SPECIAL ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Matching whole genomes to rare genetic disorders: Identification of potential causative variants using phenotype-weighted knowledge in the CAGI SickKids5 clinical genomes challenge

Lipika R. Pal[1] | Kunal Kundu[1,2] | Yizhou Yin[1] | John Moult[1,3]

[1]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland

[2]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, Maryland

[3]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland

Correspondence
Lipika R. Pal, and John Moult, Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850.
Email: lray1@umd.edu (LP) and jmoult@umd.edu (JM)

Abstract

Precise identification of causative variants from whole-genome sequencing data, including both coding and noncoding variants, is challenging. The Critical Assessment of Genome Interpretation 5 SickKids clinical genome challenge provided an opportunity to assess our ability to extract such information. Participants in the challenge were required to match each of the 24 whole-genome sequences to the correct phenotypic profile and to identify the disease class of each genome. These are all rare disease cases that have resisted genetic diagnosis in a state-of-the-art pipeline. The patients have a range of eye, neurological, and connective-tissue disorders. We used a gene-centric approach to address this problem, assigning each gene a multiphenotype-matching score. Mutations in the top-scoring genes for each phenotype profile were ranked on a 6-point scale of pathogenicity probability, resulting in an approximately equal number of top-ranked coding and noncoding candidate variants overall. We were able to assign the correct disease class for 12 cases and the correct genome to a clinical profile for five cases. The challenge assessor found genes in three of these five cases as likely appropriate. In the postsubmission phase, after careful screening of the genes in the correct genome, we identified additional potential diagnostic variants, a high proportion of which are noncoding.

KEYWORDS
CAGI5, connective-tissue disorder, diagnostic variants, eye disorder, Human Phenotype Ontology (HPO), neurological diseases, whole-genome sequencing data

1 | INTRODUCTION

Identification of the variant(s) causing a patient's clinical symptoms is one of the key challenges in rare disease diagnostics. The problem has assumed increasing urgency, with recent advances in sequencing technology and a decrease in the sequencing cost (Schwarze, Buchanan, Taylor, & Wordsworth, 2018), leading to vast amounts of data to interpret. While whole-genome sequencing data provides more comprehensive coverage than other more restricted sequencing technologies (such as gene panel data, exome sequencing data, or

chromosomal microarray data), identification of potential causative variant(s) out of the approximately four million variants found in a genome resonates with finding a needle in a haystack (Cooper & Shendure, 2011). The variant diagnostic rate (rate of causative, pathogenic, or likely pathogenic genotypes in known disease genes for children) from whole-genome sequencing data is currently only about 40% (Clark et al., 2018) with the implication that there remain substantial deficiencies in the current methodology. Many factors contribute to this shortfall, but there is a clear need for improved methods from the computational biology community. The Critical

Assessment of Genome Interpretation (CAGI; https://genomeinterpretation.org) is a platform for community experiments in genome interpretation. Typically, the experiments take the form of blinded prediction of the phenotypic impacts of genomic variation followed by an objective independent assessment of the results (Hoskins et al., 2017). The SickKids5 experiment is one such challenge (https://genomeinterpretation.org/SickKids5_clinical_genomes) and follows an earlier CAGI SickKids4 one (https://genomeinterpretation.org/content/4-SickKids_clinical_genomes). Here, we report our methods and the results obtained for this challenge and draw conclusions on directions for future improvement.

Participants were provided with a set of 24 whole-genome sequencing data and 24 clinical profiles for pediatric patients and asked to match each genome to the corresponding phenotype profile. Two general strategies have been developed: genotype to phenotype and phenotype to genotype. In a genotype to phenotype approach, a patient's clinical profile is not utilized to prioritize potential causative genes and variant(s)—rather all deleterious variants in a genome are identified from genotype data. Several variant annotation programs (such as VAAST (Hu et al., 2013), ANNOVAR (Wang, Li, & Hakonarson, 2010), SnpEff (Cingolani, Platts et al., 2012), VAT (Habegger et al., 2012), and VEP (McLaren et al., 2010)), utilize population allele frequency data and evolutionary conservation information together with appropriate disease inheritance models to prioritize disease-relevant genes and variants in a genome, without explicitly considering a specific patient phenotypic profile. Conversely, the common theme of a phenotype to genotype approach is that a set of patient-specific phenotypes, either in the form of Human Phenotype Ontology (HPO, Köhler et al., 2014) terms or other clinical descriptors, is used to generate a list of relevant genes and only variants in these genes are considered for further analysis. A number of strategies have been developed for the incorporation of patient-specific gene prioritization information. The information may come from various biomedical ontologies, including human-specific ontologies, like HPO (Köhler et al., 2014), Disease Ontology (DO, Schriml et al., 2019), Gene Ontology (GO, Blake et al., 2015; such as used in Phevor; Singleton et al., 2014) and other model organism-specific ontologies, such as Mammalian Phenotype Ontology (MPO, Smith & Eppig, 2009), Zebrafish Phenotype Ontology (ZPO, van Slyke, Bradford, Westerfield, & Haendel, 2014; used in Exomiser; Smedley et al., 2015). Several computational tools leverage gene-disease-phenotype relationships and phenotype information, for instance, phenolyzer (Yang, Robinson, & Wang, 2015) and Phenotype Driven Ranking (PDR, Krämer, Shah, Rebres, Tang, & Richards, 2017). Also, some tools extract phenotype information using keyword search from the text (used in GeneCards; Safran et al., 2010), or free-text boolean search (used in VarElect; Stelzer et al., 2016).

Prioritization of variants in prioritized genes includes evaluation of the likely impact on gene function and pruning variants on the basis of population frequency data (rare disease implies rare variants). Computational tools use a variety of methodologies to assess the likely impact of coding and noncoding variants. Evaluation of coding variants is usually divided into loss of function (frameshifts, direct splice site impact, and stop gain or loss), and missense. Many methods have been developed to estimate the effect of missense mutations based on sequence conservation properties (e.g., Sorting Intolerant From Tolerant [SIFT] (Kumar, Henikoff, & Ng, 2009), PolyPhen-2 (Adzhubei et al., 2010), SNPs3D profile (Yue & Moult, 2006), SNAP2 (Hecht, Bromberg, & Rost, 2015), and evolutionary action (Katsonis & Lichtarge, 2017)) and on protein stability, as estimated from protein structure (e.g., SNPs3D stability (Yue, Li, & Moult, 2005), Rosetta (Park et al., 2016), and FoldX (Delgado, Radusky, Cianferoni, & Serrano, 2019; Schymkowitz et al., 2005)). Some methods also include functional information (e.g., MutPred2; Pejaver, Mooney, & Radivojac, 2017). Noncoding variant analysis methods utilize features including regional purifying selection, enrichment with functional elements such as transcription factor-binding sites and DNase hypersensitivity as well as DNA-based evolutionary conservation. Example methods are PhastCons (Siepel et al., 2005), PhyloP (Pollard, Hubisz, Rosenbloom, & Siepel, 2010), and Gerp++ (Davydov et al., 2010). Features are often combined using machine-learning approaches (such as those used in Genomiser (Smedley et al., 2016) and CADD (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019)). In one analysis (Smedley & Robinson, 2015), phenotype-driven approaches were found to have substantially better performance than variant driven ones. So far, most of these methodologies have only been benchmarked against simulated data, and there has been very limited blind testing.

In the Sickkids5 challenge, participants were provided with clinical profiles in the form of a set of PhenoTips terms (Girdea et al., 2013; represented using HPO terms) and whole-genome sequencing data for the 24 pediatric patients. These are all difficult cases where the standard SickKids analysis pipeline failed to find any reportable diagnostic variants (Kasak et al., 2019). The challenge was to assign each genome to one of three disease classes (eye disorders, neurological disorders, and connective-tissue disorders) and to match each genome to the appropriate clinical profile. An additional optional part of the challenge was the identification of specific diagnostic variants for each patient. The identification of predictive secondary variants (related to the risk of other serious diseases and with no phenotypes reported in the clinical descriptions) was also optional. In the earlier SickKids4 challenge, participants were asked to identify diagnostic variants as well as predictive secondary variants from whole-genome sequencing data for 25 pediatric patients. Unlike in SickKids5, matches between the clinical profiles and the genome data were available.

Here, we report our approach and results for the SickKids5 challenge. We used a phenotype to genotype approach, selecting only clinical symptom-specific genes. For this purpose, we developed a phenotype-weighted scoring scheme to select the set of genes associated with each clinical profile. Each variant in the selected genes was assigned to one of six impact-related categories. The final selection of a genome for each clinical profile included a subjective evaluation of the match of each gene's online mendelian inheritance in man (OMIM) description (Hamosh, Scott, Amberger, Bocchini, & McKusick, 2005) with the clinical profile. The results were analyzed in a number of ways, especially the role of clear clinical documentation in developing the phenotype-weighted scoring scheme and types of prioritized variants.

2 | MATERIALS AND METHODS

2.1 | SickKids5 clinical profile data

The SickKids Genome Clinic at the Hospital for Sick Children in Toronto (http://www.sickkids.ca/) provided the clinical profiles for the challenge. The profiles for the 24 patients included an overall disease class, with six eye disorder cases, seven neurological, and 11 connective-tissue disorders cases. Additional profile information for each patient included gender, age, indication for referral and clinical symptoms in the form of a set of terms from the hierarchical HPO (Köhler et al., 2014) entered through the PhenoTips interface (Girdea et al., 2013). Inheritance information was also provided for some patients: Six were described as autosomal recessive cases and pedigree charts were given for 14 patients (including two of the six autosomal recessive cases). Ethnicity information was also provided for 19 out of 24 patients, none of whom were declared as African origin. We used in-house software to identify genomes of African origin (described in Pal, Kundu, Yin, & Moult, 2017). In the postchallenge submission phase, using the answer key, we found that one patient with declared Philippine ethnicity is genetically of African origin, and this caused a prediction error.

2.2 | SickKids5 whole-genome data: annotation of variant call format (VCF) files and quality control (QC) filters

Anonymized whole-genome data for all 24 patients were available via the CAGI SickKids5 challenge website (https://genomeinterpretation.org/SickKids5_clinical_genomes) in the form of VCF files produced by the Illumina HiSeq X system. We annotated single nucleotide variants (SNVs) and Indels in the VCF files using Varant (https://doi.org/10.5060/D2F47M2C), including region of occurrence (intron, exon, splice site, or intergenic), observed minor allele frequencies, mutation type, predicted impact on protein function (methods used in this step are listed in Section 2.5.), and associated phenotypes reported in ClinVar (Landrum et al., 2016). The RefGene (Pruitt et al., 2014) gene definition file was used for gene and transcript annotations in Varant. In addition, in-house scripts were used to annotate variants with the human gene mutation database (HGMD, Stenson et al., 2014) disease-related information and with dbscSNV (Jian, Boerwinkle, & Liu, 2014) information on potential splicing effects. We also used Annovar annotations (Wang et al., 2010) to add Genome Aggregation Database (GnomAD) frequency data (Lek et al., 2016), Eigen scores (Ionita-Laza, McCallum, Xu, & Buxbaum, 2016), and GERP++ scores (Davydov et al., 2010) information for each variant. Chromosome M was annotated and searched for pathogenic variants using MSeqDR mv (Shen et al., 2018). We used only high quality (graded "PASS" in the VCF file) variants for further analysis. We used SnpSift (Cingolani, Patel et al., 2012) to calculate Ts/Tv and Het/Hom alternate allele ratios from the VCF file data. We only considered variants for which the highest population frequency is <1% in all the referenced databases (GnomAD exomes and GnomAD genomes, ExAC database (Lek et al., 2016), and 1000 Genomes (Auton et al., 2015)).

2.3 | Method rationale

To address the challenge of matching genomes to clinical profiles and identifying the disease class of each genome, we used a phenotype to genotype approach, first identifying genes compatible with clinical profile information, and then analyzing variants in those genes. If we are able to identify an appropriate candidate causal variant (or pair if variants for a recessive trait) for a specific profile, which is taken as evidence of a genome and profile match, and will also imply the disease class of that genome. The steps in the method are: (a) collection of disease-relevant genes for a particular clinical profile from all 24 genomes (details in Section 2.4.); (b) identification of rare variants (less than 1% population frequency) in the relevant genes (as mentioned earlier in "SickKids5 whole genome data" section); (c) search for impact variants (both coding and noncoding) in the relevant genes and assignment of these to one of the six categories of impact confidence (details in Section 2.5.); (d) use of a subjective scoring scheme of the clinical profile (depending on the presumed disease class to which that particular profile belongs) to score each disease-relevant gene in a genome for each clinical profile (details in Section 2.6.); (e) selection of the top five scoring genomes for each clinical profile, and within those, selection of top five scoring genes; and (f) manual screening of the variants selected for each profile for appropriate inheritance model, ethnicity compatibility, and the full match of the OMIM disease description associated with each gene to the clinical profile (details in Section 2.7.). The workflow of the method is shown in Figure 1.

2.4 | Candidate gene list generation

For each patient, we extracted the HPO-based terms from the PhenoTips annotations provided in the clinical profile. Relevant genes for each profile were identified by matching the profile HPO terms to those associated with each gene in the HPO database (Build #139; Köhler et al., 2014) and in the dbNSFP database (version 3.5a; Liu, Wu, Li, & Boerwinkle, 2016). The latter includes genes related to phenotypes observed in humans as well as similar phenotypes included in the mouse database (Eppig et al., 2015; Georgi, Voight, & Bućan, 2013). We also used the list of 319 genes from the RetNet database (RetNet; http://www.sph.uth.tmc.edu/RetNet/; Daiger, 2004) to search for eye disorder-related variants. The gene list for secondary variants, containing 59 genes, was taken from Table 1 in the 2017 ACMG (The American College of Medical Genetics and Genomics) guidelines (Kalia et al., 2017).

2.5 | Categorization of variants according to their likely pathogenic impact

As outlined in the rationale, for each genome we identified 24 different sets of possible candidate causative variants, one for each of the clinical profiles. Only variants with less than 1% population frequency were
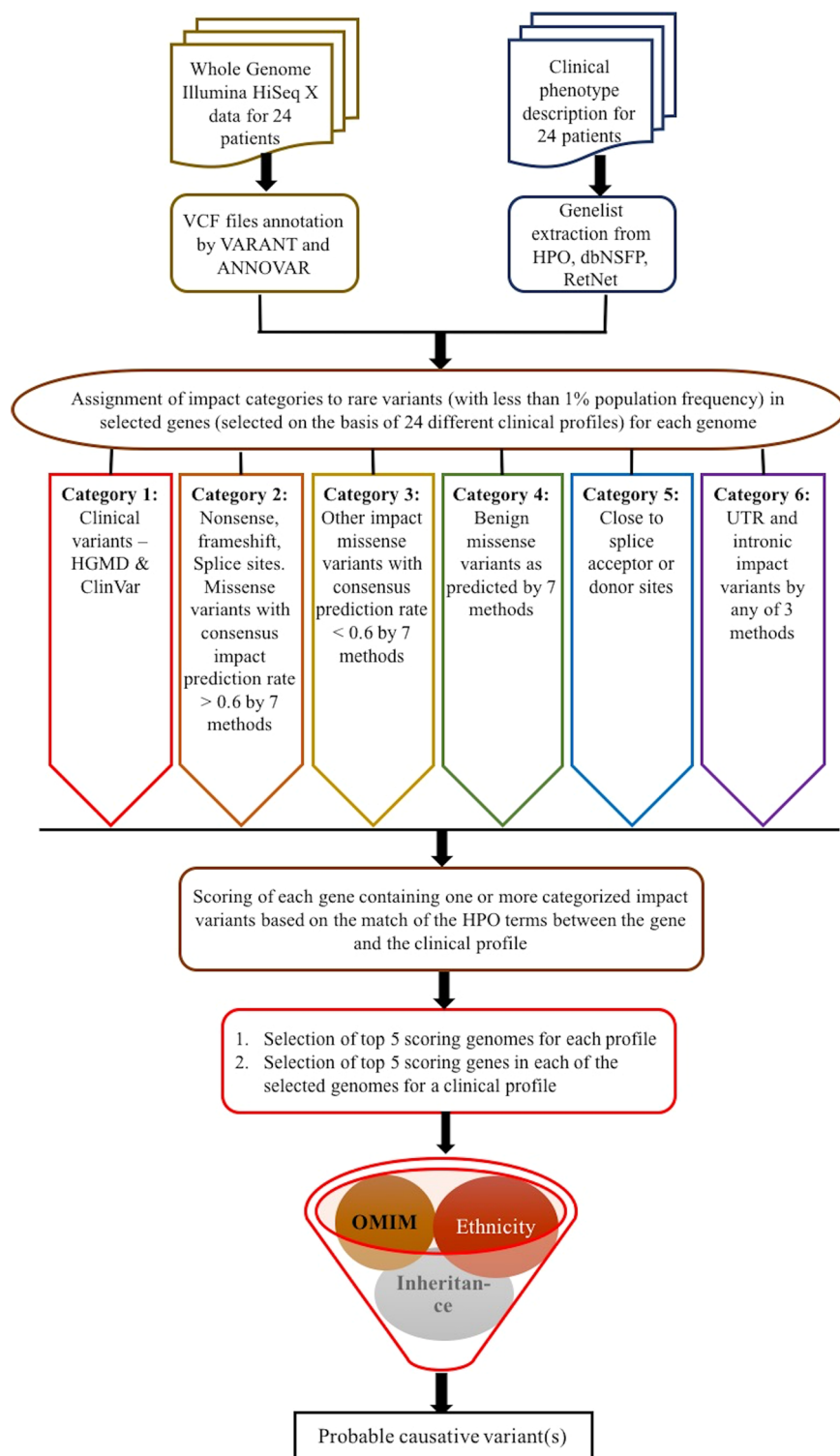
FIGURE 1 The workflow of the method for identification of probable causative variants

considered. Each selected variant was assigned to one of six categories, based on the likelihood of pathogenicity and variant type, as follows:

Category 1 (C1): Variants with HGMD annotation of either DM (disease-causing mutation) or DP (disease-associated polymorphism), and/or reported in ClinVar with pathogenic or likely pathogenic clinical significance status.

Category 2 (C2): Nonsense mutation, frameshift or nonframeshift indel, a mutation disrupting either a splice donor or acceptor site, splice altering variants (splicing consensus regions around direct splice sites) predicted by the dbscSNV (Jian et al., 2014), and missense mutations predicted as damaging by SNPs3D profile and stability methods (Yue & Moult, 2006; Yue et al., 2005), SIFT (Kumar et al., 2009), PolyPhen-2

(Adzhubei et al., 2010), Vest (Carter, Douville, Stenson, Cooper, & Karchin, 2013), REVEL (Ioannidis et al., 2016), and CADD (Kircher et al., 2014). For inclusion of a missense mutation in Category 2, at least 60% of reporting methods were required to return a prediction of deleterious. This threshold is based on calibration against HGMD (Yin, Kundu, Pal, & Moult, 2017).

Category 3 (C3): Missense mutations predicted as damaging by one or more of the above missense impact prediction methods, with the fraction of deleterious predictions <0.6.

Category 4 (C4): Benign missense mutations (zero reporting missense methods predicting deleterious).

Category 5 (C5): Variants annotated as close (within 12 bases) to a splice acceptor or splice donor site.

Category 6 (C6): Noncoding variants annotated as untranslated region (UTR) and intronic with at least one of the following conditions satisfied: CADD phred score >20 (Kircher et al., 2014), Eigen score ≥4 (Ionita-Laza et al., 2016), or Gerp++ score ≥2 (Davydov et al., 2010).

Variants in all categories were further subdivided on the basis of population frequency data:

Frequency bin 1: Novel mutations (not seen in any of 1000 Genomes, ExAC, gnomAD exomes, and gnomAD genomes databases).

Frequency bin 2: Variants with population frequency >0 and ≤0.001.

Frequency bin 3: Variants with population frequency >0.001 and ≤0.005.

Frequency bin 4: Variants with population frequency >0.005 and <0.01.

Variants were assigned to autosomal dominant, autosomal recessive, compound heterozygous, pseudo-autosomal recessive, or X-linked recessive models based on the OMIM inheritance pattern for the corresponding gene (https://www.ncbi.nlm.nih.gov/omim).

The subset of selected genes in a genome that contains one or more impact variants is then considered in the scoring of genome's match to a clinical profile.

2.6 | Gene scoring scheme for selection of genomes best matching to a clinical profile

For each clinical profile, each HPO term (T) was assigned a subjective weight (W) from 0 to 1, according to its importance (1 = most important and 0 = least important) in that profile, taking into account the presumed disease class. Usually, the most important terms were inferred from the "indication for referral" description. For example, if a connective-tissue disorder (presumed disease class of that clinical profile) is the most dominant and definitive term in the profile in the "indication for referral" description, it has scored the highest. If the seizure is also part of that profile but with the borderline occurrence, then that was assigned a lower value than would be the case if the

term occurred in a profile where the seizure is the most significant phenotype in the "indication for referral" field.

We started with the set of genes containing impact variants identified in each genome. For each clinical profile, each selected gene "i" of a genome was assigned a score GS_i based on the weights of its associated HPO terms. The score is a sum over the "n" HPO terms associated with a gene, and the weight for each term in the sum is that assigned to that HPO term in the clinical profile analysis described above.

$$GS_i = \text{PhenoTips terms} (P_i) \times \text{gender factor} (G_i)$$

$$\text{where } P_i = \sum_{j=1}^{n} W_{T_j},$$

and $G_i = 1$, if gender of phenotypic profile and genotypic profile are the same,

$G_i = 0$, if gender of phenotypic profile and genotypic profile are NOT the same.

For each clinical profile, we ranked the genomes according to the highest GS_i score of any gene. The five top-ranked genomes for each clinical profile were used for further analysis. If there are multiple genomes with the same score, more than five genomes will be considered for a particular clinical profile. For each of these top-ranked genomes, we selected the five top-scoring genes (i.e. a total of at least 25 genes per profile). There may be multiple genes with similar scores in a genome, in which case more than five may be selected. The selected genes were further filtered, removing those that do not exhibit the appropriate inheritance pattern or the appropriate ethnicity. The set of categorized variants in the remaining genes formed the set of candidate causal variants for a patient.

2.7 | Prioritized causative variants for a genome

The selection of candidate genes by weighted HPO term matching is an effective automated approach for generating an inclusive shortlist, but final gene selection required further manual pruning. For this purpose, the OMIM disease description (Hamosh et al., 2005) of each selected gene was compared with the associated clinical profiles. Most genes were eliminated on this basis. Previously selected variants in the remaining genes formed the set for final prioritization. Variants in lower frequency bins were prioritized over those with higher frequency. For example, a novel variant in a gene will be preferred over a variant in the 0.01% frequency bin in the same gene. Confidence levels of the categories were C1>C2>C3>C4>C5>C6. For example, a variant in a gene in Category 2 is preferred over a variant in Category 6 in the same gene. If the same gene is matched from two different genomes for a particular clinical profile, then we applied frequency and confidence criteria to select one of the two genomes.

2.8 | Probable regulatory effects of prioritized variants

To check for any probable regulatory effects of the prioritized variants, we noted the RegulomeDB (Boyle et al., 2012) scores, which are <4 and so possibly part of a regulatory motif. These scores were not used for the initial prioritization of variants. A RegulomeDB score of 1a–1f implies an eQTL. As all of the variants of interest are rare, none was found in this category. A score of 2a–2c implies that variants at that position may directly impact a transcription factor-binding site with subcategories (2a, 2b, and 2c) for different types of evidence. A score of 3a and 3b implies less strong evidence for impact on a transcription factor-binding site with subcategories (3a and 3b) indicating different types of evidence.

2.9 | Searching for predictive secondary variants

Here, we followed the rules in ACMG (2017; Kalia et al., 2017) to extract predictive secondary variants from 59 genes. We searched only for clinically known pathogenic and loss of function variants in those genes, as defined in Table 1 of (Kalia et al., 2017).

3 | RESULTS

3.1 | Demographics, clinical symptoms, and relevant genes

The SickKids5 challenge data consists of the whole-genome sequencing data and clinical profiles for 24 pediatric patients, of whom 11 are male and 13 female. The age range was from 5 to 19 years with an average age of 10.7 years. The challenge description included the information that there are six eye disorder cases, seven neurological disorder cases, and 11 Ehlers–Danlos syndrome connective-tissue disorder cases.

As mentioned in Section 2.1., clinical annotations of the 24 patients were provided in the form of HPO terms. Notable points are that some specific HPO terms co-occur in multiple patients, some terms occur in all three classes of disease, and complex diseases co-occur with rare disease symptoms (Figure 2). Some examples: Connective-tissue disorder patients exhibit symptoms involving a large number of organs such as the gastrointestinal tract including irritable bowel syndrome and Crohn's disease (four cases), cardio-vascular/hypertension (four cases), eye defects (four cases), developmental/motor delay (five cases), scarring of tissue (three cases), and bruising susceptibility (four cases). Similarly, neurological disorder patients often exhibit developmental delays or motor delays. Autism is manifested in one patient out of the seven neurological disorder cases. One neurological disorder patient is affected by an eye disorder as well as musculoskeletal disorders, including scoliosis and osteopenia. Similarly, an eye disorder patient is also affected by other musculoskeletal disorders, including hyper-extensibility of the joints and ear defects. Altogether, 10 of the 24 cases have symptoms in two or more disease classes.

In total, there are 213 unique HPO terms for the 24 cases. These terms were used to compile a total 6,239 potentially relevant genes from the HPO (Köhler et al., 2014) and dbNSFP (Liu et al., 2016) databases and the 319 genes in the RetNet eye disorder database (Daiger, 2004). The number of genes related to each clinical profile ranges from 350 to 4,000, with an average of 1,600. Figure S1 shows the number of genes for each case, grouped by disease class. Eye disorder patients have an average of 770 candidate genes. Neurological clinical profiles are usually associated with more genes, with an average of 2,300 genes. Connective-tissue disorder patients have the widest range, from 400 to 2,800 genes.

3.2 | SickKids5 data quality

Figure 3 shows the SickKids5 challenge data quality in terms of Ts/Tv ratio, Het/Hom alternate allele ratio, total SNV counts and rare (<1% population frequency) SNV counts. We compared these data with that for the corresponding ethnicities in the 1000 Genome set (Auton et al., 2015) and the high-quality reference Genome in a Bottle (GIAB) data (Zook et al., 2016).

In the previous SickKids challenge (https://genomeinterpretation. org/content/4-SickKids_clinical_genomes), we observed an excess of rare and novel variants for 25 patients with sequencing data provided by Complete Genomics (Pal et al., 2017), relative to 1000 Genome data. Compared with the Complete Genomics data, the CAGI5 Illumina HiSeq X data are of better quality—the data have comparable Ts/Tv ratio, Het/Hom alternate allele ratio, and total SNV counts to that of 1000 Genome data (Figure 3). Rare SNV counts in SickKids5 AFR data are comparable to that with 1000 Genome AFR data. Non-AFR rare SNV counts in SickKids5 are closer to that in GIAB than 1000 Genome EUR data. The excess of rare variants in both GIAB and SickKids5 data compared with 1000 Genome data may be due to the increasing identification of rare variants in recent years as a result of improved sequencing technologies. Nevertheless, there is a small excess of rare as well as total variant counts for SickKids5 data compared with GIAB data, not unexpected given the very high quality of the GIAB data. To investigate this discrepancy in data quality, we checked the alternate allele fraction (alt allele counts/ref allele counts) distribution for heterozygous calls in both GIAB and SickKids5 data for all "PASS" variants (Figure S2). This distribution has a broader range even within "PASS" variants for SickKids5 data compared with GIAB, indicating a higher noise level in the SickKids5 data. If we restrict this alternate allele fraction distribution in SickKids5 data to the range as observed in GIAB data, the SNV counts agree.

3.3 | Distribution of candidate variants

As described in Section 2, for each clinical profile, we identified the five or more top-scoring (best HPO term matches) genes in the five (or more) top scoring genomes. Genes in the female 13 genomes are matched to the female profiles and genes in the 11 male genomes are matched to the male profiles. An average of 35 genes per profile

FIGURE 2 Disease classes and selected PhenoTips terms for 24 cases. Each row shows the data for one patient and the total number of PhenoTips terms for each patient is given in the last column, shaded gray. Most patients have multiple symptoms and some symptoms occur in multiple patients. Further, some symptoms occur in all three classes of disease. Complex disease symptoms are also noted

were selected, resulting in a total of 342 unique genes for all 24 profiles for further analysis. There is an average of five variants in each of the five genes selected in each genome, with an average total of about 116 candidate variants per clinical profile. For eye disorder clinical profiles, we also included candidate variants in the 319 RetNet genes. For each profile, the set of candidate variants were ranked using two criteria—the impact category for a variant and its frequency bin (lower frequencies rank higher; details in Section 2.5.).

Figure 4 shows the counts of candidate variants in each category from selected genomes, for each profile. For all clinical profiles, the fraction of candidate variants in Category 6 (noncoding variants) is the highest (on average 83%, 5–10 variants per included genome), followed by the variants in C2 (loss of function [LOF] and other high impact coding variants including missense, on average 7%, 0–1 variants per genome) and then in C3 category (possibly high impact missense variants, on average 5%, 0–1 variants per genome). Where a clinical

profile contains very few HPO terms (such as eye disorder cases) with less discriminating weights among the terms the gene scoring scheme is less able to discriminate between genes in final reporting. This usually results in the inclusion of more than five genes per genome with the same score. One such eye disorder case included an average of 30 C6 candidate variants (Figure 4, last row, third column). Figure S3 shows the scores for candidate variants in the genomes selected for one clinical profile.

Comparison of each selected gene's OMIM description with the clinical profile together with filtering by ethnicity and inheritance patterns eliminated most of the genes and thus the variants within them. The remaining candidate variants were further prioritized by population frequency and confidence categories. The result is a total of 35 variants (Table S1) for all 24 clinical profiles. Figure 5 shows the distribution of these 35 prioritized variants by category and frequency bins. A total of 46% (16 out of 35) of the prioritized
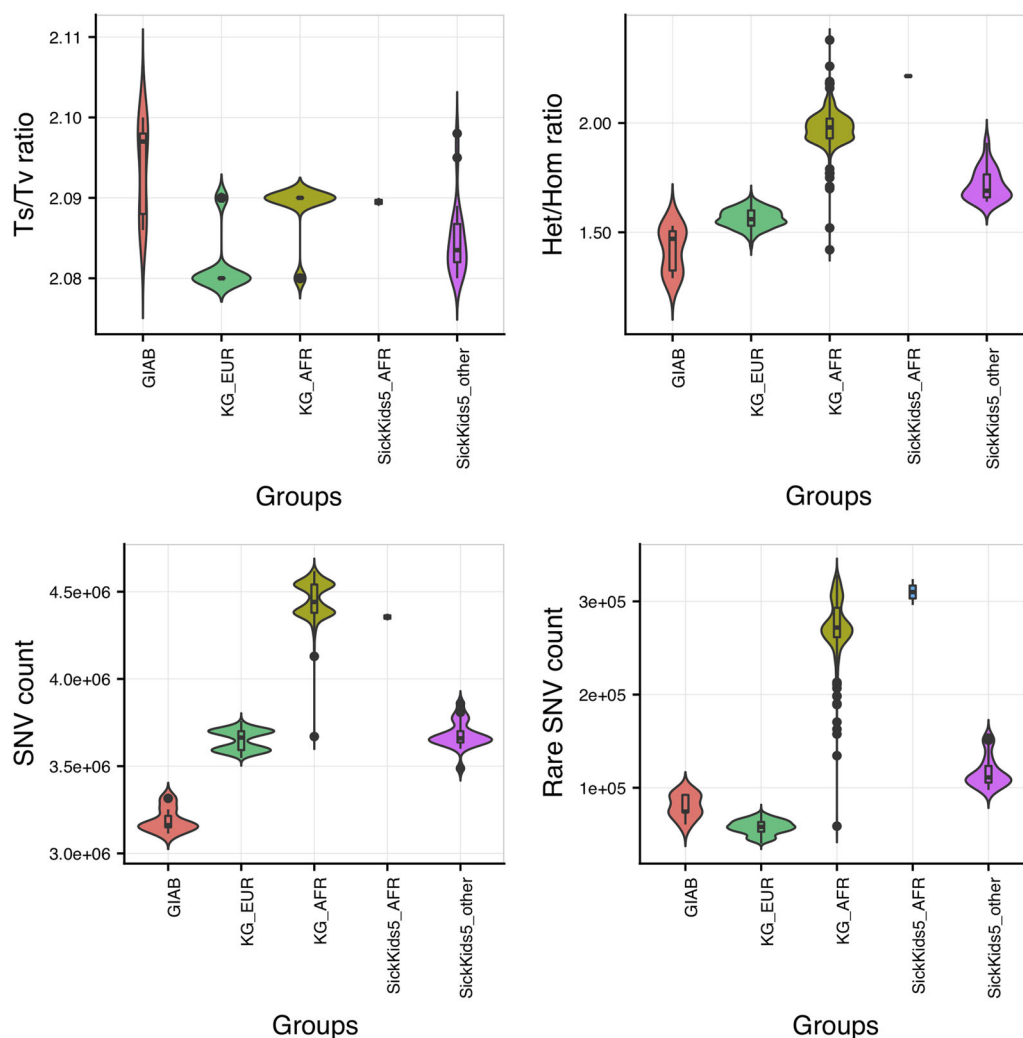
FIGURE 3 SickKids5 data quality analysis in terms of Ts/Tv ratio, Het/Hom alternate allele ratio, total SNV count, and rare (<1% population frequency) SNV counts in whole genomes. Only variants with "Pass" status are included. The 1000 Genome EUR and AFR data and GIAB data provide controls. Abbreviations used: GIAB—Genome in a bottle data, KG_EUR—1000 Genome Caucasian (EUR) data, KG_AFR—1000 Genome African (AFR) data, SickKids5_AFR—SickKids5 African (AFR) data and SickKids5_other—all other SickKids5 data excluding Africans. Although there are some differences to the GIAB data controls, generally the Sickkids data appear to be of high quality

variants are in Category 6% and 44% (7 out of 16) of these are novel —that is not seen in the 1000 Genome, ExAc or gnomAD databases. The next highest relative occurrence of novel variants (3) and total prioritized variants (8) is in the C2 category, which includes loss of function variants together with predicted high impact nonsynonymous variants. There were four prioritized variants each in the C1 and C3 categories and three variants in the C4 category.

3.4 | Molecular mechanism underlying the prioritized variants

Figure 6 shows the distribution of the 35 prioritized causative variants according to the probable underlying molecular mechanisms. A total of 46% are missense (including those occurring in categories 1, 2, 3, and 4) and 46% are intronic or UTR variants. There are three

frameshift insertion/deletion variants (9%). The missense variants have a range of impact confidence, from very high in C1 (based on clinical observation), high in C2, uncertain in C3, to predicted benign in C4. All intronic and UTR variants are predicted high impact by the Gerp++ criterion, implying conserved features at that position. Two intronic variants and one UTR variant are also predicted high impact by CADD.

Figure 7a,b shows how the distribution of variant impact categories changed from the initial set to the candidate causative variants set to the final prioritized set. There is a 99% reduction in the number of variants going from the initial set (outermost circle in Figure 7a, all rare variants present in the genomes) to the candidate causative variants set (middle circle in Figure 7a, candidate causative variants in the selected genes). The reduction is in all three variant sets (exonic, intronic, and UTR; Figure 7b),
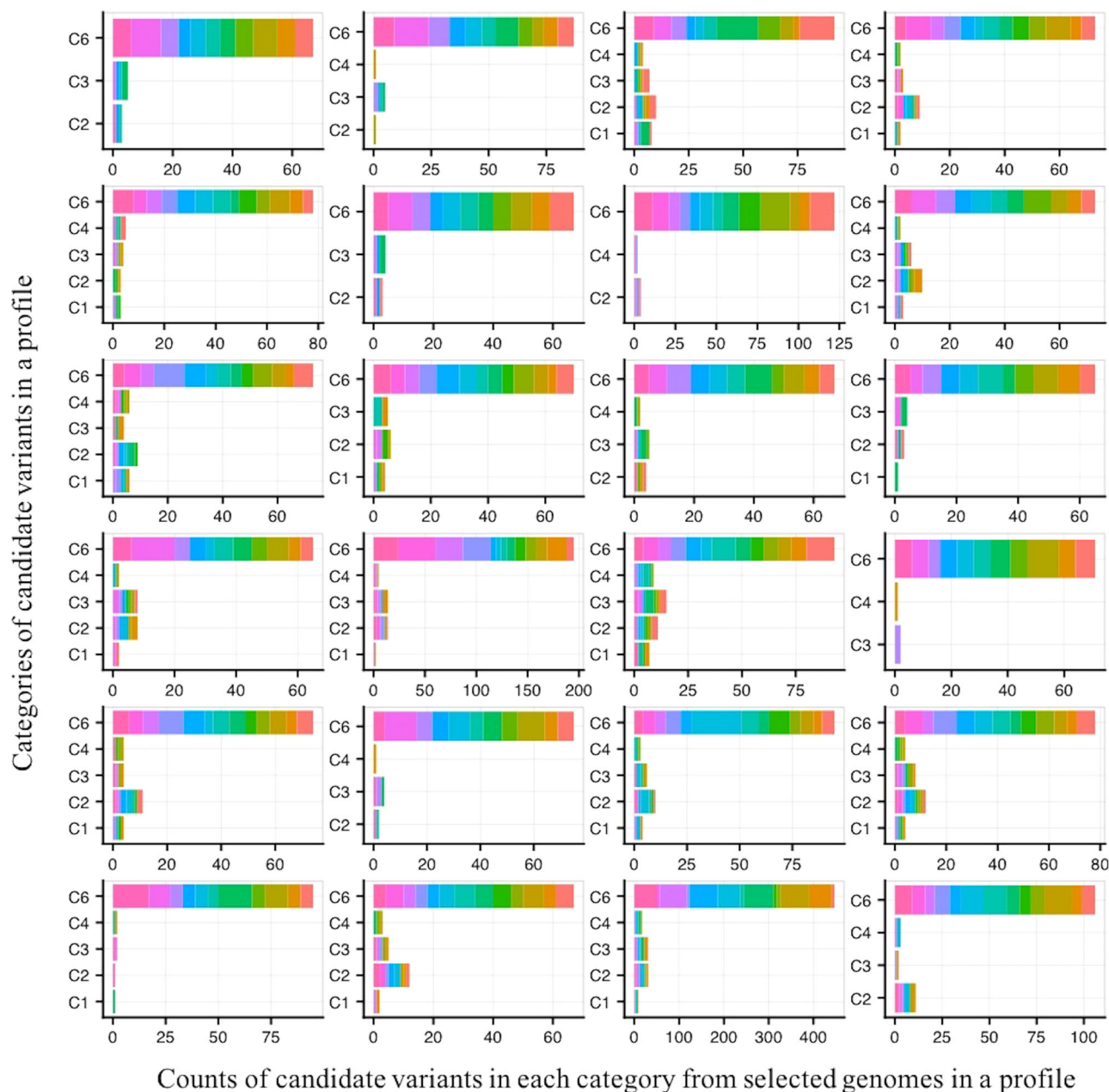
FIGURE 4 Impact distribution of selected variants for each of the 24 clinical profiles. For each profile, the number of variants falling into each of the six impact categories is shown. Variants are colored by the genome of origin. The large majority of selected variants are in the C6 category: impact variants in UTRs and introns

and is a result of only a small fraction of these variants meeting the impact selection criteria from the selected genes. Exonic variants are reduced by about 92%. The lowest range of decrease is for missense variants (from 4.34 to 3.50, reduced by 85%) and loss of function variants, for example, nonframeshift indel (from 3.32 to 2.39 on the $\log_{10}$ scale), frameshift indel (from 3.21 to 2.04), stop gain/loss (2.75–1.72). In the prioritized variants set submitted for the challenge, only missense, indels, intronic, and UTR variants were selected.

We also compared the variant compositions of the SickKids5 and SickKids4 data (Figure 7c). The initial compositions are very similar, but for the candidate and prioritized variants set, there is a dramatic shift from a large majority of exonic variants in Sickkids4 to a large majority of intronic variants in Sickkids5 set. The percent of UTR variants in the intronic and UTR sets also increased in SickKids5 set, by about two- to threefold. This is due to the introduction of intronic and UTR variant impact predictions in the SickKids5 analysis.

3.5 | Performance in CAGI5—matching of disease classes and exact matches

Table 1 shows the performance of the method in the SickKids5 challenge. Overall, we were able to identify the correct disease class

TABLE 1 Performance of the method in CAGI SickKids5 challenge

Broad disease class	Total number of cases	Number of cases with correct disease class assignment	Number of cases with exact match between clinical profile and genome	Number of cases with proposed diagnostic genes considered probable by the assessor
Connective-tissue disorder	11	6	2	1
Eye disorder	6	3	2	2
Neurological disorder	7	3	1	0

Abbreviation: CAGI, Critical Assessment of Genome Interpretation.

of a genome for 12 cases and exactly matched clinical profiles to the correct genome for five cases. Disease class matching is the most successful for connective-tissue disorders (six cases, 55%), second highest for eye disorders (three cases, 50%) and least successful (three cases, 43%) for neurological disorders. The five correct profile/genome matches are composed of two connective-tissue disorder cases, two for eye disorders, and one for a neurological disorder.

According to the data provider (who was also the challenge assessor), out of the five exact match cases, the genes carrying two eye disorder diagnostic variants and one of the genes for connective-tissue disorder diagnostic variants are possibly correct (Table 1 and Table S1). The eye disorder diagnostic variants are: (a) compound heterozygous coding-intronic variants (conserved by Gerp++ scores) in the USH2A gene—annotated for recessive retinitis pigmentosa and (b) compound heterozygous variants, one clinical missense variant and a coding-intronic variant (conserved by Gerp++ score) in the ABCA4 gene—annotated for retinitis pigmentosa, rod-cone dystrophy and other eye disorders. For the connective-tissue disorder case, we prioritized two heterozygous variants in the FBN1 gene, for an autosomal dominant inheritance pattern (information not provided in the clinical profile). One variant is in the 3′-UTR region, conserved by Gerp++ score, and the other is a novel coding-intronic variant, conserved by Gerp++ and an impact variant according to CADD. We

prioritized both variants as either might be the correct diagnostic variant, and we could not distinguish between them. The noncoding variants were checked for possible regulatory effects with RegulomeDB. The FBN1 intronic variant has a score of 3a, indicating partial evidence for transcription factor binding (the RegulomeDB annotation was not used in the variant selection procedure). Variants in the other two genes (ABCA4 and USH2A genes) have RegulomeDB scores >4 (implying lack of evidence for the variant disrupting the transcription factor-binding site).

3.6 | Illustrative example of matching a genome to a phenotypic profile

Clinical profile N is of an 11-year-old female whose indication for the referral is "Cerebral arteriovenous malformation." The "Clinical symptoms and physical findings" section for this patient also note "aortic dilation" and "joint hypermobility," both described as "borderline." In the subjective weighting of these HPO terms, we put the highest weight on "Cerebral arteriovenous malformation," a lower weight on "joint hypermobility" and related HPO terms and a further
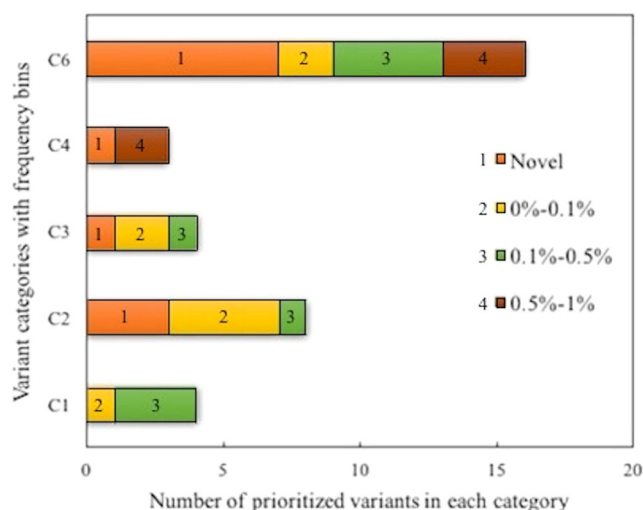
FIGURE 5 Stacked bar plot of the impact categories and frequency ranges of the 35 prioritized probable causative variants. Almost half are in the Category 6 of UTR and intronic variants
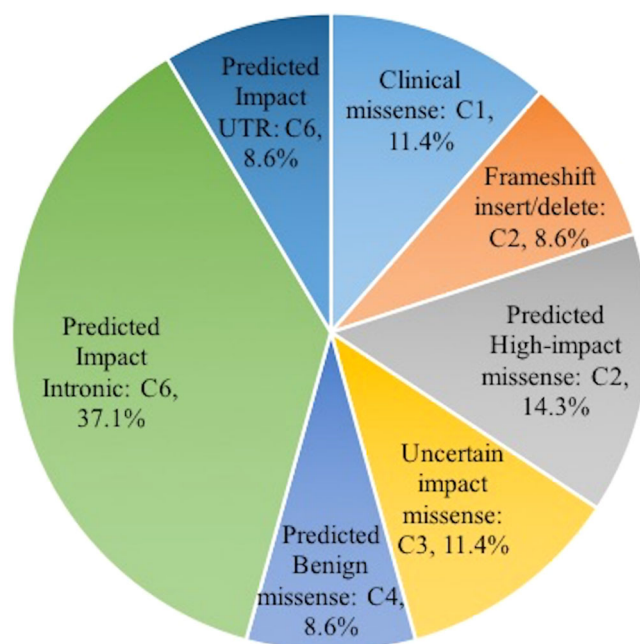
FIGURE 6 Distribution of prioritized variants by a probable molecular mechanism

(a)

(b)

Type of variants	Initial variant set	Candidate causative variants	Prioritized variant set
Exonic	49434	3908	19
Intronic	6603061	30069	13
UTR	121144	1937	3

(c)

Type of variants	Initial variant set (%)		Candidate causative variants (%)		Prioritized variant set (%)	
	SickKids4	SickKids5	SickKids4	SickKids5	SickKids4	SickKids5
Exonic	1.40	0.73	70.23	10.88	88.46	54.29
Intronic	96.33	97.48	27.58	83.73	7.69	37.14
UTR	2.27	1.79	2.19	5.39	3.85	8.57

FIGURE 7 (a) Postanalysis of the distribution of variant types ($\log_{10}$ scale) in the total set of 6,239 genes selected for the 24 SickKids5 patients. The outer most circle shows the distribution of all rare (allele frequency <1%) variants present in the genomes. The middle circle shows the distribution of candidate causative variants in the selected genes for each clinical profile for its matching genome. The innermost circle shows the distribution of final prioritized causative variants, submitted for the challenge. (b) The upper table shows the changes in variant composition at different stages of the selection process. (c) The lower table shows the comparison of variant composition (in %) between the SickKids4 and SickKids5 data at different stages of the selection process. The heat map highlights the differences in composition between the datasets. CI, coding intronic; CSA, close to splice acceptor; CSD, close to splice donor; FS, frameshift indel; NCI, noncoding intronic; NCE, noncoding exonic; NFS, nonframeshift indel; NS, missense; PGL, stop gain/loss; SA, splice acceptor; SD, splice donor; SN, synonymous; TGL, start gain/loss; UTR, UTR variants

lower weight on "aortic dilation" related terms. The least weight was set for the neurological "Headache" symptom. With these weights, we selected the top-scoring genomes for profile N. In this case, there were many equally scoring genes, resulting in all 13 female genomes being selected. Three of these genomes contained the same three highest scoring genes: *ACVRL1*, *ENG*, and *SMAD4* with matching terms for only "Headache" and "Cerebral arteriovenous malformation." The next highest scoring gene was *FBN1* with matching terms for "joint hypermobility" and "aortic dilation." *FBN1* was considered more relevant (according to the OMIM disease description for Marfan syndrome, which matches the presumed disease class of connective-tissue disorder in the profile) than any of the three top-scoring genes. So we selected all variants in that gene for further analysis. We found *FBN1* variants in a total of seven genomes. These are all category 6 variants, falling in the UTR and intronic regions. The frequency criteria were used to select the final variant. The variants span all four frequency bins (described in Section 2.5.). Two are novel, and so were given priority. One of these, in genome 056, is annotated as pathogenic by two methods (GERP++ and CADD) and

on that basis, we selected genome 056 for clinical profile N. This is a case where we successfully matched the clinical profile with the correct genome.

3.7 | Puzzling cases—limitation of phenotype-weighted scoring

One of the most critical factors in the phenotype-weighted scoring strategy is to correctly rank the importance of symptoms in a clinical profile, otherwise, the predictions will be erroneous. This information, which is usually obvious to physicians, is typically absent from the clinical profile documentation. In the SickKids5 challenge, we considered the "indication for referral" field to understand the relative importance of clinical symptoms. We failed to identify the proper disease class for one neurological disorder patient (J) because the indication for referral was "mitochondrial disorder" and this patient also has multiorgan failure, including severe eye problems, seizures, and connective-tissue disorders. For this patient, we found a ClinVar variant in chromosome M, with a disease description very

similar to that of the patient. Table S2 documents this and two other puzzling cases where high confidence loss of function variants are not causative. In one, for clinical profile "I", we found a nonframeshift-deletion variant in the *ELN* gene for one genome and in the same gene, we found a 5′-UTR variant in a different genome. According to our prioritization criteria, we selected the loss of function variant (nonframeshift-deletion) as the causative variant. However, the genome with the 5′-UTR variant in *ELN* was the correct match.

3.8 | Re-evaluation of the genome to clinical profile matches in the postsubmission phase

The availability of the answer key in the postsubmission phase allowed us to examine the genes in the correct genome more critically for the matched profile, followed by prioritizing suitable impact variant(s) as we did for the challenge. Table S3 lists 44 such prioritized variants in 26 genes for 24 SickKids5 patients. As these are all cases where the conventional bioinformatics pipeline did not identify diagnostic variants (implying no suitable clinical variant or loss of function variants or coding variants of unknown significance were found), we expected the task to be difficult. Often it seems that there are disparate symptoms that can only be accommodated by potential causative variants in two different genes, rather than one. One such example is for clinical profile F, where for the connective-tissue disorder, we identified a rare coding-intronic variant in the *COL5A2* gene. However, this patient also has very fragile skin and a food intolerance problem. We identified another novel 5′-UTR5 variant in the *PLEC* gene consistent with these additional symptoms.

Figure S4 shows that the fraction of noncoding variants (intronic and UTR) is much higher in the postsubmission analysis (46% in submitted predictions vs. 77% in postsubmission predictions) with 38% (13 out of 34 C6 category variants) being novel. Accordingly, the missense variant fraction is reduced to 20% compared with 46% in the submitted predictions.

Validation of these noncoding variants is difficult. To check for any probable regulatory effects of these variants, we noted the RegulomeDB scores <4 (see the Section 2.8.) in Table S3. Out of 44 prioritized variants, 36 returned a RegulomeDB score <4. We found three variants with a score of 2 and two variants with a score of 3 out of the 36 variants. These variants are a neurological disorder case with a score of 2a (rare—<0.05% allele frequency, an intronic variant in *CHD2*, related to myoclonic encephalopathy). Two connective-tissue disorder cases with score 2b—one is the novel 5′-UTR variant in the *PLEC* gene (described above) and another one is rare—<0.01% allele frequency—missense variant in the *TNXB* gene, predicted to be deleterious by half of the methods used (so a C3 category variant). We found one neurological variant with score 3a (novel coding-intronic variant in *ARID1B*, related to developmental delay with seizures). Another novel connective-tissue disorder variant in the *FBN1* gene with score 3a was already included at the challenge stage (described in Section 3.5.).

3.9 | Predictive secondary variants

Table S4 lists the eight predicted secondary variants we submitted, found in six SickKids5 patients. There are three novel secondary variants (a clinical missense variant in *KCNH2* for Long QT syndrome, a nonframeshift-deletion in *MSH2* for Lynch syndrome, and a nonframeshift-deletion in *BRCA2* for hereditary breast cancer) found in one neurological disorder patient (of African origin). The other predicted secondary variants are clinical variants in the *MSH6* and *MSH2* genes for Lynch syndrome and in the *LMNA* gene for hypertrophic cardiomyopathy. The same novel *MSH2* variant was found in two patients (genome 081 and 091) and according to the challenge assessor (Kasak et al., 2019), these might be sequencing errors. The alternate allele fraction (alt allele counts/ref allele counts) of these variants (Table S4) are poor, 0.36 and 0.42 for genomes 081 and 091, respectively, supporting the sequencing error hypothesis.

4 | DISCUSSION

The CAGI SickKids5 challenge provided an opportunity to assess methods for correlating whole-genome sequencing data to clinical information. Participants were asked to predict the disease class (eye, neurological, and connective-tissue disorders) of 24 undiagnosed whole genomes and to identify which genome matches to each clinical profile. To address this challenge, we developed a semi-automated gene-centric method. The method builds on one we had previously implemented for identifying causative variants based on clinical information in the CAGI4 SickKids challenge (Pal et al., 2017). The key CAGI5 innovation is the introduction of a phenotype weighting scheme to evaluate the match of gene descriptions and clinical profiles, using HPO (Köhler et al., 2014) terms. Using this approach, we were able to identify correct disease classes for 12 of the 24 genomes and to match five genomes to the correct clinical profiles. Analysis of the method's performance and results have provided a number of insights into issues related to the scoring scheme, nature of prioritized variants, the methodology used, and key factors in extracting clinical information from a whole genome.

4.1 | Phenotype-weighted scoring scheme for genes

The success of the phenotype-weighted scoring scheme depends on how effectively the clinical documentation portrays patients' symptoms. SickKids clinical profiles are constrained to terms in the HPO, and phenotypes associated with specific genes are also available in that form. Thus, the HPO-based gene by gene matching to a clinical profile provides a strategy for the selection of genes that are most likely to harbor causative variants. However, simply looking for overlap between gene and profile HPO terms is a not sensitive enough matching algorithm. Instead, we assigned a weight to each of the clinical profile terms, depending on the prominence of the term in the description (e.g., up-weighting referral terms), and

down-weighting terms that are characterized as less severe and those that do not match the disease class.

Although this approach did allow us to match a significant number of genomes to profiles, it has issues in some specific circumstances. Generally, too few terms in a clinical profile are not informative enough. For example, in one eye disorder case (case W), there were only two effective HPO terms in the profile, resulting in low discriminatory power and the selection of a large number of genes. As a consequence, there are a very large number of candidate variants (Figure 4, last row and third column). Although more terms are usually better, term combinations are of varying discriminatory power. For example, in one neurological disorder (case J), the patient has HPO terms for all three disease classes. As a result, we failed to identify the correct disease class and so did not assign appropriate weights, resulting in an erroneous choice of the genome for the profile. A limitation of the current scoring method is that it does not penalize for missing terms—ones that are not present for a gene but are in a profile or conversely ones that are not present in a profile but are there for a gene. For example, there are some genes related to eye disorders that are also related to hearing problems, and some that are not. The method as used in CAGI5 would select all these genes even if the profile includes hearing HPO terms (e.g., case X). These limitations will be addressed in future versions of the method.

4.2 | Nature of prioritized variants

In SickKids4 (Pal et al., 2017) we mostly prioritized coding variants (88% of all types of variant in SickKids4 vs. 54% in SickKids5, Figure 7c). The high proportion of noncoding candidate variants in SickKids5 is a consequence of introducing two more noncoding variant impact analysis methods, GERP++ (Davydov et al., 2010) and Eigen (Ionita-Laza et al., 2016), in addition to CADD (Kircher et al., 2014), which was also used in CAGI4. GERP++ turned out to select many more variants than CADD, whereas Eigen returned none. While CADD for coding missense variants is considered to have a reasonable performance (Anderson & Lassmann, 2018), CADD scores for noncoding variants have been found to have limited clinical utility in one study of rare noncoding variants in a hereditary cancer panel (Mather et al., 2016). There is no such benchmarking data for rare noncoding variants available for GERP++ scores. The authors of the method report an overall very low (0.86% in; Davydov et al., 2010) false-positive rate. A general problem at present is that methods for noncoding variants are less mature than those for coding. Nevertheless, noncoding variants do play a critical role in our analysis (four out of five exact match cases in our predictions were identified using noncoding variants).

4.3 | Scope for improvement

The cases in the Sickkids5 challenge were unresolved by a traditional pipeline. Although we and another group did better than random at matching genomes to profiles and the assessors considered some prioritized genes promising, most of the cases remain a mystery. And, as noted earlier, in general, rare disease pipelines have a success rate

below 50% (Clark et al., 2018). There is a number of possible explanations for the low yield of diagnostic variants, even given whole-genome sequencing data. We conclude by considering the most relevant of these, and the prospects for progress:

(A) New genes related to specific disease phenotypes are continually being discovered (Friedman et al., 2019; Guelfi et al., 2019) implying that there are many more still to be found. A strategy that might help address this problem is to consider all putative impact variants in all genes, and see if any of these genes have phenotype descriptions that offer some clue to a possible match (a genotype to phenotype approach; Hu et al., 2013; Wang et al., 2010). As more rare disease genome data accumulate, it will be possible to look for enrichment of impact variants in particular genes in the presence of particular phenotypes, and this is likely to prove a powerful approach providing a long-term solution. In the meantime, for analysis of a single genome, and even more so for the Sickkids5 challenge with 24 genomes, the large number of putative impact variants makes the strategy very difficult. If we consider only C2 variants—not clinically recognized but confidently predicted high impact, there is an average of about one in every eight genes, so that in a single full genome there will be about 3,000 variants to screen. When considering 24 genomes, there will be about 70,000 such variants. Nevertheless, it may be possible to develop a tuned version of the phenotype scoring scheme we used in the challenge to filter the variants. Consideration of knockout or knock-in data in model organisms (Smedley et al., 2015) together with such clues may be partially effective.

(B) In some complicated clinical profiles (such as for connective-tissue disorders or neurological disorders), contributions from more than one gene may be present. Indeed, one study estimates that this occurs in 5% of rare disease cases (Yang et al., 2014) and this likely is a considerable underestimate. We see evidence for the involvement of multiple genes in three cases (Table S3). For example, for clinical profile F, a connective-tissue disorder, we originally predicted a novel missense variant in *EP300* as causative, with an OMIM disease description of Rubinstein–Taybi syndrome. According to the assessor, even though we selected the correct genome, this disease description is not an adequate match to the patient's symptoms. On further inspection of the genome, knowing it is a correct match, we found an intronic variant in *COL5A2*, which is related to the classic Ehlers–Danlos syndrome (a partial match of patient's symptoms). We also found two other novel impact UTR variants in *PLEC* (the gene has an autosomal recessive inheritance pattern in OMIM), with an OMIM disease description of epidermolysis bullosa with pyloric atresia, related to the patient's fragile skin and food intolerance symptoms.

(C) As discussed above, present methods for identifying noncoding impact variants are not mature. Recent strong CAGI results for predicting which variants affect expression are encouraging in this regard (Shigaki et al., 2019), and it would be interesting to see how some of the more successful methods used there perform on the SickKids data.

(D) Nonstandard descriptions in clinical reports: an advantage of the Sickkids data is the use of HPO terms to describe patients' symptoms (Girdea et al., 2013). That greatly facilitated the identification of candidate genes, and its broader adoption by other analysis centers will improve performance. In addition, some kind of weighting scale would also help—it may be obvious to a physician that a particular HPO is not central to a patient's phenotype, but at present, that information is often not available in the record.

(E) The role of variants affecting so far poorly understood the function, particularly those that may affect chromatin structure. Examples of these have already been found in cancer (Fudenberg & Pollard, 2019; Makova & Hardison, 2015). It is not clear how well general noncoding impact methods will work on such variants, and they may be very far from genes, requiring a much larger total number of variants to be considered, with an accompanying rise in false positives. Advances in resolving three-dimensional chromatin structure and how it varies (Kishi & Gotoh, 2018; Marti-Renom et al., 2018; Qi & Zhang, 2019) hold long-term hope for progress here.

ACKNOWLEDGMENTS

DATA AVAILABILITY STATEMENT

The prediction files are available to registered users from the CAGI website (https://genomeinterpretation.org/SickKids5_clinical_genomes). Restrictions apply to the availability of the genomic data, which were used for this study under license from the Hospital for Sick Children. Genomic data are available from Dr. Stephen Meyn with the permission of the Hospital for Sick Children.

ORCID

Lipika R. Pal http://orcid.org/0000-0002-3390-110X
Kunal Kundu http://orcid.org/0000-0002-4452-4290
Yizhou Yin http://orcid.org/0000-0002-5365-2294
John Moult http://orcid.org/0000-0002-3012-2282

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature Methods, 7, 248–249. https://doi.org/10.1038/nmeth0410-248

Anderson, D., & Lassmann, T. (2018). A phenotype centric benchmark of variant prioritisation tools. NPJ Genomic Medicine, 3, 5. https://doi.org/10.1038/s41525-018-0044-9

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., ... Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., ... Westerfeld, M. (2015). Gene ontology consortium: Going forward. Nucleic Acids Research, 43, D1049–D1056. https://doi.org/10.1093/nar/gku1179

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., ... Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome Research, 22, 1790–1797. https://doi.org/10.1101/gr.137323.112

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N., & Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics, 14, S3. https://doi.org/10.1186/1471-2164-14-S3-S3

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics, 3, https://doi.org/10.3389/fgene.2012.00035

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., ... Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6, 80–92. https://doi.org/10.4161/fly.19695

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., & Kingsmore, S. F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ Genomic Medicine, 3, 16. https://doi.org/10.1038/s41525-018-0053-8

Cooper, G. M., & Shendure, J. (2011). Needles in stacks of needles: Finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics, 12(9), 628–640. https://doi.org/10.1038/nrg3046

Daiger, S. P. (2004). Identifying retinal disease genes: How far have we come, how far do we have to go? Novartis Foundation Symposium, 255, 17–27. discussion 27-36, 177–178. http://www.ncbi.nlm.nih.gov/pubmed/14750594

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Computational Biology, 6, e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Delgado, J., Radusky, L. G., Cianferoni, D., & Serrano, L. (2019). FoldX 5.0: Working with RNA, small molecules and a new graphical interface. Bioinformatics, 35, 4168–4169. https://doi.org/10.1093/bioinformatics/btz184

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., & Richardson, J. E., Mouse Genome Database Group. (2015). The Mouse Genome Database (MGD): Facilitating mouse as a model for human biology and disease. Nucleic Acids Research, 43(Database issue), D726–D736. https://doi.org/10.1093/nar/gku967

Friedman, J., Smith, D. E., Issa, M. Y., Stanley, V., Wang, R., Mendes, M. I., ... Gleeson, J. G. (2019). Biallelic mutations in valyl-tRNA synthetase gene VARS are associated with a progressive neurodevelopmental epileptic encephalopathy. Nature Communications, 10, 707. https://doi.org/10.1038/s41467-018-07067-3

Fudenberg, G., & Pollard, K. S. (2019). Chromatin features constrain structural variation across evolutionary timescales. Proceedings of the National Academy of Sciences of the United States of America, 6, 2175–2180. https://doi.org/10.1073/pnas.1808631116

Georgi, B., Voight, B. F., & Bućan, M. (2013). From mouse to human: Evolutionary genomics analysis of human orthologs of essential genes.

PLoS Genetics, 9(5), e1003484. https://doi.org/10.1371/journal.pgen.1003484

Girdea, M., Dumitriu, S., Fiume, M., Bowdin, S., Boycott, K. M., Chénier, S., … Brudno, M. (2013). PhenoTips: Patient phenotyping software for clinical and research use. *Human Mutation*, 34, 1057–1065. https://doi.org/10.1002/humu.22347

Guelfi, S., Botia, J. A., Thom, M., Ramasamy, A., Perona, M., Stanyer, L., … Matarin, M. (2019). Transcriptomic and genetic analyses reveal potential causal drivers for intractable partial epilepsy. *Brain*, 142, 1616–1630. https://doi.org/10.1093/brain/awz074

Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., … Gerstein, M. (2012). Vat: A computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*, 28 , 2267–2269. https://doi.org/10.1093/bioinformatics/bts368

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., & McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33, D514–D517. https://doi.org/10.1093/nar/gki033

Hecht, M., Bromberg, Y., & Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, 16, S1. https://doi.org/10.1186/1471-2164-16-S8-S1

Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., & Brenner, S. E. (2017). Reports from CAGI: The critical assessment of genome interpretation. *Human Mutation*, 38, 1039–1041. https://doi.org/10.1002/humu.23290

Hu, H., Huff, C. D., Moore, B., Flygare, S., Reese, M. G., & Yandell, M. (2013). VAAST 2.0: Improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genetic Epidemiology*, 37, 622–634. https://doi.org/10.1002/gepi.21743

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., … Sieh, W. (2016). REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *American Journal of Human Genetics*, 99(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Ionita-Laza, I., McCallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2), 214–220. https://doi.org/10.1038/ng.3477

Jian, X., Boerwinkle, E., & Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Research*, 42(22), 13534–13544. https://doi.org/10.1093/nar/gku1206

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., … Miller, D. T. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. *Genetics in Medicine*, 19(255), 249. https://doi.org/10.1038/gim.2016.190

Kasak, L., Hunter, J. M., Udani, R., Bakolitsa, C., Hu, Z., Adhikari, A. N., … Meyn, M. S. (2019). CAGI SickKids challenges: Assessment of phenotype and variant predictions derived from clinical and genomic data of children with undiagnosed diseases. *Human Mutation*, 40, 1191–1622. https://doi.org/10.1002/humu.23874

Katsonis, P., & Lichtarge, O. (2017). Objective assessment of the evolutionary action equation for the fitness effect of missense mutations across CAGI-blinded contests. *Human Mutation*, 38, 1072–1084. https://doi.org/10.1002/humu.23266

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315. https://doi.org/10.1038/ng.2892

Kishi, Y., & Gotoh, Y. (2018). Regulation of chromatin structure during neural development. *Frontiers in Neuroscience*, 12, https://doi.org/10.3389/fnins.2018.00874

Krämer, A., Shah, S., Rebres, R. A., Tang, S., & Richards, D. R. (2017). Leveraging network analytics to infer patient syndrome and identify causal genes in rare disease cases. *BMC Genomics*, 18, 551. https://doi.org/10.1186/s12864-017-3910-4

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7), 1073–1081. https://doi.org/10.1038/nprot.2009.86

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., … Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42, D966–D974. https://doi.org/10.1093/nar/gkt1026

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., … Maglott, D. R. (2016). ClinVar: Public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. https://doi.org/10.1038/nature19057

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation*, 37(3), 235–241. https://doi.org/10.1002/humu.22932

Makova, K. D., & Hardison, R. C. (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics*, 16(4), 213–223. https://doi.org/10.1038/nrg3890

Marti-Renom, M. A., Almouzni, G., Bickmore, W. A., Bystricky, K., Cavalli, G., Fraser, P., … Torres-Padilla, M.-E. (2018). Challenges and guidelines toward 4D nucleome data and model standards. *Nature Genetics*, 50(10), 1352–1358. https://doi.org/10.1038/s41588-018-0236-3

Mather, C. A., Mooney, S. D., Salipante, S. J., Scroggins, S., Wu, D., Pritchard, C. C., & Shirts, B. H. (2016). CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genetics in Medicine*, 18, 1269–1275. https://doi.org/10.1038/gim.2016.44

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(2070), 2069. https://doi.org/10.1093/bioinformatics/btq330

Pal, L. R., Kundu, K., Yin, Y., & Moult, J. (2017). CAGI4 SickKids clinical genomes challenge: A pipeline for identifying pathogenic variants. *Human Mutation*, 38, 1169–1181. https://doi.org/10.1002/humu.23257

Park, H., Bradley, P., Greisen, P., Liu, Y., Mulligan, V. K., Kim, D. E., … Dimaio, F. (2016). Simultaneous Optimization of biomolecular energy functions on features from small molecules and macromolecules. *Journal of Chemical Theory and Computation*, 12, 6201–6212. https://doi.org/10.1021/acs.jctc.6b00819

Pejaver, V., Mooney, S. D., & Radivojac, P. (2017). Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Human Mutation*, 38, 1092–1108. https://doi.org/10.1002/humu.23258

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20, 110–121. https://doi.org/10.1101/gr.097857.109

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., … Ostell, J. M. (2014). RefSeq: An update on mammalian reference sequences. *Nucleic Acids Research*, 42(Database issue), D756–D763. https://doi.org/10.1093/nar/gkt1114

Qi, Y., & Zhang, B. (2019). Predicting three-dimensional genome organization with chromatin states. *PLoS Computational Biology*, *15*(6), e1007024. https://doi.org/10.1371/journal.pcbi.1007024

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*, D886–D894. https://doi.org/10.1093/nar/gky1016

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., … Lancet, D. (2010). GeneCards Version 3: The human gene integrator. *Database: The Journal of Biological Databases and Curation*, *2010*, 2010. https://doi.org/10.1093/database/baq020

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., … Greene, C. (2019). Human Disease Ontology 2018 update: Classification, content and workflow expansion. *Nucleic Acids Research*, *47*, D955–D962. https://doi.org/10.1093/nar/gky1032

Schwarze, K., Buchanan, J., Taylor, J. C., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine*, *20*(10), 1122–1130. https://doi.org/10.1038/gim.2017.247

Schymkowitz, J. W. H., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F., & Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences*, *102*, 10147–10152. https://doi.org/10.1073/pnas.0501980102

Shen, L., Attimonelli, M., Bai, R., Lott, M. T., Wallace, D. C., Falk, M. J., & Gai, X. (2018). MSeqDR mvTool: A mitochondrial DNA Web and API resource for comprehensive variant annotation, universal nomenclature collation, and reference genome conversion. *Human Mutation*, *39*(6), 806–810. https://doi.org/10.1002/humu.23422

Shigaki, D., Adato, O., Adhikar, A. N., Dong, S., Hawkins-Hooker, A., Inoue, F., … Beer, M. A. (2019). Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Human Mutation*, *40*, 1280–1291. https://doi.org/10.1002/humu.23797

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*, 1034–1050. https://doi.org/10.1101/gr.3715005

Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., … Yandell, M. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics*, *94*, 599–610. https://doi.org/10.1016/j.ajhg.2014.03.010

van Slyke, C. E., Bradford, Y. M., Westerfield, M., & Haendel, M. A. (2014). The zebrafish anatomy and stage ontologies: Representing the anatomy and development of *Danio rerio*. *Journal of Biomedical Semantics*, *5*, 12. https://doi.org/10.1186/2041-1480-5-12

Smedley, D., Jacobsen, J. O. B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., … Robinson, P. N. (2015). Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nature Protocols*, *10*(12), 2004–2015. https://doi.org/10.1038/nprot.2015.124

Smedley, D., & Robinson, P. N. (2015). Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Medicine*, *7*, 81. https://doi.org/10.1186/s13073-015-0199-2

Smedley, D., Schubach, M., Jacobsen, J. O. B., Köhler, S., Zemojtel, T., Spielmann, M., … Robinson, P. N. (2016). A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *The American Journal of Human Genetics*, *99*(3), 595–606. https://doi.org/10.1016/j.ajhg.2016.07.005

Smith, C. L., & Eppig, J. T. (2009). The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *1*, 390–399. https://doi.org/10.1002/wsbm.44

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., … Lancet, D. (2016). VarElect: The phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics*, *17*, 444. https://doi.org/10.1186/s12864-016-2722-2

Stenson, P. D., Mort, M., Ball, E. V., Shaw, K., Phillips, A., & Cooper, D. N. (2014). The Human Gene Mutation Database: Building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human Genetics*, *133*(1), 1–9. https://doi.org/10.1007/s00439-013-1358-4

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*, e164. https://doi.org/10.1093/nar/gkq603

Yang, H., Robinson, P. N., & Wang, K. (2015). Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, *12*, 1–6. https://doi.org/10.1038/nmeth.3484

Yang, Y., Muzny, D. M., Xia, F., Niu, Z., Person, R., Ding, Y., … Eng, C. M. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. *Journal of the American Medical Association*, *312*(18), 1870–1879. https://doi.org/10.1001/jama.2014.14601

Yin, Y., Kundu, K., Pal, L. R., & Moult, J. (2017). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human *N*-acetyl-glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. *Human Mutation*, *38*(9), 1109–1122. https://doi.org/10.1002/humu.23267

Yue, P., Li, Z., & Moult, J. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. *Journal of Molecular Biology*, *353*(2), 459–473. https://doi.org/10.1016/j.jmb.2005.08.020

Yue, P., & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *Journal of Molecular Biology*, *356*(5), 1263–1274. https://doi.org/10.1016/j.jmb.2005.12.025

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*, 160025. https://doi.org/10.1038/sdata.2016.25

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Pal LR, Kundu K, Yin Y, Moult J. Matching whole genomes to rare genetic disorders: Identification of potential causative variants using phenotype-weighted knowledge in the CAGI SickKids5 clinical genomes challenge. *Human Mutation*. 2020;41:347–362. https://doi.org/10.1002/humu.23933