brown GR, Hem V, Katz KS, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015;43(Database issue):D36-42.
142. Stelzer G, Rosen N, Plaschkes I, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics. 2016;54:1.30.1-1.30.33.

143. Rubinstein WS, Maglott DR, Lee JM, et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. Nucleic Acids Res. 2013;41(Database issue):D925-935.
144. Smedley D, Haider S, Durinck S, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43(W1):W589-598.
145. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457-464.
146. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(Database issue):D966-974.
147. Rogers FB. Medical subject headings. Bull Med Libr Assoc. 1963;51:114-116.
148. SNOMED-CT (Systematized Nomenclature of Medicine—Clinical Terms). Copenhagen: International Health Terminology Standards Development Organisation; 2016. http://www.ihtsdo.org/snomed-ct. Accessed June 29, 2018.
149. ICD-10 (International Classification of Diseases–10). World Health Organization; 2016. http://www.who.int/classifications/icd/en/. Accessed June 29, 2018.
150. Kurbatova N, Adamusiak T, Kurnosov P, Swertz MA, Kapushesky M. ontoCAT: an R package for ontology traversal and search. Bioinformatics. 2011;27(17):2468-2470.
151. SPECIALIST Lexical Tools. Bethesda, MD: Lexical Systems Group; 2015.
152. Hastings J, de Matos P, Dekker A, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res. 2013;41(Database issue):D456-463.
153. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem. 2003;49(4):624-633.
154. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM. 1975;18(11):613-620.
155. Frauendienst-Egger G, Trefz FK. www.metagene.de: online knowledge base for inborn errors of metabolism. J Inherit Metab Dis. 2006;29(suppl 1):84.
156. Töpel T, Scheible D, Trefz F, Hofestädt R. RAMEDIS: a comprehensive information system for variations and corresponding phenotypes of rare metabolic diseases. Hum Mutat. 2010;31(1):E1081-1088.
157. van Karnebeek CDM, Houben RFA, Lafek M, Giannasi W, Stockler S. The treatable intellectual disability APP www.treatable-id.org: a digital tool to enhance diagnosis & care for rare diseases. Orphanet J Rare Dis. 2012;7:47.
158. Tebani A, Afonso C, Marret S, Bekri S. Omics-Based Strategies in Precision Medicine: Toward a Paradigm Shift in Inborn Errors of Metabolism Investigations. Int J Mol Sci. 2016;17(9).
159. Sahoo S, Franzson L, Jonsson JJ, Thiele I. A compendium of inborn errors of metabolism mapped onto the human metabolic network. Mol Biosyst. 2012;8(10):2545-2558.
160. Stark Z, Schofield D, Alam K, et al. Prospective comparison of the cost-effectiveness of clinical whole-exome sequencing with that of usual care overwhelmingly supports early use and reimbursement. Genet Med. 2017;19(8):867-874.
161. Vissers LELM, van Nimwegen KJM, Schieving JH, et al. A clinical utility study of exome sequencing versus conventional genetic testing in pediatric neurology. Genet Med. 2017;19(9):1055-1063.

162. Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat Methods. 2014;11(9):935-937.
163. Zemojtel T, Köhler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6(252):252ra123.
164. Singleton MV, Guthery SL, Voelkerding KV, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014;94(4):599-610.
165. Berg JS. Exploring the importance of case-level clinical information for variant interpretation. Genet Med. 2017;19(1):3-5.
166. Bland A, Harrington EA, Dunn K, et al. Clinically impactful differences in variant interpretation between clinicians and testing laboratories: a single-center experience. Genet Med. 2018;20(3):369-373.
167. Wilson JR. Fundamentals of ergonomics in theory and practice. Appl Ergon. 2000;31(6):557-567.
168. Gurrieri F, Tiziano FD, Zampino G, Neri G. Recognizable facial features in patients with alternating hemiplegia of childhood. Am J Med Genet A. 2016;170(10):2698-2705.
169. Concannon N, Hegarty A-M, Stallings RL, Reardon W. Coffin-Lowry phenotype in a patient with a complex chromosome rearrangement. J Med Genet. 2002;39(8):e41.
170. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognitive Psychology. 1976;8(3):382-439.
171. Gentner D, Loewenstein J, Thompson L. Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology. 2003;95(2):393-408.
172. Lewis JR. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction. 1995;7(1):57-78.
173. Shyr C, Kushniruk A, van Karnebeek CDM, Wasserman WW. Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists, and genetic counselors. Journal of American Medical Informatics Association. 2016;23:257–68.
174. Rosch E, Simpson C, Miller RS. Structural bases of typicality effects. Journal of Experimental Psychology: Human Perception and Performance. 1976;2(4):491-502.
175. Smith EE, Shoben EJ, Rips LJ. Structure and process in semantic memory: A featural model for semantic decisions. Psychological Review. 1974;81(3):214-241.
176. Hampton JA. Concepts as Prototypes. In: Psychology of Learning and Motivation. Vol 46. Elsevier; 2006:79-113.
177. Pengelly RJ, Alom T, Zhang Z, Hunt D, Ennis S, Collins A. Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. Scientific Reports. 2017;7(1).
178. Raven ME, Flanders A. Using contextual inquiry to learn about your audiences. ACM SIGDOC Asterisk Journal of Computer Documentation. 1996;20(1):1-13.
179. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, the Mouse Genome Database Group. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Research. 2018;46(D1):D836-D842.

180. Howe DG, Bradford YM, Conlin T, et al. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Research. 2012;41(D1):D854-D860.
181. Safran M, Dalah I, Alexander J, et al. GeneCards Version 3: the human gene integrator. Database. 2010;2010(0):baq020-baq020.
182. National Center for Biotechnology Information, U.S. National Library of Medicine. PubMed. https://www.ncbi.nlm.nih.gov/pubmed. Accessed June 29, 2018.
183. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Research. 2014;42(D1):D986-D992.
184. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
185. Sedlazeck FJ, Dhroso A, Bodian DL, Paschall J, Hermes F, Zook JM. Tools for annotation and comparison of structural variation. F1000Research. 2017;6:1795.
186. Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? BMJ: British Medical Journal. 2001;322(7294):1115–1117.
187. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2(11):84.
188. Tsiplova K, Zur RM, Marshall CR, et al. A microcosting and cost–consequence analysis of clinical genomic testing strategies in autism spectrum disorder. Genetics in Medicine. 2017;19(11):1268-1275.
189. Monroe GR, Frederix GW, Savelberg SMC, et al. Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. Genetics in Medicine. 2016;18(9):949-956.
190. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011;12(8):125.
191. Muir P, Li S, Lou S, et al. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biology. 2016;17(1).
192. 23andMe, Inc. 23andMe. https://www.23andme.com/en-ca/. Published 2018. Accessed June 29, 2018.
193. Ancestry. AncestryDNA. https://www.ancestry.ca/dna/. Published 2018. Accessed June 29, 2018.
194. MyHeritage Ltd. MyHeritage DNA. https://www.myheritage.com/dna. Published 2018. Accessed June 29, 2018.
195. Helix OpCo LLC. Helix. https://www.helix.com/. Published 2018. Accessed June 29, 2018.
196. Color Genomics, Inc. Color. https://www.color.com/. Published 2018. Accessed June 29, 2018.
197. Shaer O, Nov O, Westendorf L, Ball M. Communicating Personal Genomic Information to Non-experts: A New Frontier for Human-Computer Interaction. Foundations and Trends® in Human–Computer Interaction. 2017;11(1):1-62.
198. Shaer O, Nov O, Okerlund J, et al. Informing the Design of Direct-to-Consumer Interactive Personal Genomics Reports. Journal of Medical Internet Research. 2015;17(6):e146.
199. Ramos E, Weissman SM. The dawn of consumer-directed testing. American Journal of Medical Genetics Part C: Seminars in Medical Genetics. 2018;178(1):89-97.

200. Phillips AM. “Only a click away — DTC genetics for ancestry, health, love…and more: A view of the business and regulatory landscape.” Applied & Translational Genomics. 2016;8:16-22.
201. Telenti A, Steinhubl SR, Topol EJ. Rethinking the medical record. The Lancet. 2018;391(10125):1013.
202. Wiljer D, Urowitz S, Apatu E, et al. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res. 2008;10(4):e34.
203. Apple Inc. Empower your patients with Health Records on iPhone. https://www.apple.com/healthcare/health-records/. Published 2018. Accessed June 29, 2018.
204. Apple Inc. A bold way to look at your health. https://www.apple.com/ca/ios/health/. Published 2018. Accessed June 29, 2018.
205. Vasilevsky NA, Foster ED, Engelstad ME, et al. Plain-language medical vocabulary for precision diagnosis. Nature Genetics. 2018;50(4):474-476.
206. Mikk KA, Sleeper HA, Topol EJ. The Pathway to Patient Data Ownership and Better Health. JAMA. 2017;318(15):1433.
207. and The eMERGE Network, Gottesman O, Kuivaniemi H, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present and future. Genetics in Medicine. 2013;15(10):761-771.
208. Masys DR, Jarvik GP, Abernethy NF, et al. Technical desiderata for the integration of genomic data into Electronic Health Records. J Biomed Inform. 2012;45(3):419-422.
209. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nature Reviews Genetics. 2018;19(5):299-310.
210. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386).
211. Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nature Communications. 2017;8:15824.
212. Chen R, Mias GI, Li-Pook-Than J, et al. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012;148(6):1293-1307.
213. Piening BD, Zhou W, Contrepois K, et al. Integrative Personal Omics Profiles during Periods of Weight Gain and Loss. Cell Systems. 2018;6(2):157-170.e8.
214. Graham E, Lee J, Price M, et al. Integration of genomics and metabolomics for prioritization of rare disease variants: a 2018 literature review. Journal of Inherited Metabolic Disease. 2018;41(3):435-445.
215. Li G, Bankhead P, Dunne PD, et al. Embracing an integromic approach to tissue biomarker research in cancer: Perspectives and lessons learned. Briefings in Bioinformatics. June 2016:bbw044.
216. Cottret L, Frainay C, Chazalviel M, et al. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Research. 2018;46(W1):W495-W502.
217. van Karnebeek CDM, Wortmann SB, Tarailo-Graovac M, et al. The role of the clinician in the multi-omics era: are you ready? Journal of Inherited Metabolic Disease. 2018;41(3):571-582.
218. Knight W. The Machines Are Getting Ready to Play Doctor. MIT Technology Review. July 2017. https://www.technologyreview.com/s/608234/the-machines-are-getting-ready-to-play-doctor/. Accessed June 29, 2018.

219. Devlin H. London hospitals to replace doctors and nurses with AI for some tasks. The Guardian. https://www.theguardian.com/society/2018/may/21/london-hospitals-to-replace-doctors-and-nurses-with-ai-for-some-tasks. Published May 21, 2018. Accessed June 29, 2018.
220. Campbell D. The robot will see you now: how AI could revolutionise NHS. The Guardian. https://www.theguardian.com/society/2018/jun/11/the-robot-will-see-you-now-how-ai-could-revolutionise-nhs. Published June 11, 2018. Accessed June 29, 2018.
221. Sun Y. AI could alleviate China's doctor shortage. MIT Technology Review. March 2018. https://www.technologyreview.com/s/610397/ai-could-alleviate-chinas-doctor-shortage/. Accessed June 29, 2018.
222. Kenny P, Parsons T, Gratch J, Rizzo A. Virtual humans for assisted health care. In: ACM Press; 2008:1.
223. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69:S36-S40.
224. The Lancet null. Artificial intelligence in health care: within touching distance. Lancet. 2018;390(10114):2739.
225. Luxton DD. Recommendations for the ethical use and design of artificial intelligent care providers. Artificial Intelligence in Medicine. 2014;62(1):1-10.
226. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. JAMA. 2018;319(1):19.
227. Fast E, Horvitz E. Long-Term Trends in the Public Perception of Artificial Intelligence. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. 2017:963–969.
228. Koch M. Artificial Intelligence Is Becoming Natural. Cell. 2018;173(3):531-533.
229. Hood L, Auffray C. Participatory medicine: a driving force for revolutionizing healthcare. Genome Med. 2013;5(12):110.
230. Sagner M, McNeil A, Puska P, et al. The P4 Health Spectrum – A Predictive, Preventive, Personalized and Participatory Continuum for Promoting Healthspan. Progress in Cardiovascular Diseases. 2017;59(5):506-521.
231. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature Reviews Clinical Oncology. 2011;8(3):184-187.
232. Flores M, Glusman G, Brogaard K, Price ND, Hood L. P4 medicine: how systems medicine will transform the healthcare sector and society. Per Med. 2013;10(6):565-576.
233. Levy KD, Blake K, Fletcher-Hoppe C, et al. Opportunities to implement a sustainable genomic medicine program: lessons learned from the IGNITE Network. Genet Med. July 2018.
234. Gaff CL, M. Winship I, M. Forrest S, et al. Preparing for genomic medicine: a real world demonstration of health system change. npj Genomic Medicine. 2017;2(1).
235. Miniwatts Marketing Group. Internet World Stats - Usage and Population Statistics. https://www.internetworldstats.com/stats.htm. Published 2018. Accessed June 29, 2018.
236. Steinhubl SR, Topol EJ. Digital medicine, on its way to being just plain medicine. npj Digital Medicine. 2018;1(1).
237. LaRosa E, Danks D. Impacts on Trust of Healthcare AI. In: AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, New Orleans, USA, February 1-3, 2018. 2018.

Appendices

Appendix A

A.1 Walkthrough of IEMbase

When users open the application interface, the starting page presents a disclaimer. Upon agreeing to the disclaimer, users are directed to the main page, which presents a search form and the following three buttons: Browse, Search, and Mini-Expert (Figure A1). In the search form - which is also accessible by the Search button - users can type in disorder, gene, biomarker, or symptom names to look up information on a particular disorder.

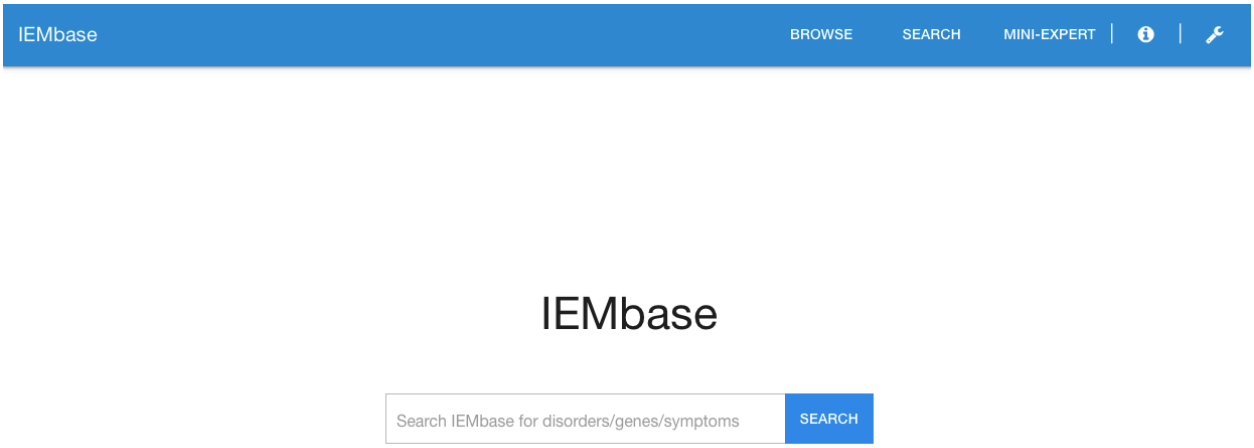


Figure A1 Screenshot of main page.

The Browse button directs to a page with a full catalog of IEMs that are currently curated on IEMbase (Figure A2). The catalog is represented as a tree, where each branch represents a disease classification used by the IEM community. Users can hide or expand the branches of the tree as they browse, and they can look up detailed information on each disorder by clicking on the disorder name (Figure A3). In addition, users can search for a particular disorder by its name using the search form located above the catalog.

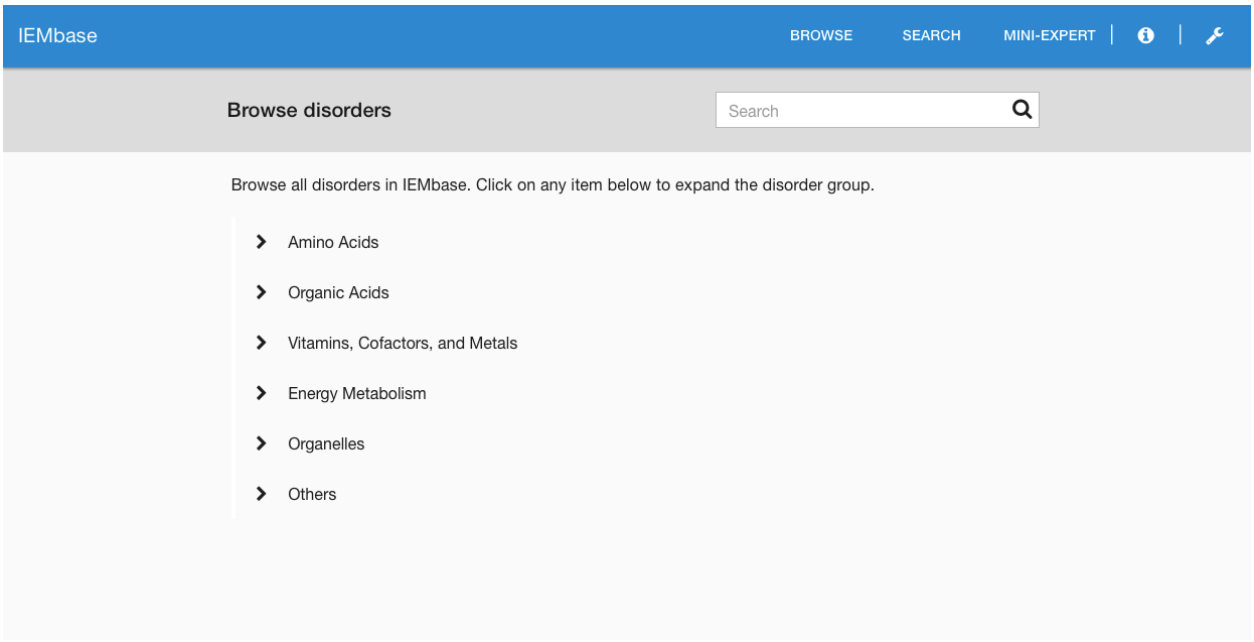


Figure A2 Screenshot of Browse page.

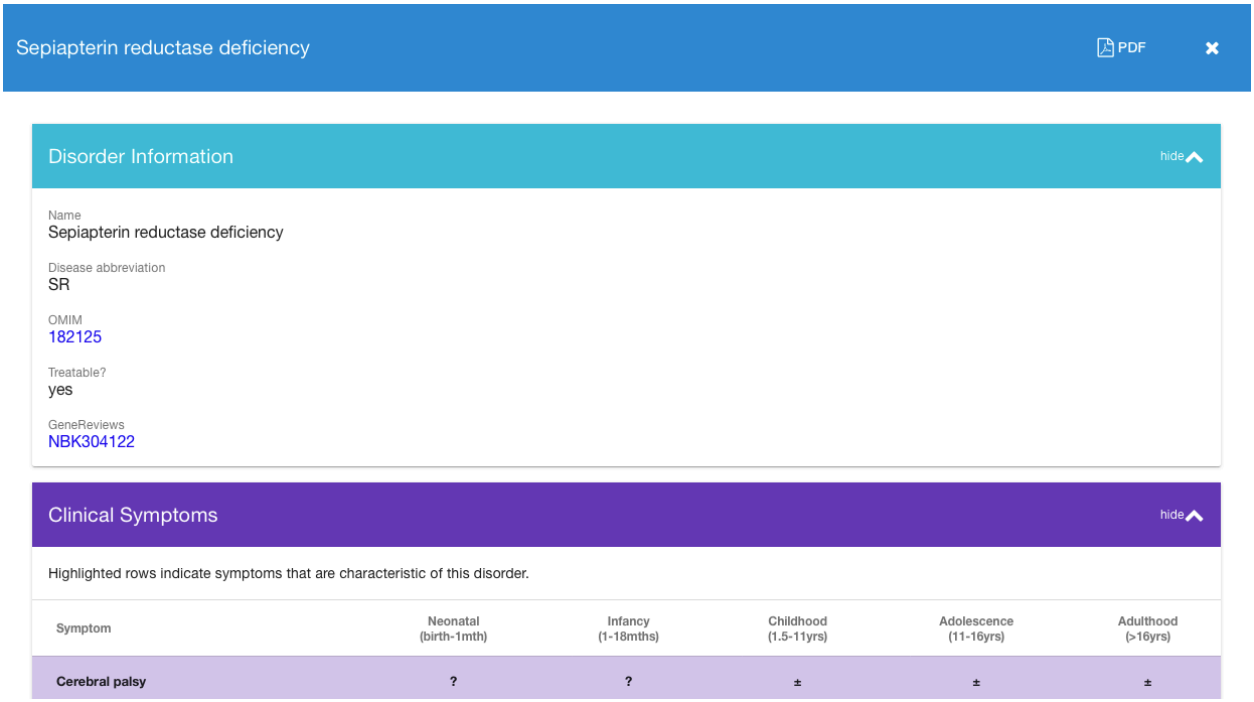


Figure A3 Screenshot of Disorder Information page.

Upon selecting the Mini-Expert option, users are directed to a page with the Input Profile form (Figure A4). In this form, users are asked to enter a list of biochemical and clinical phenotypes using a search bar. For biochemical entries, the system asks to specify relative levels as low, normal, or high. As the phenotypes are added, they will appear in the list below the search bar.

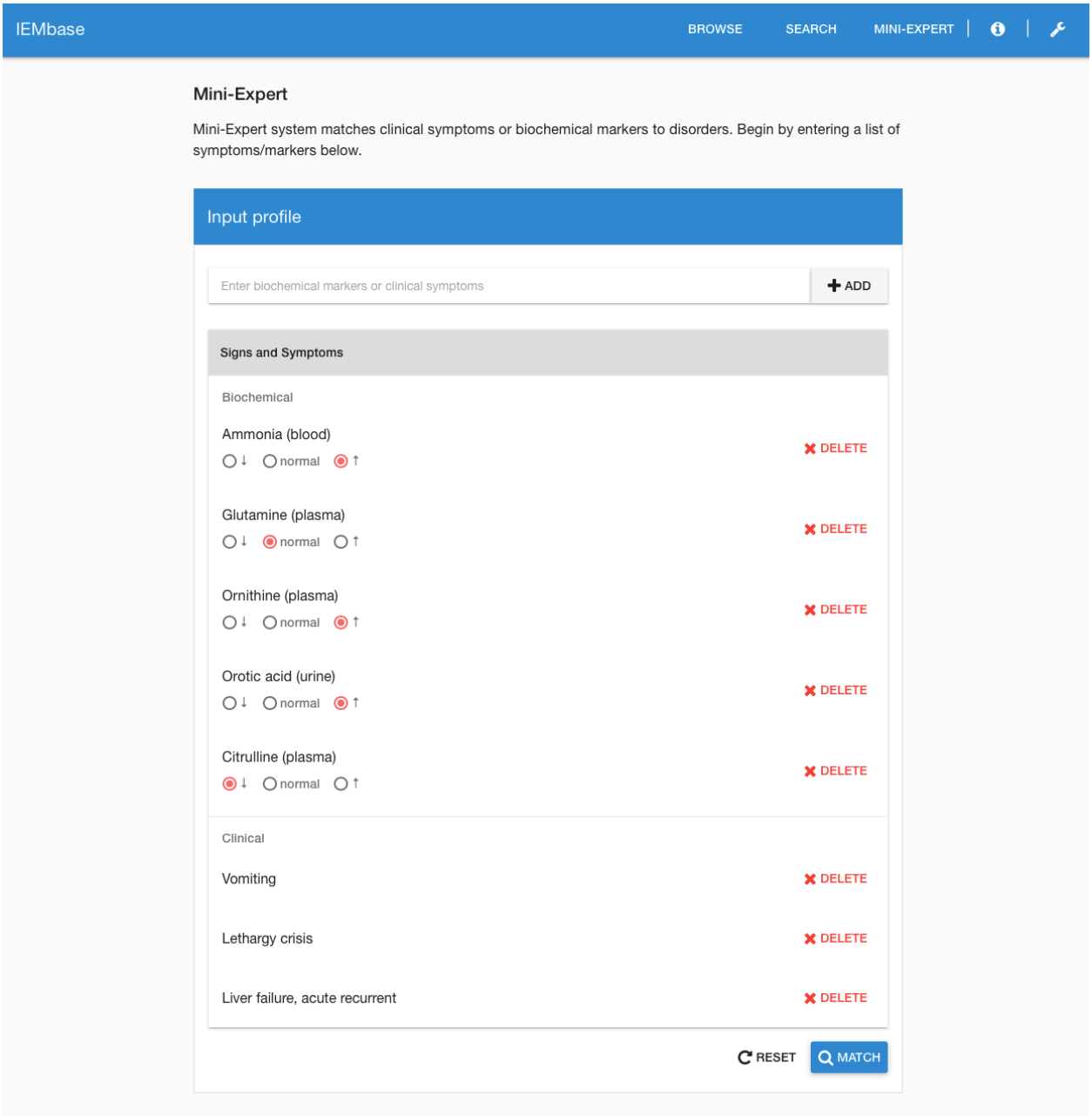


Figure A4 Screenshot of Mini-Expert Query page.

Upon submitting the phenotype list, the system returns a list of matching IEMs in the Results section, which is located below the Mini-Expert Query section (Figure A5). In the Results section, users can look up the details of each disorder in the list, build a differential diagnosis chart, or build a gene panel.

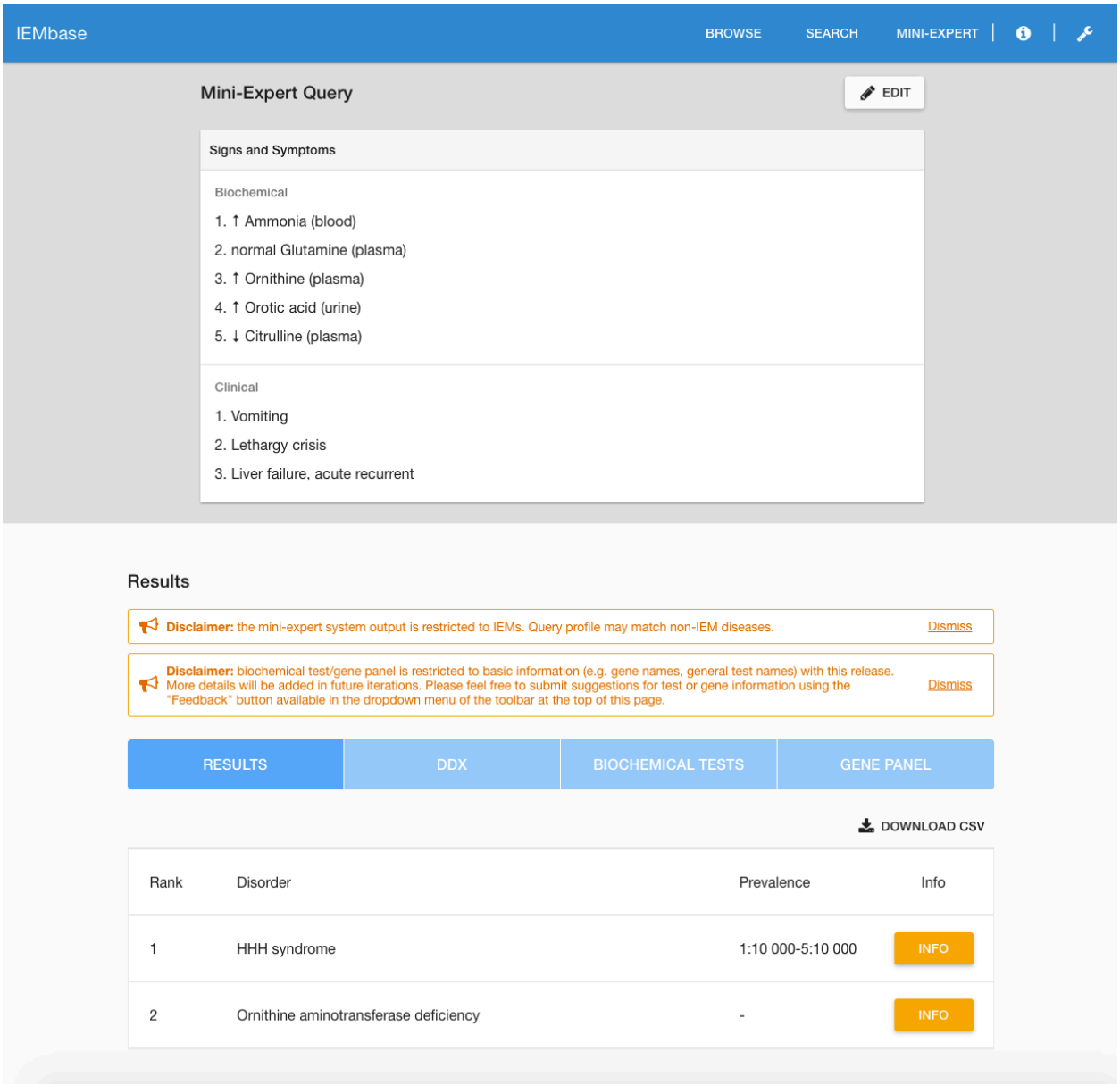


Figure A5 Screenshot of Mini-Expert Results page.

The DDx button in the Results section leads to a page where users can select multiple candidate disorders (Figure A6) and generate a differential diagnosis chart based on their selection (Figure A7). Similarly, the Gene Panel button and Biochemical Test button in the Results section direct

to respective pages where users can select multiple disorders and generate a gene panel or a biochemical test panel based on their selection (Figure A8, A9).

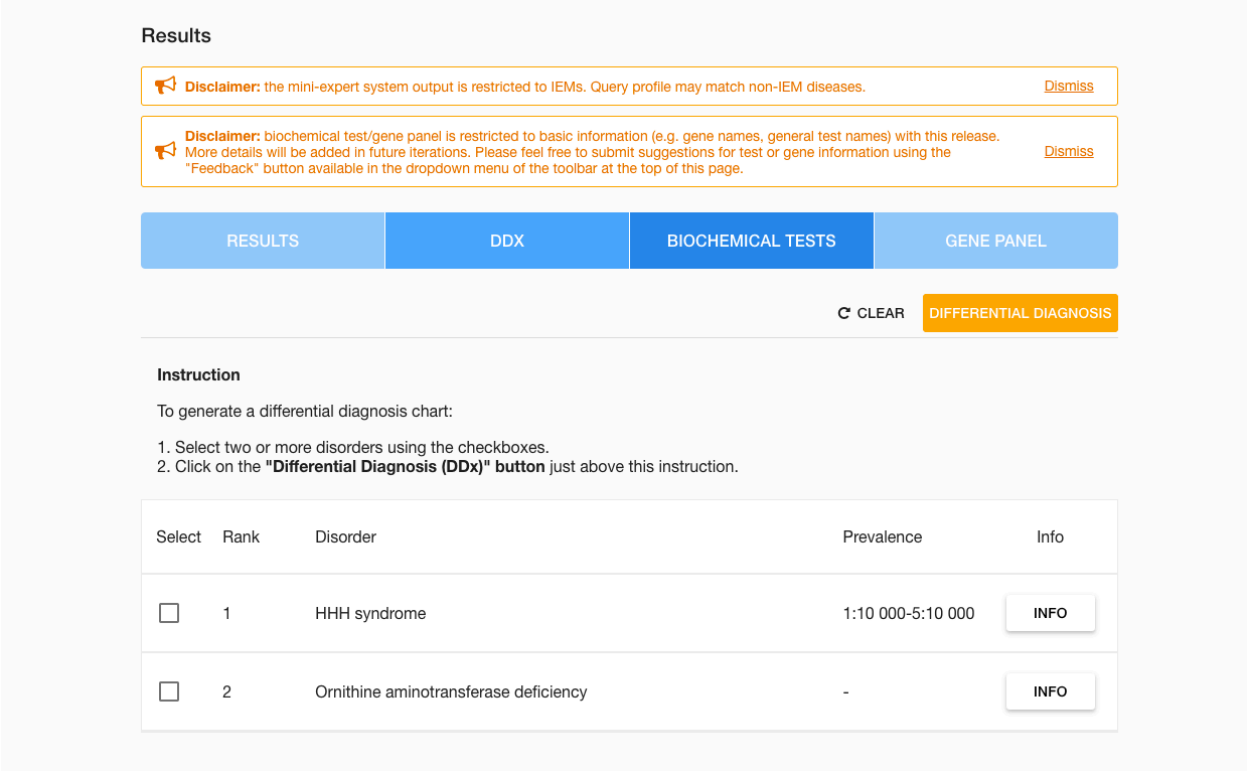


Figure A6 Screenshot of Mini-Expert DDx (Differential Diagnosis) selection page.

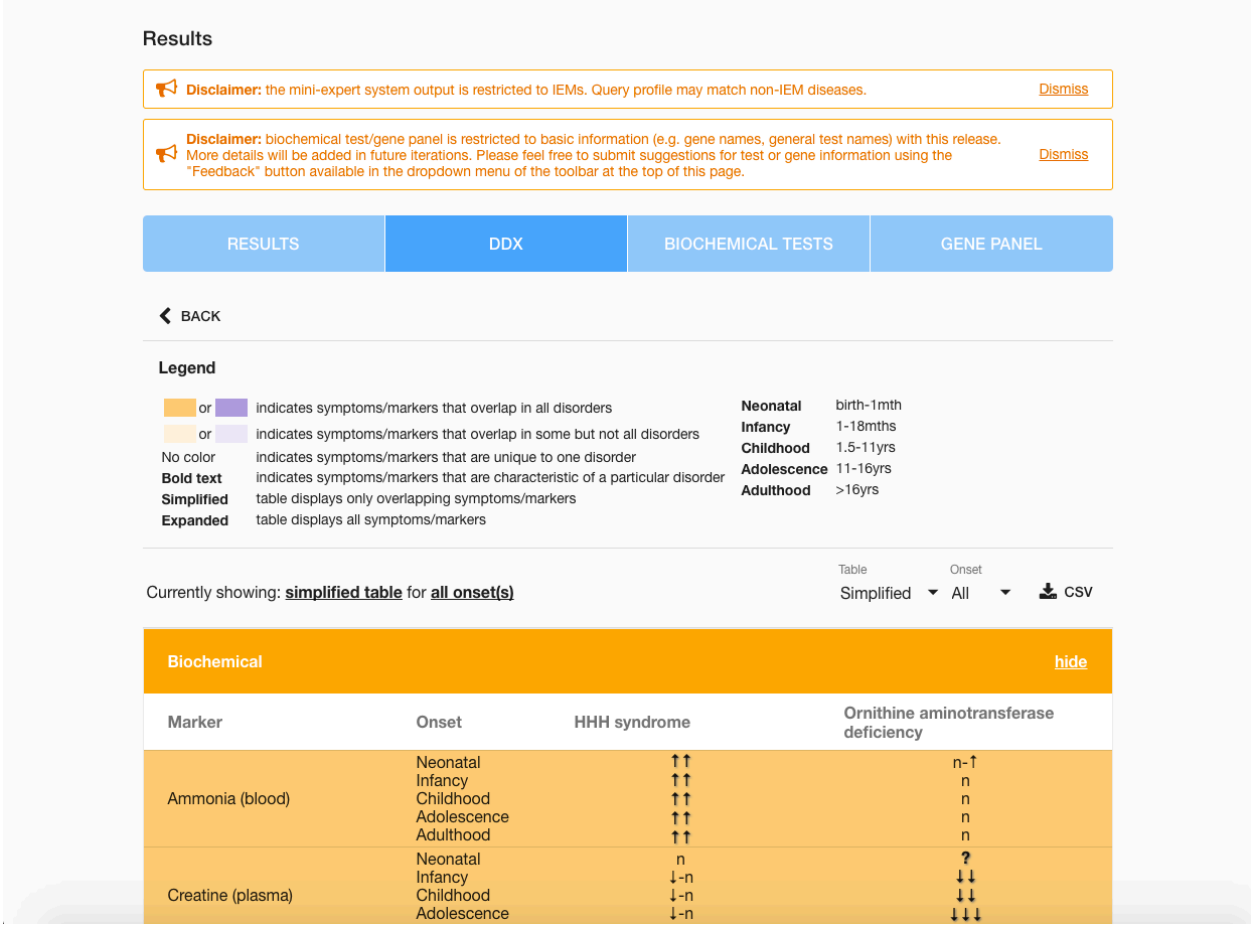


Figure A7 Screenshot of Mini-Expert DDx (Differential Diagnosis) result page.

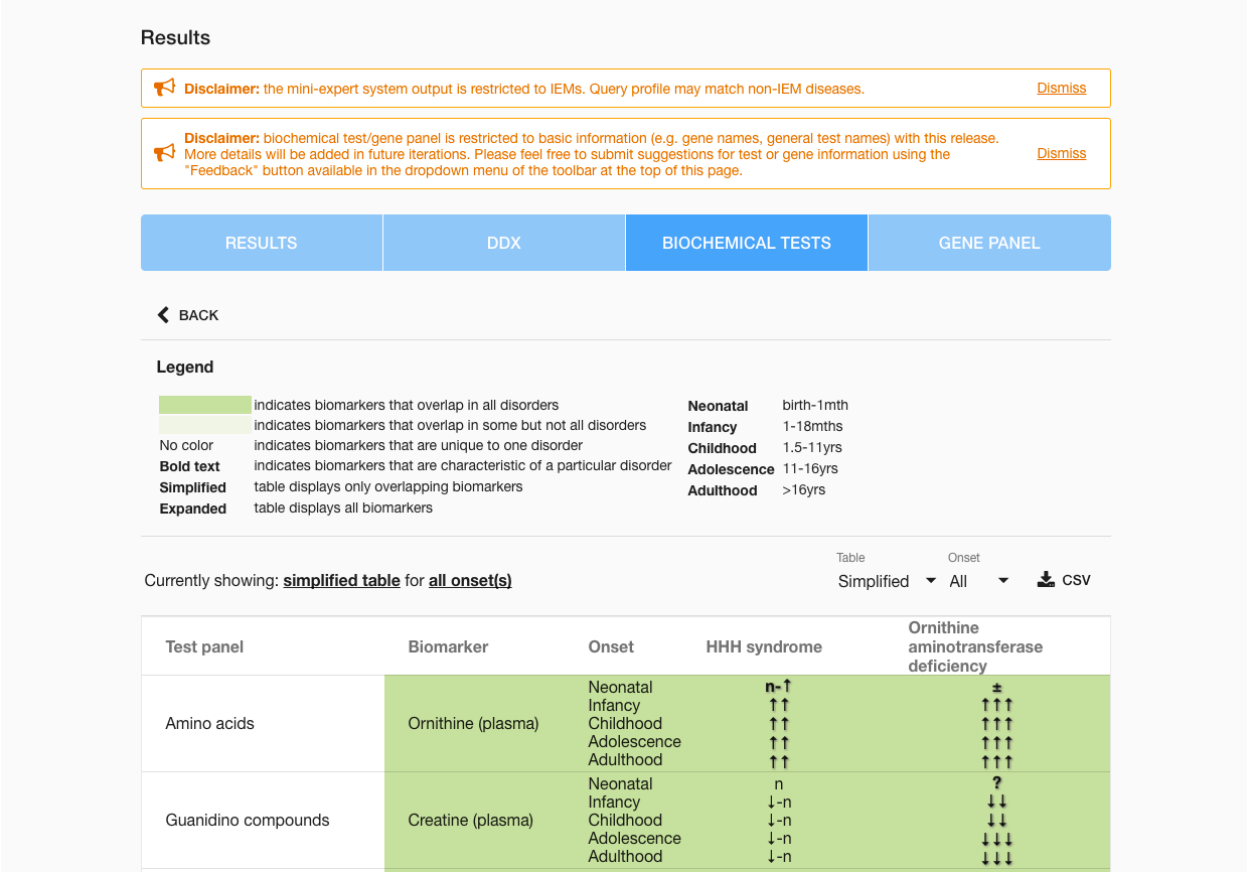


Figure A8 Screenshot of Mini-Expert Biochemical Tests page.

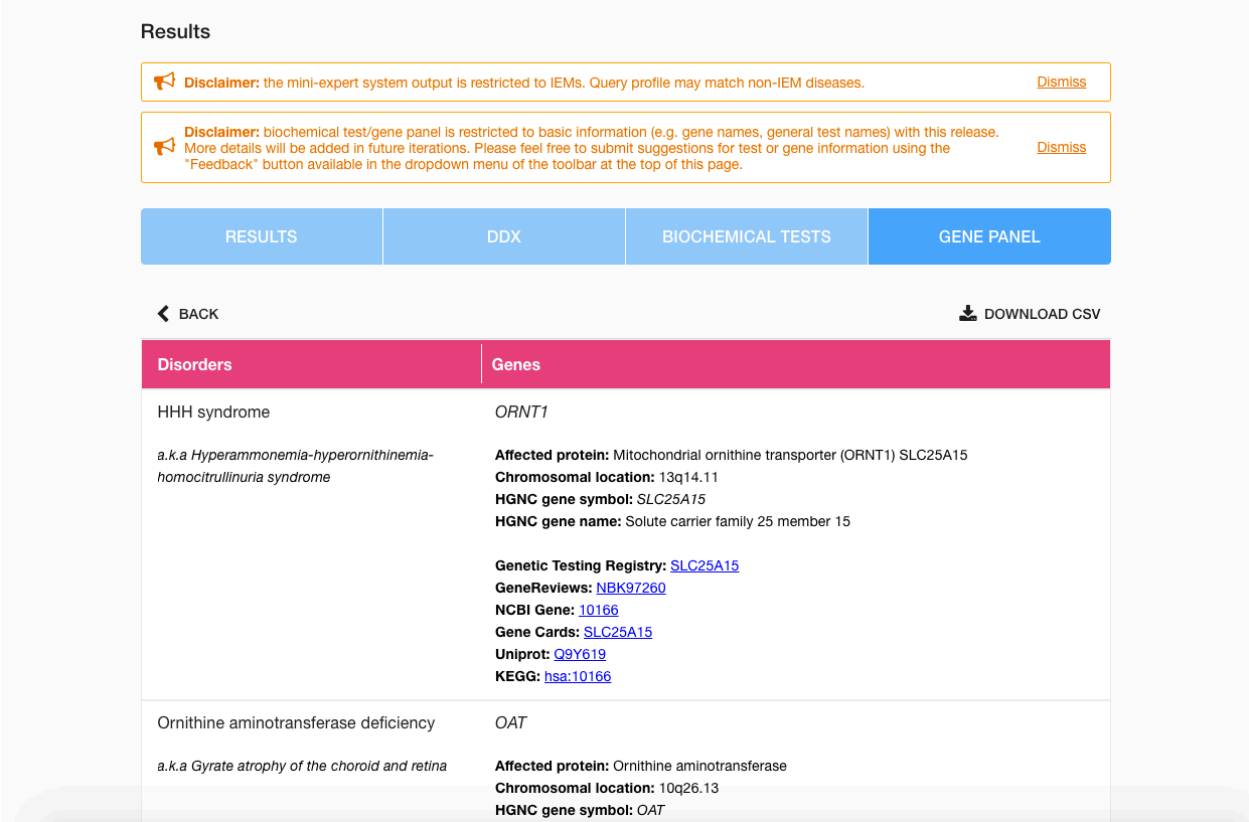


Figure A9 Screenshot of Mini-Expert Gene Panel page.

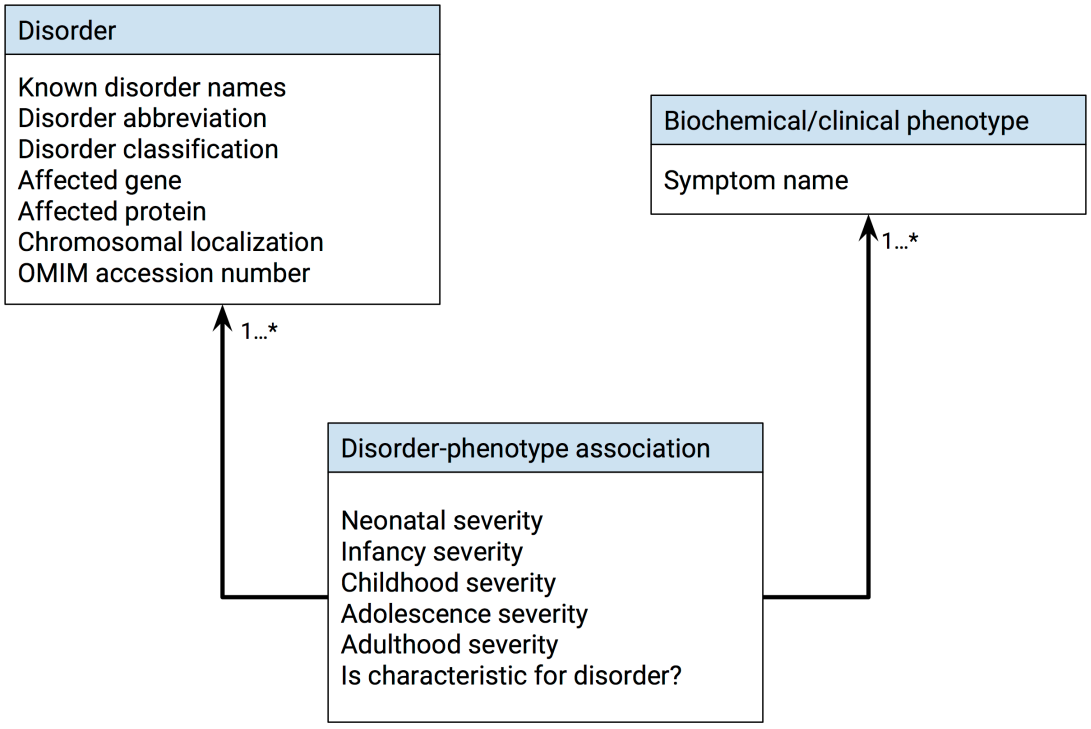


Figure A10 Knowledgebase schema.

The knowledgebase consisted of three tables which were extracted from the nascent disease database. Each table represented different data types: disorders, biochemical/clinical phenotypes, and associations between disorders and phenotypes.

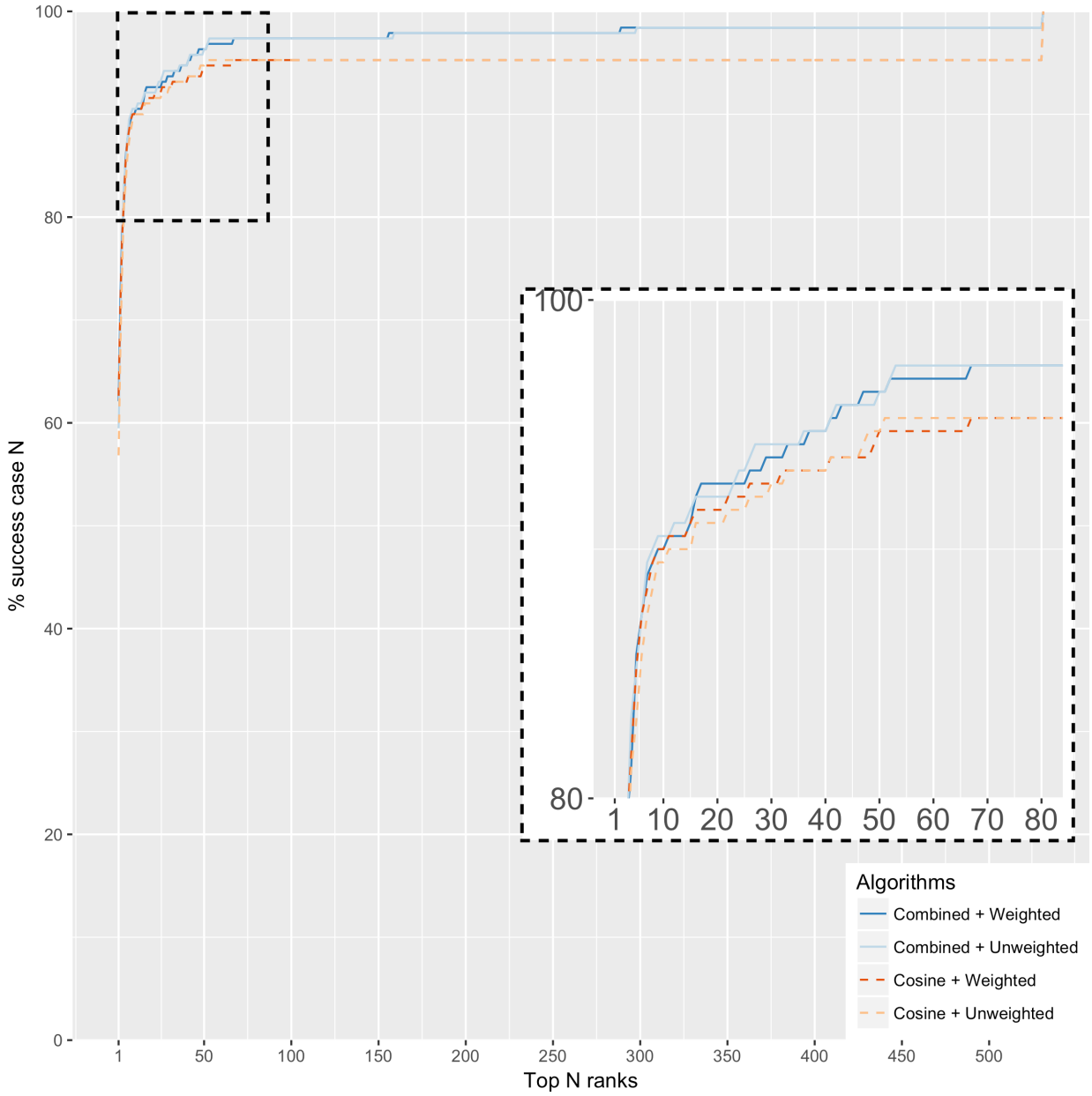


Figure A11 Mini-expert system performance evaluation results.

The performance of IEMbase's mini-expert algorithm (Combined + Weighted) was compared to three other algorithms: combined cosine similarity and semantic similarity without weights (Combined + Unweighted), cosine similarity only with weights (Cosine + Weighted), and cosine similarity only without weights (Cosine + Unweighted). There was no significant performance difference between the mini-expert system and other algorithms ($p = 0.66$ in Mini-expert vs Combined + Unweighted, $p = 1.0$ in Mini-expert vs Cosine + Weighted, $p = 0.30$ in

Mini-expert vs Cosine + Unweighted; Mann-Whitney-U). Black dotted boxes show a section of the plot between the top one candidate disorder and the top 80 candidate disorders.

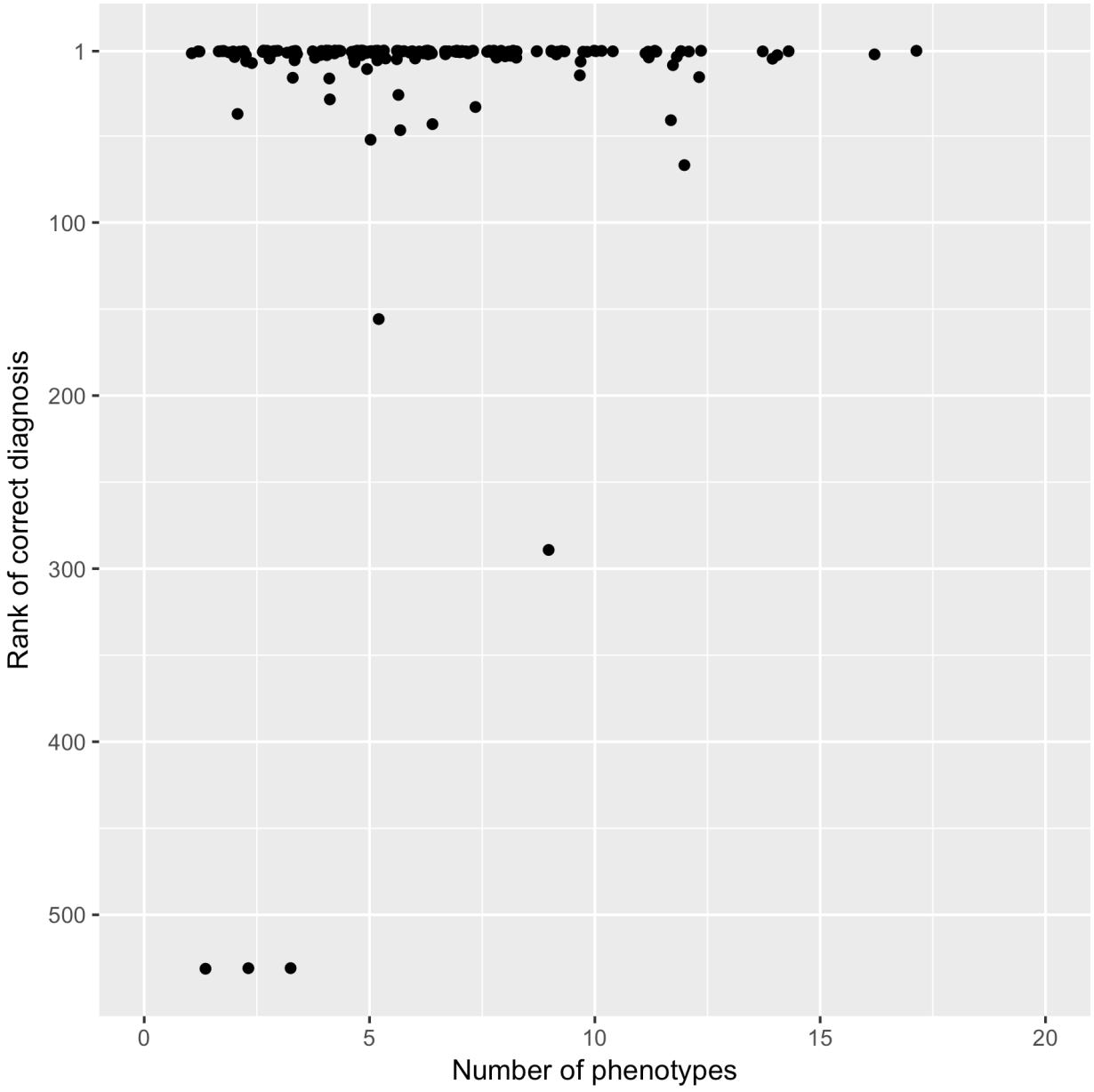


Figure A12 Scatterplot of rank of correct diagnosis against number of phenotypes specified.

Rank of correct diagnosis did not correlate with number of phenotypes specified for each case ($p = 0.69$; Spearman's rank correlation test).

Biochemical markers
1. ↑ Ammonia (blood)
2. Normal Glutamine (plasma)
3. ↑ Ornithine (plasma)
4. ↑ Orotic acid (urine)
5. ↓ Citrulline (plasma)
Clinical symptoms
1. Vomiting
2. Lethargy crisis
3. Liver failure, acute recurrent

Table A1 Case study query.

Diagnosis	Number of cases with diagnosis	Rank of actual diagnosis*
Glutaric aciduria type I	6	1, 2, 3
HHH syndrome	4	1
Tyrosinaemia type I	4	1, 2, 29
Succinic semialdehyde dehydrogenase deficiency	4	1
Fructose-1,6-bisphosphatase deficiency	4	1, 5, 15
Molybdenum cofactor deficiency A	4	1, 2, 3
Guanidinoacetate methyltransferase deficiency	3	1
Smith-Lemli-Opitz syndrome	3	1
S-adenosylhomocysteine hydrolase deficiency	3	1
Cystathionine beta-synthase deficiency	3	1, 7, 9
Suphite oxidase deficiency	3	1, 6
Nonketotic hyperglycinaemia	3	1
6-Pyruvoyl-tetrahydropterin synthase deficiency	3	1
Prolidase deficiency	3	1, 5, 7
Ornithine aminotransferase deficiency	3	1
Propionic acidemia	3	1, 3, 5
Methylenetetrahydrofolate reductase deficiency	3	1
Carnitine palmitoyltransferase 1 deficiency	3	1, 2, **
Glycerol kinase deficiency, isolated	3	2, 3
Aromatic L-amino acid decarboxylase deficiency	2	1
Ornithine transcarbamylase deficiency	2	1
Citrullinemia type I	2	1
Argininemia	2	1
Canavan disease	2	1
Fumarase deficiency	2	1
Citrullinemia type II	2	1, 289
Tyrosinaemia type II	2	1
Alkaptonuria	2	1
Hurler, Scheie disease	2	1
Refsum disease (classic, adult)	2	1, 47
Hyperprolinaemia type II	2	1
Galactosaemia	2	1, 41
Glycogen storage disease type III	2	1, 52
Lysinuric protein intolerance	2	1, 3
Maple syrup urine disease	2	1
Congenital hypophosphatasia	2	1

Methylmalonic acidemia	2	1
Alpha-amino adipic semialdehyde (AASA) dehydrogenase deficiency	2	1
Sepiapterin reductase deficiency	2	1
Biotinidase deficiency	2	1, 43
Arginine:glycine amidinotransferase deficiency	2	2
Metachromatic leukodystrophy-like disorder due to saposin B deficiency	2	2, 3
Galactokinase deficiency	2	2, 4
Multiple acyl-CoA dehydrogenase deficiency	2	3, 17
Adenosylcobalamin and methylcobalamin synthesis defect - cblC	2	5, 11
Niemann-Pick disease type C1	2	6, 16
Maternally Inherited Mitochondrial Dystonia	1	**
2-Methylbutyrylglycinuria	1	**
Adenylosuccinate lyase deficiency	1	1
Hypoxanthine guanine phosphoribosyltransferase deficiency	1	1
Argininosuccinic aciduria	1	1
3-Hydroxy-3-methylglutaryl-CoA synthase deficiency	1	1
Isolated deficiency of long-chain 3-hydroxyacyl-CoA dehydrogenase	1	1
Sterol 27-hydroxylase deficiency	1	1
Transaldolase deficiency	1	1
Ribose-5-phosphate isomerase deficiency	1	1
Acrodermatitis enteropathica	1	1
Trimethylaminuria	1	1
Gamma-glutamylcysteine synthetase deficiency	1	1
Hawkinsinuria	1	1
Pyruvate dehydrogenase complex deficiency E3	1	1
Hunter disease	1	1
Morquio A disease	1	1
Fucosidosis	1	1
Salla disease	1	1
Dihydropyrimidinase deficiency	1	1
3-Hydroxy-3-methylglutaryl-CoA lyase deficiency	1	1
Tay-Sachs disease	1	1
Farber disease	1	1
GTP cyclohydrolase I deficiency	1	1
Lysosomal acid lipase deficiency	1	1
Phosphoglycerate dehydrogenase deficiency	1	1
Hydroxyprolinemia	1	1
Glutamate formimino transferase deficiency	1	1

Hereditary fructose intolerance	1	1
Cystinuria	1	1
Glucose transporter-1 deficiency	1	1
Glycogen storage disease type I a	1	1
Glycogen storage disease type I non-a	1	1
Hartnup disorder	1	1
Isovaleric acidemia	1	1
Folate receptor alpha deficiency	1	1
Carnitine transporter deficiency	1	1
Thiamine-responsive megaloblastic anemia syndrome (SLC19A2)	1	1
Primary hyperoxaluria type I	1	1
Tyrosine hydroxylase deficiency	1	1
L-2-hydroxyglutaric aciduria	1	2
D-2-hydroxyglutaric aciduria type I	1	2
Pyruvate dehydrogenase complex deficiency E3 X	1	2
Mitochondrial trifunctional protein deficiency	1	2
Phosphoribosyl pyrophosphate synthetase 1 superactivity	1	2
Xanthine dehydrogenase deficiency	1	2
Hyperprolinemia type I	1	2
3-Hydroxy-3-methyl glutaric aciduria	1	2
2-Methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency	1	2
Zellweger spectrum disorders	1	3
Multiple sulfatase deficiency	1	3
Adenosine kinase deficiency	1	3
Hyperinsulinism of infancy	1	3
Pyruvate dehydrogenase complex deficiency E1a	1	4
Medium - chain acyl CoA dehydrogenase deficiency	1	4
Methylacetoacetyl-CoA thiolase deficiency	1	4
Transcobalamin deficiency	1	4
GM1-gangliosidosis	1	5
Krabbe disease	1	5
Adenosylcobalamin and methylcobalamin synthesis defect - cblD-MMA/HC	1	5
ATP synthase deficiency	1	5
Methylglutaconic aciduria type IV	1	5
3-Methylcrotonylglycinuria	1	6
MEGDEL Syndrome	1	7
Carnitine palmitoyltransferase 2 deficiency	1	8
Dopa-responsive dystonia	1	16

Methylcobalamin synthesis defect - cblD-HC	1	26
X-linked adrenoleukodystrophy and adrenomyeloneuropathy	1	33
Tangier disease (ABCA1)	1	37
Mitochondrial Depletion Syndrome 4A	1	67
Carbamoyl phosphate synthetase I deficiency	1	156

Table A2 Overview of disorders (n=117) investigated within the validation of 190 cases.

Cases have been selected to validate the mini-expert system using a diverse range of disorders. The selected cases cover approximately 22% of the 530 disorders in IEMbase. In the "Rank of actual diagnosis" column, multiple ranks are recorded as some cases ranked differently than one another.

* Disorders ranked over 20 are described in detail in Table A3.

** Disorder was ranked out.

Rank	Diagnosis	Comment	User specified biomarkers	User specified clinical symptoms
Ranked out	2-Methylbutyrylglycinuria	"C5 2-Methylbutyrylcarnitin" should be high, not "C4 Butyrylcarnitine", in 2-Methylbutyrylglycinuria.	↑ C4 Butyrylcarnitine (blood)	No entry
Ranked out	Carnitine palmitoyltransferase 1 deficiency	"C18:2-Acylcarnitine (dried blood spot)" should be low in carnitine palmitoyltransferase 1 deficiency.	↑ C18:2-acylcarnitine (dried blood spot) ↑ Carnitine, free (dried blood spot)	No entry
Ranked out	Maternally Inherited Mitochondrial Dystonia	There are no biomarkers specified in the database or in the literature for this disease.	↑ C16 Hexadecanoylcarnitine ↑ Carnitine, free (dried blood spot) ↑ C2 Acetylcarnitine	No entry
26	Methylcobalamin synthesis defect - cblD-HC	"Methylmalonic acid (plasma)" should be normal in cblD-HC. Megaloblastic anemia is one of the characteristic features. Different cobalamin defects within top 5.	↑ Homocysteine, total (plasma) ↑ Methylmalonic acid (urine)	Nystagmus Intellectual disability Diminished visual activity Heart Failure
29	Tyrosinaemia type I	The validator did not provide any essential biomarkers for this case.	No entry	Hepatosplenomegaly Growth retardation Renal Fanconi Syndrome Osteopenia
33	X-linked adrenoleukodystrophy and adrenomyeloneuropathy	Duplicate entries in the system caused incorrect matching of "Very long-chain fatty acids (0)", which was entered by the user. In the latest database, duplicate entries are merged to "Very-long-chain fatty acids (plasma)"	↑ Very long-chain fatty acids (O) normal Phytanic acid (plasma) normal Pristanic acid (plasma) ↑ C26:0 fatty acid (plasma)	Developmental regression Adrenal insufficiency White matter abnormalities (MRI)

37	Tangier disease (ABCA1)	"LDL cholesterol" was missing from the description of Tangier disease in the database. The latest database includes the biomarker and the system ranks Tangier disease at rank #5 after the correction.	↓ LDL cholesterol (plasma)	Splenomegaly
41	Galactosaemia	The validator may have entered in "↑ Prothrombin time" to indicate "↓ Coagulation factors (plasma)".	↓ Hemoglobin (blood) ↑ Transaminase (plasma) ↑ Bilirubin, total/direct (plasma) ↑ Prothrombin time	Fontanel enlarged Brain edema (MRI) Cataract Ascites Anemia, hemolytic Hepatomegaly Liver failure Hyperbilirubinemia, prolonged conjugated
43	Biotinidase deficiency	"↑ 3-Hydroxyisovaleric acid (urine)" was missing from the description of biotinidase deficiency in the database. The latest database includes the biomarker in the description, and biotinidase deficiency ranks at #1 for this case after the correction.	↑ Lactate (plasma) ↑ 3-Hydroxyisovaleric acid (urine)	Loss of hair Epilepsy Developmental delay Blindness
47	Refsum disease (classic, adult)	"↑ Pipecolic acid (serum)" was missing from the description of Refsum disease in the database. The latest version includes the biomarker, and Refsum disease ranks at #5 for this case after the correction.	↑ Pipecolic acid (serum)	Deafness, sensorineural Developmental delay Retinopathy Facial dysmorphism Hypotonia

52	Glycogen storage disease type III	The system does not recognize the relationship between "Transaminase (plasma)" and other enzymes specified in the description of this disease. In a future development cycle, the system will be able to make the recognition using a synonyms table.	↑ Transaminase (plasma) ↑ Creatine kinase (plasma)	Hepatomegaly Hypoglycemia, episodic Motor developmental delay
67	Mitochondrial Depletion Syndrome 4A	Only "↑ Lactate (plasma)" is associated with Mitochondrial Depletion Syndrome 4A in the database. The listed biomarkers may not be specific enough for the system to make a match.	↑ Protein (CSF) ↓ 5-Methyl-THF (CSF) ↑ Neopterin (CSF) ↑ Lactate (MRS)	Epilepsy +/- encephalopathy Developmental delay Regression, psychomotor Developmental regression Seizures, Intractable Seizures, myoclonic MR Spectroscopy brain Cerebral atrophy (MRI)
156	Carbamoyl phosphate synthetase I deficiency	The validator likely entered carbamoyl phosphate synthase I instead of carnitine palmitoyltransferase 1 deficiency as the final diagnosis - which would therefore rank at #1.	↑ Carnitine, free normal Dicarboxylic acids (urine) ↓ Long-chain acylcarnitine (DBS)	Hepatopathy Renal tubular acidosis
289	Citrullinemia type II	Biomarkers and clinical presentation are not specific enough for the system to match to a disorder	↑ Ketone, during hypoglycemia normal Lactate (plasma) normal Acylcarnitine, all (plasma) ↑ Beta-hydroxybutyrate (urine) ↑ Acetoacetate (urine) ↓ Amino acids (urine)	Hypoglycemia, episodic Abdominal pain Short stature

Table A3 Overview of cases whose diagnoses ranked out of the top 20.

	Biochemical only	Clinical only
MRR	0.70	0.29
% success at 1	60	19
% success at 5	83	38
% success at 10	89	49
% success at 20	91	55

Table A4 Mini-expert system performance using only biochemical/clinical queries.

Mean reciprocal rank (MRR) measures how close the correct match is to the top rank on average. It ranges from 0 to 1 and values close to 1 indicate that correct matches appear closer to the top on average. % success at N = % of cases with correct diagnoses within top N ranks. Cases with only biochemical phenotypes or only clinical phenotypes were removed from the set (n=172).

Appendix B

CHARGE syndrome	**Smith-Lemli-Opitz syndrome**	**Tuberous sclerosis**
1. Tellier AL, Cormier-Daire V, Abadie V, et al. CHARGE syndrome: report of 47 cases and review. Am J Med Genet 1998;76:402–9. 2. Lalani SR, Hefner MA, Belmont JW, et al. CHARGE Syndrome. 2006 Oct 2 [Updated 2012 Feb 2]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1117/ 3. Hsu P, Ma A, Wilson M, et al. CHARGE syndrome: a review. J Paediatr Child Health 2014;50:504–11.	1. Ryan AK, Bartlett K, Clayton P, et al. Smith-Lemli-Opitz syndrome: a variable clinical and biochemical phenotype. J Med Genet 1998;35:558–65. 2. Kelley RI, Hennekam RC. The Smith-Lemli-Opitz syndrome. J Med Genet 2000;37:321–35. 3. Greene C, Pitts W, Rosenfeld R, et al. Smith-Lemli-Opitz syndrome in two 46,XY infants with female external genitalia. Clin Genet 1984;25:366–72. 4. Lachman MF, Wright Y, Whiteman DA, et al. Brief clinical report: a 46,XY phenotypic female with Smith-Lemli-Opitz syndrome. Clin Genet 1991;39:136–41. 5. Haas D, Armbrust S, Haas J-P, et al. Smith-Lemli-Opitz syndrome with a classical phenotype, oesophageal achalasia and borderline plasma sterol concentrations. J Inherit Metab Dis 2005;28:1191–6. 6. Mueller C, Patel S, Irons M, et al. Normal cognition and behavior in a Smith-Lemli-Opitz syndrome patient who presented with Hirschsprung disease. Am J Med Genet A 2003;123A:100–6.	1. Roach ES, Sparagana SP. Diagnosis of tuberous sclerosis complex. J Child Neurol 2004;19:643–9. doi:10.1177/08830738040190090301 2. Northrup H, Koenig MK, Pearson DA, et al. Tuberous Sclerosis Complex. 1999 Jul 13 [Updated 2015 Sep 3]. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1220/ 3. Northrup H, Krueger DA, International Tuberous Sclerosis Complex Consensus Group. Tuberous sclerosis complex diagnostic criteria update: recommendations of the 2012 Iinternational Tuberous Sclerosis Complex Consensus Conference. Pediatr Neurol 2013;49:243–54. doi:10.1016/j.pediatrneurol.2013.08.001 4. Teplick JG. Tuberous Sclerosis: Extensive Roentgen Findings Without the Usual Clinical Picture: A Case Report 1. Radiology 1969;93:53–5. doi:10.1148/23.1.53

	7. Nowaczyk MJ, Whelan DT, Hill RE. Smith-Lemli-Opitz syndrome: phenotypic extreme with minimal clinical findings. Am J Med Genet 1998;78:419–23. 8. Langius FAA, Waterham HR, Romeijn GJ, et al. Identification of three patients with a very mild form of Smith-Lemli-Opitz syndrome. Am J Med Genet A 2003;122A:24–9.	5. Rott H-D, Lemcke B, Zenker M, et al. Cyst-like cerebral lesions in tuberous sclerosis. Am J Med Genet 2002;111:435–9. doi:10.1002/ajmg.10637 6. Kaufmann R, Kornreich L, Goldberg-Stern H. Unusual clinical presentation of tuberless tuberous sclerosis complex. J Child Neurol 2009;24:361–4. doi:10.1177/0883073808325659 7. Han X, Zheng L, Zheng T. Onychogryphosis in tuberous sclerosis complex: an unusual feature. Anais Brasileiros de Dermatologia 2016;91:116–8. doi:10.1590/abd1806-4841.20164720 8. Fox J, Ben-Shachar S, Uliel S, et al. Rare familial TSC2 gene mutation associated with atypical phenotype presentation of Tuberous Sclerosis Complex. Am J Med Genet A 2017;173:744–8. doi:10.1002/ajmg.a.38027 9. McGrae JD, Hashimoto K. Unilateral facial angiofibromas--a segmental form of tuberous sclerosis. Br J Dermatol 1996;134:727–30.

Table B1 List of publications and clinical resources reviewed for simulated clinical scenario development.

Scenarios	Scenario-described phenotype	Symptom-based workflow participant-specified phenotype	# of times phenotype was selected***	Prototype-based workflow participant-specified phenotype	# of times phenotype was selected***
Scenario 1 Smith-Lemli-Opitz syndrome (MIM 270400) Symptom n** = 4 Prototype n** = 4	2nd-3rd toe syndactyly	2-3 toe syndactyly	3	2-3 toe syndactyly	4
		Syndactyly	1		
	Anteverted nares	Anteverted nares	4	Anteverted nares	4
	Broad nasal bridge	Narrow nasal bridge	1	Wide nasal bridge	3
		Wide nasal bridge	3	Depressed nasal bridge	1
	Developmental delay	Global developmental delay	3	Global developmental delay	1
		Neurodevelopmental delay	1		
	Feeding difficulties and failure to thrive @ 3 months	Feeding difficulties	2	Feeding difficulties	1
		Feeding difficulties in infancy	1		
		Failure to thrive	1	Failure to thrive	2
		Failure to thrive in infancy	1		
	Hypotonia	Generalized hypotonia	3	Generalized hypotonia	2
		Muscular hypotonia	1		
	Irritable	Irritability	1	-	-
	Low-set ears	Low-set ears	3	Low-set ears	3
				Posteriorly rotated ears	1
	Microcephaly	Microcephaly	4	Microcephaly	3
	Micrognathia	Micrognathia	4	Micrognathia	2
	Postaxial polydactyly	Postaxial polydactyly	4	Postaxial hand polydactyly	2
				Postaxial polydactyly	1
				Postaxial foot polydactyly	1
	Ptosis	Ptosis	3	Ptosis	3
	*Extra**	-	-	Autosomal recessive inheritance	1
				Ambiguous genitalia	1
				Hypospadias	1
				Elevated 7-dehydrocholesterol	1

Scenarios	Scenario-described phenotype	Symptom-based workflow participant-specified phenotype	# of times phenotype was selected***	Prototype-based workflow participant-specified phenotype	# of times phenotype was selected***
				Hypocholesterolemia	1
				Intellectual disability	2
Scenario 2 Smith-Lemli-Opitz syndrome (MIM 270400) Symptom n** = 5 Prototype n** = 2	Feeding difficulties @ 3 months	Feeding difficulties	2	Gastroesophageal reflux	1
		Feeding difficulties in infancy	1	Poor suck	1
				Vomiting	1
	Broad nasal bridge	Wide nasal bridge	2	Wide nasal bridge	1
		Wide nasal ridge	1		
	Developmental delay	Global developmental delay	2	-	-
	Finger clinodactyly	Finger clinodactyly	4	Finger clinodactyly	1
	Micrognathia	Micrognathia	4	Micrognathia	1
	Mild hypotonia	Generalized hypotonia	3	Generalized hypotonia	2
		Muscular hypotonia	1		
		Central hypotonia	1		
	Mild ptosis	Ptosis	3	Ptosis	2
	Minimal cutaneous 2nd-3rd toe syndactyly	2-3 toe syndactyly	5	2-3 toe syndactyly	2
	*Extra**	Weak voice	1	Autosomal recessive inheritance	1
		Hypertonia	1	Intellectual disability	1
Scenario 3 Tuberous sclerosis 1 (MIM 191100)	Brain MRI: cortical sclerotic tubers	Cortical tubers	5	Cortical tubers	1
	Epileptic seizure	Seizures	4	Infantile spasms	2
	Hypomelanotic macules on the chest	Hypomelanotic macule	5	Cafe-au-lait spot	1
				Hypomelanotic macule	1

Scenarios	Scenario-described phenotype	Symptom-based workflow participant-specified phenotype	# of times phenotype was selected***	Prototype-based workflow participant-specified phenotype	# of times phenotype was selected***
Symptom n** = 4 Prototype n** = 4	Hypsarrhythmia	Hypsarrhythmia	4	-	-
	Renal cysts	Renal cyst	4	Renal cyst	2
		Multiple renal cysts	1		
	Skin papules on the side of nose	Skin-colored papule	1	Adenoma sebaceum	2
		Papule	1	Subcutaneous nodule	1
		Facial papilloma	1		
	*Extra**	Hypermelanotic macule	1	Subependymal nodules	1
Scenario 4 Tuberous sclerosis 1 (MIM 191100) Symptom n** = 5 Prototype n** = 2	Epileptic seizure	Seizures	1	Seizures	1
		Status epilepticus	1	Infantile spasms	1
	Hypsarrhythmia	Hypsarrhythmia	4	Hypsarrhythmia	1
				Cardiac rhabdomyoma	1
				Wolff-Parkinson-White syndrome	1
	Intellectual disability	Intellectual disability	4	Intellectual disability	4
	Renal cysts	Renal cyst	3	Renal cyst	3
		Multiple renal cysts	1	Renal angiomyolipoma	1
	Skin papules on the side of nose	Papule	1	Adenoma sebaceum	3
		Skin-colored papule	1	Subcutaneous nodule	1
				Papule	1
	*Extra**	-	-	Autosomal dominant inheritance	1

Table B2 Qualitative summary of phenotypes with counts.

*Extra refers to phenotypes that were not described in scenarios

** n = number of participants assigned to scenario

*** Counts how many times the phenotype was selected by participants. If a participant changed phenotypes multiple times, each selected phenotype was counted

Appendix C

C.1 Development of contextual interview template

An interview template was developed by (a) reviewing common WES/WGS analysis practices that were reported in literature and (b) brainstorming interview focuses.

For (a), a small-scale literature review was conducted on papers relevant to clinical exome/genome sequencing. The review focused on rare disease literature because most of our prospective participants were expected to be from that domain at the time. It was also restricted to papers published since 2015 to consider only recent bioinformatics practices. Papers were searched on PubMed (accessed on Jan 12, 2018) using the following terms: (((genome sequencing OR exome sequencing) AND human AND (rare OR genetic) AND (disease OR disorder) NOT cancer NOT bacteria NOT virus)) AND ("2015/01/01"[Date - Publication] : "2018/01/31"[Date - Publication]). Among the search result, 270 papers were randomly selected. Their titles and abstracts were screened for relevance. 102 relevant papers were identified and their full texts were reviewed to extract the names of computational analysis/information visualization tools used as well as the context of using such tools. The extracted information was categorized by the context of use, and these categories were incorporated into questions on the characteristics of routine WES/WGS analyses. The list of reviewed papers is provided in Appendix C.2.

(b) was conducted by following a contextual inquiry procedure described by Raven and Flanders.[5] JJYL brainstormed open-ended questions on sticky notes. The questions were then grouped by similar themes and a generalized heading was created for each group. Duplicate questions were discarded. After the exercise, four groups (or topics) were identified: characteristics of routine WES/WGS analyses, context of using information visualization during routine analyses, perception of current visualization tools, and suggestions for new visualization. The questions and their topics were added to the interview template. CDMvK and WWW reviewed the template for flow and quality.

[5] Raven ME, Flanders A. Using contextual inquiry to learn about your audiences. ACM SIGDOC Asterisk Journal of Computer Documentation. 1996;20(1):1-13.

C.2 List of papers reviewed for contextual interview template development

1. Aggarwal A, Rodriguez-Buritica DF, Northrup H. Wiedemann-Steiner syndrome: Novel pathogenic variant and review of literature. Eur J Med Genet. 2017;60(6):285-288.
2. Alkelai A, Olender T, Haffner-Krausz R, et al. A role for TENM1 mutations in congenital general anosmia. Clin Genet. 2016;90(3):211-219.
3. Al-Maawali A, Dupuis L, Blaser S, et al. Prenatal growth restriction, retinal dystrophy, diabetes insipidus and white matter disease: expanding the spectrum of PRPS1-related disorders. Eur J Hum Genet. 2015;23(3):310-316.
4. Al-Mubarak B, Abouelhoda M, Omar A, et al. Whole exome sequencing reveals inherited and de novo variants in autism spectrum disorder: a trio study from Saudi families. Sci Rep. 2017;7(1):5679.
5. Astuti GDN, van den Born LI, Khan MI, et al. Identification of Inherited Retinal Disease-Associated Genetic Variants in 11 Candidate Genes. Genes (Basel). 2018;9(1).
6. Bashamboo A, Bignon-Topalovic J, Moussi N, McElreavey K, Brauner R. Mutations in the Human ROBO1 Gene in Pituitary Stalk Interruption Syndrome. J Clin Endocrinol Metab. 2017;102(7):2401-2406.
7. Bayram Y, White JJ, Elcioglu N, et al. REST Final-Exon-Truncating Mutations Cause Hereditary Gingival Fibromatosis. Am J Hum Genet. 2017;101(1):149-156.
8. Brady PD, Van Esch H, Fieremans N, et al. Expanding the phenotypic spectrum of PORCN variants in two males with syndromic microphthalmia. Eur J Hum Genet. 2015;23(4):551-554.
9. Bravo-Gil N, Méndez-Vidal C, Romero-Pérez L, et al. Improving the management of Inherited Retinal Dystrophies by targeted sequencing of a population-specific gene panel. Sci Rep. 2016;6:23910.
10. Butcher NJ, Merico D, Zarrei M, et al. Whole-genome sequencing suggests mechanisms for 22q11.2 deletion-associated Parkinson's disease. PLoS ONE. 2017;12(4):e0173944.
11. Cabezas OR, Flanagan SE, Stanescu H, et al. Polycystic Kidney Disease with Hyperinsulinemic Hypoglycemia Caused by a Promoter Mutation in Phosphomannomutase 2. J Am Soc Nephrol. 2017;28(8):2529-2539.
12. Cai N, Bigdeli TB, Kretzschmar WW, et al. 11,670 whole-genome sequences representative of the Han Chinese population from the CONVERGE project. Sci Data. 2017;4:170011.
13. Casey JP, McGettigan PA, Healy F, et al. Unexpected genetic heterogeneity for primary ciliary dyskinesia in the Irish Traveller population. Eur J Hum Genet. 2015;23(2):210-217.
14. Castro-Sánchez S, Álvarez-Satta M, Tohamy MA, Beltran S, Derdak S, Valverde D. Whole exome sequencing as a diagnostic tool for patients with ciliopathy-like phenotypes. PLoS ONE. 2017;12(8):e0183081.
15. Chatzispyrou IA, Alders M, Guerrero-Castillo S, et al. A homozygous missense mutation in ERAL1, encoding a mitochondrial rRNA chaperone, causes Perrault syndrome. Hum Mol Genet. 2017;26(13):2541-2550.
16. Chelban V, Patel N, Vandrovcova J, et al. Mutations in NKX6-2 Cause Progressive Spastic Ataxia and Hypomyelination. Am J Hum Genet. 2017;100(6):969-977.
17. Chiu C-Y, Su S-C, Fan W-L, et al. Whole-Genome Sequencing of a Family with Hereditary Pulmonary Alveolar Proteinosis Identifies a Rare Structural Variant Involving CSF2RA/CRLF2/IL3RA Gene Disruption. Sci Rep. 2017;7:43469.

18. Choi B-O, Nakhro K, Park HJ, et al. A cohort study of MFN2 mutations and phenotypic spectrums in Charcot-Marie-Tooth disease 2A patients. Clin Genet. 2015;87(6):594-598.
19. Choi HJ, Lee JS, Yu S, et al. Whole-exome sequencing identified a missense mutation in WFS1 causing low-frequency hearing loss: a case report. BMC Med Genet. 2017;18(1):151.
20. Chong JX, Caputo V, Phelps IG, et al. Recessive Inactivating Mutations in TBCK, Encoding a Rab GTPase-Activating Protein, Cause Severe Infantile Syndromic Encephalopathy. Am J Hum Genet. 2016;98(4):772-781.
21. Choudhury A, Ramsay M, Hazelhurst S, et al. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. Nat Commun. 2017;8(1):2062.
22. Chow Y-P, Abdul Murad NA, Mohd Rani Z, et al. Exome sequencing identifies SLC26A4, GJB2, SCARB2 and DUOX2 mutations in 2 siblings with Pendred syndrome in a Malaysian family. Orphanet J Rare Dis. 2017;12(1):40.
23. Cohen I, Staretz-Chacham O, Wormser O, et al. A novel homozygous SLC25A1 mutation with impaired mitochondrial complex V: Possible phenotypic expansion. Am J Med Genet A. 2018;176(2):330-336.
24. Coll M, Striano P, Ferrer-Costa C, et al. Targeted next-generation sequencing provides novel clues for associated epilepsy and cardiac conduction disorder/SUDEP. PLoS ONE. 2017;12(12):e0189618.
25. C Yuen RK, Merico D, Bookman M, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602-611.
26. Degn M, Dauvilliers Y, Dreisig K, et al. Rare missense mutations in P2RY11 in narcolepsy with cataplexy. Brain. 2017;140(6):1657-1668.
27. Dinckan N, Du R, Petty LE, et al. Whole-Exome Sequencing Identifies Novel Variants for Tooth Agenesis. J Dent Res. 2018;97(1):49-59.
28. Dougherty M, Lazar J, Klein JC, et al. Genome sequencing in a case of Niemann-Pick type C. Cold Spring Harb Mol Case Stud. 2016;2(6):a001222.
29. Duncan JL, Biswas P, Kozak I, et al. Ocular Phenotype of a Family with FAM161A-associated Retinal Degeneration. Ophthalmic Genet. 2016;37(1):44-52.
30. Eichstaedt CA, Song J, Viales RR, et al. First identification of Krüppel-like factor 2 mutation in heritable pulmonary arterial hypertension. Clin Sci. 2017;131(8):689-698.
31. Einarsdottir E, Grauers A, Wang J, et al. CELSR2 is a candidate susceptibility gene in idiopathic scoliosis. PLoS ONE. 2017;12(12):e0189591.
32. Ellard S, Kivuva E, Turnpenny P, et al. An exome sequencing strategy to diagnose lethal autosomal recessive disorders. Eur J Hum Genet. 2015;23(3):401-404.
33. Falkenberg KD, Braverman NE, Moser AB, et al. Allelic Expression Imbalance Promoting a Mutant PEX6 Allele Causes Zellweger Spectrum Disorder. Am J Hum Genet. 2017;101(6):965-976.
34. Frosk P, Arts HH, Philippe J, et al. A truncating mutation in CEP55 is the likely cause of MARCH, a novel syndrome affecting neuronal mitosis. J Med Genet. 2017;54(7):490-501.
35. Gao W, Chen C, Zhou T, et al. Rare coding variants in MAPK7 predispose to adolescent idiopathic scoliosis. Hum Mutat. 2017;38(11):1500-1510.
36. Guella I, McKenzie MB, Evans DM, et al. De Novo Mutations in YWHAG Cause Early-Onset Epilepsy. Am J Hum Genet. 2017;101(2):300-310.

37. Gueneau L, Fish RJ, Shamseldin HE, et al. KIAA1109 Variants Are Associated with a Severe Disorder of Brain Development and Arthrogryposis. Am J Hum Genet. 2018;102(1):116-132.
38. Gui H, Schriemer D, Cheng WW, et al. Whole exome sequencing coupled with unbiased functional analysis reveals new Hirschsprung disease genes. Genome Biol. 2017;18(1):48.
39. Guo Y, Hwang L-D, Li J, et al. Genetic analysis of impaired trimethylamine metabolism using whole exome sequencing. BMC Med Genet. 2017;18(1):11.
40. Guo Y, Prokudin I, Yu C, et al. Advantage of Whole Exome Sequencing over Allele-Specific and Targeted Segment Sequencing in Detection of Novel TULP1 Mutation in Leber Congenital Amaurosis. Ophthalmic Genet. 2015;36(4):333-338.
41. Gupta S, Chaurasia A, Pathak E, et al. Whole exome sequencing unveils a frameshift mutation in CNGB3 for cone dystrophy: A case report of an Indian family. Medicine (Baltimore). 2017;96(30):e7490.
42. Gustafson K, Duncan JL, Biswas P, et al. Whole Genome Sequencing Revealed Mutations in Two Independent Genes as the Underlying Cause of Retinal Degeneration in an Ashkenazi Jewish Pedigree. Genes (Basel). 2017;8(9).
43. Habib AM, Matsuyama A, Okorokov AL, et al. A novel human pain insensitivity disorder caused by a point mutation in ZFHX2. Brain. 2018;141(2):365-376.
44. Hartley T, Wagner JD, Warman-Chardon J, et al. Whole-exome sequencing is a valuable diagnostic tool for inherited peripheral neuropathies: Outcomes from a cohort of 50 families. Clin Genet. 2018;93(2):301-309.
45. Hayer SN, Deconinck T, Bender B, et al. STUB1/CHIP mutations cause Gordon Holmes syndrome as part of a widespread multisystemic neurodegeneration: evidence from four novel mutations. Orphanet J Rare Dis. 2017;12(1):31.
46. Hengel H, Magee A, Mahanjah M, et al. CNTNAP1 mutations cause CNS hypomyelination and neuropathy with or without arthrogryposis. Neurol Genet. 2017;3(2):e144.
47. Holm I, Spildrejorde M, Stadheim B, Eiklid KL, Samarakoon PS. Whole exome sequencing of sporadic patients with Currarino Syndrome: A report of three trios. Gene. 2017;624:50-55.
48. Horai M, Mishima H, Hayashida C, et al. Detection of de novo single nucleotide variants in offspring of atomic-bomb survivors close to the hypocenter by whole-genome sequencing. J Hum Genet. 2018;63(3):357-363.
49. Hotchkiss L, Donkervoort S, Leach ME, et al. Novel De Novo Mutations in KIF1A as a Cause of Hereditary Spastic Paraplegia With Progressive Central Nervous System Involvement. J Child Neurol. 2016;31(9):1114-1119.
50. Jacquinet A, Brown L, Sawkins J, et al. Expanding the FANCO/RAD51C associated phenotype: Cleft lip and palate and lobar holoprosencephaly, two rare findings in Fanconi anemia. Eur J Med Genet. 2018;61(5):257-261.
51. Karimzadeh P, Naderi S, Modarresi F, et al. Case reports of juvenile GM1 gangliosidosisis type II caused by mutation in GLB1 gene. BMC Med Genet. 2017;18(1):73.
52. Keser V, Khan A, Siddiqui S, et al. The Genetic Causes of Nonsyndromic Congenital Retinal Detachment: A Genetic and Phenotypic Study of Pakistani Families. Invest Ophthalmol Vis Sci. 2017;58(2):1028-1036.

53. Khateb S, Kowalewski B, Bedoni N, et al. A homozygous founder missense variant in arylsulfatase G abolishes its enzymatic activity causing atypical Usher syndrome in humans. Genet Med. January 2018.
54. Khera AV, Won H-H, Peloso GM, et al. Association of Rare and Common Variation in the Lipoprotein Lipase Gene With Coronary Artery Disease. JAMA. 2017;317(9):937-946.
55. Kim J-H, Ko YJ, Kim J, et al. Genetic investigation of bisphosphonate-related osteonecrosis of jaw (BRONJ) via whole exome sequencing and bioinformatics. PLoS ONE. 2015;10(2):e0118084.
56. Kolanczyk M, Krawitz P, Hecht J, et al. Missense variant in CCDC22 causes X-linked recessive intellectual disability with features of Ritscher-Schinzel/3C syndrome. Eur J Hum Genet. 2015;23(5):633-638.
57. König E, Volpato CB, Motta BM, et al. Exploring digenic inheritance in arrhythmogenic cardiomyopathy. BMC Med Genet. 2017;18(1):145.
58. Krenn M, Zulehner G, Hotzy C, et al. Hereditary spastic paraplegia caused by compound heterozygous mutations outside the motor domain of the KIF1A gene. Eur J Neurol. 2017;24(5):741-747.
59. Kruszka P, Tanpaiboon P, Neas K, et al. Loss of function in ROBO1 is associated with tetralogy of Fallot and septal defects. J Med Genet. 2017;54(12):825-829.
60. Kuo DS, Sokol JT, Minogue PJ, et al. Characterization of a variant of gap junction protein α8 identified in a family with hereditary cataract. PLoS ONE. 2017;12(8):e0183438.
61. Lescai F, Als TD, Li Q, et al. Whole-exome sequencing of individuals from an isolated population implicates rare risk variants in bipolar disorder. Transl Psychiatry. 2017;7(2):e1034.
62. Lipstein N, Verhoeven-Duif NM, Michelassi FE, et al. Synaptic UNC13A protein variant causes increased neurotransmission and dyskinetic movement disorder. J Clin Invest. 2017;127(3):1005-1018.
63. Lucas SEM, Zhou T, Blackburn NB, et al. Rare, Potentially Pathogenic Variants in ZNF469 Are Not Enriched in Keratoconus in a Large Australian Cohort of European Descent. Invest Ophthalmol Vis Sci. 2017;58(14):6248-6256.
64. Lutter M, Bahl E, Hannah C, et al. Novel and ultra-rare damaging variants in neuropeptide signaling are associated with disordered eating behaviors. PLoS ONE. 2017;12(8):e0181556.
65. Meldau S, De Lacy RJ, Riordan GTM, et al. Identification of a single MPV17 nonsense-associated altered splice variant in 24 South African infants with mitochondrial neurohepatopathy. Clin Genet. 2018;93(5):1093-1096.
66. Merello E, Tattini L, Magi A, et al. Exome sequencing of two Italian pedigrees with non-isolated Chiari malformation type I reveals candidate genes for cranio-facial development. Eur J Hum Genet. 2017;25(8):952-959.
67. Minnerop M, Kurzwelly D, Wagner H, et al. Hypomorphic mutations in POLR3A are a frequent cause of sporadic and recessive spastic ataxia. Brain. 2017;140(6):1561-1578.
68. Mitropoulos K, Merkouri Papadima E, Xiromerisiou G, et al. Genomic variants in the FTO gene are associated with sporadic amyotrophic lateral sclerosis in Greek patients. Hum Genomics. 2017;11(1):30.
69. Moccia A, Srivastava A, Skidmore JM, et al. Genetic analysis of CHARGE syndrome identifies overlapping molecular biology. Genet Med. January 2018.

70. Nafisinia M, Sobreira N, Riley L, et al. Mutations in RARS cause a hypomyelination disorder akin to Pelizaeus-Merzbacher disease. Eur J Hum Genet. 2017;25(10):1134-1141.
71. Nuglozeh E. Whole-Exomes Sequencing Delineates Gene Variants Profile in a Young Saudi Male with Familial Hypercholesterolemia: Case Report. J Clin Diagn Res. 2017;11(6):GD01-GD06.
72. Okur V, Cho MT, Henderson L, et al. De novo mutations in CSNK2A1 are associated with neurodevelopmental abnormalities and dysmorphic features. Hum Genet. 2016;135(7):699-705.
73. Pal LR, Kundu K, Yin Y, Moult J. CAGI4 SickKids clinical genomes challenge: A pipeline for identifying pathogenic variants. Hum Mutat. 2017;38(9):1169-1181.
74. Park K, Seltzer LE, Tuttle E, Mirzaa GM, Paciorkowski AR. PLXNA1 developmental encephalopathy with syndromic features: A case report and review of the literature. Am J Med Genet A. May 2017.
75. Phadke SR, Kar A, Bhowmik AD, Dalal A. Complex Camptosynpolydactyly and Mesoaxial synostotic syndactyly with phalangeal reduction are allelic disorders. Am J Med Genet A. 2016;170(6):1622-1625.
76. Renkema GH, Visser G, Baertling F, et al. Mutated PET117 causes complex IV deficiency and is associated with neurodevelopmental regression and medulla oblongata lesions. Hum Genet. 2017;136(6):759-769.
77. Riedhammer KM, Siegel C, Alhaddad B, et al. Identification of a Novel Heterozygous De Novo 7-bp Frameshift Deletion in PBX1 by Whole-Exome Sequencing Causing a Multi-Organ Syndrome Including Bilateral Dysplastic Kidneys and Hypoplastic Clavicles. Front Pediatr. 2017;5:251.
78. Rios JJ, Delgado MR. Using whole-exome sequencing to identify variants inherited from mosaic parents. Eur J Hum Genet. 2015;23(4):547-550.
79. Ritelli M, Morlino S, Giacopuzzi E, et al. A recognizable systemic connective tissue disorder with polyvalvular heart dystrophy and dysmorphism associated with TAB2 mutations. Clin Genet. 2018;93(1):126-133.
80. Salih M, Gautschi I, van Bemmelen MX, et al. A Missense Mutation in the Extracellular Domain of αENaC Causes Liddle Syndrome. J Am Soc Nephrol. 2017;28(11):3291-3299.
81. Sampaio-Silva J, Batissoco AC, Jesus-Santos R, et al. Exome Sequencing Identifies a Novel Nonsense Mutation of MYO6 as the Cause of Deafness in a Brazilian Family. Ann Hum Genet. 2018;82(1):23-34.
82. Spiegler S, Rath M, Hoffjan S, et al. First large genomic inversion in familial cerebral cavernous malformation identified by whole genome sequencing. Neurogenetics. 2018;19(1):55-59.
83. Srour M, Shimokawa N, Hamdan FF, et al. Dysfunction of the Cerebral Glucose Transporter SLC45A1 in Individuals with Intellectual Disability and Epilepsy. Am J Hum Genet. 2017;100(5):824-830.
84. Stutterd C, Diakumis P, Bahlo M, et al. Neuropathology of childhood-onset basal ganglia degeneration caused by mutation of VAC14. Ann Clin Transl Neurol. 2017;4(12):859-864.
85. Tang F, Ma D, Wang Y, et al. Novel compound heterozygous mutations in the OTOF Gene identified by whole-exome sequencing in auditory neuropathy spectrum disorder. BMC Med Genet. 2017;18(1):35.
86. Tsai P-C, Soong B-W, Mademan I, et al. A recurrent WARS mutation is a novel cause of autosomal dominant distal hereditary motor neuropathy. Brain. 2017;140(5):1252-1266.

87. Upadia J, Oakes J, Hamm A, Hurst ACE, Robin NH. Foramen magnum compression in Coffin-Lowry syndrome: A case report. Am J Med Genet A. 2017;173(4):1087-1089.
88. Vissers LELM, Bonetti M, Paardekooper Overman J, et al. Heterozygous germline mutations in A2ML1 are associated with a disorder clinically related to Noonan syndrome. Eur J Hum Genet. 2015;23(3):317-324.
89. Wang J, Liu Y, Liu F, et al. Loss-of-function Mutation in PMVK Causes Autosomal Dominant Disseminated Superficial Porokeratosis. Sci Rep. 2016;6:24226.
90. Weiss K, Wigby K, Fannemel M, et al. Haploinsufficiency of ZNF462 is associated with craniofacial anomalies, corpus callosum dysgenesis, ptosis, and developmental delay. Eur J Hum Genet. 2017;25(8):946-951.
91. Whitford W, Hawkins I, Glamuzina E, et al. Compound heterozygous SLC19A3 mutations further refine the critical promoter region for biotin-thiamine-responsive basal ganglia disease. Cold Spring Harb Mol Case Stud. 2017;3(6).
92. Xi J, Yan C, Liu W-W, et al. Novel SEA and LG2 Agrin mutations causing congenital Myasthenic syndrome. Orphanet J Rare Dis. 2017;12(1):182.
93. Xiao X, Cao Y, Zhang Z, et al. Novel Mutations in PRPF31 Causing Retinitis Pigmentosa Identified Using Whole-Exome Sequencing. Invest Ophthalmol Vis Sci. 2017;58(14):6342-6350.
94. Xie H, Li X, Peng J, et al. A complex intragenic rearrangement of ERCC8 in Chinese siblings with Cockayne syndrome. Sci Rep. 2017;7:44271.
95. Yang L, Li Z, Mei M, et al. Whole genome sequencing identifies a novel ALMS1 gene mutation in two Chinese siblings with Alström syndrome. BMC Med Genet. 2017;18(1):75.
96. Yang Z, Li M, Hu X, et al. Rare damaging variants in DNA repair and cell cycle pathways are associated with hippocampal and cognitive dysfunction: a combined genetic imaging study in first-episode treatment-naive patients with schizophrenia. Transl Psychiatry. 2017;7(2):e1028.
97. Yıldız Bölükbaşı E, Mumtaz S, Afzal M, Woehlbier U, Malik S, Tolun A. Homozygous mutation in CEP19, a gene mutated in morbid obesity, in Bardet-Biedl syndrome with predominant postaxial polydactyly. J Med Genet. 2018;55(3):189-197.
98. You G, Zu B, Wang B, Wang Z, Xu Y, Fu Q. Exome Sequencing Identified a Novel FBN2 Mutation in a Chinese Family with Congenital Contractural Arachnodactyly. Int J Mol Sci. 2017;18(4).
99. Zehavi Y, Mandel H, Zehavi A, et al. De novo GRIN1 mutations: An emerging cause of severe early infantile encephalopathy. Eur J Med Genet. 2017;60(6):317-320.
100. Zeissig Y, Petersen B-S, Milutinovic S, et al. XIAP variants in male Crohn's disease. Gut. 2015;64(1):66-76.
101. Zhang T, Hou L, Chen DT, McMahon FJ, Wang J-C, Rice JP. Exome sequencing of a large family identifies potential candidate genes contributing risk to bipolar disorder. Gene. 2018;645:119-123.
102. Zhou T, Souzeau E, Sharma S, et al. Whole exome sequencing implicates eye development, the unfolded protein response and plasma membrane homeostasis in primary open-angle glaucoma. PLoS ONE. 2017;12(3):e0172427.

C.3 Contextual interview template

Interview rule: do not have to ask all the questions on the interview template.

Introduction (max. 20 min)

- Brief on interview process + videotaping
- Explain about the consent form
- Answer questions re: interview + consent
- Instruct participants to not disclose/display patient information (name, age etc)
- Ask about routine analysis
- What is the goal of your analysis?
- Could you describe your routine analysis in steps?
    - *Alt:* how you conduct your analysis?
    - *Alt:* could you describe your analysis pipeline?
- How many exome/genome cases have you analyzed to date?
- What types of analysis do you conduct?
    - *e.g.* trio, singleton, cohort
- What type of data do you use in your analysis?
    - *e.g.* sequencing quality, variant quality, coverage analysis, functional annotation (synonymous/nonsynonymous), variant frequency in population db
    - *Alt:* what kind of tools or programs do you use for your analyses?
    - *Follow-up:* do you have any particular order in looking at this data?
- Which task takes the most time during analyses?
- Do you prefer to work with any particular format of data?
    - *e.g.* Excel spreadsheet, SQL

Observation (est. 1 – 2 hrs; 2 hrs max)

- Ask for clarification
    - Why do you use tool X?
    - What does the tool X do?
    - What data are you looking at?

Follow-up interview (est. 40 min – 1.5 hrs)

- Go over observation
    - Uses visualization
        - I saw that you use X in Y context. Am I correct?
            - *Follow-up:* how useful do you find X?
            - *Follow-up:* How easy was it for you to learn how to use X?
            - *Follow-up:* How did you hear about X?
            - *Follow-up:* What are some challenges or usage barriers that you encountered while using X?
            - *Follow-up:* Where else do you use X for? (i.e. do you use X in a different context?)

                - Are there any visualization tools that you regularly use but you didn't use during today's analysis?
        - Does not use visualization
            - I saw that you don't use visualizations. Could you tell me why?
                - *Alt:* why do you prefer tabular data?
            - If you don't use visualizations, how useful do you find the current set of tools that you use?
            - Are there any tools that you regularly use but didn't use during today's analysis?
- Other visualizations/tools: show a catalogue of commonly used data types. Ask if there are any additional data types that are used.
    - Sequencing quality
    - Variant quality
    - Coverage analysis (e.g. coverage of targeted regions/consensus coding sequence)
    - Relatedness assessment (e.g. KING)
    - Functional annotation (e.g. synonymous/nonsynonymous, nonsense/missense, frameshift)
    - Location of variant (e.g. overlap with disease-associated region, within exon/intron, exon-intron boundaries)
    - Variant frequency in population databases (e.g. dbSNP, 1000 Genomes, gnomAD)
    - Variant frequency in in-house databases
    - in-silico functional prediction (e.g. SIFT, PolyPhen2, CADD)
    - Nucleotide conservation (e.g. GERP, PhyloP)
    - Splice-site prediction (e.g. NNSPLICE, MaxEntScan)
    - Inheritance model
    - Known gene-disease association
    - Human Phenotype Ontology-based phenotype/gene similarity
    - Overlapping or similar phenotypes in disease databases (e.g. OMIM, DECIPHER)
    - Presence and designation in disease-focused variation databases (e.g. ClinVar, LOVD)
    - Interaction with known disease-associated gene
    - Other
- Ask additional questions
    - Re: previous search
        - Have you looked for data visualizations for your analysis in the past? If so what are they? If not, why?
    - Show a catalogue of visualization tools
        - Integrative Genomics Viewer (IGV)
        - UCSC Genome Browser
        - Protein structure visualizations (e.g. Chimera)
        - Network visualizations (e.g. Cytoscape, GeneMANIA)
        - Phenotype-driven visual prioritization tools (e.g. OMIM Explorer)
        - Phenotype comparison visualizations (e.g. PhenoBlocks)
        - Custom R visualizations

        - Visualization features within population databases (e.g. read data browser in GnomAD/ExAC, graphical sequence viewer in dbSNP)
        - Visualization features within sequence databases (e.g. graphical sequence viewer in NCBI Gene, Feature viewer in UniProt)
        - Visualization features within disease databases (e.g. protein browser or phenotype browser in DECIPHER)
        - Visualization features within commercial variant analysis tools (e.g. Alamut Visual, SnapGene)
- Ask the following
    - Have you tried or heard about X?
        - Did try or hear about X
            - How useful did you find it to be?
            - How easy was for you to learn?
            - If you tried or heard about it but don't use it in your analysis, could you tell me why?
        - Never heard of X
            - Explain what X does
            - How useful do you think X would be?
- Suggestions for future visualization
    - You said that you use data X, Y, Z for your analysis. Is there any tasks or data types that you think a visualization would be helpful for your analysis?
        - *Follow-up:* how would it be helpful?
        - *Follow-up:* is there any tasks or data types that you don't think a visualization would be helpful or necessary for your analysis?
        - *Follow-up:* why would it not be helpful?
    - Can you think of any analysis tools that you would find more useful if it incorporated a visualization?

C.4 Online survey questionnaire

Routine exome/genome analysis
In the following section, we would like to ask basic questions about your exome/genome analyses for rare disease diagnosis.

1. How many rare disease cases with exome/genome sequencing data have you analyzed to date?
    1. Less than or equal to 50 cases
    2. 51 – 100 cases
    3. 101 – 200 cases
    4. More than 200 cases
    5. Other
    6. N/A

2. What types of analysis do you conduct? Please select all that apply.
    1. Singleton
    2. Trio
    3. Cohort
    4. Other
    5. N/A

3. What types of sequencing data do you work with? Please select all that apply.
    1. Whole exome sequencing
    2. Whole genome sequencing
    3. Other
    4. N/A

4. In which setting are the analyses conducted? Please select all that apply.
    1. Clinical
    2. Research
    3. Other
    4. N/A

5. Below are types of data that are commonly used in exome/genome analyses. Please **select all** types of data that you use in your analyses.
    1. Sequencing quality
    2. Variant quality
    3. Coverage analysis (e.g. coverage of targeted regions/consensus coding sequence)
    4. Relatedness assessment (e.g. KING)
    5. Functional annotation (e.g. synonymous/nonsynonymous, nonsense/missense, frameshift)
    6. Location of variant (e.g. overlap with disease-associated region, within exon/intron, exon-intron boundaries)
    7. Variant frequency in population databases (e.g. dbSNP, 1000 Genomes, gnomAD)
    8. Variant frequency in in-house databases
    9. *in-silico* functional prediction (e.g. SIFT, PolyPhen2, CADD)
    10. Nucleotide conservation (e.g. GERP, PhyloP)

11. Splice-site prediction (e.g. NNSPLICE, MaxEntScan)
12. Inheritance model
13. Known gene-disease association
14. Human Phenotype Ontology-based phenotype/gene similarity
15. Overlapping or similar phenotypes in disease databases (e.g. OMIM, DECIPHER)
16. Presence and designation in disease-focused variation databases (e.g. ClinVar, LOVD)
17. Interaction with known disease-associated gene
18. Other: tell us any data that you routinely use but are not listed above
19. N/A

Data visualizations that are currently used during routine analyses
In the following section, we would like to ask about data visualization tools or analysis tools with data visualization features that **you currently use** for your routine analyses.

6a. Below are data visualization tools or types of data visualizations that are commonly used in exome/genome analyses.

Please **select all** tools or visualizations that **you currently use for your routine analyses**. If applicable, please provide the name of the tool in the textbox provided underneath the appropriate category.

1. Integrative Genomics Viewer (IGV)
2. UCSC Genome Browser
3. Protein structure visualizations (e.g. Chimera)
4. Network visualizations (e.g. Cytoscape, GeneMANIA)
5. Phenotype-driven visual prioritization tools (e.g. OMIM Explorer)
6. Phenotype comparison visualizations (e.g. PhenoBlocks)
7. Custom R visualizations
8. Visualization features within population databases (e.g. read data browser in gnomAD/ExAC, graphical sequence viewer in dbSNP)
9. Visualization features within sequence databases (e.g. graphical sequence viewer in NCBI Gene, Feature viewer in UniProt)
10. Visualization features within disease databases (e.g. protein browser or phenotype browser in DECIPHER)
11. Visualization features within commercial variant analysis tools (e.g. Alamut Visual, SnapGene)
12. Other: tell us about any tools or data visualizations that you use but are not listed above
13. I do not use any data visualization tools or data visualizations for my routine analyses
14. N/A

6b. (Show if participant answered that they currently use some visualizations; show only the tools they selected in Q6a) Below are the data visualizations that you have selected in the previous question. For **each** tool or visualization, in what context do you use it? (e.g. I use IGV for manual inspection of mapped reads)

1. Integrative Genomics Viewer (IGV)
2. UCSC Genome Browser
3. Protein structure visualizations (e.g. Chimera)

4. Network visualizations (e.g. Cytoscape, GeneMANIA)
5. Phenotype-driven visual prioritization tools (e.g. OMIM Explorer)
6. Phenotype comparison visualizations (e.g. PhenoBlocks)
7. Custom R visualizations
8. Visualization features within population databases (e.g. read data browser in gnomAD/ExAC, graphical sequence viewer in dbSNP)
9. Visualization features within sequence databases (e.g. graphical sequence viewer in NCBI Gene, Feature viewer in UniProt)
10. Visualization features within disease databases (e.g. protein browser or phenotype browser in DECIPHER)
11. Visualization features within commercial variant analysis tools (e.g. Alamut Visual, SnapGene)
12. Other

6c. (Show if participant answered that they do not use any visualizations) If you don't use any visualizations, could you describe why?

Data visualizations that are available but are not currently used during routine analyses
In the following section, we would like to ask about data visualization tools or analysis tools with data visualization features that **you have tried or may know about, but do not use** for your routine analyses.

7a. Have you ever tried or looked for data visualizations to use for your routine analyses?
1. Yes
2. No
3. Other
4. N/A

7b. (Show if participant answered yes to Q7a; eliminate answers that participant selected in Q6a) Below are data visualization tools or types of data visualizations that are commonly used in exome/genome analyses, aside from the ones which you indicated that you currently use for your analyses.

Please **select all** tools or visualizations that you **have tried or know about but do not currently use** for your routine analyses. If applicable, please provide the name of the tool in the textbox provided underneath the appropriate category.
1. Integrative Genomics Viewer (IGV)
2. UCSC Genome Browser
3. Protein structure visualizations (e.g. Chimera)
4. Network visualizations (e.g. Cytoscape, GeneMANIA)
5. Phenotype-driven visual prioritization tools (e.g. OMIM Explorer)
6. Phenotype comparison visualizations (e.g. PhenoBlocks)
7. Custom R visualizations
8. Visualization features within population databases (e.g. read data browser in gnomAD/ExAC, graphical sequence viewer in dbSNP)

9. Visualization features within sequence databases (e.g. graphical sequence viewer in NCBI Gene, Feature viewer in UniProt)
10. Visualization features within disease databases (e.g. protein browser or phenotype browser in DECIPHER)
11. Visualization features within commercial variant analysis tools (e.g. Alamut Visual, SnapGene)
12. Other: tell us about any tools or visualizations that you have encountered but are not listed above
13. N/A

7c. (Show if participant answered no to Q7a) If you have never tried or looked for data visualizations, could you describe why?

7d. (Show if participants selected any options in Q7b; show only the tools selected in Q7b) Below are the data visualizations that you have selected in the previous question. For **each** tool or visualization, could you describe why they are not used?

1. Integrative Genomics Viewer (IGV)
2. UCSC Genome Browser
3. Protein structure visualizations (e.g. Chimera)
4. Network visualizations (e.g. Cytoscape, GeneMANIA)
5. Phenotype-driven visual prioritization tools (e.g. OMIM Explorer)
6. Phenotype comparison visualizations (e.g. PhenoBlocks)
7. Custom R visualizations
8. Visualization features within population databases (e.g. read data browser in gnomAD/ExAC, graphical sequence viewer in dbSNP)
9. Visualization features within sequence databases (e.g. graphical sequence viewer in NCBI Gene, Feature viewer in UniProt)
10. Visualization features within disease databases (e.g. protein browser or phenotype browser in DECIPHER)
11. Visualization features within commercial variant analysis tools (e.g. Alamut Visual, SnapGene)
12. Other: tell us about any visualizations that you have encountered but are not provided above

Suggestions for future data visualizations

In the following section, we would like to ask about improvement ideas or feature suggestions for data visualization tools for exome/genome analyses.

8a. (Show based on participants answer to Q5) Below are the types of data which you indicated that you commonly use for your routine analyses.

For **each** data type, please indicate whether you think a visualization would be **helpful or not helpful** for your analyses. (Display options as table)

1. Sequencing quality
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
2. Variant quality
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
3. Coverage analysis (e.g. coverage of targeted regions/consensus coding sequence)
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
4. Relatedness assessment (e.g. KING)
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
5. Functional annotation (e.g. synonymous/nonsynonymous, nonsense/missense, frameshift)
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
6. Location of variant (e.g. overlap with disease-associated region, within exon/intron, exon-intron boundaries)
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
7. Variant frequency in population databases (e.g. dbSNP, 1000 Genomes, gnomAD)
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
8. Variant frequency in in-house databases
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
9. *in-silico* functional prediction (e.g. SIFT, PolyPhen2, CADD)
   i. Visualization would be helpful
   ii. Visualization would not be helpful
   iii. N/A
10. Nucleotide conservation (e.g. GERP, PhyloP)
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
11. Splice-site prediction (e.g. NNSPLICE, MaxEntScan)
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A

12. Inheritance model
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
13. Known gene-disease association
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
14. Human Phenotype Ontology-based phenotype/gene similarity
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
15. Overlapping or similar phenotypes in disease databases (e.g. OMIM, DECIPHER)
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
16. Presence and designation in variation databases (e.g. ClinVar, LOVD)
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
17. Interaction with known disease-associated gene
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A
18. Other: <show what participant wrote in Q5>
    i. Visualization would be helpful
    ii. Visualization would not be helpful
    iii. N/A

8b. (Show if participants indicated as helpful in Q8a; Show only the options selected for Q8a)
For **each** of the data types which you indicated that a visualization would be **helpful** for your analyses, could you describe why? Also, what kind of visualization would make it helpful?
Sequencing quality

1. Variant quality
2. Coverage analysis (e.g. coverage of targeted regions/consensus coding sequence)
3. Relatedness assessment (e.g. KING)
4. Functional annotation (e.g. synonymous/nonsynonymous, nonsense/missense, frameshift)
5. Location of variant (e.g. overlap with disease-associated region, within exon/intron, exon-intron boundaries)
6. Variant frequency in population databases (e.g. dbSNP, 1000 Genomes, gnomAD)
7. Variant frequency in in-house databases
8. *in-silico* functional prediction (e.g. SIFT, PolyPhen2, CADD)
9. Nucleotide conservation (e.g. GERP, PhyloP)
10. Splice-site prediction (e.g. NNSPLICE, MaxEntScan)
11. Inheritance model
12. Known gene-disease association

13. Human Phenotype Ontology-based phenotype/gene similarity
14. Overlapping or similar phenotypes in disease databases (e.g. OMIM, DECIPHER)
15. Presence and designation in variation databases (e.g. ClinVar, LOVD)
16. Interaction with known disease-associated gene
17. Other: <show what participant wrote in Q5>

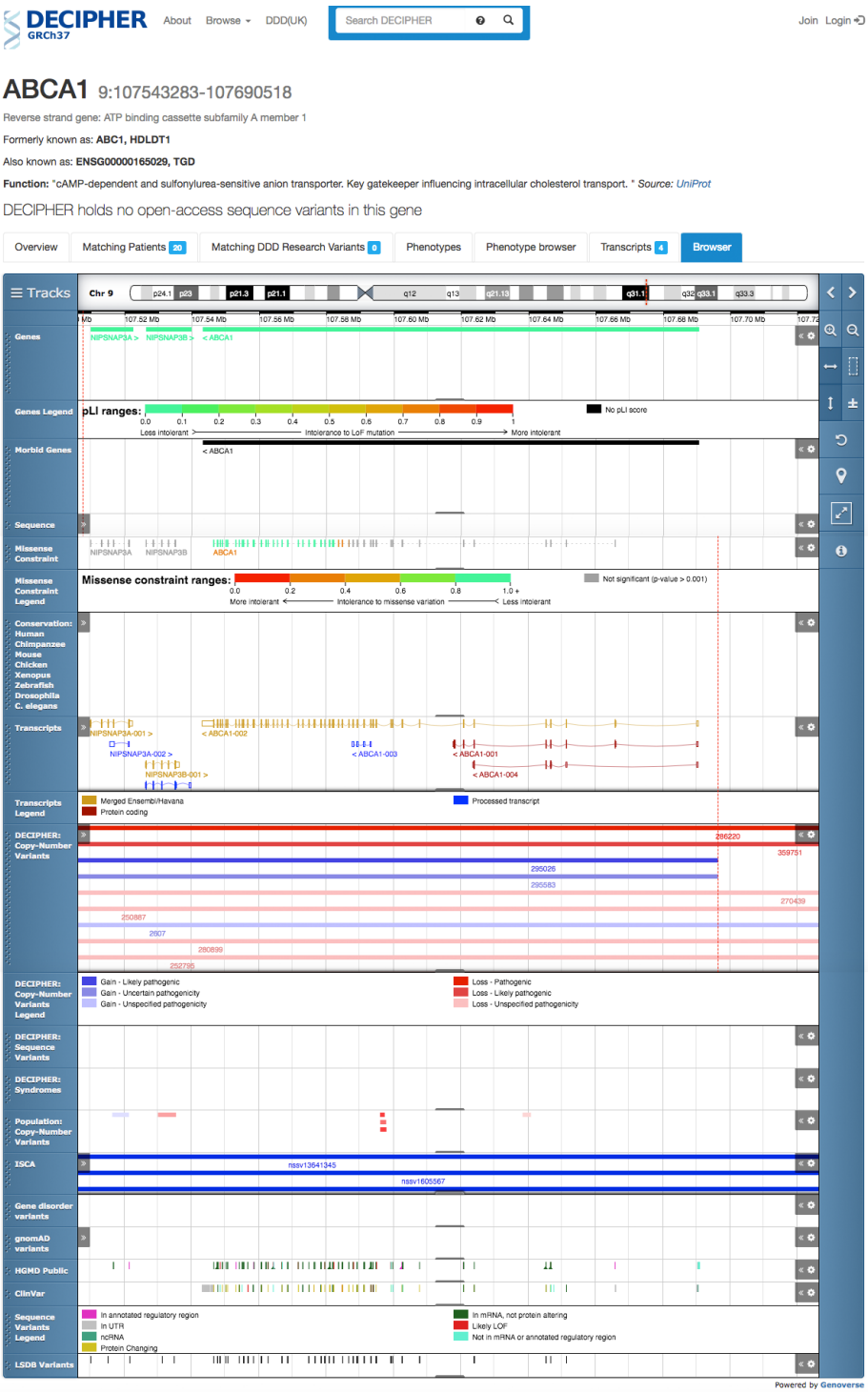
8c. (Show if participants indicated as not helpful in Q8a; Show only the options selected for Q8a) For **each** of the data types which you indicated that a visualization would be **not helpful** for your analyses, could you describe why?

1. Sequencing quality
2. Variant quality
3. Coverage analysis (e.g. coverage of targeted regions/consensus coding sequence)
4. Relatedness assessment (e.g. KING)
5. Functional annotation (e.g. synonymous/nonsynonymous, nonsense/missense, frameshift)
6. Location of variant (e.g. overlap with disease-associated region, within exon/intron, exon-intron boundaries)
7. Variant frequency in population databases (e.g. dbSNP, 1000 Genomes, gnomAD)
8. Variant frequency in in-house databases
9. *in-silico* functional prediction (e.g. SIFT, PolyPhen2, CADD)
10. Nucleotide conservation (e.g. GERP, PhyloP)
11. Splice-site prediction (e.g. NNSPLICE, MaxEntScan)
12. Inheritance model
13. Known gene-disease association
14. Human Phenotype Ontology-based phenotype/gene similarity
15. Overlapping or similar phenotypes in disease databases (e.g. OMIM, DECIPHER)
16. Presence and designation in variation databases (e.g. ClinVar, LOVD)
17. Interaction with known disease-associated gene
18. Other: <show what participant wrote in Q5>

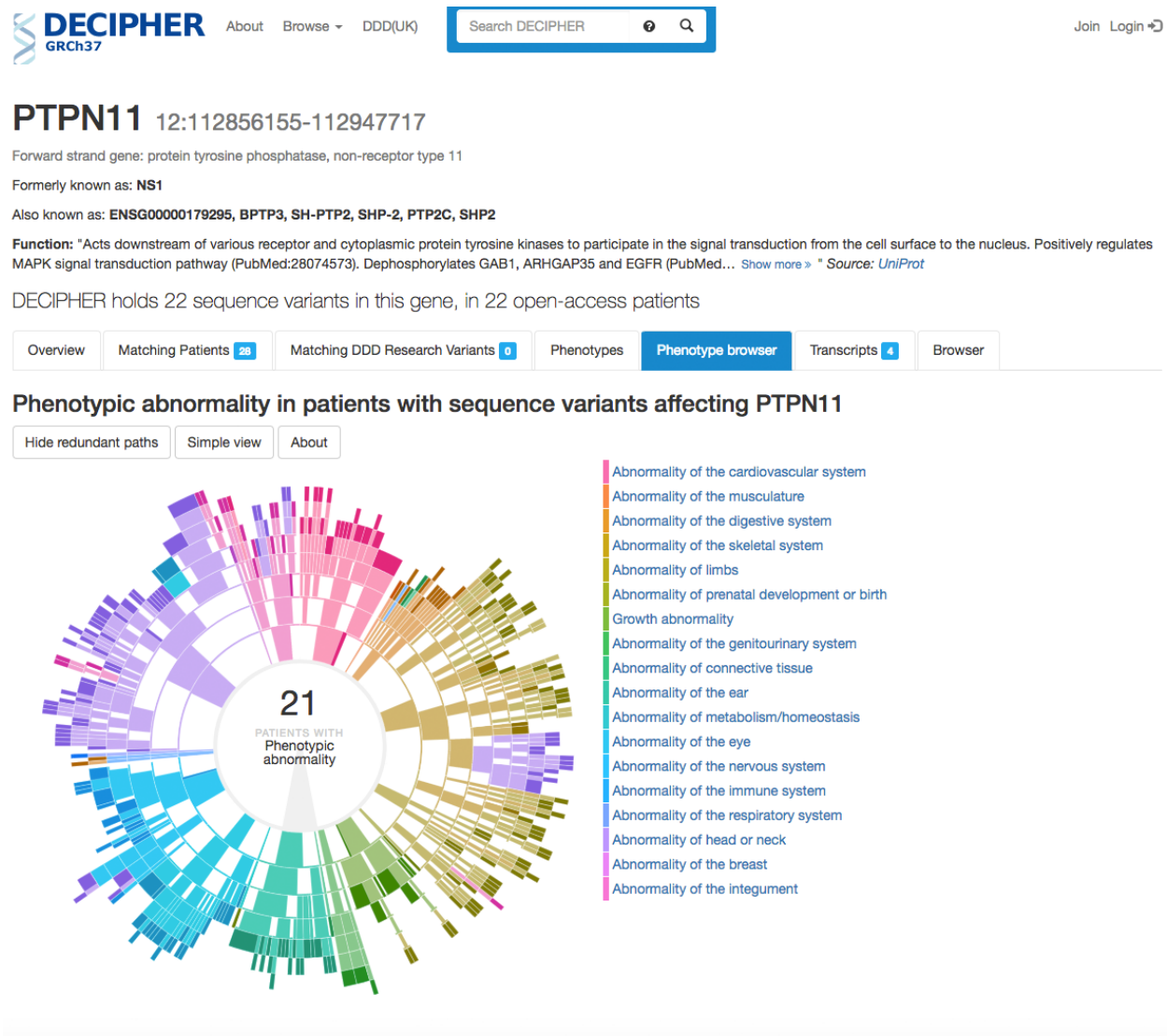
9. Please share new visualization ideas if you are willing.

C.5 Screenshots of information visualization tools captured in this study (excluding non-open-access/offline tools)

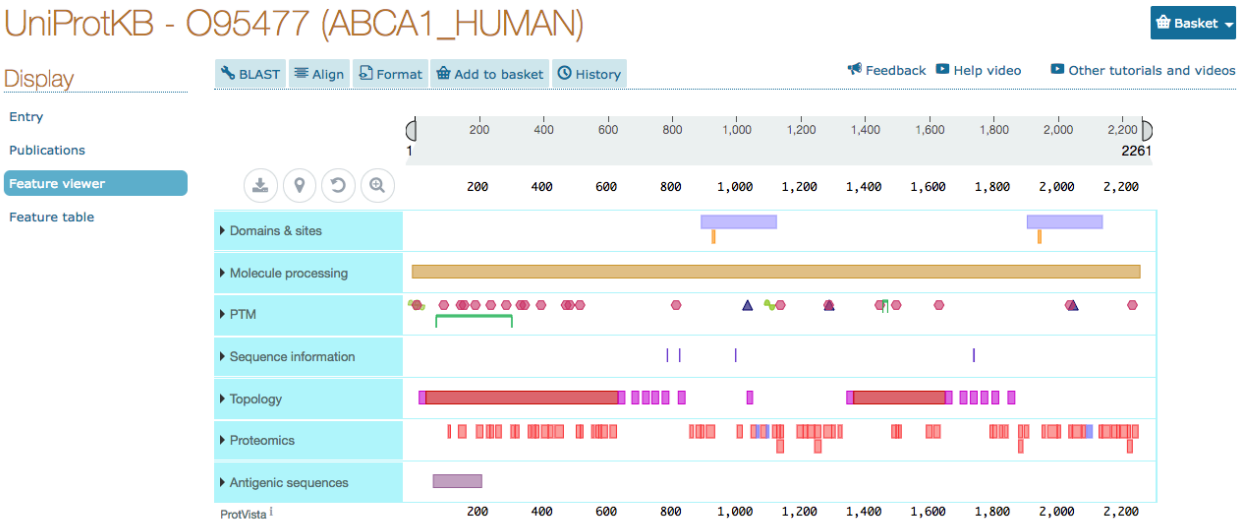
1a. Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER): genome browser


DECIPHER
GRCh37
About
Browse
DDD(UK)
Search DECIPHER
Join
Login
ABCA1 9:107543283-107690518
Reverse strand gene: ATP binding cassette subfamily A member 1
Formerly known as: ABC1, HDLDT1
Also known as: ENSG00000165029, TGD
Function: "cAMP-dependent and sulfonylurea-sensitive anion transporter. Key gatekeeper influencing intracellular cholesterol transport. " Source: UniProt
DECIPHER holds no open-access sequence variants in this gene
Overview
Matching Patients 20
Matching DDD Research Variants 0
Phenotypes
Phenotype browser
Transcripts 4
Browser
Tracks
Chr 9
Genes
Genes Legend
pLI ranges:
No pLI score
Less intolerant
Intolerance to LoF mutation
More intolerant
Morbid Genes
Sequence
Missense Constraint
Missense Constraint Legend
Missense constraint ranges:
Not significant (p-value > 0.001)
More intolerant
Intolerance to missense variation
Less intolerant
Conservation:
Human
Chimpanzee
Mouse
Chicken
Xenopus
Zebrafish
Drosophila
C. elegans
Transcripts
Transcripts Legend
Merged Ensembl/Havana
Protein coding
Processed transcript
DECIPHER: Copy-Number Variants
DECIPHER: Copy-Number Variants Legend
Gain - Likely pathogenic
Gain - Uncertain pathogenicity
Gain - Unspecified pathogenicity
Loss - Pathogenic
Loss - Likely pathogenic
Loss - Unspecified pathogenicity
DECIPHER: Sequence Variants
DECIPHER: Syndromes
Population: Copy-Number Variants
ISCA
nssv13641345
nssv1605567
Gene disorder variants
gnomAD variants
HGMD Public
ClinVar
Sequence Variants Legend
In annotated regulatory region
In UTR
ncRNA
Protein Changing
In mRNA, not protein altering
Likely LOF
Not in mRNA or annotated regulatory region
LSDB Variants
Powered by Genoverse

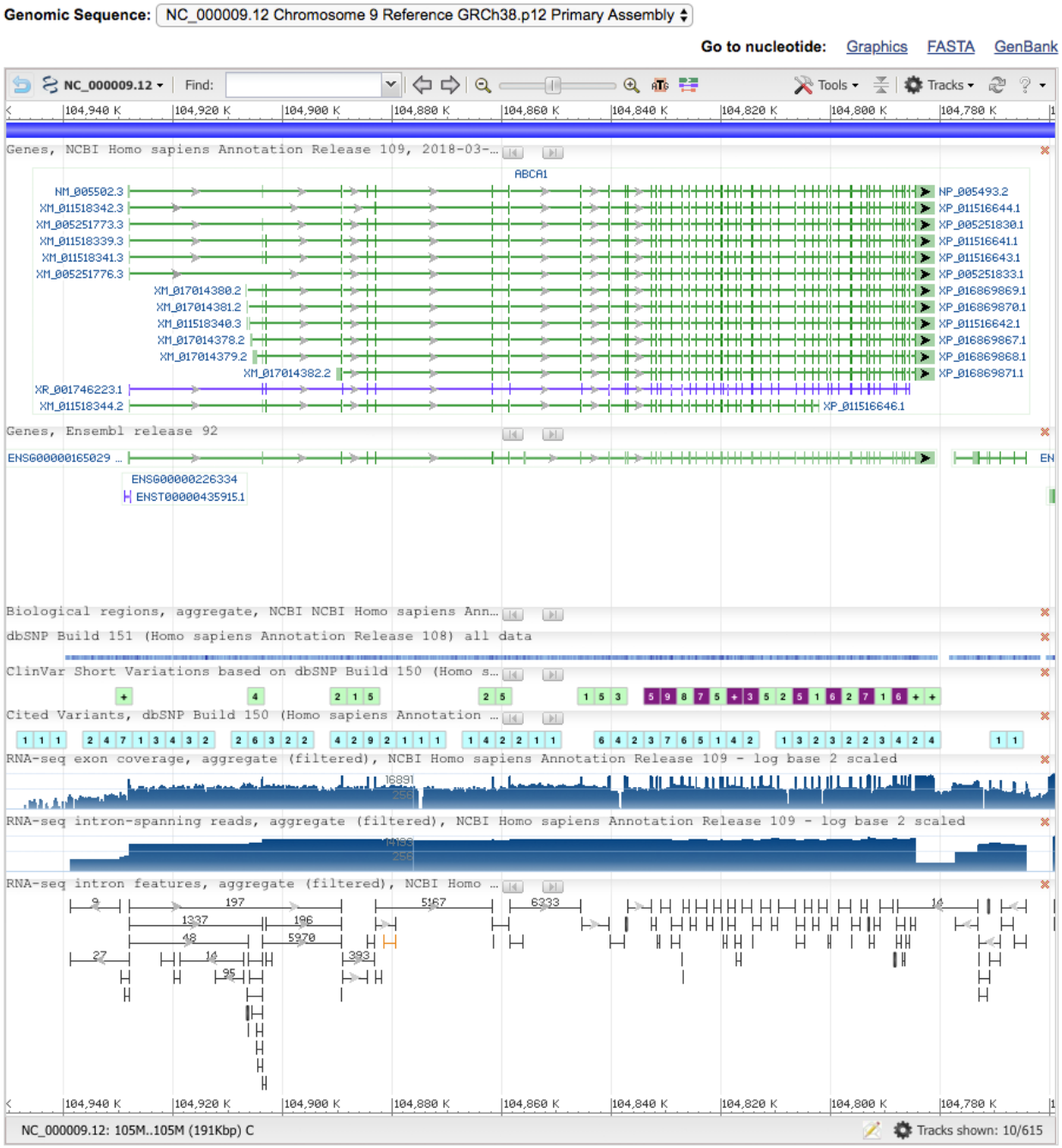
1b. Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER): phenotype browser



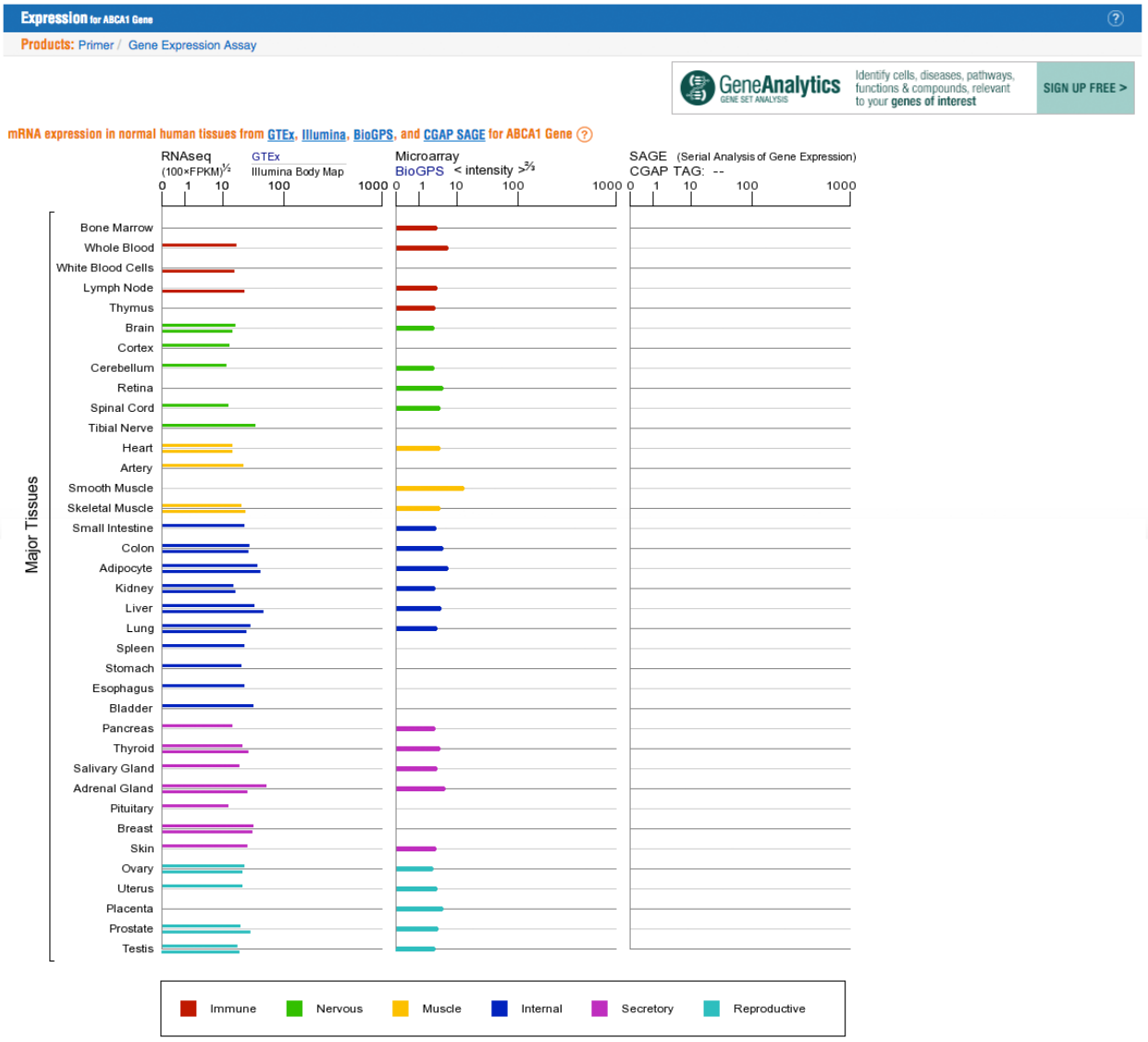
2. UniProt: feature viewer



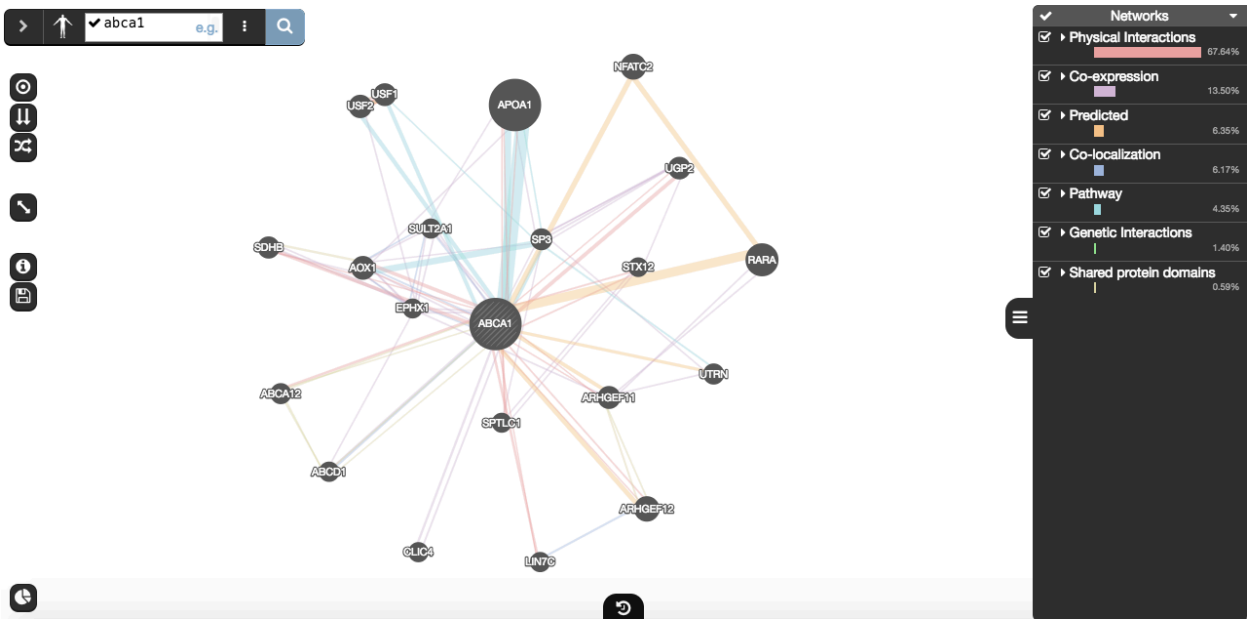
3. National Center for Biotechnology Information (NCBI) Gene: graphical sequence viewer



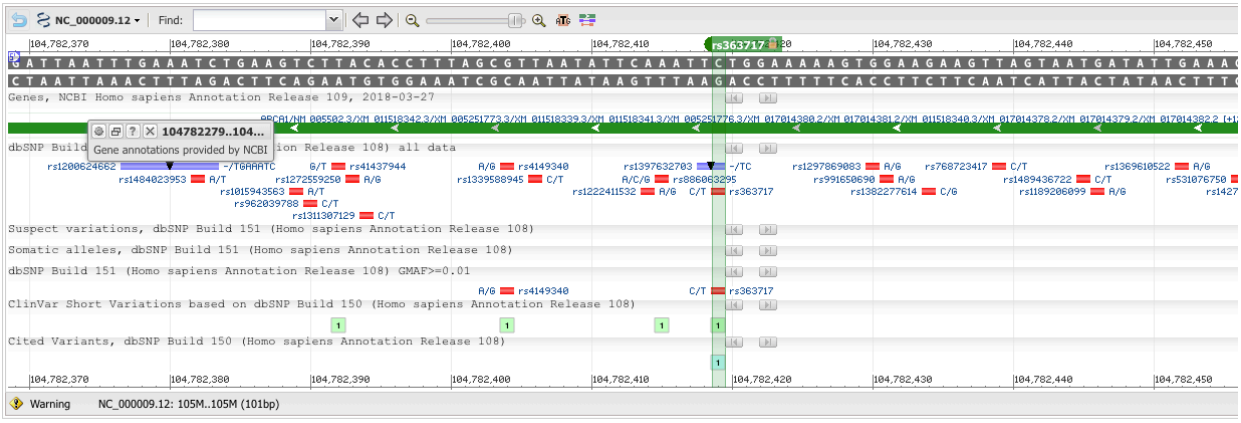
4. GeneCards: mRNA expression plot



5. GeneMANIA: pathway/protein-protein interaction visualization

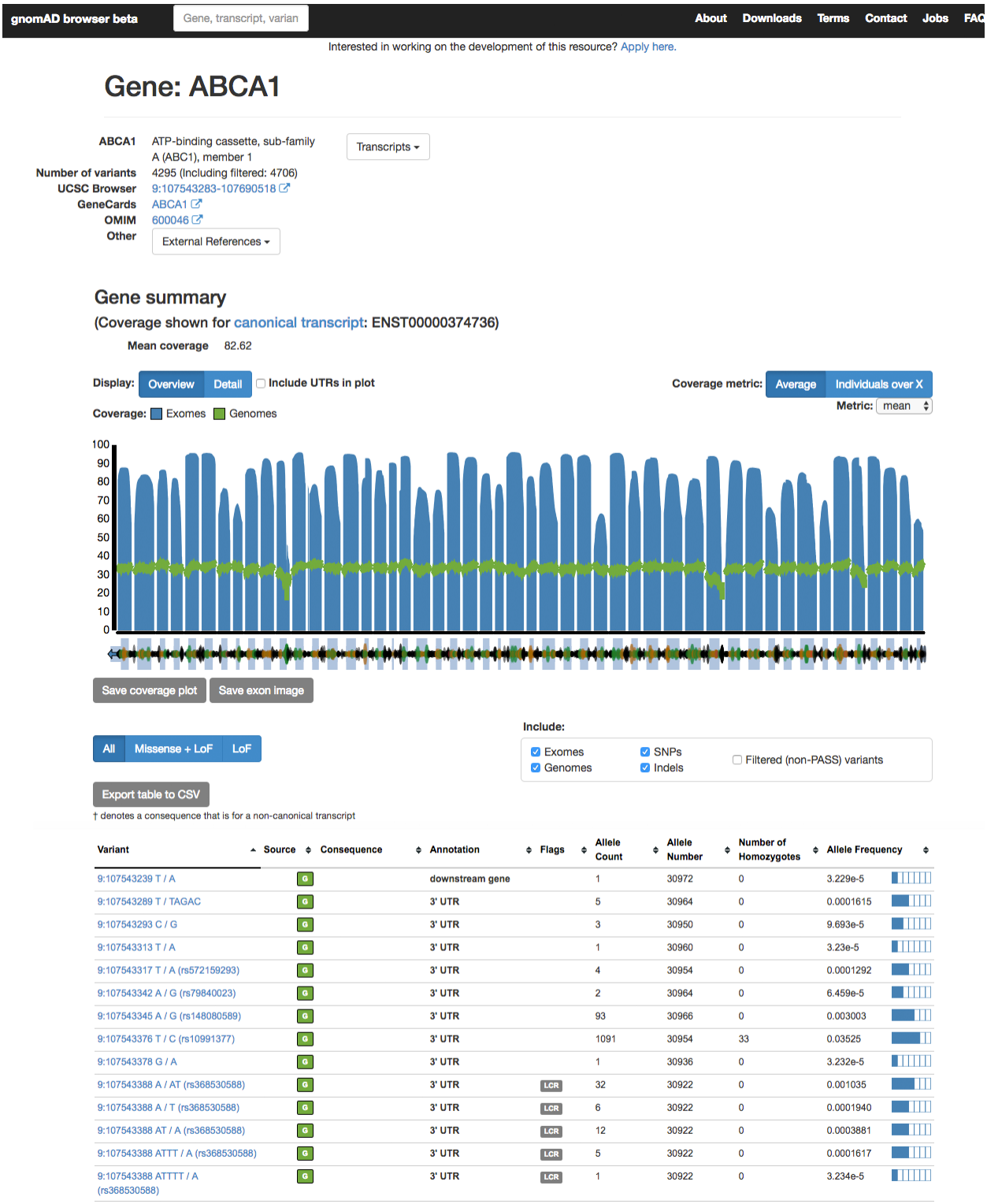


6. Single Nucleotide Polymorphism database (dbSNP): Graphical sequence viewer

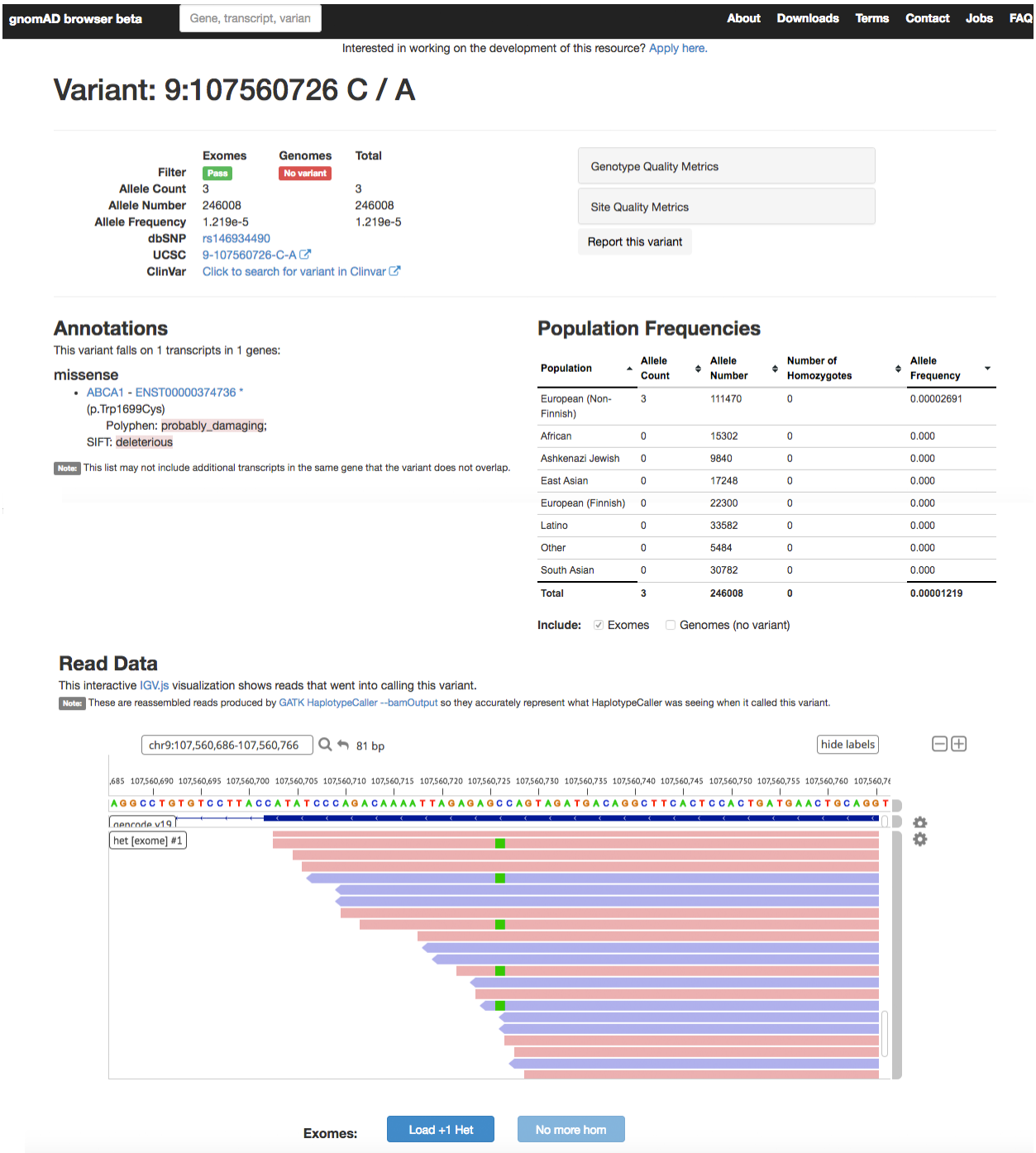


GeneView via direct blast against RefSeq sequences (used when no gene model is available): N/A

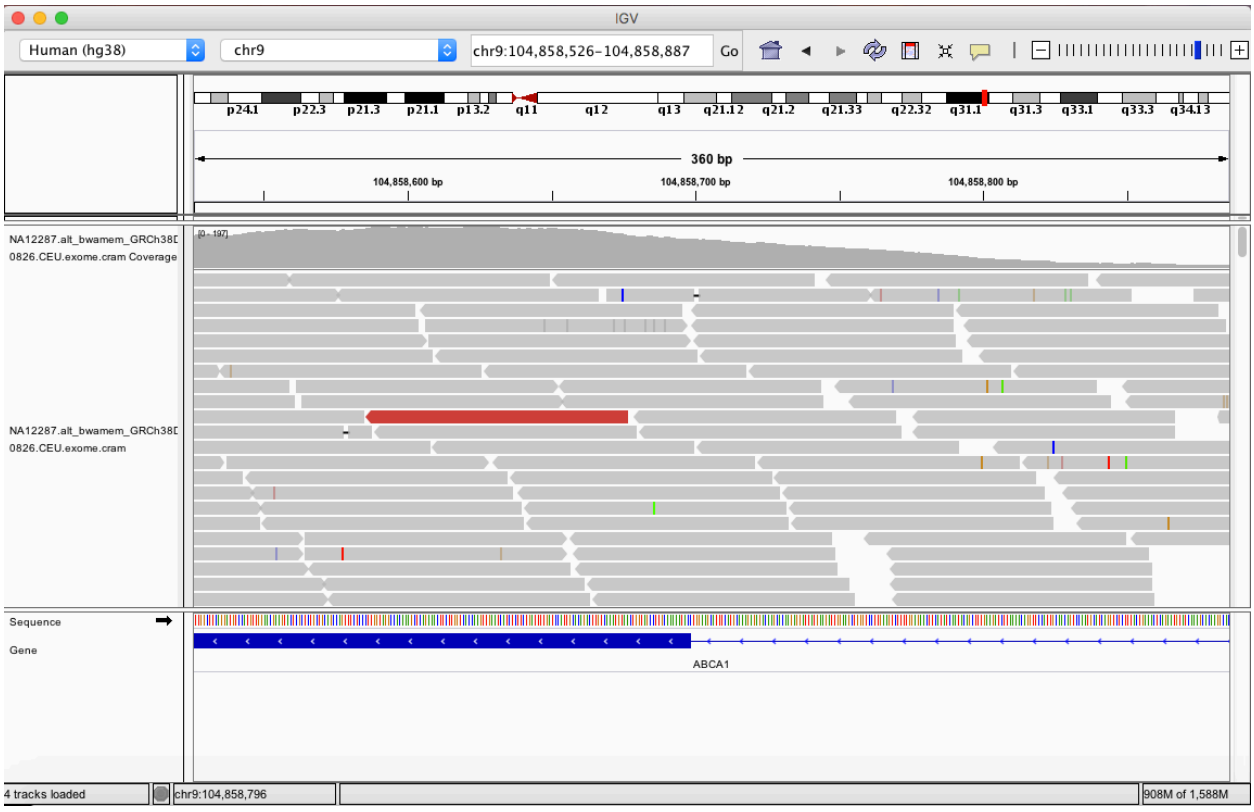
7a. Genome Aggregation Database (gnomAD): illustrated gene summary



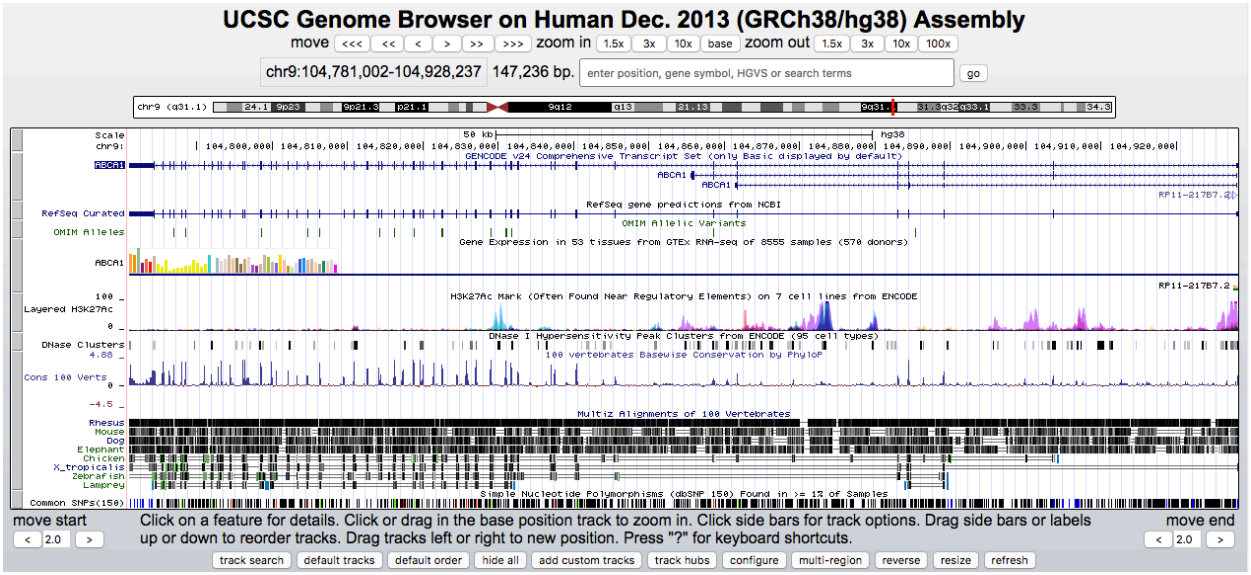
7b. Genome Aggregation Database (gnomAD): read data browser



8. Integrative Genomics Viewer (IGV): read alignment visualization



9. University of California Santa Cruz (UCSC) Genome Browser: arbitrary/custom annotation visualization



Commonly used data types	Reasons for considering visualization would be helpful	Reasons for considering visualization would be not helpful
Sequencing quality **Variant quality**	*"Seeing the number of mismatches in a read with a variant is helpful"* *"I like context/adjacency when assessing things"* *"Easier to understand using visualization than going through values/technical details"*	*"One should have this before variant interpretation and if passing then doesn't need to be looked at further."*
Coverage analysis	*"Seeing how coverage in the regions compared to overall is helpful"* *"Probably would be the most useful to have this (coverage information) visualized and searchable for genes in general."*	-
Functional annotation	*"We currently use Alamut for this; however, it would be helpful to have visualizations of protein structure changes."*	*"A number or yes/ no is ok"* *"Statement of annotation works just fine no need to complicate it with colors or charts"* *"Easy to look up within data. Visualization would not add more to already present data."*
Location of variant	*"You can, to some extent, see this in UCSC. It would be nice to see if specific variants still overlap with disease associated regions."* *"Good way to visualize what is present in the region"*	-
Variant frequency in population databases **Variant frequency in in-house databases**	-	*"All I need to see are numbers"* *"% or statement of AF is sufficient"* *"Easy to look up within data. Visualization would not add more to already present data."*
***in-silico* functional prediction**	-	*"A visualization is not necessary - setting thresholds with the numbers given from these predictors is all I need to use"* *"Statement of prediction is sufficient"*

		"Easy to look up within data. Visualization would not add more to already present data."
Nucleotide conservation	*"We have this visualization in Alamut and can be helpful for interpreting novel variants."*	*"Easy to look up within data. Visualization would not add more to already present data."*
Splice-site prediction	*"We have this visualization in Alamut and can be helpful for interpreting splicing since you can see the 5 splice predictors at the same time."*	-
Inheritance model	-	*"Statement is sufficient"* *"This is something that you need to look up - not sure how a visualization will help."*
Known gene-disease association	*"This could be useful to determine if there are multiple diseases associated with the same gene, or multiple phenotypes described for the same gene."*	*"Statement is sufficient"* *"This is something that you need to look up - not sure how a visualization will help."*
Human Phenotype Ontology-based phenotype/gene similarity	-	*"Numbers are useful for this. Could a gene similarity correlation be done?"* *"This is something that you need to look up - not sure how a visualization will help."*
Overlapping or similar phenotypes in disease databases	*"It would be nice to have both OMIM and DECIPHER data together in one platform to visualize the data."* *"Usually phenotypes/diseases are listed - it'd be easier to see diseases (sorted by similarity e.g. neuro, musculoskeletal, cardio) in a visualization instead."*	-
Presence and designation in disease-focused variation databases	-	*"This doesn't need to be visualized. What would be nice is if all ClinVar entries for a specific variant were contained in one link/page."* *"Statement is sufficient"* *"Easy to look up within data. Visualization would not add more to already present data."*

Table C1 Explanations provided by online survey participants (n = 17) regarding why the commonly used data types would be helpful or not helpful to visualize.

Observation	Recommendation	Applies to	Benefit
After variant filtering, participants used custom rules for quickly deciding if a filtered variant should be further assessed	Support creation of custom decision rules and color-code variant annotations accordingly (similar to conditional formatting function in spreadsheet software)	Tools that support browsing of filtered variants Tools/resources that support browsing of a table of genomic features	Reduce the cognitive burden of assessing multiple numeric/textual values
Participants used variant annotations in two ways: (a) They examined numeric/textual values (b) They assessed the information in combination with other layers of information in a genome browser	Support simultaneous presentation of a genome browser and a table of values	Tools that support browsing of filtered variants Tools/resources that present a summary of variants/genomic features within a regional context (e.g. list of known pathogenic variants within a gene)	Ease the cognitive transition between consideration of single evidence and multi-layered evidence
Participants' analyses were two-tiered: (a) first-tier: disease-associated variants (b) second-tier: all other variants	Present each tier in a separate view or panel, (similar to a worksheet in spreadsheet software). Within each view, support custom curation of analysis tools/features that are frequently used during the respective tier of analysis	Tools that support browsing of filtered variants Tools/resources that curate multiple sources of information	Enhances utility of curated information and software features as visual presentation aligns with the context of analysis
Participants take notes on each variant examined	Support creation of notes on each variant/gene	Tools that support browsing of filtered variants	Helps keeping track of the analyses

Table C2 Information visualization design recommendations extracted from observation of interview participants (n = 6).