Title

Computational methods to improve clinical variant classification for the diagnosis of rare genetic disorders

Author

Sharo, Andrew George

Publication Date

2021

Peer reviewed|Thesis/dissertation

Computational methods to improve clinical variant classification for the diagnosis of rare genetic disorders

by

Andrew George Sharo

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Steven E. Brenner, Chair

Professor Nilah Ioannidis

Professor Priya Moorjani

Professor Daniel Rokhsar

Fall 2021

Computational methods to improve clinical variant classification for the diagnosis of rare genetic disorders

Andrew George Sharo

Abstract

Computational methods to improve clinical variant classification for the diagnosis of rare genetic disorders

by

Andrew George Sharo

Doctor of Philosophy in Biophysics

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Steven E. Brenner, Chair

The term ‘rare disease’ may at first suggest a problem that, if addressed, would benefit few and lead only to obscure scientific discoveries. Nothing could be further from the truth. Since before the discovery of the structure of DNA, rare disease research has enabled essential biological insights. These insights are surpassed only by the clinical innovations that were developed to treat rare disease, which benefit not only those living with rare disease but also millions of individuals living with common diseases. In the past decade, whole genome sequencing has revolutionized the diagnosis and care of individuals with rare genetic disease. However, at least half of individuals do not reach a conclusive diagnosis after whole genome sequencing. Structural variants (SVs; genomic variants longer than 50 base pairs) are the genetic cause of a portion of these unresolved cases. As sequencing methods using long reads become more accessible and structural variant detection algorithms improve, clinicians and researchers are gaining access to thousands of reliable SVs of unknown disease relevance. To address this emerging need, I developed StrVCTVRE to distinguish pathogenic SVs from benign SVs that overlap exons. StrVCTVRE performs accurately across a wide SV size range on independent test sets, which will allow clinicians and researchers to eliminate about half of SVs from consideration while retaining a 90% sensitivity. I anticipate clinicians and researchers will use StrVCTVRE to prioritize SVs in patients where no SV is immediately compelling, empowering deeper investigation into novel SVs to resolve cases and understand new mechanisms of disease.

To illustrate the value of StrVCTVRE, I next applied it to a cohort of 50 probands with undiagnosed rare disease. Linked-read sequencing and optical mapping were performed for each proband, mother, and father in this cohort. I investigated the diagnostic value of these two methods by comparing them to short-read sequencing. Clinical analysis and validation discovered 11 diagnostic or candidate SVs in this cohort. Analysis of optical mapping and linked-read sequencing data were each able to detect all 11 SVs. Analysis of short-read sequencing data could detect only 7 out of 11 (64%) of these SVs. After prioritizing the SVs in each case with StrVCTVRE, I considered the number of SVs a clinical researcher would need to manually investigate to find the diagnostic or candidate SV. This number of SVs was

surprisingly consistent across methods, and this can be attributed to the greater sensitivity of newer methods and the poor specificity of older methods. While newer methods detect more SVs with greater specificity, I found that they have not been carefully calibrated in several measures that are clinically important, including SV type, zygosity, and endpoint accuracy. These are mostly algorithmic limitations and should improve as these methods mature.

An important limitation of SV classification is the relatively few SVs that have been cataloged as pathogenic, compared to the number of cataloged single nucleotide variants (SNVs). To investigate how the accuracy of cataloged variants has changed over time, I shifted my focus to SNVs. Curated databases of pathogenic SNVs assist clinicians and researchers to interpret genetic testing results and classify novel variants. Yet these databases contain errors. Several studies have sought to identify cataloged variants that are misclassified, but none have recorded how variant misclassification has changed over time. Using archives of ClinVar and HGMD, I investigated how variant misclassification has changed over six years across different ancestry groups. I considered a class of disorders that are often highly penetrant with neonatal phenotypes—inborn errors of metabolism (IEMs) screened in newborn screening—as a model system. I used samples from the 1000 Genomes Project (1KGP) to identify individuals with genotypes that were annotated as pathogenic. Due to the rarity of IEMs, nearly all annotated pathogenic genotypes indicate likely variant misclassification. While the accuracy of both ClinVar and HGMD have improved over time, HGMD variants currently imply two orders of magnitude more affected individuals in 1KGP than ClinVar variants. I observed that African ancestry individuals have a significantly increased chance of being incorrectly predicted to be affected by a screened IEM when HGMD variants are used. However, this African ancestry bias was no longer significant once common variants were removed in accordance with recent variant interpretation guidelines.

Table of Contents

Acknowledgements

I would like to thank my advisor Steven Brenner for his insight, advice, and loyalty throughout my PhD. Thank you for providing opportunities to work on interesting and exciting research projects. I would also like to thank Priya Moorjani, Dan Rokhsar, Nilah Ioannidis, Nick Ingolia, and Sandrine Dudoit for their thoughtful feedback throughout my graduate studies. This work would not be possible without insightful comments from all the members of the Brenner lab, but especially Aashish Adhikari, Gaia Andreoletti, Tina Bakolitsa, John-Marc Chandonia, Jingqi Chen, Anna Desai, Courtney French, Zhiqiang Hu, Jennifer Lin, James Lloyd, Roshni Patel, Paulameena Shultes, and Yaqiong Wang. A special thanks to Robin Peters, Kenna Fowler, Ana Gonzalez, and Arthur Lei for their administrative support. I would also like to thank Susan Marqusee, Ahmet Yildiz, Jim Hurley, and (most of all!) Kate Chase for their support through the Biophysics Graduate Group.

I owe a huge debt to all of my collaborators, but in particular Aashish Adhikari, Russ Altman, Dario Boffelli, Julia Brown, Renata Gallagher, Zhiqiang Hu, Megan Koleske, Pui-Yan Kwok, Greg McInnes, Monica Penon, Hazel Perry, Jennifer Puck, Joseph Shieh, Uma Sundaram, Shamil Sunyaev, Karen Wong, and Yangyun Zou.

I would like to thank all of my housemates over the years for their social and emotional support, in particular Jude Berry, Bonnie Betts, Katie Brown, Max Cassowary, Tania Compos, Leo Chia, Garrick Dodson, Kitty Jones, Min Kim, Cassie King, Christina Liu, Sara Muniz, Paul Picklesimer, Rohan Prasad, Corey Rowland, Alvina Tahir, Almira Tanner, Hsin-Yeh Tsai, Cristobal Van Breen, Erica Wilson, Dean Wyrzykowski, and Jinni Zou. I would like to thank the tireless friends I made in activism, including Lewis Bernier, Carla Cabral, Matthew Dempsky, Alona Duhan, Molly Flanagan, Jon Frohnmeyer, Sepi Hosseini, Jake Hobbes, Wayne Hsiung, Matt Johnson, Samer Masterson, Ronnie Rose, Jennifer Rosie, Amanda Ruberg, Alicia Santurio, Trevor Slack, and Wilson Wong,

I would like to acknowledge the support of the NSF through a graduate research fellowship (DGE 1752814) and the NIH through a Molecular Biophysics Training Grant.

Finally, I would like to thank my mom, without whom none of this would be possible.

In memory of my friends Diane and Dan Sorbi

Diane was a fearless advocate for animals

Dan was always by her side

A rare disease took Diane too soon

Dan died from a broken heart

Their bravery inspires me everyday

Chapter 1: Introduction

In the United States, rare diseases are legally defined as those that affect fewer than 200,000 Americans[1]. Since epidemiological data for most rare diseases are unavailable, their true burden is difficult to estimate, but there could be as many as 25 million Americans living with a rare disease[2]. Since most rare diseases present in childhood[3], lifetime care costs for these individuals can amount to several millions of dollars per person, a significant healthcare burden[4]. Additional costs come from undiagnosed genetic diseases, which, in the absence of treatment, can increase demand for expensive procedures, such as organ transplants[5]. A large portion of rare diseases are caused by one or more genetic variants in a single gene, yet the gene underlying more than 3,000 of these diseases remain unknown[6]. As scientists who are often publicly funded, we should aim to do research that both deepens our understanding of biology and broadly improves society. Studying rare disease does both.

Historically, insights from rare disease research have resulted in treatments that have improved the lives of not only those living with rare disease but also millions living with common diseases[7]. In the 1970s researchers sought a treatment for familial hypercholesterolemia (FH), a rare disease in which individuals are born with excessive levels of LDL cholesterol and suffer early heart attacks. Pioneering work by Japanese biochemist Akira Endo into a class of cholesterol-lowering drugs called statins provided a potential treatment. The first clinical trials involving statins treated individuals with FH[8]. As the safety and efficacy of statins became clear, they were used more broadly. Statins are now taken by millions of people worldwide to treat high cholesterol. Consider also the rare disease hypophosphatasia, which results in brittle bones early in life. It is treated with bisphosphonates, a class of drugs that inhibit cells that break down bone tissue. Clinical work to understand the mechanism of bisphosphonates in treating hypophosphatasia has led to clinical innovation in the treatment of common bone-mineralization diseases[9], such as osteoporosis, which affects millions worldwide. Innovation in the treatment of rare diseases continue today. Recent gene editing treatments using CRISPR-Cas9 are almost exclusively focused on the treatment of rare diseases such as sickle cell anemia[10] and transthyretin amyloidosis[11]. These technologies are poised to advance our treatment of many rare diseases and even common diseases, as long as the genetic causes of every disease are well understood.

From the earliest days of our understanding of heredity, researchers have sought to understand the transmission and cause of rare diseases. In 1902, English physician Archibald Garrod discovered the first disease to segregate according to Mendelian rules, a rare disease called alkaptonuria[12]. In 1956, just a few years after the discovery of the DNA double helix, an MIT professor named Vernon Ingram discovered the first amino-acid substitution associated with a disease—the variant in hemoglobin that causes sickle cell anemia[13]. This was perhaps the birth of molecular medicine. Today, the decreasing cost of sequencing technologies has begun to again transform our ability to identify the molecular causes of rare diseases. Many rare diseases can reliably be predicted by phenotypes and a diagnosis confirmed through sequencing to identify the disease-causing variant. Many of these causes are so well characterized that carrier screening can be used to reduce the incidence of a growing number of rare diseases[14]. For some of the most frequent rare diseases, such as cystic fibrosis, we are beginning to see treatments that are tailored to an individual's pathogenic variants—precision medicine[15].

Despite this progress, there remains a minor but substantial fraction of individuals with rare disease who are undiagnosed. These families sometimes undergo a diagnostic odyssey,

traveling from specialist to specialist to find a diagnosis. The frustration, uncertainty, and time off work burdens these families with significant stress and financial cost, while the healthcare system must absorb the financial cost of additional clinician-hours and diagnostic tests. Discerning the precise genotype that explains the clinical phenotypes can inform disease management and may give the family a more confident prognosis. Once diagnosed, families may be able to learn about the disease progression in older children with the same disease, as well as disease-specific support groups, both of which reduce isolation. Of these unresolved cases, about 30-50% can be resolved through DNA sequencing[16]. Given that these cases should have a genetic etiology, that leaves a large fraction for which we will need novel methods to resolve.

Structural variants (SVs; genomic variants longer than 50 base pairs) are the genetic cause of a portion of unresolved rare disease cases. As sequencing methods using long reads become more accessible and SV detection algorithms improve, clinicians and researchers are gaining access to thousands of reliable SVs of unknown disease relevance. Methods to predict the pathogenicity of these SVs are required to realize the full diagnostic potential of long-read sequencing. To address this emerging need, in chapter 2 I introduce StrVCTVRE, a method to distinguish pathogenic SVs from benign SVs that overlap exons. Using a random forest classifier, I integrated features that capture gene importance, coding region, conservation, expression, and exon structure. I found that features such as expression and conservation are important but are absent from SV classification guidelines. I leveraged multiple resources to construct a size-matched training set of rare, putatively benign and pathogenic SVs. StrVCTVRE performs accurately across a wide SV size range on independent test sets, which will allow clinicians and researchers to eliminate about half of SVs from consideration while retaining a 90% sensitivity. I anticipate clinicians and researchers will use StrVCTVRE to prioritize SVs in patients where no SV is immediately compelling, empowering deeper investigation into novel SVs to resolve cases and understand new mechanisms of disease.

DNA sequencing provides a molecular diagnosis in less than half of undiagnosed rare disease cases[16]. Up to 10% of these unresolved clinical cases are caused by pathogenic structural variants (SVs)[17]. To routinely resolve such cases, clinicians and researchers are beginning to detect SVs using a diverse group of long DNA molecule methods. In chapter 3, I compare optical mapping, linked-read sequencing, and short-read sequencing in their ability to detect structural variants and resolve cases in a clinical diagnostic setting. Clinical analysis and validation discovered 11 SVs that were plausibly pathogenic in this cohort. All 11 SVs were detected through analysis of optical mapping and linked-read sequencing. Analysis of short-read sequencing could only detect 7 out of 11 (64%) of these SVs. Next, I developed SV prioritization recommendations that are applicable across these methods for filtering variants based on SV quality, rarity, type, size, and predicted impact. With this prioritization framework in place, I considered the number of SVs a clinical researcher would need to manually investigate to find the diagnostic or candidate SV, a measure of practical clinical interest. When appropriately prioritized, SVs detected by these methods are clinically manageable to investigate, with most diagnostic or candidate variants detected within the top 5 SVs. The number of SVs to investigate was surprisingly consistent across methods, and this can likely be attributed to the greater sensitivity of newer methods and the poor specificity of older methods. While newer methods detect more SVs with greater specificity, I found that they have not been carefully calibrated in several measures that are clinically important, including SV type, zygosity,

and endpoint accuracy. These are mostly algorithmic limitations and should improve as these methods mature.

Pathogenic variants, such as those identified in chapter 3, are often deposited in curated databases of variants. These databases assist clinical researchers to interpret genetic testing results and classify novel variants. Yet these databases also contain errors. Several studies have sought to identify cataloged variants that are misclassified, but none have recorded how variant misclassification has changed over time. Using archives of ClinVar and HGMD, in chapter 4 I investigated how variant misclassification has changed over six years across different ancestry groups. I considered a class of disorders that are often highly penetrant with neonatal phenotypes—inborn errors of metabolism (IEMs) screened in newborn screening—as a model system. I used samples from the 1000 Genomes Project (1KGP) to identify individuals with pathogenic genotypes that were annotated as pathogenic. Due to the rarity of IEMs, we would expect less than one individual in 1KGP have an IEM, thus nearly all annotated pathogenic genotypes indicate likely variant misclassification. While the accuracy of both ClinVar and HGMD have improved over time, HGMD variants currently imply two orders of magnitude more affected individuals than ClinVar variants. After investigating misclassified variants that have since been reclassified, I found that variant interpretation guidelines and allele frequency databases of genetically diverse samples are important factors in reclassification. I observed that African ancestry individuals have a significantly increased chance of being incorrectly predicted to be affected by a screened IEM when HGMD variants are used. However, this African ancestry bias was no longer significant once common variants were removed in accordance with recent variant interpretation guidelines. I discovered that ClinVar variants classified as Pathogenic or Likely Pathogenic are reclassified 12-fold more often than DM or DM? variants in HGMD, which has likely resulted in ClinVar's lower false positive rate. Finally, I found that ClinVar variants common in European and South Asian individuals were more likely to be reclassified to a lower confidence category, perhaps reflecting the greater chance that these variants will be annotated by multiple submitters.

Finally, in Chapter 5 I offer some concluding thoughts and perspectives on promising future directions.

References

1 Rare Disease Act of 2002, Pub. L. No. 107-280, 116 Stat. 1988.

2 Schieppati, A., Henter, J.-I., Daina, E. & Aperia, A. Why rare diseases are an important medical and social issue. *The Lancet* **371**, 2039-2041 (2008).

3 Smith, L. D., Willig, L. K. & Kingsmore, S. F. Whole-exome sequencing and whole-genome sequencing in critically ill neonates suspected to have single-gene disorders. *Cold Spring Harbor Perspectives in Medicine* **6**, a023168 (2016).

4 Henrard, S. *et al.* The health and economic burden of haemophilia in Belgium: a rare, expensive and challenging disease. *Orphanet Journal of Rare Diseases* **9**, 39 (2014).

5 Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233-1239 (2018).

6 Chong, J. X. *et al.* The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *The American Journal of Human Genetics* **97**, 199-215 (2015).

7 Taylor, A. How studying a disease that affects hundreds of people could save millions of lives. *The Conversation* (2016).

8 Endo, A. A gift from nature: the birth of the statins. *Nat. Med.* **14**, 1050-1052 (2008).

9 Whyte, M. P. Atypical femoral fractures, bisphosphonates, and adult hypophosphatasia. *J. Bone Miner. Res.* **24**, 1132-1134 (2009).
10 Frangoul, H. *et al.* CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. *New Engl. J. Med.* **384**, 252-260 (2021).
11 Gillmore, J. D. *et al.* CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis. *New Engl. J. Med.* **385**, 493-502 (2021).
12 Garrod, A. E. About alkaptonuria. *Medico-chirurgical transactions* **85**, 69 (1902).
13 Ingram, V. M. Sickle-cell anemia hemoglobin: the molecular biology of the first "molecular disease"—the crucial importance of serendipity. *Genetics* **167**, 1-7 (2004).
14 Kraft, S. A., Duenas, D., Wilfond, B. S. & Goddard, K. A. The evolving landscape of expanded carrier screening: challenges and opportunities. *Genet. Med.* **21**, 790-797 (2019).
15 Manfredi, C., Tindall, J. M., Hong, J. S. & Sorscher, E. J. Making precision medicine personal for cystic fibrosis. *Science* **365**, 220-221 (2019).
16 Clark, M. M. *et al.* 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** (2018).
17 Holt, J. M. *et al.* Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing. *BioRxiv*, 627661 (2019).

Chapter 2: StrVCTVRE: A supervised learning method to predict the pathogenicity of human genome structural variants*

Background

Whole genome sequencing (WGS) can identify causative variants in clinical cases that elude other diagnostic methods[2]. As the price of WGS falls and it is used more frequently, researchers and clinicians will increasingly observe structural variants (SVs) of unknown significance. SVs are a heterogeneous class of genomic variants that include copy number variants such as duplications and deletions, rearrangements such as inversions, and mobile element insertions. While a typical short-read WGS study finds 5,000–10,000 SVs per human genome, long-read WGS is able to identify more than 20,000 with much greater reliability[3-5]. This is two orders of magnitude fewer than the ~3 million single nucleotide variants (SNVs) identified in a typical WGS study. Still, despite their relatively small number, SVs play a disproportionately large role in genetic disease and are of great interest to clinical geneticists and researchers[6,7].

SVs are of clinical interest because they cause many rare diseases. Most SVs identified by WGS are benign, but on average, a given SV is more damaging than an SNV due to its greater size and ability to disrupt multiple exons, create gene fusions, and change gene dosage. In a study of 119 probands who received a molecular diagnosis from short-read WGS, 13% of cases were caused be an SV[8]. Similarly, an earlier study that found 7% of congenital scoliosis cases are caused by compound heterozygotes comprised of at least one deletion[9]. Yet, since SVs continue to be challenging to identify and analyze, these figures may underestimate the true causal role that SVs play in rare disease. Indeed, in some rare diseases, the majority of cases are caused by SVs. For example, deletions cause most known cases of Smith-Magenis syndrome, and duplications cause most known cases of Charcot-Marie-Tooth disease type 1A[10]. This suggests that for rare disorders, SVs constitute a minor yet appreciable fraction of pathogenic variants.

To continue discovering SVs which cause disease, researchers face a daunting challenge: prioritizing and analyzing the tens of thousands of SVs found by WGS. Best practices for SV prioritization are evolving, and generally mirror steps used to prioritize SNVs. Few SV-tailored impact predictors have been developed, but a small number of published studies have focused on identifying pathogenic SVs from WES[11,12] and WGS[8,13,14] and have identified a handful of important steps. Removing low-quality SV calls is essential, as short-read SV callers rarely achieve precision above 80% for deletions and 50% for duplications, even at low recall[15]. Most studies also remove SVs seen at high frequency in population databases or internal controls[7,16]. Moreover, many studies only investigate SVs that overlap an exonic region, as non-coding SVs remain particularly difficult to interpret. Depending on its sensitivity, a pathogenic SV discovery pipeline may produce tens to hundreds of rare exon-altering SVs per proband to be investigated. These values are consistent with a recent population-level study that estimates SVs comprise at least 25% of all rare predicted loss-of-function events per genome[17]. Prioritizing SVs will be necessary for the majority of probands, as shown by a study of nearly

* This chapter was primarily written by Andrew Sharo, with contributions from Zhiqiang Hu, Shamil Sunyaev, and Steven Brenner. It was adapted from a preprint deposited to biorxiv[1]. Andrew Sharo performed al the work described, with advice from other authors. This work is included with permission from the authors.

500 unresolved cases that found one or more SVs that warranted further investigation in 60% of cases[8]. Clinically validating all SVs of uncertain significance in a genome is currently infeasible, and cohort size for rare diseases will likely never reach a scale sufficient to statistically associate these SVs with disease. Therefore, computational tools are needed to prioritize and predict the pathogenicity of rare SVs.

Among methods that consider SVs, several annotate the features of SVs but very few prioritize SVs by pathogenicity. General-purpose annotation frameworks such as Ensembl's Variant Effect Predictor (VEP)[18] and SnpEff[19] both annotate SVs with broad consequences based on sequence ontology terms (e.g., transcript_ablation), which we found are not sufficient for effective prioritization. One standalone annotator, SURVIVOR_ant[20], annotates SVs with genes, repetitive regions, SVs from population databases, and user defined features. This and similar tools put the onus on researchers to provide informative features and determine how to consider these features in combination, a difficult challenge. A complementary approach is to annotate SVs using cataloged SVs known to be pathogenic or benign. One such SV annotator, AnnotSV[21], classifies SVs into five classes based on their overlap with known pathogenic or benign SVs and genes known to be associated with disease or predicted to be intolerant to variation. This approach can be successful when a disease-causing SV has previously been seen in another proband and was cataloged as pathogenic, but we show it has limitations when a disease-causing SV is novel. In contrast, SNVs can be effectively prioritized by methods such as Revel[22] and VEST[23] that integrate diverse annotations to provide a quantitative score. Similarly powerful methods are needed to predict SV pathogenicity.

In order to provide a summary pathogenicity score to prioritize rare SVs genome-wide, a predictor must address two questions. The first question is whether a gene is likely associated with a Mendelian phenotype. This relationship can be predicted through gene importance features. The second question is whether an SV impacts gene function, which requires considering intragenic features. Although these are two separate questions, for convenience researchers often combine them into a single summary score. Few methods provide such a summary score for SV pathogenicity. One standalone impact predictor, SVScore[24], calculates the deleteriousness of all possible SNVs within each SV (using CADD[25] scores by default), while considering SV type and gene truncation. SVScore then generates a summary score by aggregating across these CADD scores (mean of the top 10% by default), and this approach has shown promise in identifying SVs that are under purifying selection and thus likely deleterious[24]. Another stand-alone predictor, SVFX[26], integrates multiple features into a summary score, but focuses on somatic SVs in cancer and germline SVs in common diseases so we do not discuss it further.

In this manuscript, we introduce StrVCTVRE (Structural Variant Classifier Trained on Variants Rare and Exonic), a method that generates a summary pathogenicity score for exon-altering SVs. We anticipate clinicians and researchers will use StrVCTVRE to prioritize rare SVs associated with Mendelian phenotypes. Since nearly all pathogenic SVs are rare (minor allele frequency (MAF) < 1%), the salient challenge in resolving undiagnosed cases is to distinguish rare pathogenic SVs from rare benign SVs[17]. Existing SV predictors have been trained and assessed on common benign SVs[24,26], so they may rely on features that instead separate common SVs from rare SVs and may not be optimal for this clinical question[27]. StrVCTVRE is the first method trained to distinguish benign rare SVs from pathogenic rare SVs. StrVCTVRE is available at https://compbio.berkeley.edu/proj/strvctvre.

Results

StrVCTVRE design and assessment

StrVCTVRE is implemented as a random forest, in which many decision trees 'vote' for whether a given SV is pathogenic. The StrVCTVRE score reflects the fraction of decision trees that 'voted' that the SV is pathogenic. The decision trees are shaped by a learning algorithm, in which each tree sees thousands of example SVs from a training dataset of known pathogenic and benign SVs, and the decision nodes are optimized for accuracy. To promote diverse trees, each node of the decision tree uses only a random subset of the features. Finally, StrVCTVRE is assessed on a held-out test dataset and independent test datasets.

Characterization of StrVCTVRE features

To classify SVs, StrVCTVRE employs 17 features in five categories: gene importance, conservation, coding sequence, expression, and exon structure of the disrupted region (see Methods, Table 1.1 for details). We assessed gene importance using two features that summarize the degree of depletion of predicted loss-of-function (pLoF) variants in healthy individuals: pLI[28] and LOEUF[16]. Although LOEUF is effectively an updated, continuous version of pLI, and the two are highly correlated, we found better performance when both were included rather than just one. To explicitly capture when an important gene is highly impacted by an SV, we included two additional features: pLI of a highly impacted gene and LOEUF of a highly impacted gene. We define a gene as highly impacted when an SV overlaps the APPRIS[29] principal start codon or 50% of CDS. To specifically model coding sequence (CDS) disruptions, we used three coding features: percentage of the CDS overlapped by the SV, distance from the CDS start to the nearest position in the SV, and distance from the CDS end to the nearest position in the SV. We included a single conservation feature, phyloP of 100 vertebrates[30], by considering the average of the 400 most conserved sites in the SV. PhyloP produced the best classification among the conservation features we investigated (see Methods) and was the most informative conservation feature in a rare missense variant classifier[22]. To infer expression impacts from the SV, we included the average expression across all tissues for each exon in the SV, the proportion of gene transcripts that included each exon in the SV, and the overlap with known topologically associating domain (TAD) boundaries. To model potential differences that drive the pathogenicity of deletions and duplications, we included as a feature whether an SV is a deletion or duplication. The remaining features were related to the structure of exons in the SV including the number of exons in a disrupted gene, the number of exons disrupted, whether any affected exons were constitutive, whether all disrupted exons could be skipped in frame, and the order of the exon in the transcript. When multiple exons or genes were disrupted, we typically took the value of the most severely impacted one, as appropriate (see Methods). Missing or non-applicable feature data were replaced by the median value of each feature.

Correlation and relative importance of SV features in StrVCTVRE

Clusters emerged when we calculated these features for our SV training set, computed the correlation between each feature, and clustered by correlation (Fig. 2.1a). The most prominent cluster (labeled i) contains gene importance, conservation, CDS, and one exonic feature, with most correlations above Spearman's $\rho = 0.6$. Since both gene importance of highly impacted gene features are present in this cluster, the other features in this cluster may also capture when an important gene is highly disrupted. A smaller cluster (labeled ii) included the remaining gene importance features, pLI and LOEUF. Expression features and deletion/duplication status

were the features least correlated with all other features (all $\rho \leq 0.26$). This low correlation suggests that these features capture unique information, which is unsurprising for deletion/duplication status. But given the relative importance of some expression features (Fig. 2.1b), our results suggest expression data contains both orthogonal and valuable information for determining SV pathogenicity. The two features capturing gene importance of a highly impacted gene were the features most correlated with each other ($\rho = 0.97$), indicating that pLI and LOEUF are generally interchangeable for assessing the importance of highly disrupted genes.

By training on thousands of example SVs, StrVCTVRE discovers which features are useful for discriminating between pathogenic and benign SVs (Fig. 2.1b). Using Gini importance (see Methods), we found gene importance features were most useful to StrVCTVRE. This was followed by a group of features with similar importance that include the number of exons in a gene, conservation, CDS features, exon expression, and gene importance of a highly impacted gene. The value of these features is largely intuitive; gene importance, CDS, and conservation features are expected to be helpful to assess pathogenicity. In contrast, we suspect number of exons in gene is highly ranked due to sampling bias. We found that many well-studied pathogenic genes have numerous exons (DMD, NF1, BRCA2), and these genes have many representative SVs in our dataset even after removing near-duplicates (Methods). This may lead StrVCTVRE to have improved performance on these known clinically relevant genes, but reduced performance genome-wide (discussed further below). Surprisingly, several exonic features had relatively low importance, which may have been caused by the sparsity of SVs in our dataset that alter just a single exon. The low importance of TAD boundaries is counter to findings from a recent cancer SV impact predictor[31] and may reflect StrVCTVRE's focus on SVs that impact exons. Additionally, the low importance of deletion/duplication status suggests that on average, for exon-altering deletions and duplications, the region altered by an SV is more important than whether there was a gain or loss of genome content.

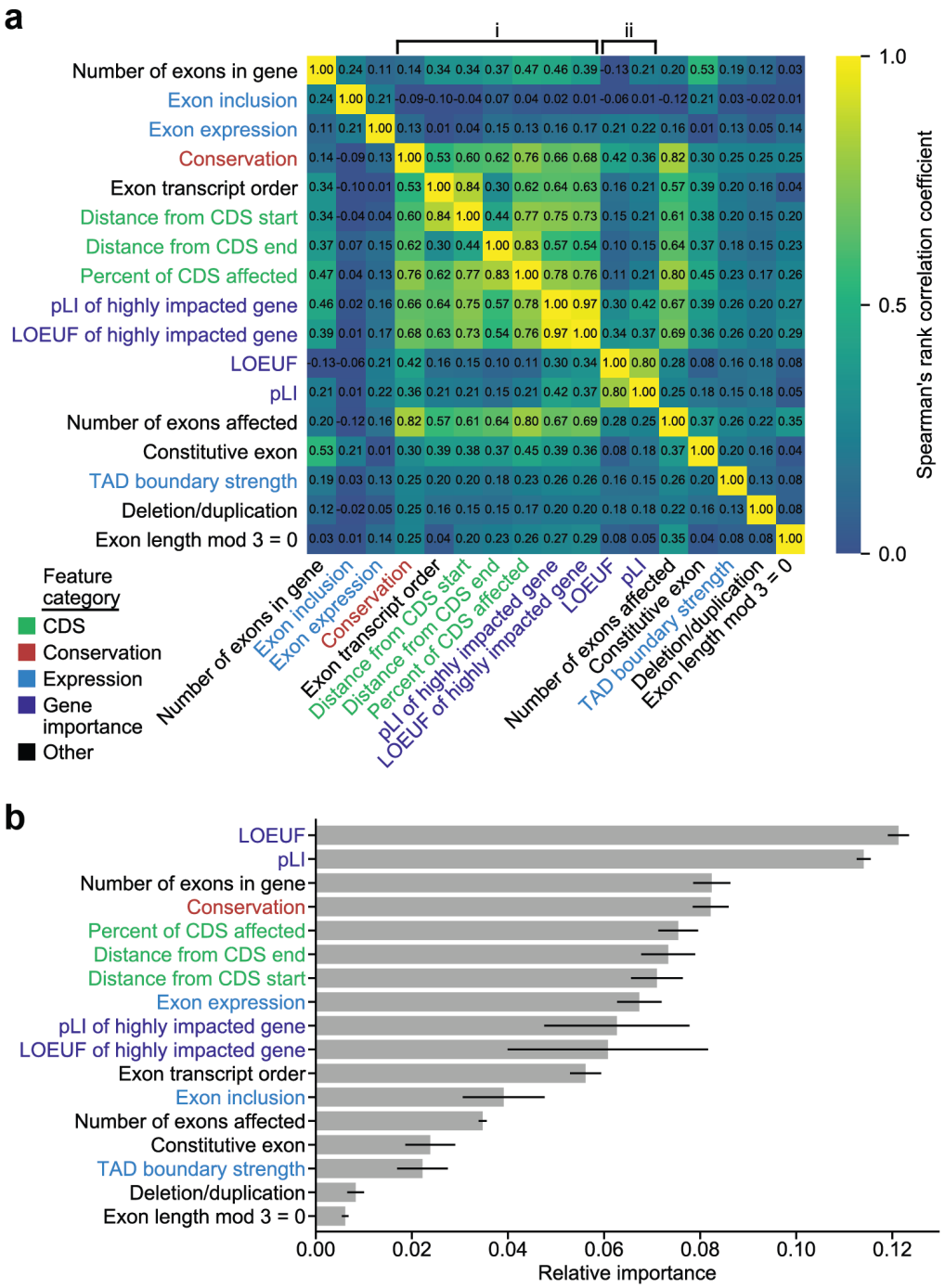

Figure 2.1 By considering feature clustering and importance, we can identify features providing unique and predictive information. a Correlation matrix of StrVCTVRE features in training data. Features were ordered by hierarchical clustering, and some values were reversed to reduce negative correlation between features. Values represent Spearman's rank correlation between features. Text is colored by feature category. b Feature importance of StrVCTVRE features. Gray bars indicate feature importance, estimated using mean decrease in impurity (Gini importance). Black lines indicate 95% confidence intervals. Note that exon expression had high

importance yet was uncorrelated with all other features, suggesting it captures unique and predictive information.

Characterization of StrVCTVRE training and held-out test sets

A total of 7,263 pathogenic or likely pathogenic deletions and 4,551 pathogenic or likely pathogenic duplications were collected from ClinVar[32], a public database of variants cataloged by academic institutions and clinical laboratories. We restricted our data to deletions and duplications, as they are the only SV types with more than 500 pathogenic examples in ClinVar. Additionally, deletions and duplications constitute the vast majority (> 95%) of rare gene-altering SVs[7]. A set of primarily benign SVs (described in greater detail below) were collected from ClinVar, gnomAD-SVs[17], and a recent great ape sequencing study[33]. Because these ape SVs were mapped to the human genome, they may be biased towards more conserved genomic regions. We retained only rare (MAF < 1% in general population) SVs in order to match the challenge faced by SV discovery pipelines. Indeed, 92% of SVs identified through cohort sequencing are rare[17], so the salient challenge is to distinguish rare pathogenic SVs from rare benign SVs. Existing SV predictors have been trained and assessed on common benign SVs[24,31], which may cause them to instead rely on features that separate common from rare SVs and result in lower accuracy in clinical use[27].

By training on rare SVs, we intend to achieve better accuracy in the challenge faced in pathogenic SV discovery. To create a rare benign dataset that matches the size range of our pathogenic dataset, we included SVs observed as homozygous at least once in great apes but rare in humans, which we assume should be mostly benign in humans due to our recent shared ancestry with great apes. Our benign dataset also included unlabeled rare SVs from gnomAD-SVs. Although we expect a small fraction of these unlabeled SVs are pathogenic, we made two assumptions that mitigated this issue: (1) pathogenic SVs have been depleted by selection so the large majority of unlabeled SVs are benign, and (2) the fraction of truly pathogenic SVs in the pathogenic and benign training sets is sufficiently different for StrVCTVRE to learn important distinguishing features. By including these additional data sources, we brought the ratio of pathogenic to benign SVs closer to 1:1 in our training set, even at small sizes. This would have been impossible with ClinVar data alone due to the dearth of small benign SVs in ClinVar.

To assess the appropriateness of including SVs from apes and gnomAD in our benign dataset, we explored how performance and feature importance changed with these data included. One predictor was trained only on ClinVar SVs, and a second predictor was trained on ClinVar SVs, ape SVs, and gnomAD SVs (altogether 3.8x more SVs than ClinVar alone). Using leave-one-chromosome-out cross validation, we found both training sets performed similarly (Fig. 2.2a), supporting our theory that the selected rare unlabeled gnomAD SVs and great ape SVs are sufficiently depleted in pathogenic SVs to be used as a training set of rare, benign SVs. Additionally, the predictor trained on all data showed a distribution of feature importance that is more evenly distributed among feature categories and possibly more robust. This includes a decrease in usefulness of gene importance features, which are likely to be overrepresented in ClinVar data, and an increase in importance in CDS features, which are an important line of evidence for assessing SV pathogenicity[34].

Before training, all data were extensively cleansed to remove duplicate records within and between datasets, remove common SVs, and remove SVs larger than 3 Mb (see Methods). Pathogenic deletions and duplications were found to have a large size bias, likely due to the

sensitivity of detection methods to specific size ranges. To avoid training on this acquisition bias, putatively benign SVs were sampled to match the pathogenic SV size distribution (Fig. 2.3). Specifically, in our training data we included only pairs of pathogenic and benign SVs that were of similar size and the same type (deletion or duplication). Using this matching strategy, we were able to include nearly all pathogenic deletions and duplications below 1 Mb. By incorporating ape and gnomAD SVs, we were able to include pathogenic SVs below 10 kilobases (kb), a range nearly absent in ClinVar benign SVs. In the benign training set, 26% of deletions and 75% of duplications came from ClinVar benign or likely benign SVs.

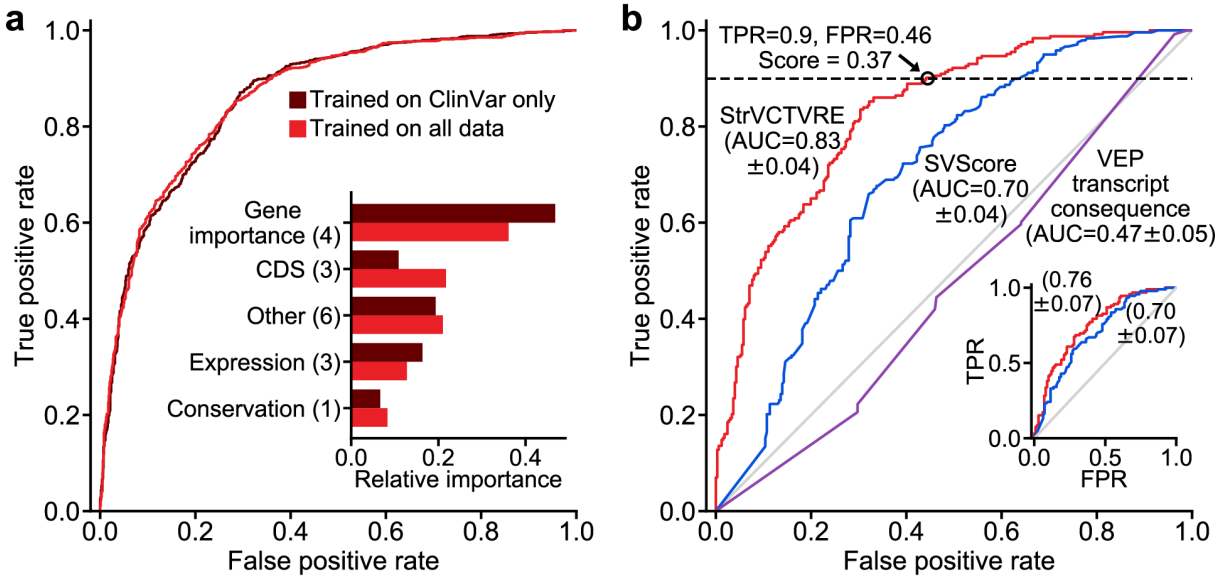

Figure 2.2 By training on multiple datasets, StrVCTVRE learned diverse feature importances and performed well on a held-out ClinVar test set. **a** Receiver operating characteristic comparing StrVCTVRE models trained on two different benign datasets: ClinVar in dark red, and all data (ClinVar, SVs common to apes but not humans, and rare gnomAD SVs) in medium red. When tested only on ClinVar data, performance does not significantly differ between the two training sets. However the feature importances (inset) of the classifier trained on all data (medium red) were more evenly distributed among feature categories. This suggests that unlabeled rare SVs and common ape SVs are a suitable benign training set. **b** Receiver operating characteristic comparing StrVCTVRE (red) to other methods on a held-out test set comprised of ClinVar SVs on chromosomes 1, 3, 5, and 7. Black circle indicates a StrVCTVRE score of 0.37, which we refer to as the ClinVar 90% sensitivity threshold. Inset shows performance on the same held-out test, modified so that each gene is overlapped by a maximum of 1 SV. AUC with 95% confidence interval is in parentheses.

To accurately assess StrVCTVRE's performance, we used a held-out test set of ClinVar SVs on chromosomes 1, 3, 5, and 7 (~20% of the total ClinVar dataset). Only ClinVar SVs were used for testing since it is the highest-confidence dataset. The training set consisted of SVs from all three data sources on all remaining chromosomes. The training set consisted of 2,463 pathogenic SVs and 2,372 benign SVs, and the test set consisted of 244 pathogenic SVs and 334 benign SVs. The test set is of reduced size because pathogenic and benign SVs in the test set were matched on length. None of the SVs in the test set were used to develop the trained algorithm.

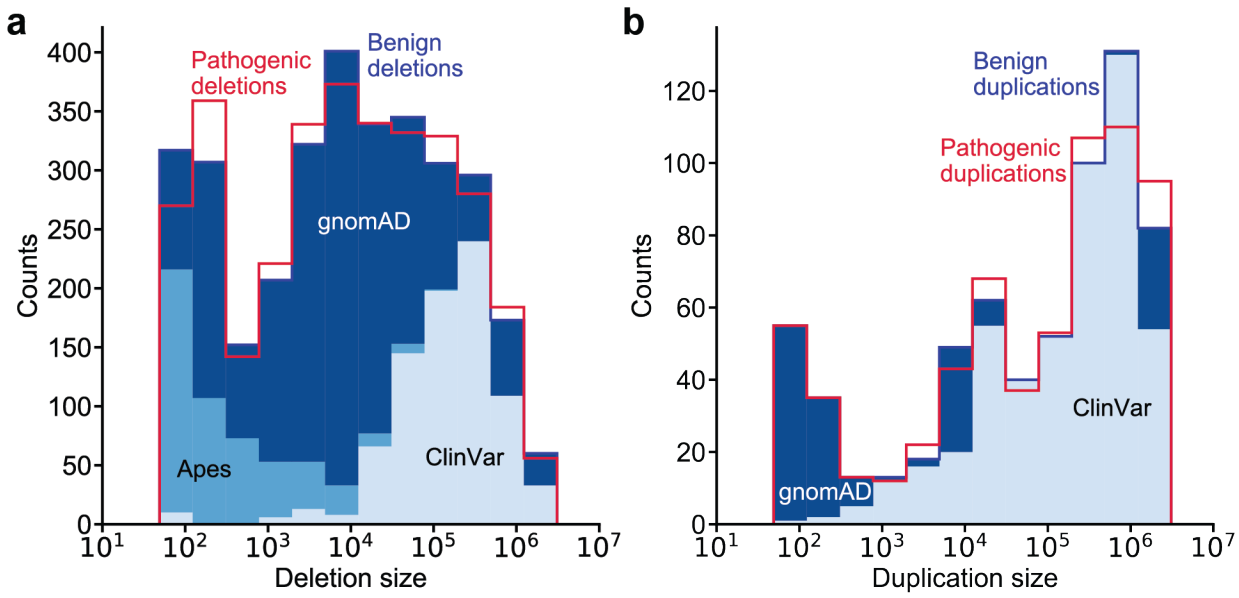

Figure 2.3 Benign training SVs (blue-shaded histograms) closely match the size distribution of pathogenic training SVs (red histogram outlines) and were drawn from multiple datasets. Histogram of pathogenic and benign (**a**) deletions and (**b**) duplications. **a** Benign deletions are composed of 26% ClinVar, 16% apes, and 58% gnomAD. **b** Benign duplications are composed of 75% ClinVar and 25% gnomAD. We were able to include more small pathogenic SVs in our training data by using apes and gnomAD SVs. Pathogenic SVs are composed entirely of ClinVar Pathogenic and Likely Pathogenic SVs and thus only histogram outlines are shown.

StrVCTVRE eliminates more than half of benign SVs from consideration at 90% sensitivity

In discriminating between pathogenic and putatively benign ClinVar SVs in the test dataset, StrVCTVRE performed substantially better than published methods. Performance was measured using the area under the receiver operating characteristic curve (AUC). The AUC for StrVCTVRE was 0.83 (95% CI: 0.79 – 0.87). By comparison, SVScore had an AUC of 0.70 (95% CI: 0.66-0.74). StrVCTVRE improved notably in the classification of large duplications and deletions (> 1 MB), a regime in which SVScore by default classifies all SVs as pathogenic (lower left corner of Fig. 2.2b). We also evaluated the predictive ability of transcript consequence reported by VEP (AUC = 0.47; 95% CI: 0.42 – 0.52), and we found it performed no better than random. This poor performance was largely due to VEP annotating more benign SVs than pathogenic SVs with its most deleterious sequence ontology term, transcript ablation. The poor performance of transcript consequence from VEP reinforces the known limitations of prioritizing variants using sequence ontology terms in isolation. As we intend StrVCTVRE to be used to prioritize SVs seen in clinical cases, it needs to perform well in clinically relevant regimes. Clinicians must limit cases in which pathogenic variants are misclassified as benign (false negatives), which requires strong performance at high sensitivity[35]. When compared to existing methods, StrVCTVRE makes substantial improvements in the high-sensitivity regime, as it is able to capture 90% of pathogenic SVs at a 46% false positive rate (black circle, Fig. 2.2b). StrVCTVRE scores range from 0 to 1, with higher scores indicating a greater likelihood of pathogenicity. In Fig. 2b, 90% sensitivity is reached at a StrVCTVRE score of 0.37, which suggests that when used on a collection of SVs called from a clinical cohort, this threshold may

identify 90% of pathogenic SVs while reducing the candidate SV list by 54%. We refer to this StrVCTVRE score as the ClinVar 90% sensitivity threshold.

We observed apparent clustering in the ClinVar data that led to additional analysis. Genes that are well-studied are overlapped by multiple pathogenic SVs catalogued in ClinVar. This resulted in several genes that were over-represented in our test set. Since SVs that overlap the same gene tend to be mostly pathogenic or mostly benign, this results in clustered test data, which may lead to higher variance in AUC performance. While this may yield improved performance for genes of particular interest, it may hide possible deficits in genome-wide performance. To address this, we randomly generated a test dataset in which each gene is overlapped by at most one SV (Fig. 2.2b inset). We found that the StrVCTVRE AUC was reduced when applied to this dataset, but StrVCTVRE was able to identify pathogenic SVs better than or equal to SVScore at all sensitivities. On this dataset, StrVCTVRE shows a sensitivity of 90% at a false positive rate of 59%.

StrVCTVRE sensitivity threshold is validated on recent clinical SVs

To assess the accuracy of our ClinVar 90% sensitivity threshold and evaluate whether StrVCTVRE performs well on clinical data, we evaluated our method on a set of SVs identified by researchers at the Broad Institute Center for Mendelian Genomics (CMG). These SVs were recently identified through exome sequencing of patient cohorts with undiagnosed neuromuscular or retinal degeneration disorders[36-40]. Clinical researchers determined these rare SVs were disease-causing or likely disease-causing. To avoid overlap between these CMG clinical SVs and StrVCTVRE training SVs, we used a leave-one-chromosome-out approach, in which 24 separate StrVCTVRE classifiers were developed, one for each chromosome. For example, CMG clinical SVs on chromosome 1 were predicted by a StrVCTVRE classifier trained on chromosomes 2, 3, 4, etc. The CMG clinical SVs consisted of 32 deletions and 2 duplications, were located on 14 chromosomes, and had a median size of 12kb. At the ClinVar 90% sensitivity threshold (StrVCTVRE score >0.37), StrVCTVRE identified 31 of 34 disease-causing SVs (91%) as potentially pathogenic.

Performance of StrVCTVRE on an independent test set from DECIPHER

All held-out test SVs, and a large fraction of training SVs, come from a single database: ClinVar. To independently test StrVCTVRE, we collected pathogenic and benign SVs from DECIPHER, a public database to which clinical scientists submit SVs seen in patients with developmental disorders[41]. Because there is some overlap between training ClinVar SVs and DECIPHER SVs, we tested on DECIPHER using a leave-one-chromosome-out approach, as described above. Additionally, to ensure this DECIPHER test set is independent from our ClinVar test set, we considered only DECIPHER SVs with a reciprocal overlap of less than 10% with any SV used in training or testing StrVCTVRE. This strategy effectively removes any concerns of training and testing on the same or similar SVs. This test set included only DECIPHER variants with the highest classification confidence (Set 1, described below). Because StrVCTVRE was trained on SVs less than 3 Mb, and few benign SVs larger than 3 Mb have been observed[42], all SVs larger than 3 Mb were scored as pathogenic (given a score of 1). Compared to its performance on the ClinVar test set, StrVCTVRE performed similarly well on the DECIPHER test set, although performance varied across SV size (Fig. 2.4a). On large SVs (> 500 kb), StrVCTVRE performed very well (AUC = 0.91; 95% CI: 0.88 – 0.94; N=297), partially because most of the SVs larger than 3 Mb are correctly predicted as pathogenic. StrVCTVRE also performed very well (AUC =

0.89; 95% CI: 0.81 – 0.97, N=116) on small SVs (< 30 kb), although this is tempered somewhat by the relatively few small SVs in the DECIPHER dataset. StrVCTVRE performed well (AUC = 0.80; 95% CI: 0.72 – 0.88,N=545) on mid-length SVs, identifying pathogenic SVs significantly better than SVScore.

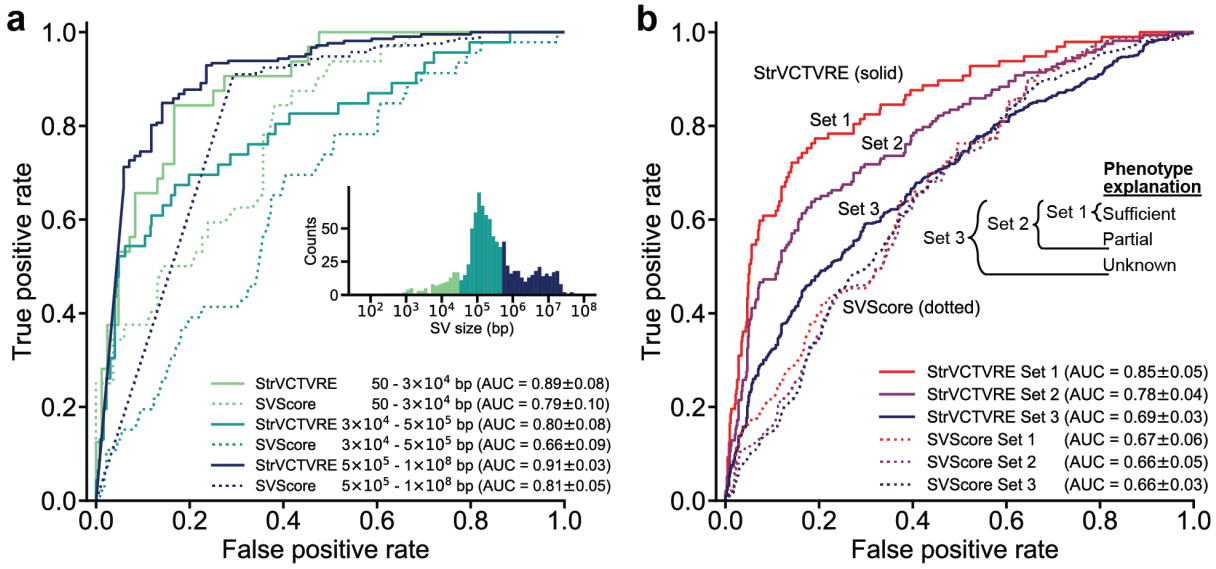

Figure 2.4 a Across three size ranges, StrVCTVRE accurately classified variants in an independent test set. In this ROC comparison of StrVCTVRE (solid line) and SVScore (dotted line), three size ranges of SVs were considered. StrVCTVRE performed very well on large and small SVs, while performing well on mid-sized variants. **b** When presented with data that are more reliably classified, StrVCTVRE's performance improved. ROC plot showing StrVCTVRE's performance increased as SV contribution to proband phenotype increases from set 3 (includes less confidently classified SVs) to set 2 and from set 2 to set 1 (most confidently classified SVs). The performance of SVScore did not significantly differ between the sets.

StrVCTVRE performance is higher when assessed on more reliably classified data

We expect that some DECIPHER pathogenic SVs are in reality benign. SVs that better explain patient phenotype are more likely to be pathogenic. To investigate the effect of SV pathogenicity on predictor performance, we grouped DECIPHER SVs into 3 sets. Set 1 consisted of SVs that sufficiently explain the proband phenotype, and these should be reliably pathogenic. Set 2 included SVs that partially explain the proband phenotype and Set 1 SVs. Set 3 included SVs with unknown contribution to proband phenotype and Set 2 SVs, and therefore their pathogenicity is less certain. StrVCTVRE was tested using a leave-one-chromosome-out approach, and DECIPHER SVs were filtered based on overlap with training and testing data as described above. We found a consistent trend towards more accurate StrVCTVRE classification in sets that were more enriched for pathogenic SVs (Fig. 2.4b). However, the same trend was not observed for SVScore. Since StrVCTVRE's performance improves on presumably more reliably classified data, we have reason to believe StrVCTVRE is making meaningful classifications.

StrVCTVRE eliminates the most benign SVs seen in 221 individuals

Typically, patients with a rare disorder caused by homozygous SVs have one or two pathogenic SVs in their genome, and the remaining SVs are benign. An ideal impact predictor would prioritize the pathogenic homozygous SVs and eliminate from consideration as many of the benign SVs as possible. To evaluate StrVCTVRE's performance in this scenario, we applied it to SVs called in 2,504 genomes identified by the 1000 Genomes Project phase 3[43] (1KGP). Because 1KGP should be depleted in individuals with severe rare disorders, we treated each genome as if it came from a proband with a rare disorder whose pathogenic SVs have been removed. 221 of these genomes had 1 or more homozygous rare exon-altering SVs, almost all of which should be benign. For each genome, we recorded the fraction of putatively benign SVs that were correctly identified as benign by StrVCTVRE and SVScore (Fig. 2.5a). Since many genomes had just one homozygous exon-altering SV, the distribution is bimodal at 0 and 1. We used our leave-one-chromosome-out predictors (e.g., predicting on 1KGP SVs on chromosome 1 and training StrVCTVRE on all other chromosomes) to score each SV. At the ClinVar 90% sensitivity threshold (StrVCTVRE score >0.37), on average StrVCTVRE identified 59% of the putatively benign SVs in each genome as benign, compared to 43% when SVScore was used at the same sensitivity (Wilcoxon paired-rank p = 8.06e-6). In a clinical setting, StrVCTVRE may classify more benign SVs as benign than SVScore, allowing clinicians and researchers to eliminate the most benign homozygous SVs from consideration.

StrVCTVRE performance is reliable even on SVs that do not overlap cataloged pathogenic SVs.

Since probands with the same disorder often have SVs altering the same genome element, and recurrent pathogenic de novo SVs are known to occur[44], one strategy used to prioritize SVs is to annotate them with overlapping SVs of known pathogenicity. AnnotSV is a popular method to identify pathogenic SVs based on their overlap with both cataloged pathogenic SVs in the National Center for Biotechnology Information's dbVar. Because it considers catalogued SVs, AnnotSV would likely perform very well for a proband whose disease-causing SV overlaps a cataloged pathogenic dbVar SV. Yet, many probands have disease-causing SVs that are not cataloged. To address these novel SVs, AnnotSV also considers SV overlap with genes associated with disease or predicted to be intolerant to variation, and it uses manually determined decision boundaries to score SVs (e.g., an SV overlapping a gene with pLI > 0.9 is scored as likely pathogenic). To compare the performance of AnnotSV with machine learning SV impact predictors on novel SVs, we created a dataset of Set 3 DECIPHER SVs that do not overlap dbVar SVs used by AnnotSV, and we recorded the prediction accuracy of each method (Fig. 2.5b). AnnotSV performed notably worse on these uncatalogued SVs. We tested StrVCTVRE (using the leave-one-chromosome-out approach) and SVScore on these uncatalogued SVs, and both methods showed significant predictive power, which we attribute to their consideration of features beyond gene intolerance (such as conservation and expression features) and their use of methods that learn decision boundaries based on training data, rather than manually determined boundaries.

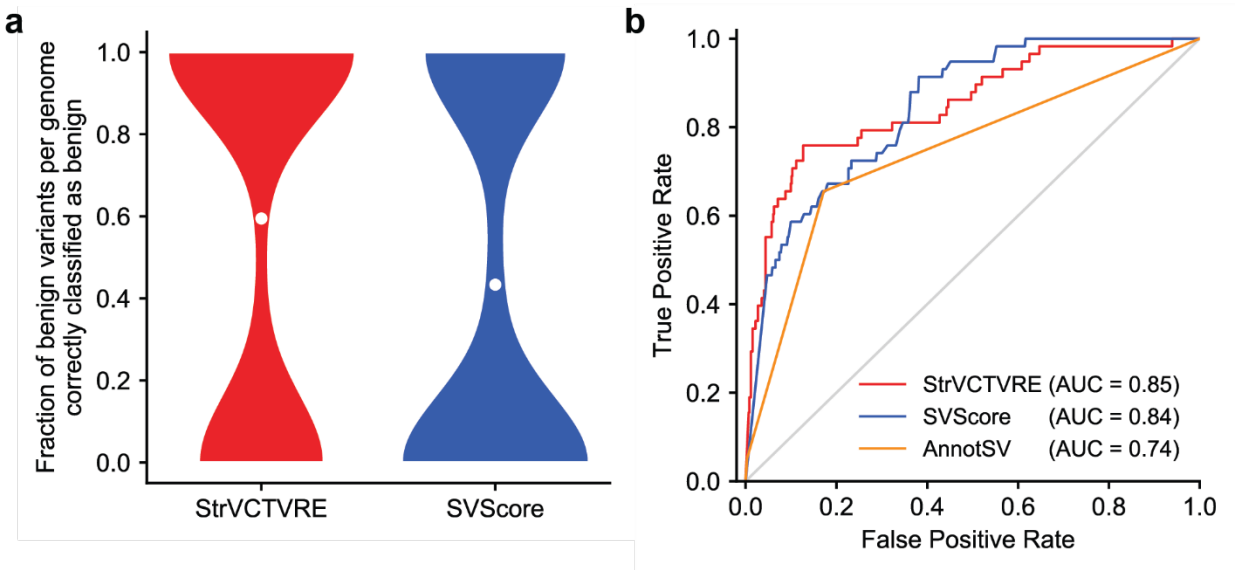

Figure 2.5 a StrVCTVRE eliminated a significantly larger fraction of benign SVs from consideration than SVScore. When tested on rare exonic SVs from the genomes of 221 putatively healthy individuals, StrVCTVRE was able to correctly classify 59% of putatively benign variants in each genome. White dots represent mean values. For both methods, the threshold for variant consideration was at the ClinVar 90% sensitivity (Fig. 2.2b). **b** ROC comparing two machine-learning methods with diverse features (StrVCTVRE and SVScore) to one method (AnnotSV) that uses limited features and manually determined decision boundaries. AnnotSV ranks an SV as 'pathogenic' or 'likely pathogenic' when the SV overlaps a catalogued pathogenic SV, known disease gene, or gene predicted to be intolerant to variation. To generate this figure, all SVs overlapping any of AnnotSV's catalogued pathogenic SVs were removed from the DECIPHER Set 3 dataset, and the remaining SVs were used for testing. AnnotSV performs relatively poorly on these novel variants. In contrast, the machine learning methods perform better, possibly because they use more diverse features and have decision boundaries trained on real data. StrVCTVRE scores were generated using a leave-one-chromosome-out approach.

Interpreting StrVCTVRE scores

StrVCTVRE scores range from 0 to 1, reflecting the proportion of decision trees in the random forest that classify an SV as pathogenic. Note that StrVCTVRE scores are not probabilities. Although we used the ClinVar 90% sensitivity threshold for evaluation, we advise against using StrVCTVRE scores as a threshold. We instead recommend that greater consideration be given to SVs with greater StrVCTVRE scores. However, thresholds are currently required for computational tools when SVs are classified using the guidelines for sequence variant interpretation recommended by the American College of Medical Genetics and Genomics (ACMG; criteria PP3, BP4)[34,45]. Within the ACMG framework, StrVCTVRE can be used as supporting evidence since it uses multiple lines of computational data. We suspect that higher levels of evidence (e.g., moderate) may be achievable, as shown by Tavtigian et al.[46] However, when using StrVCTVRE at higher levels of evidence, users should be careful not to also count other ACMG criteria that StrVCTVRE already incorporates, which could lead to double counting. Alternatively, to resolve concerns of double counting, StrVCTVRE can be used just to prioritize

variants, but not used as evidence. Users then can manually classify SVs of interest using the full ACMG criteria.

Discussion

As genome sequencing becomes more accessible, clinicians and researchers face a challenge in identifying pathogenic SVs in the thousands identified by sequencing. The ACMG recently offered guidelines for classifying SVs, acknowledging that classification is complex and many pathogenic SVs will be classified as variants of uncertain significance due to incomplete knowledge[34]. SV impact predictors can address this challenge, but few SV impact predictors exist. Although SVs comprise a significant fraction of the loss-of-function variants that cause rare disease, fewer than 10,000 pathogenic SVs have been cataloged in ClinVar. These SVs have distinct biases towards certain genes and lengths, which leads to acquisition bias that hinders predictor development. Additionally, it is not clear which features are most useful when classifying SVs and how to address the large size range of SVs. StrVCTVRE is the first method to address these problems by predicting the impact of exon-altering deletions and duplications in rare genetic disorders. We overcame data limitations and bias by combining SVs from multiple data sources as well as matching pathogenic and benign SVs by size. Since clinicians and researchers must recognize SVs that cause disease among dozens of rare exon-altering SVs detected in a proband, we trained only on rare SVs.

Determining whether a single SV is pathogenic requires consideration of numerous features in combination, as demonstrated by the recent ACMG SV guidelines. Independent of these guidelines, our method identified important features in cataloged SVs. Our findings reinforce clinical guidelines, while also highlighting new areas to explore. Both StrVCTVRE and the ACMG guidelines found gene importance and CDS disruptions to be critical for SV interpretation. Additionally, StrVCTVRE highlighted two features not discussed in the guidelines: conservation and expression. We found exon expression in particular is both predictive and poorly correlated with all other features, suggesting it captures distinctive information for determining pathogenicity. More widespread consideration of expression features could be beneficial for SV classification. StrVCTVRE additionally identified features that are not useful to classify exon-altering SVs, such as TAD boundary strength and whether there is a copy gain or loss. This is consistent with the ACMG guidelines, which do not consider TAD boundaries and provide very similar scoring metrics for both copy gain and loss.

Since SVs range from 50 bp to > 10 Mb, it is challenging to accurately classify SVs across this range. Benign SVs in ClinVar are mainly > 10 kb, but accurate classification of SVs < 10 kb requires training on benign SVs from the same size range. We accomplished this by training on small benign SVs from great apes and gnomAD. When tested on an independent test set, StrVCTVRE performed well at all size ranges. To be helpful in a clinical setting, a method must perform well at moderately high sensitivity. StrVCTVRE satisfies this requirement and was able to remove 57% of homozygous SVs from consideration at a sensitivity of 90% in the 1KGP dataset. This 90% sensitivity threshold was validated using a dataset of recent SVs observed to cause neuromuscular and retinal degeneration disorders. Overall, we found StrVCTVRE outperforms SVScore in most tasks, even though SVScore's underlying approach , CADD, was trained on > 1,000-fold more variants. Additionally, whereas StrVCTVRE was often assessed using a leave-one-chromosome-out approach, SVScore could not be readily modified and thus had the benefit of possibly training on data that overlapped the testing SVs.

StrVCTVRE is accessible as a downloadable command line program (see Data Availability). Whereas SVScore requires users to download an 80 gigabyte (Gb) CADD file, StrVCTVRE only requires a 9 Gb phyloP file. Because there are an intractably large number of possible SVs, each SV must be scored anew (unlike SNVs for which scores can be pre-computed), and this requires efficient scoring methods. StrVCTVRE runs rapidly and annotates 100,000 gnomAD SVs in three minutes, while SVScore annotates the same SVs in 24 hours.

Following existing predictors, StrVCTVRE predicts the pathogenicity of an SV in isolation. Yet human biology complicates this picture through zygosity and dominance. Since zygosity is not reported for most SVs in ClinVar, StrVCTVRE is zygosity-naïve. Additionally, StrVCTVRE's pathogenic training dataset consists largely of SVs in genes predicted to lead to dominant disorders. When tested on sets of predicted dominant or recessive SVs, StrVCTVRE performs similarly on both. Researchers who suspect a recessive mode of inheritance may need to consider StrVCTVRE scores in tandem with impact predictor scores for SNVs in trans in the same gene. Although genes vary in their tolerance of SVs and dominance, we believe a whole genome approach will be necessary to identify all pathogenic SVs, including those SVs disrupting genes not currently associated with disease. To identify new disease genes, it may be helpful to consider StrVCTVRE scores in tandem with one of the many methods that assess the match between patient phenotype and known/predicted phenotypes for an affected gene[47-49].

A method can only be as good as its training data. SV impact predictors are limited by the relatively small number of identified pathogenic and putatively benign SVs, as well as the over-representation of certain genes in the dataset. While pathogenic ClinVar variants are commonly used to train variant impact predictors, they are known to include misclassified variants[50]. We know of no characterization of the accuracy of SVs in ClinVar, but work investigating pathogenic SNVs suggest at least 90% are pathogenic based on reclassification rates[51]. 70% of our pathogenic training SVs have at least 1 review star in ClinVar, indicating they have supporting evidence which further bolsters our confidence in these data. Nonetheless, data limitations almost certainly curtail the ultimate performance of our approach. StrVCTVRE is unable to classify inversions and insertions due to limited data; however, these have been shown to contribute to a minority of the pLoF events caused by SVs[17]. We are hopeful that additional clinical sequencing studies will identify a more diverse range of SVs, which will be cataloged in open resources such as ClinVar and leveraged to develop more accurate models. We look forward to greater non-coding genome annotations, which will expand our understanding and cataloging of pathogenic noncoding SVs, which remain vexing to classify.

Much of the focus in SV algorithms has been on methods to accurately detect SVs. These methods have left clinicians and researchers awash with SVs not previously known. As experimental methods and algorithms advance, SV detection will improve, but SV interpretation will continue to be challenging. StrVCTVRE advances the clinical evaluation of SVs. During genome sequencing analysis, some cases contain an SV that matches a cataloged pathogenic SV or satisfies the conditions for pathogenicity set forth in the ACMG SV guidelines. However, these SVs are often not obvious, and StrVCTVRE can be used to quickly bring these SVs to attention. In the many cases in which no SV is immediately promising, StrVCTVRE aids clinicians and researchers in identifying compelling SVs for manual investigation. Then, if a case remains unresolved by manual investigation, SVs highlighted by StrVCTVRE that are in novel disease genes can be directed to experimental exploration. This will empower researchers to

identify novel disease genes where haploinsufficiency and triplosensitivity were not previously known causes of disease. Adoption of structural variant impact predictors will enable clinicians and researchers to make the most of these new data to improve both patient care and our understanding of basic biology.

Methods

Training, validation, and test datasets

All SVs were retrieved in GRCh38 or converted using the University of California, Santa Cruz (UCSC) liftover tool[52].

All ClinVar SVs[32] were downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/variant_summary.txt.gz on January 21, 2020. SVs were retained if they fulfilled all the following requirements: clinical significance of pathogenic, likely pathogenic, pathogenic/likely pathogenic, benign, likely benign, or benign/likely benign; not somatic in origin; type of copy number loss, copy number gain, deletion, or duplication; > 49 bp in size; at least 1 bp overlap with an exon.

Great ape SVs[33] mapped to GRCh38 were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/dgva/estd235_Kronenberg_et_al_2017/vcf/ on April 8, 2019. Deletions were retained if they were absent in humans and homozygous in exactly one of the following species: chimpanzee, gorilla, or orangutan. Only exon-altering deletions > 49 bp were retained. These deletions are subsequently referred to as *apes.*

gnomAD 2.1.1 SVs[17] (build GRCh37) were downloaded from https://storage.googleapis.com/gnomad-public/papers/2019-sv/gnomad_v2.1_sv.sites.vcf.gz on June 28, 2019. Only duplications and deletions were retained that were exon-altering, >49 bp, and PASS Filter. gnomAD SVs were divided into three categories: SVs with a global minor allele frequency (MAF) > 1% (*gnomAD common*), SVs with a global MAF < 1% with at least one individual homozygous for the minor allele (*gnomAD rare benign*), and SVs with a global MAF < 1% with no individuals homozygous for the minor allele (*gnomAD rare unlabeled*).

Database of Genomic Variants[42] release 2016-05-15 of GRCh38 “DGV Variants” was downloaded from http://dgv.tcag.ca/dgv/app/downloads on April 08, 2019. MAF of each deletion was calculated as ‘observedlosses’ / (2 * ’samplesize’). MAF of each duplication was calculated as ‘observedgains’ / (2 * ’samplesize’). Only exon-altering SVs > 49 bp were retained. Those SVs with a MAF greater than 1% are subsequently referred to as *DGV common.*

DECIPHER CNVs (build GRCh37) were downloaded from http://sftpsrv.sanger.ac.uk/ on Jan 27, 2020. Only exon-altering SVs > 49 bp with pathogenicity of “pathogenic”, “likely pathogenic”, “benign”, or “likely benign” were retained. We only considered benign or likely benign SVs without “Full” or “Partial” contribution to disease phenotype. These benign and likely benign SVs were included in all 3 sets. Set 1 pathogenic SVs consisted of pathogenic or likely pathogenic SVs with “Full” contribution to disease phenotype (referred to as “sufficient” in this manuscript). Set 2 SVs consisted of pathogenic or likely pathogenic SVs with “Full” or “Partial” contribution. Set 3 SVs consisted of pathogenic or likely pathogenic SVs with “Full”, “Partial”, or “Unknown” contribution. Identical SVs with conflicting pathogenicity were removed. SVs were then sorted by size (ascending) and SVs with a reciprocal overlap >90% were removed, keeping only the first SV.

1KGP merged SVs[43] were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/supporting/GRCh38_positions/ on Oct 22, 2019. Only exon-altering deletions and duplications with a global allele frequency less than 1% were used for testing in Fig. 2.5a.

We used exon boundaries from Ensembl biomart[53], genes v96, GRCh38.p12, limited to genes with HGNC Symbol ID(s) and APPRIS annotation[29]. For each gene, a single principal transcript was used, based on the highest APPRIS annotation. For transcripts that tied for highest APPRIS annotation, the longest transcript was used. Exon overlap was determined using bedtools intersect.

Extensive deduplication of data was performed as follows. Deletions and duplications were considered separately. Benign SVs (n=23,239) were ordered (ClinVar benign, ClinVar likely benign, apes, gnomAD rare benign, gnomAD rare unlabeled) and duplicates (reciprocal overlap of 90% or greater) were removed, keeping the first appearance of an SV. This removed 577 SVs from ClinVar benign/likely benign, 5 SVs from apes, and 408 SVs from gnomAD. The retained data are subsequently referred to as *benign.* To deduplicate pathogenic SVs (n=8,378), deletions and duplications were considered separately. Exact matches between ClinVar pathogenic and ClinVar likely pathogenic were removed from likely pathogenic. SVs were then sorted by size, ascending. SVs with > 90% reciprocal overlap were removed, keeping the smallest SV. This removed 2,421 pathogenic SVs. The retained data are subsequently referred to as *pathogenic.* Next, exact matches between the benign and pathogenic datasets were removed from both datasets. Finally, duplicates between pathogenic and benign (reciprocal overlap of 90% or greater) were removed from the pathogenic dataset. This removed 3 benign SVs and 82 pathogenic SVs.

Data were processed as follows to ensure we trained only on rare SVs. Pathogenic and benign SVs that exactly matched a DGV common SV were removed. Pathogenic and benign SVs with reciprocal overlap > 90% with an SV in gnomAD common were removed. This removed 30 benign SVs and 1 pathogenic SV. SVs between 50 bp and 3 Mb were retained, all others were removed.

We found some evidence of acquisition bias in ClinVar data due to the SV size sensitivity of different SV detection methods. To ensure StrVCTVRE was not learning on this acquisition bias, the size distribution of benign and pathogenic SVs were matched using the following procedure. After filtering as described above, benign SVs were organized into five tiers: ClinVar likely benign; ClinVar benign; apes; gnomAD rare benign; and gnomAD rare unlabeled. Each pathogenic SV was then matched by size and type (DEL or DUP) to a benign SV, iterating through each tier. Specifically, each pathogenic SV of size N seeks a benign SV of the same type in the bin $[N - (N/\alpha + 20), N + (N\alpha + 20)]$ where α = $\sqrt[101]{10^6}$ (this bin size derived from Ganel et al.[24]). A pathogenic SV first seeks a benign SV in the first benign tier. If matched, the pathogenic and benign SVs are included in the training set, and the benign SV cannot match any further pathogenic SVs. If no match is found in the first benign tier, the same process is repeated while progressing through further benign tiers. Pathogenic SVs that do not find a match in any benign tier are not included in the final training set. This process was continued for all pathogenic SVs and the resulting data are shown in Fig. 2.3.

After SVs were annotated with features (see below), we identified groups of SVs with identical features, considering pathogenic and benign SVs separately. We removed all but one of these

feature-identical SVs in order to avoid overfitting. This removed 37 SVs from the pathogenic training set and 31 SVs from the benign training set. For feature-identical SVs that were present in both the pathogenic and the benign datasets, all feature-identical SVs were removed. This removed 13 SVs.

Structural variant impact predictors

VEP[18] v96 was downloaded from https://github.com/Ensembl/ensembl-vep on April 16, 2019, and used to annotate SVs with transcript consequence sequence ontology terms. SVScore[24] v0.6 was downloaded from https://github.com/lganel/SVScore on June 16, 2019. It was run using CADD[25] v1.3, downloaded from https://cadd.gs.washington.edu/download on June 16, 2019, using default settings. AnnotSV[21] v2.3.2 was downloaded from https://github.com/lgmgeo/AnnotSV on Feb 27, 2020. AnnotSV was run using human annotation and default settings.

Structural variant features

All gene and exon boundaries used to determine features came from Ensembl Genes v96 as described above. Each SV was annotated with the following 17 features:

Feature category	Feature description	Data type	Aggregation method for multiple genes
CDS	Fraction of CDS adjacent to start codon that is not disrupted by SV	float	min
CDS	Fraction of CDS adjacent to stop codon that is not disrupted by SV	float	min
CDS	Fraction of CDS overlapping SV	float	max
Conservation	Average phyloP score of the 400 most conserved overlapping nucleotides	float	NA
Expression	Exon expression (see Methods)	float	NA
Expression	Exon inclusion (see Methods)	float	NA
Expression	TAD boundary strength (according to Gong et al[54])	float	max
Gene importance	LOEUF of gene	float	min
Gene importance	LOEUF of gene where stop codon overlaps SV or >50% of CDS overlaps SV	float	min
Gene importance	pLI of gene	float	max
Gene importance	pLI of gene where start codon overlaps SV or >50% of CDS overlaps SV	float	max
Other	All overlapped exons can be skipped in frame	boolean	NA
Other	Any overlapped exon is constitutive	boolean	NA
Other	Minimum exon transcript order*	integer	min
Other	Number of exons in canonical transcript of gene	integer	min
Other	Number of exons SV overlaps by 1 or more bp	integer	max
Other	SV is deletion or duplication	boolean	NA

Table 1.1: Features used in StrVCTVRE. *exon transcript order was defined as the number of exons preceding a given exon in a gene.

Expression features were derived from transcript data downloaded from the GTEx Portal v7[55]. Exon expression was calculated for each nucleotide as the sum of the transcripts per million (TPM) of fragments that map to that nucleotide. Exon inclusion estimated the proportion of transcripts generated by a gene that include a given nucleotide and was calculated for each nucleotide as the TPM of fragments that map to that nucleotide, divided by the sum of TPM that map to the gene containing that nucleotide. For both features, adjacent base pairs with the same value were merged together into genomic intervals. For SVs that overlapped more than one of these genomic intervals, exon expression was calculated by averaging the 400 highest exon expression genomic intervals contained in that SV. The same was done for exon inclusion. All GTEx tissues were used in this analysis.

To determine which conservation feature to use, we assessed the accuracy of both PhastCons[56] and PhyloP[30] in discriminating between pathogenic and benign SVs using the average of the highest-scoring 200, 400, 600, 800, and 1000 nucleotides. The test set consisted of 200 small (< 800 bp) SVs randomly selected from our pathogenic and benign SV training datasets (as described above). We found the mean PhyloP score of the 400 most conserved nucleotides in an SV was among the highest accuracy predictors. For both conservation and expression features, if the total overlap between the SV and all exons was less than 400 intervals, then the values of the overlapped intervals were averaged together to calculate the feature. Median imputation was used to fill in missing feature annotations.

In Fig. 2.1a, features were clustered by correlation using the linkage and fcluster functions from the SciPy[57] v 1.1.0 hierarchical clustering package. The input to this figure were the features for all SVs used as training data. Values for some features were reversed to ensure most matrix correlations are positive.

Random forest classification

StrVCTVRE was implemented as a random forest classifier in Python with scikit-learn[58] v0.17, using class RandomForestClassifier. A grid search was performed to find the optimal hyperparameters by using a leave-one-chromosome-out cross validation strategy and validation only on ClinVar data, as described previously. The hyperparameters searched included: the max depth of a tree (5, 10, 15, No limit), max features considered at each split (1, 2, 3, 4), the minimum samples at each leaf node (1, 2, 4), the minimum samples required to split a node (2, 4), the number of trees generated (500, 1000, 3000), and whether to use out-of-bag samples to estimate accuracy (True, False). Several combinations of features performed similarly well, and we chose one that performed well while unlikely to over-fit to the training data—max depth: 10, max features considered at each split: 1, minimum samples at each leaf node: 2, minimum samples required to split a node: 4, number of trees: 1,000, out of bag samples: False. Feature importance used in figures is also known as Gini importance[59], and was calculated using the feature_importances_ attribute of RandomForestClassifier.

Figures

In Fig. 2.1b, 95% confidence intervals were derived by generating 1,000 random forest predictors.

In Fig. 2.2a, the data were generated by using a leave-one-chromosome out approach that included all chromosomes besides chromosomes 1, 3, 5, and 7 (e.g., SVs in chromosome 2 were assessed using training data from chromosomes 4, 6, 8, 9, 10, etc.).

In Fig. 2.2b, to create the inset testing set, we began with the benign and pathogenic datasets as described above, and only retained ClinVar SVs from each dataset. Next, we removed any SVs larger than 3MB, and for both the benign and pathogenic dataset, we randomly sampled SVs without replacement, such that SVs were retained if they did not overlap any of the same genes as a previously sampled SV. This resulted in a reduced dataset for both pathogenic and benign SVs, in which every gene was overlapped by at most a single SV. Pathogenic and benign SVs from these reduced datasets were then matched by size as described above, and only results from testing on SVs on chromosomes 1, 3, 5, and 7 are shown in the Fig. 2.2b inset.

In Fig. 2.2b, 4a, and 4b, AUC 95% confidence intervals were derived by calculating the AUC standard error following Hanley and McNeil[60].

In Fig. 2.4b, 90% sensitivity thresholds were derived from StrVCTVRE and SVScore performance on the ClinVar held-out test set (dotted line, Fig. 2.2b).

Method availability

The StrVCTVRE command line tool can be downloaded from https://compbio.berkeley.edu/proj/strvctvre.

References

1 Sharo, A. G., Hu, Z. & Brenner, S. E. StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants. *BioRxiv* (2020).

2 Clark, M. M. *et al.* 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** (2018).

3 Lappalainen, T., Scott, A. J., Brandt, M. & Hall, I. M. Genomic analysis in the age of human genome sequencing. *Cell* **177**, 70-84 (2019).

4 Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372** (2021).

5 Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779-786 (2021).

6 Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics* **14**, 125 (2013).

7 Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83-89 (2020).

8 Holt, J. M. *et al.* Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing. *BioRxiv*, 627661 (2019).

9 Wu, N. *et al.* TBX6 null variants and a common hypomorphic allele in congenital scoliosis. *New Engl. J. Med.* **372**, 341-350 (2015).

10 Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437-455 (2010).
11 Ascari, G. *et al.* Long-Read Sequencing to Unravel Complex Structural Variants of CEP78 Leading to Cone-Rod Dystrophy and Hearing Loss. *Frontiers in cell and developmental biology* **9** (2021).
12 Zampaglione, E. *et al.* Copy-number variation contributes 9% of pathogenicity in the inherited retinal degenerations. *Genet. Med.* **22**, 1079-1087 (2020).
13 Wright, C. F. *et al.* Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med.* **20**, 1216 (2018).
14 Sanchis-Juan, A. *et al.* Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short-and long-read genome sequencing. *Genome medicine* **10**, 95 (2018).
15 Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology* **20**, 117 (2019).
16 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
17 Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020).
18 McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biology* **17**, 122 (2016).
19 Cingolani, P. *et al.* 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 (2012).
20 Sedlazeck, F. J. *et al.* Tools for annotation and comparison of structural variation. *F1000Research* **6** (2017).
21 Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572-3574 (2018).
22 Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
23 Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**, 1-16 (2013).
24 Ganel, L., Abel, H. J., Consortium, F. & Hall, I. M. SVScore: an impact prediction tool for structural variation. *Bioinformatics* **33**, 1083-1085 (2017).
25 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-D894 (2019).
26 Kumar, S., Harmanci, A., Vytheeswaran, J. & Gerstein, M. B. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome biology* **21**, 1-21 (2020).
27 Li, M.-X. *et al.* Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.* **9**, e1003143 (2013).
28 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (2016).
29 Rodriguez, J. M. *et al.* APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.* **46**, D213-D217 (2018).

30 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110-121 (2010).
31 Kumar, S., Harmanci, A., Vytheeswaran, J. & Gerstein, M. B. SVFX: a machine-learning framework to quantify the pathogenicity of structural variants. (2019).
32 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-D1067 (2018).
33 Kronenberg, Z. N. *et al.* High-resolution comparative analysis of great ape genomes. *Science* **360**, eaar6343 (2018).
34 Riggs, E. R. *et al.* Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.*, 1-13 (2019).
35 Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581 (2016).
36 Donkervoort, S. *et al.* Biallelic loss of function variants in SYT2 cause a treatable congenital onset presynaptic myasthenic syndrome. *Am. J. Med. Genet. A* **182**, 2272-2283 (2020).
37 Töpf, A. *et al.* Sequential targeted exome sequencing of 1001 patients affected by unexplained limb-girdle weakness. *Genet. Med.* **22**, 1478-1488 (2020).
38 Ravenscroft, G. *et al.* Neurogenetic fetal akinesia and arthrogryposis: genetics, expanding genotype-phenotypes and functional genomics. *J. Med. Genet.* (2020).
39 Zampaglione, E. *et al.* The Importance of Automation in Genetic Diagnosis: Lessons from Analyzing an Inherited Retinal Degeneration Cohort with the Mendelian Analysis Toolkit (MATK). *medRxiv* (2021).
40 Wahlster, L. *et al.* Familial thrombocytopenia due to a complex structural variant resulting in a WAC-ANKRD26 fusion transcript. *J. Exp. Med.* **218**, e20210444 (2021).
41 Firth, H. V. *et al.* DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics* **84**, 524-533 (2009).
42 MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986-D992 (2014).
43 Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75 (2015).
44 Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11. 23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011).
45 Richards, 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. *Genet. Med.* **17**, 405 (2015).
46 Tavtigian, S. V. *et al.* Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genet. Med.* **20**, 1054-1060 (2018).
47 Köhler, S. *et al.* Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics* **85**, 457-464 (2009).
48 Singleton, M. V. *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *The American Journal of Human Genetics* **94**, 599-610 (2014).

49 Zemojtel, T. *et al.* Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Science translational medicine* **6**, 252ra123-252ra123 (2014).
50 Shah, N. *et al.* Identification of misclassified ClinVar variants via disease population prevalence. *The American Journal of Human Genetics* **102**, 609-619 (2018).
51 Harrison, S. M. & Rehm, H. L. Is 'likely pathogenic'really 90% likely? Reclassification data in ClinVar. *Genome medicine* **11**, 1-4 (2019).
52 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
53 Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884-D891 (2021).
54 Gong, Y. *et al.* Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nature communications* **9**, 542 (2018).
55 Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580 (2013).
56 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050 (2005).
57 Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261-272 (2020).
58 Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825-2830 (2011).
59 Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. (Springer Science & Business Media, 2009).
60 Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36 (1982).

Chapter 3: Assessing the clinical value of detecting structural variants with optical mapping and linked-read sequencing to diagnose rare monogenic disorders*

Background†

Sequencing methods have revolutionized the clinical care of rare diseases, but serious challenges remain. Short-read sequencing of exomes and genomes resolves 26 to 40% of cases that could not be diagnosed by standard methods, leading to improved care and cost savings[2,3]. Yet, this leaves many cases unresolved. Some of these unresolved cases may be caused by structural variants (SVs) that either could not be detect or were of unknown significance. SVs include deletions, duplications, and rearrangements (such as translocations, inversions, and insertions) larger than 50bp. Long-read sequencing routinely identifies more than 20,000 SVs in each genome[4,5]. For many rare diseases, SVs cause a minor but considerable fraction of disease. In one study of 119 probands with undiagnosed diseases, 13% of resolved cases were caused by an SV[6].

SVs remain difficult to detect. Unlike SNVs, for which best practices are well established and methods are accurate, there are several diverse techniques used to detect SVs, few of which are accurate alone. These techniques also vary widely in cost. Established cytogenic methods such as oligo array CGH and SNP array are able to accurately detect large SVs, but without good breakpoint accuracy. A single test that offers both SNV and SV detection is attractive, since it reduces the number of samples needed from a patient and reduces time spent ordering and analyzing tests. Such a combined test is possible with sequencing-based methods such as short-read, linked-read, and long-read sequencing. To detect SVs from short-read sequencing, computational methods use clues from unexpected read pair orientation or distance, changes in read depth, and single reads that span an SV breakpoint. However, these techniques are hindered by the repetitive nature of as much as 5% of the human genome. Worse yet, it is exactly in these repetitive areas of the genome in which breakpoints for many SVs lie[7]. Linked-read whole genome sequencing from 10x Genomics uses barcodes to computationally assemble short reads into long reads, and has been used successfully to resolve clinical diagnoses in prenatal testing[8]. Specifically, the process begins with high molecular weight genomic DNA. A specialized library preparation is used to barcode short reads that originate from the same DNA molecule. These short reads are then sequenced on a short-read sequencer. During mapping, barcodes from each read allow for the construction of pseudo-long reads, called linked reads. SNVs can be accurately called from these data as well.

Long-read methods, available from Oxford Nanopore and PacBio, sequence long-reads (>10kb), most of which can be unambiguously mapped to the genome. SVs are then spanned by or entirely contained within the reads themselves, which leads to more accurate calls and precise breakpoints[9]. However, accurate SNV calling requires high-depth long reads because the per-base error rate is high. These methods are thus prohibitively expensive in most clinical

* The majority of this chapter was written by Andrew Sharo. Parts of this chapter, where indicated, were adapted from a published article[1]. Andrew Sharo performed the work described in those sections he wrote, while the sections written by others describe work that was performed primarily by others. This work is included with permission from the authors.

† Parts of the first two paragraphs in this section were adapted from a published article[1] and primarily written by Joseph Shieh, Monica Penon-Portmann, and Karen Wong.

contexts. Optical mapping by Bionano Genomics, a non-sequencing-based method to call SVs, has been used in clinical SV discovery, such as identifying SVs in a Duchenne muscular dystrophy cohort[10]. Starting with high molecular weight genomic DNA, custom restriction enzymes are used to nick the DNA, which is repaired with a fluorescent nucleotide. This gives a sequence-specific pattern of fluorescent nucleotides for each DNA molecule, which is optically recorded by linearizing DNA molecules in nanometer-scale channels. These overlapping fluorescent patterns are then aligned to each other to create contigs. These contigs can then be compared to a reference genome pattern to detect SVs. There are two important caveats: SVs breakpoints are not able to be determined more precisely than within several kb, and SVs smaller than 1 kb are difficult to detect with confidence. These limitations stem from the fluorescent nucleotides, which are located approximately every 10kb in the genome and for which the location can only be accurately determined within several hundred base pairs. However, among the SVs that can be detected, the sensitivity and specificity of calls is generally quite high[11].

Detecting SVs is only half the battle. Identifying one or two disease-causing SVs out of the thousands of SVs present in an individual is a monumental task. The American College of Medical Genetics and Genomics has released guidelines for clinicians to determine the pathogenicity of SVs, but they are demanding and cannot reasonably be done for more than a few top candidates in clinical setting with limited time[12]. Thus, there are several steps to address this challenge by removing SVs from consideration that are likely not pathogenic. A crucial first step is to remove low quality SVs. All methods have some fraction of false positive calls (that is, variants that are detected but do not exist in the genome), which can be eliminated by considering only those SVs with quality indicators above a certain threshold or which pass certain filters. Generally, these thresholds are specific to each method, and we explore one way to calibrate these quality thresholds in our analysis. Next, it is important to remove common SVs (>1%) which occur too frequently to cause a rare disease. Just like SNVs, many SVs are population specific, so reference SVs must come from a diverse cohort. gnomAD SVs provide one source of diverse SV allele frequencies[13]. At the same time, it is equally important to use a set of SVs that were called with the same pipeline to remove SVs that are systematic false positives due to errors in alignment or calling. Below, we quantify the extent of systematic false positives across methods.

This process leaves a set of rare, high-quality SVs which could include tens to hundreds of SVs depending upon the methods used. Many of these SVs will be intronic or intergenic. These SVs are typically more difficult to interpret, and they are expected to be more benign on average than SVs that affect coding regions. As a first pass, a clinician may consider only SVs that overlap a coding region. This step highlights the importance of accurate breakpoint prediction to distinguish between coding and non-coding SVs, which we compare between methods below. Additionally, researchers predict that the majority of rare gene-altering SVs are deletions and duplications[14], so we may consider these SVs first. Below we investigate the accuracy of methods to distinguish deletions and duplications. A handful of methods exist to prioritize deletions and duplications based on overlap with genomic features[15-17] or overlap with known pathogenic SVs[18]. By investigating these prioritized SVs and considering the match between patient phenotypes and the phenotypes associated with the affected gene, a clinical researcher may identify one or more SVs that explain the patient phenotypes.

Here we compare the diagnostic value of optical mapping and linked-read sequencing to detect diagnostic and candidate SVs in rare monogenic disorders. We also consider the diagnostic value of short-read sequencing that is derived from trimmed linked-read data (see Methods). We use Smoove, the successor to Lumpy[19], which integrates multiple pieces of evidence (split reads, paired reads, and read depth) to call and genotype SVs[20], to identify SVs from short-read data. Independent assessments have identified Smoove as a top-performing method[21]. Linked-read sequencing and optical mapping are expected to detect more SVs with greater accuracy. However, due to the difficulty of prioritizing SVs, it is not obvious that this will translate to greater clinical value. Indeed, a recent study comparing short-read and linked-read sequencing for germline SV detection in a clinical setting found no improvement[22].

Methods*

DNA extraction and preparation

High molecular-weight DNA was extracted and isolated using the Bionano Prep Blood Isolation Kit following the manufacturer protocol (Bionano Genomics). Bionano optical mapping libraries were prepared following the manufacturer protocol (Bionano Genomics). 10x Genomics linked-read sequencing libraries were built as published[23] using the GemCode platform (10x Genomics).

Optical mapping and linked-read data generation and processing

Optical mapping on the Bionano Irys and Saphyr platforms was used to produce de novo assemblies and identify SVs and rearrangements. DNA was labeled using Nick, Label, Repair and Stain (NLRS) and/or Direct Label and Staining Technologies (DLS). The first uses a nicking endonuclease that recognizes a specific 6-7 base pair sequence and creates a single-strand nick, filled with fluorescent nucleotides. The second uses a single direct-labeling enzymatic reaction to attach a fluorophore to a specific 6-basepair DNA sequence motif. Labeled DNA libraries were loaded onto the Bionano Genomics Irys™ Chip or Saphyr™ Chip, linearized and visualized using the Irys™ or Saphyr™ system, which detects the fluorescent labels along each molecule. Single molecule maps were assembled de novo into genome maps using Bionano Solve with the default settings[24]. Genome assembly and alignment was performed using IrysView/IrysSolve software. For optical mapping, we performed embedding of cells, long DNA extraction and Chip run over a total 3.25 days.

Linked-read sequencing data was obtained from 10x Genomics libraries sequenced to ~60X coverage using an Illumina sequencer. Reads were aligned to GRCh38 using LongRanger and SVs were identified using the callers integrated in the 10x pipeline including GATK Haplotype caller for SNPs and indels. SNPs and indels were kept for analysis if the minor allele frequency is ≤5% as reported in the gnomAD database.

Optical mapping SV filtering

We considered all SVs present in the bionano Smap files. Translocations were identified as SVs for which ‘RefcontigID1’ and ‘RefcontigID2’ did not match. Only SVs greater than 50 bp and with ‘Present_in%_of_BNG_control_samples’ ≤ 0.5 were retained. Only SVs with

* The sections titled ‘DNA extraction and preparation’, ‘Optical mapping and linked-read data generation and processing’, and ‘Approvals and phenotypic assessment’ were adapted from a published article[1] and primarily written by Joseph Shieh, Monica Penon-Portmann, and Karen Wong.

'Fail_assembly_chimeric_score' of 'pass' or 'not_appliable' and with 'Found_in_self_molecules' of 'yes' were considered further, as recommended by nanotatoR[25]. To generate Fig. 3.3, we removed deletions with a 'Confidence' below 0.99.

Linked-read LongRanger SV filtering

We removed SVs for which length and start position were perfect multiples of 10,000 since these were found to be largely false positives called exclusively based on sequence depth. Deletions and duplications with a cohort frequency greater than 0.5% were removed. For rarity filtering, SVs were considered equivalent if they had a reciprocal overlap ≥ 80% and their corresponding breakpoints were within 10,000 bp of each other. Translocations were assumed the same if their corresponding breakpoints were within 100bp. SVs were removed if their filter was anything other than 'PASS'. To generate Fig. 3.3, duplications with a quality below 3 were removed, and deletions with a quality below 4 were removed.

Short-read Smoove SV filtering

To create our short-read dataset, we used the FASTX-Toolkit (RRID:SCR_005534) to trim the first 24 bases from the forward reads of the linked-read fastq files. This step removes the barcode information which is used to assemble reads into long contigs. Next, we used bwa-mem[26] (v 0.7.10-r789) 'mem' command with default parameters to align the trimmed reads to GRCh38. We used Samblaster[27] to remove duplicates and add mate tags, with a maxSplitCount of 2 and minNonOverlap of 20. We first called SVs in 122 samples using Smoove v0.2.5 downloaded on April 6, 2020. Smoove 'call' was run with all default parameters as well as excluding intervals defined in the file available at https://github.com/hall-lab/speedseq/blob/master/annotations/exclude.cnvnator_100bp.GRCh38.20170403.bed. For each sample, to remove common SVs, we removed SVs with more than 1 heterozygote and any number of homozygotes in the cohort. For filtering, SVs were considered equivalent if they had a reciprocal overlap ≥ 80% and their corresponding breakpoints were within 10,000 bp of each other. Translocations were assumed the same if their corresponding breakpoints were within 100bp. SVs with an MSHQ below 4 were removed, except for those with an MSHQ of -1. To generate Fig. 3.3, deletions with quality below 266.6 and duplications with quality below 20.32 were removed.

Linked-read Smoove SV filtering

Filtering steps were identical to short-read Smoove SV filtering (above) except SVs were called directly from the LongRanger bam files using Smoove. To generate Fig. 3.3, deletions with quality below 251.44 and duplications with quality below 29.38 were removed.

SV prioritization

For all methods, our final step was to remove exon-affecting SVs, defined as any SV that has at least 1bp of overlap with an exon as defined by Ensembl biomart[28], genes v96, GRCh38.p12, limited to genes with HGNC Symbol ID(s) and APPRIS annotation[29]. For transcripts that tied for highest APPRIS annotation, the longest transcript was used. Exon overlap was determined using bedtools intersect[30]. StrVCTVRE[15] v.1.6, downloaded on May 7, 2020, was used to prioritize exon-affecting deletions and duplications by pathogenicity.

Calculating 90% sensitivity threshold of each method

For cases 1903, 2203, 3403, 4203, and 5104 we identified the 90% sensitivity threshold for the quality feature using the following method. Because optical mapping calls are known to be reasonably accurate for deletions and duplications larger than 1,000 bp, they were treated as a quasi-truth set. Considering deletions called by optical mapping, we removed SVs for which the size inferred by 'RefStartPos' and 'RefEndPos' did not match size given by 'Size' or for which size was less than 1kb. Due to the limited resolution of SV breakpoints called by optical mapping, the following method was devised. For SVs discovered in our linked-read LongRanger analysis, we retained optical mapping SVs for which ≥1% was overlapped by a linked-read SV, and ≥50% of the same linked-read SV was overlapped by the same optical mapping SV. SVs were retained only if the optical mapping SV size was between 85% and 115% of the linked-read SV size. SVs less than 1kb were removed. The same process was repeated for SVs identified in our linked-read Smoove and short-read analysis. Then, all the SVs that were in common between these three sets were retained. This was performed for all five cases. These SVs were concatenated together across the five methods. Then, for each method, the quality threshold was identified for which 90% of these common SVs were detected. This quality threshold was used at the 90% sensitivity threshold. This process was repeated for duplications.

gnomAD comparison

gnomAD 2.1 SVs[13] aligned to GRCh37 were downloaded on Dec 11, 2019. We used the University of California, Santa Cruz liftover tool[31] to convert SVs to GRCh38. Deletions and duplications with a popmax allele frequency greater than 0.5% and a PASS Filter were retained. For each method, these SVs were then used to removed common SVs using exactly the same procedure as for cohort SVs. To generate Fig. 3.4, for each case, we summed the number of deletions and duplications to investigate when gnomAD is used and divided by the sum of deletions and duplications to investigate when our cohort was used. Across the five cases, we plotted the mean fold increase with a 95% confidence interval, calculated as 1.96 times the std deviation computed with the students t test.

Breakpoint accuracy

We manually determined the difference between true SV start and predicted SV start as well as true SV end and predicted SV end for the diagnostic or candidate SVs in cases 1903, 2203, 3403, and 4203. No significant accuracy differences were observed between deletions and duplications, or start and stop. True breakpoints in 5104 were not able to be determined, so the case was not included. Mean distance is shown for 4 samples, considering both start and end, for a total of 8 measurements. 95% confidence interval is calculated as described above in gnomAD comparison.

Approvals and phenotypic assessment

The study was approved by the Institutional review board of Children's Hospital Oakland and University of California, San Francisco (UCSF), Committee for Human Subjects Research. Recruitment was from UCSF Benioff Children's Hospital Medical Genetics and Genomics clinics. In recruiting patients, we focused on cases of two types: cases in which whole-exome sequencing had not returned a diagnostic variant; and sporadic cases from the pediatric population that are suspected to have a genetic basis, but fall into no clear syndrome and have no clear candidate target for conventional genetic diagnosis. Individuals with undiagnosed conditions and unaffected parents were offered testing and underwent an informed consent process prior to blood draw. The nature and possible risks of the study were explained in the

consent process. Phenotypic evaluation was performed by clinical review by at least two genetics professionals, and human phenotype ontology terms were curated for each case.

Results*

Optical mapping and linked-read methods detected more diagnostic and candidate SVs than short-read sequencing

We analyzed 50 undiagnosed cases to determine the diagnostic yield of three methods: optical mapping, linked-read sequencing, and short-read sequencing. During recruitment, clinicians proposed unsolved cases for genomic sequencing, and cases were included only if prior testing was negative and there was no clear further specific test. Of the 50 cases, 23 previously had a negative commercial trio whole-exome sequencing, and 42 previously had a negative microarray. Our initial SV pathogenicity assessment integrated two complementary methods (linked-read sequencing and optical mapping) and detected deletions, duplications, translocations, inversions, insertions, and complex SVs.

Our pipeline identified 6 diagnostic SVs and 5 candidate SVs (Table 3.1). 14 diagnostic SNVs were identified as well, which gives a total diagnostic yield of 40% (20 out of 50 cases). Diagnostic SVs were found in 12% of cases (6 out of 50). 4 of the 6 diagnostic SVs were not discovered in a prior trio exome analysis, and all were missed in a prior microarray analysis.

In a 9-month-old female with craniosynostosis and syndactyly, we found a rare 32kb heterozygous de novo intronic duplication within *NHEJ1* (Fig. 3.1a,b; case 1703). Similar cases have been described under the name chromosome 2q35 duplication syndrome, but this duplication narrows the critical region of the *NHEJ1* intron that is important for the condition[32,33] (Fig. 3.1c). The duplication affects an enhancer for the Indian Hedgehog (IHH) gene, located within the third intron of *NHEJ1*. This SV was detected by both optical mapping and linked-read sequencing but could not be detected by short-read sequencing.

Considering all diagnostic and candidate SVs, both linked-read and optical mapping were able to identify 11 of 11 SVs. When cases were re-analyzed with short-read sequencing, we were able to identify 7 out of 11 SVs. There was no obvious explanation why the diagnostic or candidate SVs in these particular 4 cases (1703, 2303, 2803, 5103) were missed by short-read sequencing. We suspect random differences in read coverage played an important role. Consider, two siblings were identified to have identical biallelic diagnostic deletions in *TANGO2* (cases 5103 and 5104). Short-read sequencing was able to identify the deletions in one sibling (5104), but did not identify the deletions in the other sibling (5103).

* Parts of the section 'Optical mapping and linked-read methods detected more diagnostic and candidate SVs than short-read sequencing' were adapted from a published article[1] and primarily written by Joseph Shieh, Monica Penon-Portmann, and Karen Wong.

Case	SV type, zygosity	SV size	Disrupted element	Condition	Prior trio exome	Prior microarray	Inheritance	Location
Diagnostic cases								
0703	Translocation, Heterozygous	NA	*AGBL4*	Neuroblastoma, GDD	+	+	De novo	chr1: 49,553,197; chr9: 29,096,677
1703	Duplication, Heterozygous	32kb	*NHEJ1* and *IHH* enhancer	2q35 duplication syndrome	-	+	De novo	chr2: 219,102,941-219,134,976
v2303	Duplication, Heterozygous	302kb	4p16.3	4p16.3 duplication	-	+	De novo	chr4: 447,742-749,750
4203	Deletion, Heterozygous	1.5kb	*WAC*	DeSanto-Shinawi syndrome	+	+	De novo	chr10: 28,615,989-28,617,469
5103 & 5104	Deletion, Homozygous	36kb	*TANGO2*	*TANGO2* metabolic encephalopathy	+	+	Compound heterozygous	chr22: 20,039,637-20,075,714; chr22: 20,041,469-20,075,432
Candidate cases								
1903	Duplication, Heterozygous	515kb	*SOS1, L2HGDH*	Methylmalonic acidemia, mental retardation, siderius	+	-	Paternal	chr14: 49,926,135-50,440,756
2203	Duplication, Heterozygous	440kb	*XIAP, STAG2*	Wide mouth, facial asymmetry, profound hearing loss	-	+	De novo	chrX: 123,760,207-124,200,537
2803	Deletion, Heterozygous	4Mb	2q32.1	TAPVR, abnormality of the Eustachian tube, 2-3 toe syndactyly	-	+	De novo	chr2: 182,906,065-186,543,239
3403	Duplication, Heterozygous	1.5Mb	*ADPRHL2, KIAA0319L*	Speech delay, hypotonia, coarse facial features, thorax asymmetry, juvenile rheumatoid arthritis	-	+	De novo	chr1: 35,213,729- 36,753,629
4603	Translocation, Heterozygous	NA	*ASB1, TYW1B*	GDD, multipole basal ganglia strokes, microcephaly, synophrys, epicanthus	-	+	Maternal	chr2: 238,439,967; chr7: 72,774,107

Table 3.1 Clinical and molecular features of SVs discovered in diagnostic and candidate cases. Genome coordinates refer to human genome build GRCh38. This table was adapted from a published article[1].

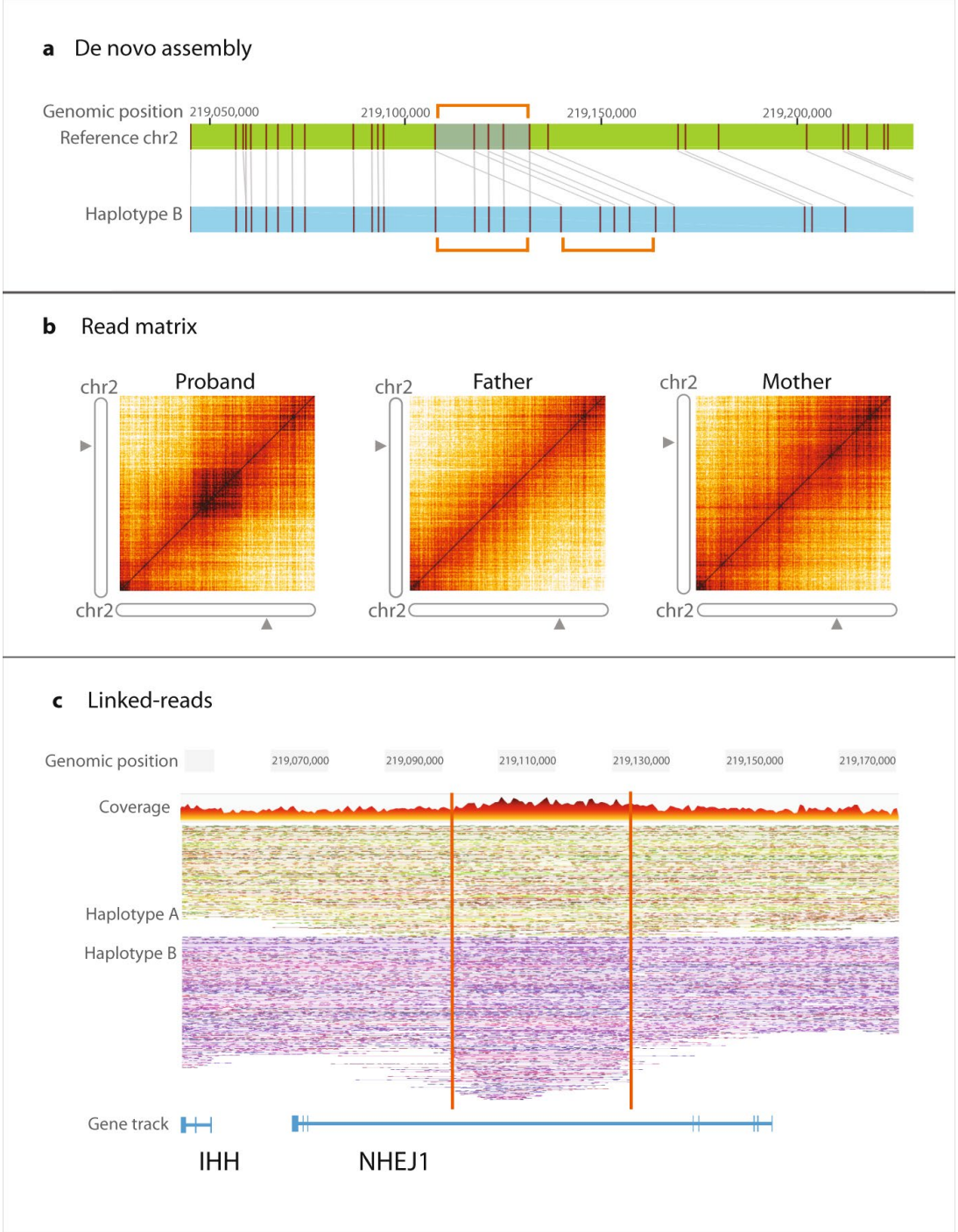

Figure 3.1 Heterozygous, intronic tandem duplication (32 kb) in NHEJ1**.** The affected region includes an IHH upstream enhancer and narrows the diagnostic interval for this condition. **a** Optical mapping assembly (light blue) and its alignment to reference (green). The labeled motifs in the reference genome (vertical maroon lines) are duplicated in the de novo assembly and their orientation demonstrates the duplication occurred adjacent to the original sequence, in tandem. **b** A matrix view of linked reads. The dark orange square in the left panel (proband), illustrates a higher density of barcode overlap in the read matrix compared to either parent, indicating the variant likely occurred de novo. **c** Phased haplotypes generated using linked-read

data. Haplotype B, in purple, contains the intronic region with a greater depth of linked reads due to sequence duplication. This figure was adapted from a published article[1].

Optical mapping detects the largest number of rare exonic deletions and duplications

To compare the three methods, we next consider the seven cases (0703, 1903, 2203, 3403, 4203, 4603, 5104) in which the diagnostic or candidate SV was detected by all methods. For each case, we calculated the number of duplications, deletions, and translocations with a quality greater than the diagnostic or candidate SV quality (Fig. 3.2). Since our linked-read analysis only reported duplications larger than 30kb, here we include for each method only those duplications larger than 30kb. We found that optical mapping detected a median of 11-fold more confident duplications than linked-read sequencing and 3-fold more than short-read sequencing (Fig. 3.2a). When only rare duplications are considered (rarity determined as described below), optical mapping continues to call more duplications, but linked-read and short-read sequencing detect a similar number of SVs (Fig. 3.2d). Given that our linked-read analysis is expected to be more accurate than our short-read analysis, this may reflect the fact that some of the confident short-read duplications are actually systematic false-positives that are called in many samples and eliminated when common duplications are removed. Among SVs that are of primary clinical interest, rare exonic duplications, we find that optical mapping confidently calls a median of 5, which is more than either linked-read (median of 1) or short-read (median of 0) sequencing (Fig. 3.2g).

We noticed a very similar pattern among deletions, in which optical mapping calls the most confident deletions, with a median of 8 rare exonic deletions compared to 1 for linked-read sequencing and 2 for short-read sequencing (Fig. 3.2h). We note that for all methods only deletions larger than 1kb are considered, due to optical mapping's technical limit on SV resolution. The opposite trend was observed for translocations. We found that our linked-read analysis predicted a median of 10-fold more confident translocations than optical mapping, and our short-read analysis predicted a median of 80-fold more confident translocations than optical mapping (Fig. 3.2c). However, once only rare, genic (overlapping an exon or intron) translocations were considered, the number of predicted translocations were similar across methods (Fig. 3.2i). Given that we expect most genomes have very few, if any, translocations, these data suggest that both linked-read and short-read sequencing predict a large number of confident translocations, but that these are systematic errors which are removed when only rare SVs are considered. Overall, we found that optical mapping identifies a greater number of confident, rare, exonic duplications and deletions than linked-read or short-read sequencing, and that all methods identify a similar number of confident, rare, genic translocations. Although this provides a helpful comparison of methods, this analysis is limited because rarely are SVs prioritized by quality. Instead, researchers typically define a quality threshold based on desired tradeoffs between sensitivity and specificity, which we investigate next.

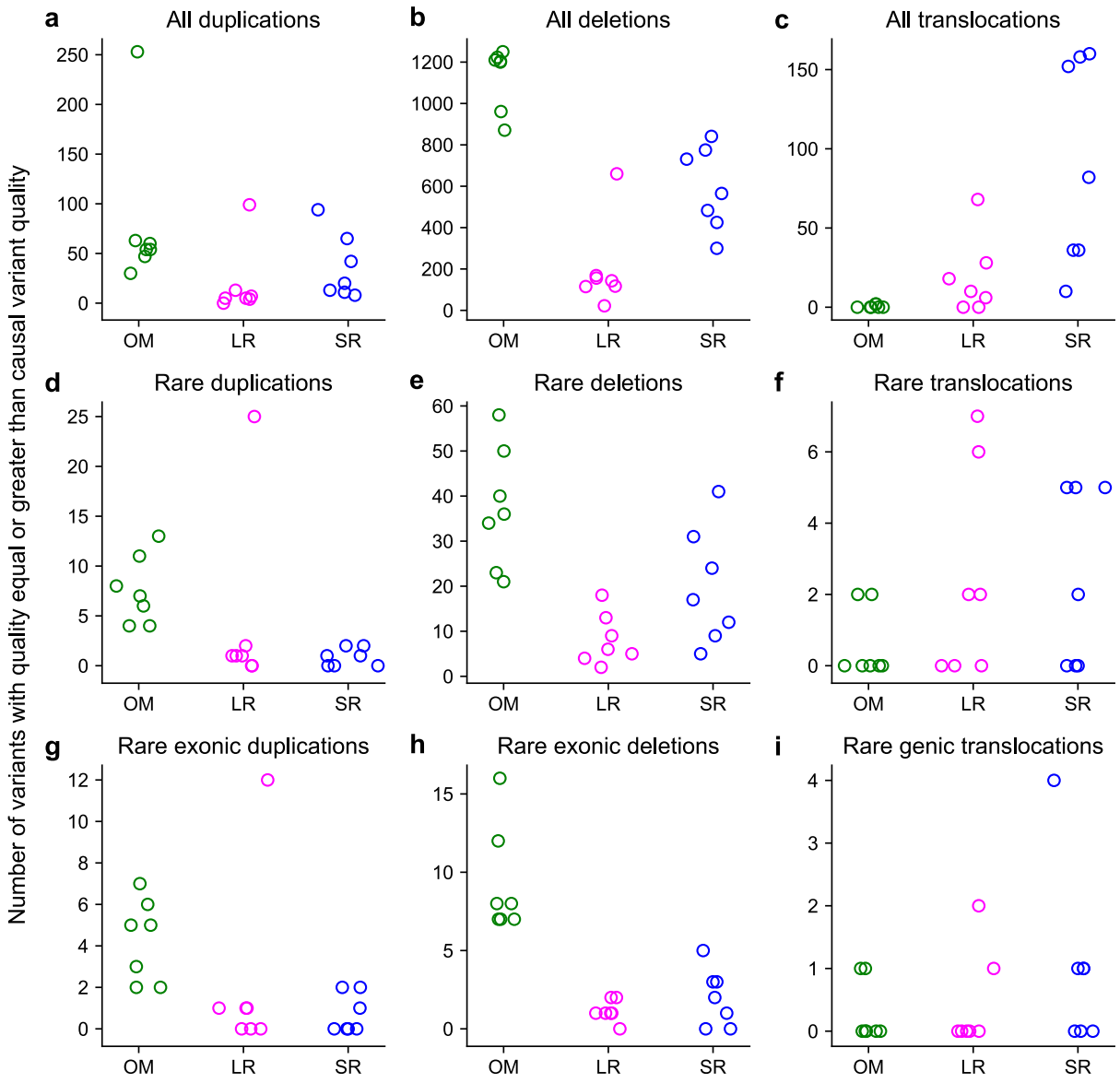

Figure 3.2 High-quality variants called by each method. The number of duplications (left column), deletions (middle column), and translocations (right column) called by optical mapping (green), linked-read sequencing (magenta), and short-read sequencing (blue). In addition to all variants (top row), we show the number of rare variants (middle row) and rare exonic variants (bottom row). Generally, optical mapping identifies a greater number of high-quality duplications and deletions than linked-read sequencing or short-read sequencing. For each case, we show the number of variants with quality equal to or greater than the causal variant. Note that y-axis scale changes between panels.

All methods required a similar number of SVs to be investigated when appropriate filtering strategies were used

We took the following steps to prioritize SVs by pathogenicity. First, we removed low quality SVs. Since a sensitive analysis that does not miss any potentially pathogenic SVs is paramount, for each method we identified a 90% sensitivity threshold. There are some fundamental differences between SVs identified by each method (e.g., linked-read sequencing did not report duplications less than 30kb), so this is not a global 90% sensitivity but rather a 90% sensitivity for the subset of SVs that could be detected by all methods. Recently developed methods (such as optical mapping and linked-reads) can provide challenges in this domain, since best practices are not always well-established, and often independent assessments of SV quality are unavailable. Additionally, peculiarities of a particular protocol for DNA extraction, library preparation, sequencing, and data analysis can create unexpected results. Quality thresholds can readily be determined when one or more standard genomes (for which SVs are already well characterized, such NA12878[34]) are included in the analysis[35]. Depending on the number of batches in which genomes are processed, it may be helpful to include a standard genome in multiple batches to identify batch effects. A standard genome was not included in our data, so we developed an alternative strategy to identify a 90% sensitivity quality threshold. For the following analysis, we considered only the five cases (1903, 2203, 3403, 4203, 5104) for which diagnostic or candidate deletions or duplications were detected by all methods. For each case, we identified the subset of deletions and duplications that were detected by all three methods (see Methods). Presumably, the SVs in this subset are nearly all true positives (that is, they do exist in the genome). For each method, we then used this subset to determine an appropriate quality threshold, which we set such that 90% of the subset SVs were retained. We used these thresholds to remove low-quality SVs from each method's call set.

Next, we filtered SVs by rarity. Since these disorders are expected to be very rare (incidence of <1 in 100,000), We retained only SVs with an allele frequency below 0.5%. There are two potential sources which can be used to identify common SVs: cohort samples and population databases. Cohort samples have the advantage of removing systematic errors in SV calling, which are often absent in population databases. However, depending on the size of the cohort, it may not be sufficient for filtering. Additionally, if a cohort contains samples from different ancestry groups, a diverse population database such as gnomAD SVs may enable filtering of SVs that are common in a single ancestry. In our analysis, we found that filtering SVs by cohort SVs was more valuable than filtering by gnomAD SVs, and that filtering by both did not improve upon filtering by cohort alone. This is possibly due to the inclusion of parents in our sampling. Additionally, parent SVs allowed us to identify de novo SVs. Of our 11 cases, we found that 7 were caused by de novo SVs. Although de novo SVs are enriched for disease-causing SVs, they are also often false positives, and so to remove false positive calls we only flagged an SV as de novo if it did not match any SV called in the parents of any quality.

With this set of high-quality, rare SVs, we next considered only exon-affecting deletions and duplications, as they are expected to constitute the vast majority (>95%) of rare gene-altering SVs[14]. Additionally, deletions are called more accurately than other events[21], and deletions and duplications are more clinically interpretable than SVs such as inversions, insertions, and translocations. Finally, we prioritized these exon-affecting deletions and duplications using StrVCTVRE, an SV impact predictor developed specifically to prioritize exon-affecting deletions and duplications for rare disease. In addition to our three methods, we also ran Smoove directly

on the linked-read alignments. This allowed us to de-couple the limitations of linked-read chemistry from the limitations of the LongRanger computational method that called linked-read SVs.

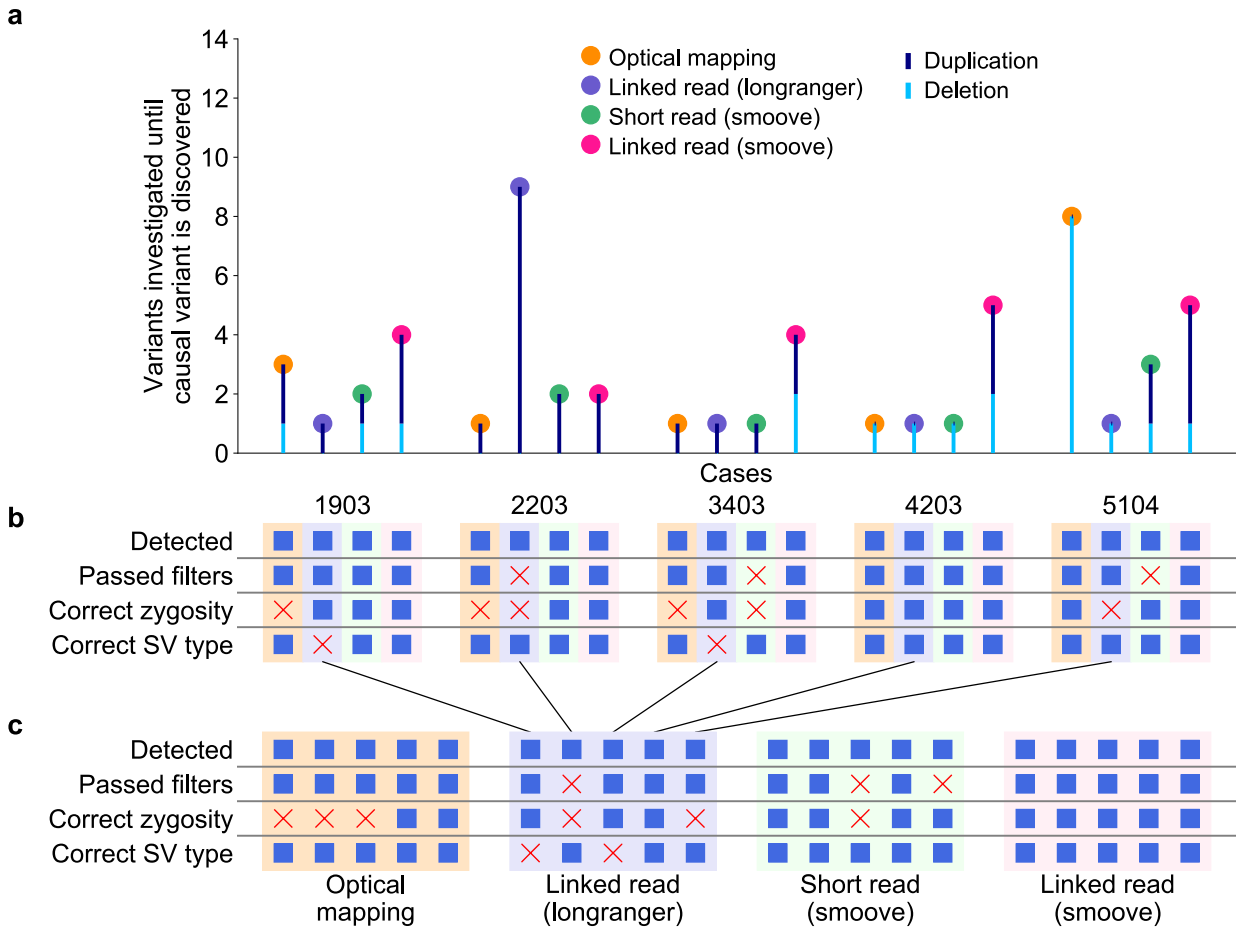

Figure 3.3 Comparison of variants investigated, and errors made by each method across cases. **a** The number of variants that would need to be investigated until the causal variant is discovered. Variants were prioritized by StrVCTVRE score after filtering for quality and rarity (see Methods). Values are grouped by case, dot color indicates method, and line color indicates the number of deletions and duplications. **b** Grouped by case, the types of errors made in each case, with column color indicating method. Blue squares indicate correct classifications, while red X's indicate errors or missing information. **c** The same information as **b** but grouped by method, which shows some clustering in errors by method.

For each of the five cases described above, we prioritized rare, exon-affecting deletions and duplications by StrVCTVRE score. We then calculated how many SVs would need to be investigated until the diagnostic or candidate SV was discovered, assuming the SVs were investigated in order of decreasing predicted pathogenicity (Fig. 3.3a). We found that there was little difference between the three methods in the number of SVs investigated. On average, short-read sequencing required the fewest SVs to be investigated, but this is offset by the fact that it did not identify the diagnostic or candidate SV in 4 of 11 cases. We found that 63% of the

SVs to be investigated were duplications. This overrepresentation could be due to the greater challenge in detecting duplications, thus requiring a lower quality threshold to achieve 90% sensitivity. There were two cases (3404 and 4203) for which all three methods prioritized the diagnostic or candidate SV as the top SV. When we applied Smoove to the linked-read alignment, we found it generally resulted in more SVs investigated than LongRanger, demonstrating that new sequencing methods do require bespoke methods to optimize SV calling. However, in addition to minimizing the number of SVs investigated, there are several additional features that are important in a clinical setting.

Each method is prone to errors in SV type, zygosity, filtering, or breakpoints

Methods varied in their ability to accurately report clinically important features of SVs. In every case, except for 4203, at least two methods incorrectly reported some aspect of the SV (Fig. 3.3b). These errors did not obviously group by case, suggesting that they are not necessarily caused by the complexity of the individual SV. When grouped by method, these errors were more clustered, suggesting that they represent systematic or sporadic limitations of each method (Fig. 3.3c). Out of 5 SVs, the LongRanger algorithm was unable to call SV type in two large (~500kb) duplications. It is clear this was an algorithmic limitation, as Smoove was able to correctly call both SVs as duplications when run on the same alignment data. Clinicians use SV type to interpret the pathogenicity of an SV. For example, duplications of some regions may be benign, yet deletions of the same interval may be pathogenic[12]. In some cases, deletions and duplications of the same region are known to cause different disorders[36]. Balanced rearrangements, such as inversions, are generally less clinically tractable without experimental studies. For these reasons, unambiguous, accurate reporting of SV type is a valuable feature of SV detection methods, which linked-reads sometimes falls short of.

All methods misreported the zygosity of at least one SV in the five cases, and more recent methods had more incorrect reports. Optical mapping was unable to report zygosity for any of the diagnostic or candidate duplications. It is unclear if this is a limitation of the optical mapping chemistry or algorithm. Linked-read sequencing incorrectly reported a candidate heterozygous 440kb duplication as homozygous, which our Smoove analysis revealed to be an algorithmic limitation. Short-read sequencing incorrectly reported a candidate heterozygous 1.5Mb duplication as homozygous. In the absence of other pathogenic variants, an SV in a homozygous state may be of much greater clinical interest than were it heterozygous, due to haploinsufficiency/triplosensitivity that varies across the genome. In these particular cases, all incorrectly reported SVs were heterozygous duplications in copy-sensitive genomic regions, and thus their initial reporting as unknown zygosity or homozygous did not significantly change their clinical interest. This may also reflect the generally greater difficulty of accurately determining the zygosity of duplications compared to deletions. Although many of our cases were caused by dominant SVs, accurate zygosity would be particularly important when recessive conditions are expected.

Most SV methods, in addition to a quality score, provide one or more filters to aid users to remove likely false positives. We found that one candidate SV identified by linked-read LongRanger was marked with the ‘LOWQ’ filter. Given the small number of SVs evaluated, it is difficult to determine if the LongRanger LOWQ filter is too aggressive or if this is a rare incident. We also found two SVs that did not meet the recommended Smoove filtering for MSHQ (see Methods). This appears to be a limitation of short-read alignment accuracy, since we found the same SVs passed filters when Smoove was run on the linked-read alignment. Were these filters

used strictly, both linked-read and short-read sequencing would have missed clinically-relevant SVs.

Breakpoint accuracy varied greatly across methods due to both experimental and algorithmic limitations. Optical mapping measures fluorescent nucleotides that are an average of 10kb apart in the genome, thus the breakpoint resolution of this method is inherently limited. Indeed, across five cases, we found the breakpoints reported by optical mapping were a mean of 4kb from the actual SV breakpoints, with all differences under 9kb (Fig. 3.4a). In one case (4203), if optical mapping had been the sole method used, it would have been impossible to tell whether the diagnostic deletion affected an exonic region, or whether it was completely intronic (Fig. 3.4b). This uncertainty would have made this deletion much less compelling, since it may have been assumed to be intronic in the absence of further evidence. We found breakpoints called by LongRanger were an average of 10 bp from the true breakpoints. To determine if this was an algorithmic limitation, we checked the breakpoints called by Smoove run on the linked-read alignments, and we found the calls were accurate within less than 1 bp, confirming that the LongRanger algorithm was limiting accuracy. We suspect this difference in breakpoint accuracy called on the same alignments is due to the way each algorithm treats of split reads. We found that breakpoints called from short-read alignments were accurate within less than 1bp on average.

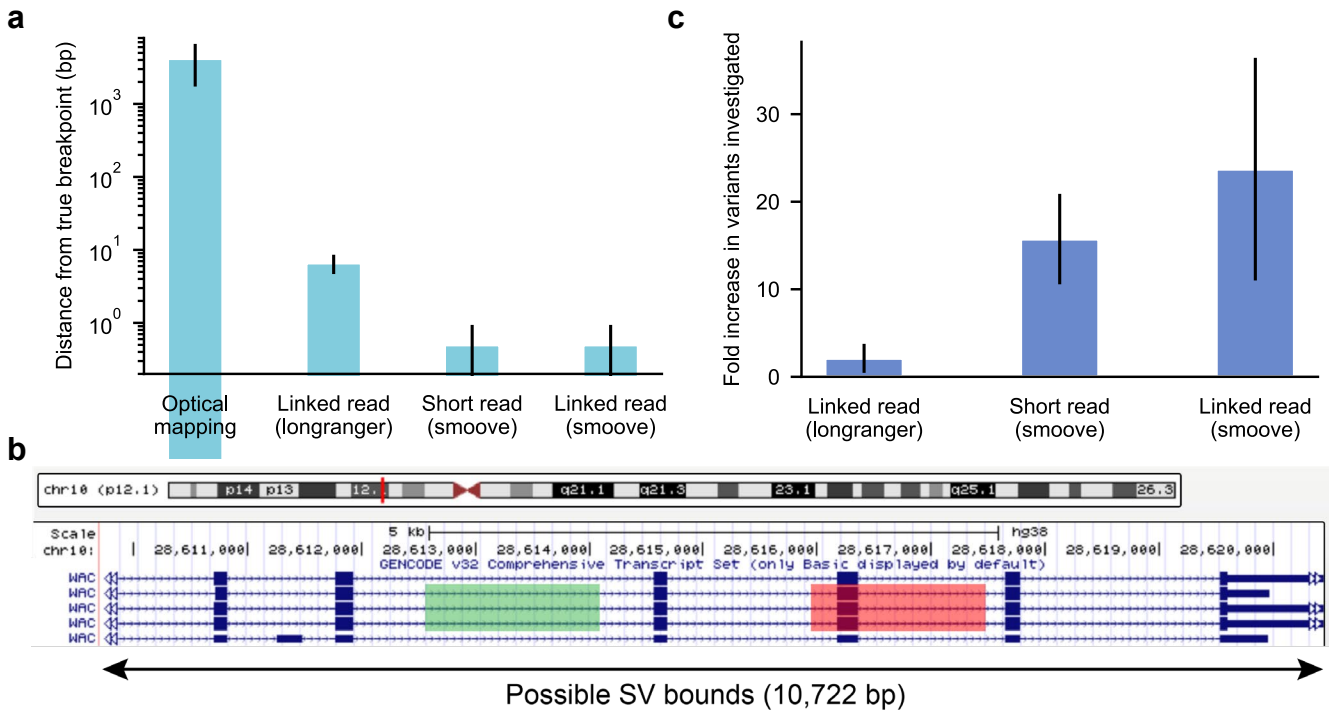

Figure 3.4 a Average distance between the breakpoint identified by each method and the true breakpoint. Error bars show 95% confidence interval. **b** For case 4203, optical mapping indicated that a 1.5kb deletion was located within a 10.7kb region. The red rectangle shows the true location of the variant, but it could easily have been entirely intronic (green rectangle) which would have resulted in lower clinical interest. **c** When gnomAD SVs are used to remove common variants, we find that the number of variants investigated until the causal variant is reached doubles for linked-read sequencing and is significantly greater in other methods.

Filtering by cohort identified rare SVs more reliably than population databases

We had a relatively large cohort (~130 individuals) in which to identify and remove common SVs. However, in many studies that may not be the case. gnomAD SVs are a large, publicly available population database which can be compared against to identify common SVs when a large cohort is not available. To evaluate the appropriateness of using gnomAD SVs in place of cohort SVs to remove common SVs, we re-calculated the number of SVs that would need to be investigated until the diagnostic or candidate SV was found (Fig. 3.3a) when only gnomAD is used to remove common SVs. We found that on average this doubled the number of linked-read sequencing SVs to be investigated (Fig. 3.4c). The number of short-read SVs to be investigated increased 16-fold, and SVs called by Smoove from linked-read alignments increased even more, although this was not significantly greater. We suspect the notable increase in SVs to be investigated in both Smoove methods reflects systematic false positive SV calls, which are not able to be filtered by gnomAD. The even greater increase in SVs called by Smoove from linked-read alignments, although not significant, suggests that bespoke methods are needed to avoid excessive false positives from novel methods.

Discussion*

In genomic medicine, rare disease diagnostics has traditionally been limited by the variants that can be detected and our ability to interpret those variants. Here, we investigate the use of optical mapping and linked-read sequencing to identify SVs implicated in rare genetic diseases. We also outline our pipeline for prioritizing SVs by pathogenicity. In 4 of 11 cases, we find that these methods detect diagnostic or candidate SVs that are missed by short-read sequencing. While SV detection from long DNA technologies can improve the detection of diagnostic SVs, it is not without its limitations.

Optical mapping identified the greatest number of confident, rare, exonic duplications and deletions. The major clinical limitations we identified were its inability to determine zygosity in duplications, and its poor resolution of SV breakpoints. Although this resolution is rarely an issue for large SVs, we observed one 1.5kb diagnostic SV which would have been much less clinically compelling if optical mapping was the only method used. Linked-read sequencing identified all diagnostic and candidate SVs, but it was hampered by algorithmic limitations in calling SV zygosity, SV type, and breakpoint resolution. These finding reinforce known limitations of novel methods: despite intrinsic advantages over existing methods, it may take years for the corresponding algorithms to reach a high level of performance[37]. As of January 2020, linked-read sequencing is no longer available from 10x genomics due to a patent infringement case, but we anticipate that our findings will be broadly relevant to sequencing methods that use long DNA molecules.

As expected, we found evidence that short-read sequencing identifies a greater number of false positive SVs. However, it seems that many of these SVs are systematic errors that occur in multiple samples, or occur in regions with poor genome mappability which tend to be non-exonic[38,39]. Indeed, many of these putative false positives are removed when only rare exon-affecting SVs are considered. As a connected issue, we also found that short-read sequencing filters meant to reduce false positive calls instead removed one diagnostic SV and one

* Parts of the first, sixth, seventh, eight, and ninth paragraphs in this section were adapted from a published article[1] and primarily written by Joseph Shieh, Monica Penon-Portmann, and Karen Wong.

candidate SV. On the other hand, short-read sequencing was the only method that identified breakpoints with single base pair accuracy.

Once SVs have been prioritized by an automated pipeline, a clinical researcher must manually investigate the top candidates to potentially identify a diagnostic SV. We anticipate that our description of SV prioritization may be valuable to others faced with a similar challenge. One step that could be added is to use one of the many phenotype-to-gene methods to prioritize genes based on association with proband phenotypes[40,41]. We chose not to take this step to be able to discover novel disease-gene associations.

Overall, we found that all three methods required a similar number of SVs to be considered before the diagnostic or candidate SV is uncovered. This analysis is limited by the accuracy of the SV prioritization method we used, StrVCTVRE. Additionally, it is not possible to know whether the SVs prioritized above the diagnostic or candidate SV are actually false positives or instead true positives that happened to be prioritized above the diagnostic or candidate SV. For these reasons, this is not a perfect metric, but it remains clinically relevant. A further limitation is we discovered diagnostic and candidate SVs using a combination of linked-read sequencing and optical mapping. This is an important limitation to our analysis, as it may have biased our results to favor optical mapping and linked-read sequencing. For example, there may be diagnostic SVs which could not be detected by these methods but would be identified by short-read sequencing such as duplications between 100 bp and 1kb. This limitation is further compounded by the fact that our short-read sequencing data was derived from our linked-read sequencing data.

For individuals with undiagnosed conditions, optical mapping and linked-read sequencing together encompass what is currently provided by the combination of karyotyping, microarray testing, and short-read WGS. By identifying novel SVs and phasing variants, these methods provide additional diagnostic information beyond current clinical tests, despite the limitations that we have described. One notable advantage is that these methods bypass the need for additional time and blood for testing. These strengths make the technologies suitable for early implementation in diagnostic evaluations, particularly if a specific genetic condition or type of variant is not immediately suspected. Thanks to the long-range phasing offered by these methods, they can also be particularly effective when parents are not available for testing. This may be useful in intensive care units or other settings where rapid diagnosis is vital to clinical care[42,43].

The number of diagnostic cases attributable to SVs was striking in our study, as 43% of exome-negative cases (3 out of 7 cases) that received a diagnostic variant were solved by identifying an SV or rearrangement. We also identified at least one highly probable SV or SNV candidate in more than half of the remaining undiagnosed patients. These cases do not meet diagnostic criteria due to the following reasons. Most SVs are not recurrent and thus do not share identical breakpoints. As a consequence, SVs overlapping similar regions do not always produce the same phenotype. Furthermore, unless a critical region can be established or a syndrome is associated with a very distinct phenotype, it is often unclear whether an SV is diagnostic even if it is de novo. This is made worse by sparse and inconsistent SV databases, in stark contrast to SNV databases. To resolve the remaining cases, researchers will need to discover genotype correlations or perform functional testing.

These methods have additional limitations. Even with the use of long DNA molecules averaging 200–300 kb in our optical mapping experiments, they are not long enough to resolve the large, near-identical segmental duplications in some of the most complex regions of the human genome. Thus, a small number of these complex regions remain inaccessible despite using long-range sequencing and mapping technologies[23]. Additionally, the current human reference genome is a set of composite haplotypes generated from 8 anonymous DNA donors[44]. As such, there are functionally important sequences found in many people around the world but missing from the reference genome[45,46]. Since the reference genome serves as the benchmark for all analyses, missing sequences are never assessed, and variants in these regions are undiagnosable.

Emerging long DNA sequencing methods have the potential to comprehensively identify genetic variants in undiagnosed patients and provide promising new diagnostic possibilities. If the limitations that we identify are addressed, these methods could improve diagnostics for direct clinical care. In this study, we found diagnostic SVs in 6 cases out of 50 families and candidate SVs in an additional 5. Although our ability to assess the impact of these candidate SVs will remain a rate limiting step for some time, these data will still be valuable for later reanalysis. Data reanalysis is becoming a successful strategy to identify variants that underlie disease in a patient's genome[47]; as our understanding of deleterious SVs grows, it will be increasingly possible to revisit previously acquired data and assign pathogenicity to previously detected SVs. By detecting a more extensive set of SVs, optical mapping and linked-read sequencing increase the likelihood that future reanalysis will be productive.

References

1 Shieh, J. T. *et al.* Application of full-genome analysis to diagnose rare monogenic disorders. *NPJ genomic medicine* **6**, 1-10 (2021).

2 Clark, M. M. *et al.* 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** (2018).

3 Stavropoulos, D. J. *et al.* Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *NPJ genomic medicine* **1**, 1-9 (2016).

4 Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372** (2021).

5 Audano, P. A. *et al.* Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663-675. e619 (2019).

6 Holt, J. M. *et al.* Identification of Pathogenic Structural Variants in Rare Disease Patients through Genome Sequencing. *BioRxiv*, 627661 (2019).

7 Sharp, A. J., Cheng, Z. & Eichler, E. E. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 407-442 (2006).

8 Hui, W. W. *et al.* Universal haplotype-based noninvasive prenatal testing for single gene diseases. *Clin. Chem.* **63**, 513-524 (2017).

9 Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347-1355 (2020).

10 Barseghyan, H. *et al.* Next-generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. *Genome Medicine* **9**, 90 (2017).

11 Chan, S. *et al.* in *Copy Number Variants* 193-203 (Springer, 2018).

12 Riggs, E. R. *et al.* Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genet. Med.*, 1-13 (2019).
13 Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020).
14 Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83-89 (2020).
15 Sharo, A. G., Hu, Z. & Brenner, S. E. StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants. *BioRxiv* (2020).
16 Kleinert, P. & Kircher, M. CADD-SV–a framework to score the effects of structural variants in health and disease. *bioRxiv* (2021).
17 Zhang, L. *et al.* X-CNV: genome-wide prediction of the pathogenicity of copy number variations. *Genome Medicine* **13**, 1-15 (2021).
18 Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572-3574 (2018).
19 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology* **15**, R84 (2014).
20 Pedersen, B. S., Layer, R. M. & Quinlan, A. R. smoove: structural-variant calling and genotyping with existing tools. (2020).
21 Sarwal, V. *et al.* A comprehensive benchmarking of WGS-based structural variant callers. *bioRxiv* (2020).
22 Uguen, K. *et al.* Genome sequencing in cytogenetics: Comparison of short-read and linked-read approaches for germline structural variant detection and characterization. *Molecular genetics & genomic medicine* **8**, e1114 (2020).
23 Levy-Sakin, M. *et al.* Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature communications* **10**, 1-14 (2019).
24 Demaerel, W. *et al.* The 22q11 low copy repeats are characterized by unprecedented size and structural variability. *Genome Res.* **29**, 1389-1401 (2019).
25 Bhattacharya, S., Barseghyan, H., Délot, E. C. & Vilain, E. nanotatoR: A tool for enhanced annotation of genomic structural variants. *BMC Genomics* **22**, 1-16 (2021).
26 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
27 Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503-2505 (2014).
28 Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884-D891 (2021).
29 Rodriguez, J. M. *et al.* APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res.* **46**, D213-D217 (2018).
30 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
31 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
32 Will, A. J. *et al.* Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat. Genet.* **49**, 1539-1545 (2017).
33 Klopocki, E. *et al.* Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *The American Journal of Human Genetics* **88**, 70-75 (2011).
34 Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3**, 1-26 (2016).
35 Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555-560 (2019).

36 Neira-Fresneda, J. & Potocki, L. Neurodevelopmental disorders associated with abnormal gene dosage: Smith–Magenis and Potocki–Lupski syndromes. *Journal of Pediatric Genetics* **4**, 159-167 (2015).
37 Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29**, 635-645 (2019).
38 Derrien, T. *et al.* Fast computation and applications of genome mappability. *PloS one* **7**, e30377 (2012).
39 Lee, H. & Schatz, M. C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28**, 2097-2105 (2012).
40 Singleton, M. V. *et al.* Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *The American Journal of Human Genetics* **94**, 599-610 (2014).
41 Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12**, 841 (2015).
42 Wang, H. *et al.* Clinical utility of 24-h rapid trio-exome sequencing for critically ill infants. *NPJ genomic medicine* **5**, 1-6 (2020).
43 Clark, M. M. *et al.* Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science translational medicine* **11** (2019).
44 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849-864 (2017).
45 Kehr, B. *et al.* Diversity in non-repetitive human sequences not found in the reference genome. *Nat. Genet.* **49**, 588-593 (2017).
46 Wong, K. H., Levy-Sakin, M. & Kwok, P.-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature communications* **9**, 1-9 (2018).
47 James, K. N. *et al.* Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. *NPJ genomic medicine* **5**, 1-8 (2020).

Chapter 4: Individuals with pathogenic genotypes reveal differences in ClinVar and HGMD variant classification over six years*

Background

Rare genetic diseases may affect as many as 1 in 20 Americans[1], but a definitive diagnosis is sometimes elusive[2]. In the past decade, exome and genome sequencing have improved the diagnostic rate for unresolved rare genetic diseases by 3 to 4-fold over previously established methods[2-4]. Identifying the causal variant(s) through sequencing can inform disease management by altering treatment, predicting disease progression, and informing risk to other family members including future births[5,6]. However, identifying causal variants can be challenging. Clinicians must objectively weigh many sources of evidence to determine if a variant explains the proband phenotypes. Indeed, the majority of individuals with a suspected rare genetic disease remain undiagnosed after exome or genome sequencing[2,7].

To standardize the interpretation of variants, in 2015 the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) developed guidelines to unify norms across clinical laboratories[8]. Since then, a growing number of laboratories have adopted these guidelines[9]. As familiarity with the guidelines has grown, variant interpretation concordance across laboratories has increased from 71% in 2016 to 84% in 2020[10,11]. These variant interpretation guidelines draw from several specialized research areas including population genetics, human gene isoforms, protein structure and function, and computational predictions of variant impact. While these specialties have all made important contributions to variant interpretation, perhaps no resource has been more valuable than the creation of diverse databases of allele frequencies, which are used to identify variants that are too common to cause a rare disease. In 2012, the Exome Sequencing Project created the first large-scale database of exonic allele frequencies that included samples from both European Americans and African Americans[12]. In 2015, phase 3 of the 1000 Genomes Project (1KGP) became available, providing genome-wide alleles from thousands of global genomes[13]. This was quickly followed by progressively larger and more diverse databases, including ExAC[14], gnomAD[15], and ALFA[16]. Here we investigate trends in the accuracy of variant interpretation since 2014, during which these allele frequency resources have grown tremendously.

Researchers communicate variant interpretations through published articles and submissions to variant databases. Until recently, variants were annotated in locus-specific databases (LSDBs) that typically collected variants in a single gene. In an effort to standardize content and improve ease of access, many LSDBs used the same software, the Leiden Open Variation Database[17], and the Human Genome Variation Society collected LSDBs to form a databases of LSDBs[18]. Authoritative reference resources such as OMIM[19] and GeneTests[20] often included additional variant information. Following calls to harmonize these resources into a single common database[21], today there are two leading genome-wide variant databases of clinical interest: ClinVar and the Human Gene Mutation Database (HGMD). In 2013, the NIH created ClinVar, a free-to-access database (maintained by ClinGen) that accepts submissions from clinical

* This chapter was primarily written by Andrew Sharo, with contributions from Yangyun Zou, Aashish Adhikari, and Steven Brenner. It was adapted from a manuscript in preparation. Andrew Sharo performed the work described in those sections he wrote, while the sections written by others describe work that was performed primarily by others. This work is included with permission from the authors.

laboratories, research groups, and specialized databases. As of 2020, 8,000+ people access ClinVar each day, and it currently contains pathogenic interpretations of nearly 130,000 variants[22,23]. ClinVar labels disease-causing variants as either Pathogenic (P) or Likely Pathogenic (LP). By definition, P indicates a 99% chance of pathogenicity, and LP indicates a 90% chance of pathogenicity. Although these definitions provide a built-in threshold for false positives, for the purposes of this paper, we highlight all variants that could be incorrectly classified as pathogenic.

HGMD began in 1996, is privately funded through subscriptions, and is curated directly from published literature by dedicated staff. It contains pathogenic interpretations of nearly 300,000 variants. A free version of HGMD is available that is several years out of date. HGMD labels disease-causing variants as either disease-causing (DM) or likely disease causing (DM?). ClinVar and HGMD attempt different strategies to reach the same goal: accurate variant annotation. ClinVar receives variants primarily from clinicians and laboratory staff who often use standardized interpretation guidelines to identify pathogenic variants in a clinical context. ClinVar annotates variants as either pathogenic or benign. HGMD curates information directly from research and clinical articles, which may include experimental assays of variant function[24]. These databases are rapidly growing. Since 2017, the number of ClinVar variants has doubled, and HGMD variants have grown by 50%.

Several studies have attempted to assess the accuracy of cataloged variants using large sequencing cohorts of healthy individuals[25-29]. Two of the earliest studies searched for variants annotated as pathogenic in individuals sequenced in a population database created by the 1,000 Genomes Project (1KGP)[13]. These researchers identified individuals in 1KGP who were homozygous for one or more recessive variants annotated as pathogenic (henceforth, 'indicated affected individuals'). Surprisingly, these two studies found that most individuals harbored multiple homozygous variants that were catalogued by HGMD to cause early-onset disease. However, individuals in 1KGP were all over 18 years of age and healthy enough to sign a consent form. Certainly, 1KGP individuals are not expected to be enriched for disease, yet these studies found that the implied rates of disease were higher in 1KGP than the known disease prevalence. There are two plausible explanations for this discrepancy. The first is that many benign variants were misclassified as pathogenic, which the authors concluded[27,28]. An alternative explanation is that some Mendelian diseases have been underdiagnosed. While this is true for some disorders[30], we analyze a subset that are screened for at birth and are likely not substantially underdiagnosed (see Methods). With this modification, we believe that most, and likely all, of indicated affected individuals are not affected by a disease, and rather the annotated pathogenic variants they harbor were misclassified. A similar approach has also been used to investigate ClinVar variants, which a 2018 study showed imply disease prevalence much higher than recorded prevalence for several clinically actionable or rare disorders[25]. Using orthogonal methods, researchers have identified variant features that are associated with correct classification. Specifically, they have found that recently curated variants, with lower minor allele frequency (MAF), with multiple concordant submissions, and submitted by clinical researchers are more likely to be correctly classified[29,31].

Since many variants are found principally in a single ancestral population, misclassification can lead to racial disparities in variant interpretation and clinical care. Indeed, one study determined that variants erroneously associated with sudden heart failure were found at higher allele frequency in Black Americans than white Americans[32]. Fortunately, these misclassified variants

were eventually corrected. However, until erroneously annotated variants are corrected, which may take years, probands who harbor these variants may undergo inappropriate medical care. Furthermore, misclassified variants can have effects beyond the clinical care of individuals with those variants, since cataloged pathogenic variants can influence novel variant interpretation. In the ACMG/AMP variant interpretation guidelines, two categories of evidence that support pathogenicity rely directly on cataloged variants: the same amino acid change as a cataloged pathogenic variant (PS1) and a different amino acid change at the same residue as a cataloged pathogenic variant (PM5). Misclassified variants can also have indirect effects through the ACMG/AMP guidelines' consideration of variant impact predictors, which contribute supporting evidence (PP3, BP4). Since many variant impact predictors are trained or are validated on cataloged variants [33-36], their predictions may be influenced by misclassified variants. In the worst case, a researcher following the ACMG/AMP guidelines may be misled by misclassified variants to incorrectly classify a novel variant, either by using misclassified variants as direct evidence (PS1, PM5) or indirectly though variant impact predictors that trained on misclassified variants (PP3, BP4). Such an event would propagate existing variant misclassifications and possibly reinforce racial disparities.

Variant databases have taken different approaches to address misclassifications. ClinVar introduced a star system to indicate the review status of a variant interpretation, in which a variant gains credibility when assertion criteria are provided, multiple submitters concur, or an interpretation comes from experts in the field who follow gene-specific classification guidelines[37]. Wright et al. found that variants annotated as pathogenic with more review stars were more likely to be truly pathogenic[29]. ClinGen has also supported the formation of expert panels—composed of healthcare professionals with expertise relevant to a disease gene—which can provide high-confidence variant interpretations and resolve conflicting variant interpretations. As of December 2020, ClinVar contains just 36 genes in which 10 or more variants are reported as reviewed by an expert panel, out of more than 3,000 genes associated with a monogenic disorder by OMIM[19]. Although expert panels are promising, they have so far contributed to a small fraction of ClinVar variant reclassifications. HGMD curators reclassify variants based on new published evidence such as functional studies or population frequency, and their reclassification rate has been reported as similar to that of ClinVar[24,38]. Here, we consider whether these reclassification efforts, in concert with improved resources, have reduced the number of apparently misclassified variants over time. We consider variants in a subset of well-studied genes with highly penetrant phenotypes.

Inborn errors of metabolism (IEMs) are a group of rare, primarily recessive or X-linked, monogenic disorders caused by defects in a metabolic enzyme or its cofactors. Newborns in most developed countries are screened for IEMs using blood metabolites. Untreated, many of these screened IEMs are highly penetrant and lead to metabolite accumulation that often causes irreversible disability or death. They are thus a model system for identifying false positives in variant databases, as they should not be present as pathogenic genotypes in healthy individuals. While many screened IEMs are debilitating or fatal in childhood unless treated, there are notable exceptions. For example, our screened IEMs include Short Chain Acyl CoA Dehydrogenase Deficiency (SCADD; associated with *ACADS*) and Hyperprolinemia type I (HPI; associated with *PRODH*), both of which are often asymptomatic in newborns who screen positive for metabolite levels indicative of the disease[40,41]. Additionally, our screened IEMs include Ornithine transcarbamylase deficiency (OTCD; associated with *OTC*) and Glutaric

Acidemia Type II (GAII; primarily associated with *ETFDH*), both of which are often seen in a late-onset form which may not result in outward symptoms until 40 years of age[42,43].

Because screened IEMs are systematically identified in the population, their maximum possible incidence is generally known, and there has been greater opportunity to identify and catalog the genetic variants that cause these diseases. Indeed, one recent study found potential benefit to screening newborns for IEMs using exome sequencing alongside mass spectrometry, the current standard for screening[39]. However, these researchers found it necessary to manually curate dozens of variants cataloged in ClinVar or HGMD for which the MAF was higher than expected for a rare disorder. Out of 60 variants with MAF > 0.1%, they deemed 41 were not reportable due to insufficient published evidence for pathogenicity.

Variants with a MAF greater than expected from disease incidence are addressed in the 2015 ACMG/AMP variant interpretation guidelines[8] under the BA1 evidence for benign variants. These guidelines recommend that a MAF >5% in 1KGP, ExAC (now superseded by gnomAD), or the Exome Sequencing Project (ESP) may be considered stand-alone evidence that the variant is benign. In 2018, the guidelines for this classification were updated by Ghosh et al. to recommend that a MAF >5% in any continental population dataset of at least 2,000 alleles (with some additional constraints) is stand-alone evidence the variant is benign[44]. We have investigated how implementing the original vs revised guidelines impacts our results.

Here, we investigate how the degree of variant misclassification has changed over time in ClinVar and HGMD, using screened IEMs as a model system. Building on previously developed methods[27,28], we used samples in the 1000 Genomes Project (1KGP) to identify individuals who harbor genetic variants that have been listed in ClinVar or HGMD as pathogenic. We identified more individuals than expected from screened IEM incidence, an indication of the specificity of each database. We investigated how the number of these likely false positive individuals indicated by ClinVar and HGMD changed over time, and we considered whether certain ancestry groups were over-represented. Since we do not measure false negatives, we cannot assess the sensitivity of each database even though the balance between specificity and sensitivity is an important tradeoff to consider. We looked in detail at variants that were misclassified and what led to their eventual reclassification. Additionally, we probed overall trends of reclassification in ClinVar and HGMD, identifying surprising trends in the reclassification of confidently classified variants to uncertain classifications. Finally, we replicated our findings using samples from gnomAD, which includes 63,269 genomes.

Methods

Identifying putatively affected individuals in 1KGP

We used GRCh38 genotypes from 1KGP phase 3[13] VCF files (downloaded on 14 November, 2019) to identify individuals who harbor genotypes annotated as pathogenic (defined as homozygous, hemizygous, or compound heterozygous) but who likely do not suffer from a screened IEM. Ancestry was determined by superpopulation membership, as listed by the International Genome Sample Resource[45]. We created a curated list of 80 genes, associated with 48 IEMs screened by the California newborn screening program[46] (henceforth, screened IEMs). These screened IEMs include some disorders where a large fraction of affected individuals is asymptomatic. In our analysis below, we identified several ClinVar variants in *PRODH*, associated with HPI. This condition is characterized by elevated levels of proline, and it is sometimes considered benign and asymptomatic[41]. However, there are reports of

individuals with HPI who have severe neurological impairment[47]. Additionally, recent long-term follow-up of patients with HPI suggests it results in impaired social skills, and there is evidence that deletions containing *PRODH* (and possibly variants in *PRODH*) contribute to schizophrenia risk[48-50]. Given the possible clinical phenotypes associated with this gene, we retained it in our analysis.

The population incidence of screened IEMs is approximately 1 in 3,200[51]. Thus, if the individuals sequenced in 1KGP were a random sample with unknown health status at birth, we would expect less than 1 individual to have a screened IEM. Given that most of the indicated affected individuals lived in countries without newborn screening programs before 1990, they are unlikely to have been screened and treated early enough to prevent irreversible damage.

ClinVar GRCh38 variants were obtained from VCF files (downloaded on 8 January, 2021) from the NCBI ClinVar FTP site[37]. VCF files were gathered from both archives 1.0 and 2.0 (starting with clinvar_20140401.vcf.gz and ending with clinvar_20201226.vcf.gz). Bcftools norm[52] was used to left-align and normalize indels. Only variants within our list of 80 genes were considered further. Variants that were listed as only somatic or variants with null alt alleles were not considered further. For ClinVar archive 1.0 variants, variants were assigned clinical significance using the following categories: '0': VUS, '2': Benign (B), '3': Likely benign (LB), '4': Likely pathogenic (LP), and '5': Pathogenic (P). Variants were inferred to have conflicting interpretations when they had interpretations in two or more of the following three categories: B or LB, VUS, P or LP. Due to inconsistencies in review star annotation in archive 1.0 files before June 15, 2015, 'not' was assigned 0 review stars, 'single' was assigned as 0or1 review stars (see below for details), 'conf' was assigned as 1 review star, and 'mult' was assigned as 2 review stars. For archive 1.0 files after June 15, 2015, 'no_assertion_criteria_provided', 'no_assertion_provided', 'not', 'no_criteria', and 'no_assertion' were grouped as 0 review stars; 'criteria_provided', 'conf', and 'single' were grouped as 1 review star; '_multiple_submitters', '_no_conflicts', and 'mult' were grouped as 2 review stars. For all archive 1.0 files, review stars were assessed manually for variants with an inferred pathogenic genotype in 1KGP. For archive 2.0 variants, "no_assertion_criteria_provided", "No_assertion_provided", and "no_interpretation_for_the_single_variant" were grouped as 0 review stars, "criteria_provided", "_single_submitter", and "_conflicting_interpretations" as 1 review star, "_multiple_submitters", "_no_conflicts", and "reviewed_by_expert_panel" as 2+ review stars. In calculating indicated affected individuals for each year, we reported the maximum number of individuals with an inferred pathogenic genotype at any time in that year. In our analysis of 1KGP affected individuals, ClinVar submissions were removed from consideration if the submitted condition was a not a screened IEM (e.g., Schizophrenia). Submissions for which the condition was "not provided" were included in our analysis. For all other analyses, it was not feasible to check the submitted condition of variants.

HGMD variants were obtained from privately archived versions of HGMD 2014.1 and 2016.2, and a recently accessed version of 2020.3 through Qiagen Digital Insights HGMD Professional. Only SNVs classified at least once as 'DM' or 'DM?' within our list of 80 screened IEM genes were considered further.

In our analysis using the 2015 BA1 guidelines, variants with a global MAF > 5% in 1KGP, the Exome Sequencing Project (ESP6500SI-V2), or gnomAD v2.1 exomes were removed from consideration. In our analysis using the 2018 BA1 guidelines, variants with a global MAF > 5% in 1KGP or ESP, or a MAF >5% in any gnomAD exome continental population were removed.

Ensembl Variant Effect Predictor with custom annotations was used to annotate the 1KGP VCF with all features. For rapid I/O of VCFs, we used cyvcf2[53]. To identify when the ancestry composition of indicated affected individuals (aggregated across all screened IEMs) was significantly different from the ancestry composition of 1KGP or gnomAD, we first performed a two-sided Fisher's exact test on a 5 x 2 contingency table that included the five continental populations (African, Latino, East Asian, European, South Asian), using fisher.test in the R 'stats' package[54]. When the expected count for every population was greater than 40, we instead performed a Pearson's Chi-squared test using chisq.test to reduce computation time. For those global analyses that showed significant deviation from the 1KGP database ancestry composition, we performed individual tests to identify the significantly skewed population. These individual tests were performed using a one-sided Fisher's exact test on a 2 x 2 contingency table as described above. To correct for multiple tests, we used a 5% significance threshold with Bonferroni correction for 222 tests, yielding a p-value threshold of 2.2×10^{-4}. We determined 222 tests by calculating the total number of tests performed across all figures (including supplementary figures), which were typically 1 Fisher's exact test per bar, with an additional 5 tests per bar when the Fisher's exact test was significant. Bars that had zero height were not tested. Odds ratios and 95% confidence intervals were determined using two-sided Fisher's exact tests as described above.

To confirm that the inferred pathogenic genotypes we observed in 1KGP were not sequencing errors, we attempted to confirm the quality of all variants that comprised these genotypes. Specifically, we downloaded whole genome and deep exome sequencing BAM alignment files of select individuals with homozygous, hemizygous, or compound heterozygous inferred pathogenic genotypes. Most of these alignments were improved by quality-control steps including marking duplicates, local realignment around indels, and base quality recalibration, especially for the Illumina sequencing data. Next, we detected variants and calculated genotypes for each sample at specific sites based on both low-coverage genome sequencing data (<5×per site per individual) and high-coverage exome sequencing information (at least >20×per site per individual) using 'UnifiedGenotyper' from the Genome Analysis Toolkit (GATK 3.4-0) under a multi-sample calling strategy[55,56]. Variant Quality Score Recalibration (VQSR) was conducted to evaluate variant quality by GATK 3.4-0. Finally, we obtained variant and genotype information of select individuals and their site-specific genotype quality parameters such as genotype quality (GQ) to validate the quality of the called genotypes. We used GQ≥30 (p-value of 0.001) as our threshold for high quality genotype calls. Genotypes of some individuals were re-confirmed based on high-coverage whole genome sequencing by Complete Genomics[57]. Thanks to the recent availability of high-coverage whole genome sequencing of all 1KGP samples from the New York Genome Center[58], the remaining inferred pathogenic genotypes were confirmed using these data. Two inferred pathogenic genotypes in PRODH were not able to be reconfirmed due to poor sequencing quality in the gene.[*]

To infer screened IEM incidence from 1KGP, for each IEM gene *g*, we summed the allele frequencies of all annotated pathogenic variants in *g*, which we call p_g. Genes were then divided into two categories: X chromosome and autosomal. The disease incidence for all X-linked disorders was calculated as $\sum_{g \in X} p_g(1 - p_g) + {p_g}^2$ where X is the set of all X-linked screened

[*] This paragraph was primarily written by Yangyun Zou.

IEM genes. For all autosomal genes, the incidence was calculated as $\sum_{g \in A} {p_g}^2$ where A is the set of all autosomal screened IEM genes . We repeated this process for each population using the 1KGP population-specific allele frequency as well. In Figs. 4.3,4.4, the height of each bar represents the incidence inferred using the database-wide allele frequency, while the proportion of the bar comprised by each ancestry is based on the relative disease incidence calculated using the population-specific allele frequency. The same process was repeated for our gnomAD analysis.

Variant reclassification in ClinVar and HGMD

ClinVar and HGMD variants were filtered as described above. VEP[59] was used to annotate each variant with its gnomAD v2.1 exomes MAF. In order to identify reclassifications, for each time point available for ClinVar, each variant was classified into one of the following categories: B/LB 3 stars, B/LB 2 stars, B/LB 1 star, B/LB 0 stars, VUS, Conflicting, P/LP 0 stars, P/LP 1 star, P/LP 2 stars, P/LP 3 stars. At each time point available for HGMD, each variant was classified into one of the following categories: DM, DM?, DFP, DP, R. Variants that were removed from the database were classified as R. Variants classified in any other category (such as 'not provided') and all ClinVar variants prior to June 15, 2015 were not considered. To create figures for the Results section 'Comparison of variant reclassification between ClinVar and HGMD', for each variant we considered only its first category chronologically (typically its category when first entered into the database) and its last category chronologically.

Next, for each ClinVar variant we used gnomAD v2.1 exomes to determine the ancestry group in which it occurs at the highest allele frequency. To reduce bias from the unequal number of individuals in each ancestry group in gnomAD, all allele frequencies below 6.152×10^{-5} (the smallest possible allele frequency in African ancestry, which has the smallest number of individuals in gnomAD) were set to zero. Next, each variant was assigned to the ancestry with the highest allele frequency. Variants with zero allele frequency in all ancestries were not considered further.

For each variant, we recorded all reclassifications it underwent. To avoid classifications without stars, only ClinVar reclassifications after June 15, 2015 were considered. ClinVar GRCh38 VCF files (as described above) were used to identify reclassifications. Reclassifications were considered every month. Since more recent ClinVar VCFs were archived weekly, these were downsampled to approximate monthly archives. The removal of a ClinVar variant from the database was not considered a reclassification. If a variant re-entered into ClinVar under a new classification, it was considered reclassified.

Variant reclassifications were grouped into two categories: increasing confidence and decreasing confidence. Increasing confidence was defined as Conflicting or VUS to P/LP or B/LB with any number of stars. Decreasing confidence was defined as: P/LP or B/LB with any number of stars to Conflicting or VUS. Variants were grouped by these categories, colored by assigned ancestry (see above), and visualized using Floweaver[60], resulting in Fig. 4.6C,E.

To correct for bias caused by the possible overrepresentation of some ancestries in ClinVar, for each ancestry we calculated the number of variants in each classification category. The number of variants per category was were calculated for every month, yielding a measure we call variant-months. A variant-month is a measure of both the number of variants and how long they

have been in ClinVar. For example, 2 variants classified in ClinVar for a month is 2 variant-months, and 1 variant classified in ClinVar for 2 months is also 2 variant-months. For each ancestry, we analyzed its assigned variants to determine how many variant-months were catalogued for each category between June 15, 2015 and Dec 31, 2020. The differences in variant-months between ancestries reflects differences in genetic diversity as well as ClinVar submission bias. These variant-months are used to normalize comparisons across ancestries which we report in reclassifications per variant-month. In normalizing a reclassification category (increasing confidence or decreasing confidence), we divide the number of reclassifications by the variant-months of the source category. For example, if we wanted to compare increasing confidence across ancestries, then for each ancestry we would calculate the number of reclassifications with increasing confidence among variants assigned to that ancestry and divide that by the variant-months of the source category, in this case VUS and Conflicting variants. 95% confidence intervals were calculated for each ancestry group as +/- 1.96*sqrt(p*(1-p)/n) where p is reclassified variants / variant-months of source variants and n is variant-months of source variants.

Results

Individuals affected by ClinVar variants

We analyzed ClinVar screened IEM variants submitted between April 2014 and December 2020, and first examined a Select subset based on review stars (see Methods). This Select subset included P variants with 1 or more review stars (indicating the submitter included assertion criteria), which consisted of 2,118 variants in 2020 (Fig. 4.1A). In accordance with the 2015 ACMG/AMP BA1 guidelines, we removed variants with a MAF that reached the threshold for classification as stand-alone benign (global MAF > 5% in 1KGP, gnomAD, or ESP). This resulted in the removal of a single variant with 1 review star and a global MAF of ~5%. We later discuss applying the 2018 BA1 guidelines. To identify individuals who harbored inferred pathogenic genotypes of these Select ClinVar variants, we used the 1KGP database. 1KGP includes 2,504 individuals that are drawn approximately evenly from 5 continental populations (Fig. 4.1B). We considered all individuals who were homozygous, hemizygous, or compound heterozygous for one or more Select variants to be indicated affected. We found a single indicated affected individual, with South Asian ancestry, who was homozygous for a P variant (ACADS:c.1108A>G) added to ClinVar in 2015, which was re-classified as Conflicting by 2017 (Fig. 4.1C; Table 1). There have since been zero indicated affected individuals through 2020.

In addition to Select ClinVar variants, clinicians often consider and report P and LP variants with 0 review stars (no assertion criteria) but give them appropriately lower credence. To analyze these variants, we next considered the Full dataset of ClinVar screened IEM variants, which included P and LP variants with any number of review stars. We removed from consideration variants that fulfilled the 2015 BA1 criteria. This eliminated six variants from 2014 to 2020, with a median MAF in 1KGP of 12%. We searched for individuals in 1KGP who were indicated affected (Fig. 4.1D). In 2014, there were 8 indicated affected individuals, which increased to 9 in 2015, and declined to just 1 by 2020 (reclassification causes discussed below). 11 variants played a role in the genotypes of these indicated affected individuals. We also considered whether P or LP variants led to a larger number of indicated affected individuals. However, due to the relatively small fraction of variants that are classified as LP, the results of considering only P variants were nearly identical to considering both P and LP. We did not observe any

statistically significant skew in the ancestries of the 1KGP individuals who were indicated affected.

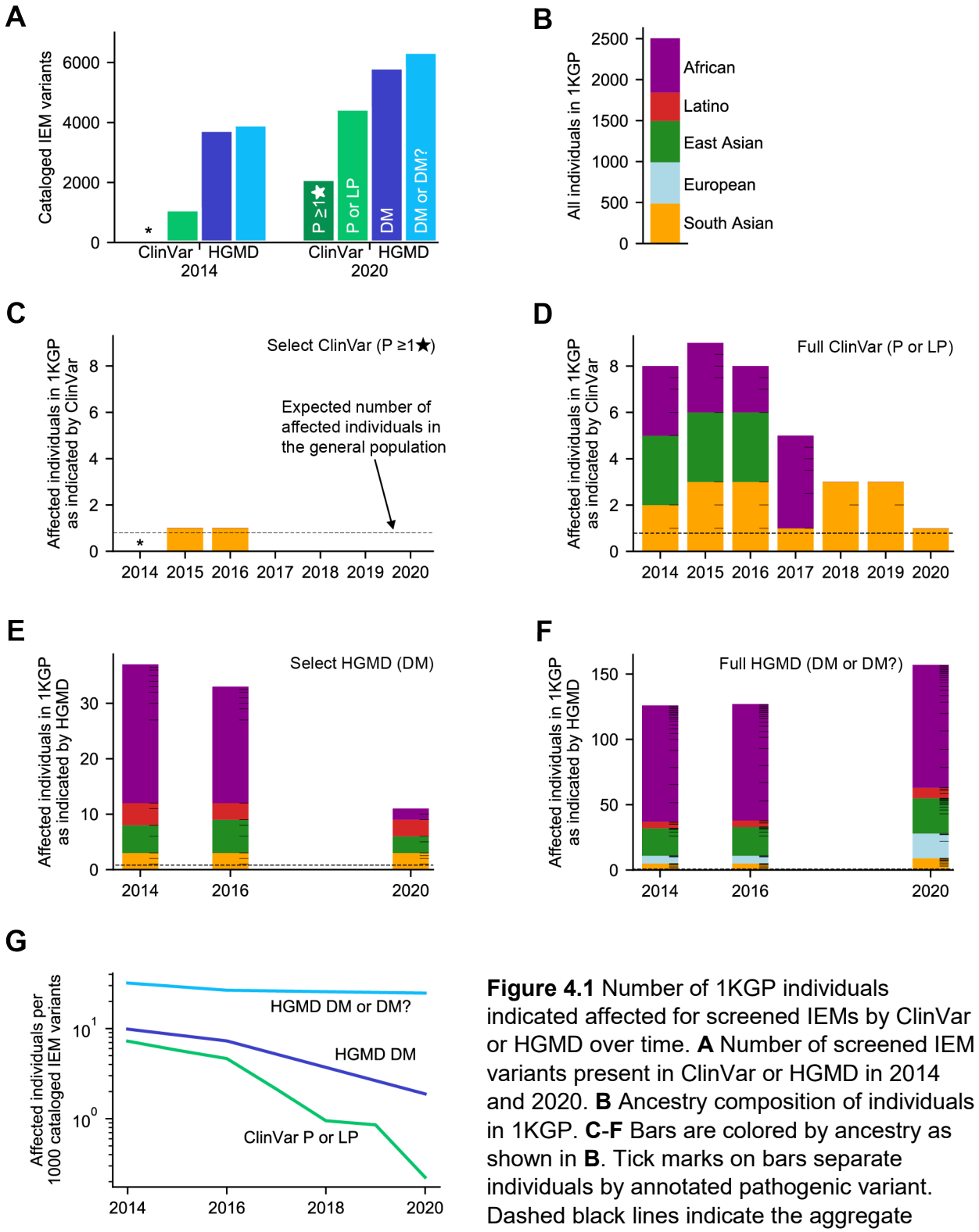

Figure 4.1 Number of 1KGP individuals indicated affected for screened IEMs by ClinVar or HGMD over time. **A** Number of screened IEM variants present in ClinVar or HGMD in 2014 and 2020. **B** Ancestry composition of individuals in 1KGP. **C**-**F** Bars are colored by ancestry as shown in **B**. Tick marks on bars separate individuals by annotated pathogenic variant. Dashed black lines indicate the aggregate population incidence of screened IEMs. The number of 1KGP individuals with a pathogenic

genotype for a variant in **C** Select ClinVar variants annotated as pathogenic, defined as variants with a P interpretation with at least 1 review star. Variants that also have conflicting interpretations (with VUS or B/LB) with 1 or more review stars are removed. **D** Full ClinVar variants annotated as pathogenic, defined as variants with a P or LP interpretation. Variants that also have conflicting interpretations (with VUS or B/LB) are removed. **E** Select HGMD variants, defined as variants classified as DM. 2014, 2016, and 2020 are shown because they are the years for which we have archived HGMD data. **F** Full HGMD variants, defined as variants classified as DM or DM?. **G** The number of affected individuals relative to the number of variants classified in each variant set. This approximates a false positive rate, which has fallen over time for each database. *Data not available because existing review star framework was not in place until 2015.

Individuals affected by HGMD variants

Similar to our ClinVar analysis, we first examined a Select subset of HGMD variants. This subset included HGMD DM (disease-causing) variants in a screened IEM gene, which consisted of 5,833 variants in 2020. We removed one variant that met the 2015 BA1 criteria in 2014 and 2016 with a MAF of 20%. We removed an additional variant in 2016 with a MAF of 50%. We investigated individuals in 1KGP who harbored Select HGMD variants, and we found 37 indicated affected individuals in 2014, caused by 16 variants (Fig. 4.1E). Repeating this analysis with Select HGMD classifications from December 2020, we found 11 indicated affected individuals in 1KGP (70% reduction from 2014) due to 9 variants (reclassification causes discussed below). 3 of these 9 variants were added to HGMD after 2014.

To gain a larger picture of potential variant misclassification in HGMD, we next considered the Full dataset of HGMD variants annotated to likely cause disease, which included DM and DM? variants (henceforth, Full HGMD variants). We removed 5 variants that met the 2015 BA1 criteria in 2014, 4 variants in 2016, and 7 variants in 2020. We investigated individuals in 1KGP who harbored Full HGMD variants. In 2014 there were 126 indicated affected individuals in 1KGP due to 20 DM and 12 DM? variants (Fig. 4.1F). This increase in the number of DM variants compared to our Select analysis is due entirely to compound heterozygotes consisting of one DM variant and one DM? variant. Unexpectedly, we found indicated affected individuals increased over time, with 157 individuals in 2020, due to 17 DM and 27 DM? variants. These include 7 DM? and 4 DM variants that were added to HGMD since 2014.

These indicated affected individuals are not only a barometer for changes in potentially misclassified variants, but they can also inform whether particular ancestry groups are more likely to be affected by variant misclassifications. Considering Select HGMD variants in 2014, African ancestry individuals were significantly more likely to be indicated affected (Fig. 4.1E). While 26.4% of individuals in 1KGP are of African ancestry (Fig. 4.1B), 25 out of 37 (67.6%) of indicated affected individuals had African ancestry, which is significantly more than expected by chance ($p < 10^{-6}$) and indicates an odds ratio of 5.8 for African ancestry individuals (95% CI: 2.8-12.8). By 2020, no populations were significantly skewed. Notably, in 2014, 2016, and 2020, no European ancestry individuals were indicated affected. When considering Full HGMD variants, we found that 89 out of 126 indicated affected individuals in 2014 and 94 out of 157 indicated affected individuals in 2020 were of African ancestry (both $p<10^{-15}$) (Fig. 4.1F). This translates to an odds ratio of 6.7 (95% CI: 4.5-10.2) in 2014 and 4.2 (95% CI: 3.0-5.9) in 2020.

Chr	Position	Ref	Alt	Gene	cDNA, protein	# hom or hemi in 1KGP	# comp het in 1KGP	1KGP sample ID(s)	First pathogenic submission: submitter, date, interpretation, evidence	Consensus interpretation as of Dec 2020	Submitted Interpretations
9	130458549	G	T	ASS1	c.323G>T R108L	1	1	NA19030 NA19395	OMIM, April 2014, Pathogenic NC, Heterozygous variant in affected individual	Conflicting Interpretations of Pathogenicity	1 Pathogenic NC 1 VUS 4 Likely benign 1 Benign 1 Benign NC
12	109561798	C	T	MMAB	c.403G>A p.A135T	1	0	HG03169	GeneReviews, February 2016, Pathogenic NC, Seen in affected individuals	Conflicting Interpretations of Pathogenicity	1 Pathogenic NC 1 VUS 1 Benign 1 Benign NC
12	120739317	A	G	ACADS	c.1108A>G p.M370V	1	0	NA20878	GeneDx, August 2015, Pathogenic, clinical testing	Conflicting Interpretations of Pathogenicity	2 VUS 1 Likely Benign
X	38367361	G	A	OTC	c.148G>A p.G50R	1	0	NA21124	GenMed Metabolism Lab, April 2014, Pathogenic NC, Identified in late onset individual	Pathogenic, 0 stars	2 Pathogenic NC

Table 4.1: Subset of ClinVar variants seen in a pathogenic genotype in 1KGP. See Table S6 for full list of ClinVar variants. NC indicates no assertion criteria were provided by the submitter.

Chr	Position	Ref	Alt	Gene	cDNA, protein	# hom or hemi in 1KGP	# comp het in 1KGP	1KGP sample ID(s)	Variant-specific assay	Outcome	PubMed ID	HGMD 2014	HGMD 2020
12	120739317	A	G	ACADS	c.1108A>G p.M370V	1	0	NA20878	In vitro activity in mouse mitochondria	Very mild effect on protein misfolding	18523805	DM	DM
21	43060481	G	A	CBS	c.1105C>T p.R369C	0	2	HG02645 NA20289	Yeast model	No effect on enzyme function	9361025	DM	DM
X	38381417	C	T	OTC	c.374C>T p.T125M	1	0	NA19117	Biochemical test	OTC liver activity < 1%	8807340	DM	DM

Table 4.2: Subset of HGMD variants seen in a pathogenic genotype in 1KGP. See Table S7 for full list of HGMD variants.

Unlike the ancestry skew observed in Select HGMD variants, the ancestry skew in Full HGMD variants has persisted over time.

Accuracy per variant across datasets

Since each ClinVar or HGMD dataset contains a different number of cataloged IEM-associated variants (Fig. 4.1A), we developed a metric to enable a comparison of classification accuracy across datasets. For each available year, we calculated the number of indicated affected individuals in 1KGP divided by the number of cataloged variants. Although we cannot be certain that no individual in 1KGP has a screened IEM, this metric is a proxy for the false positive rate per variant for each database. In 2014, the Full ClinVar dataset indicated 7.3 affected individuals per 1,000 cataloged P or LP variants (Fig. 4.1G). By 2020, this false positive rate had decreased by 97%. We could not determine a meaningful false positive rate for the Select ClinVar dataset due to the several years with zero affected individuals. For Select HGMD variants, the false positive rate decreased by 81%, with most of this decrease occurring between 2016 and 2020 (Fig. 4.1G). For Full HGMD variants, the false positive rate decreased 26% from 2014 to 2020 (Fig. 4.1G). It may seem surprising that the false positive rate of Full HGMD variants is decreasing given the increase in affected individuals over time (Fig. 4.1F). However, this decrease is due to the ~60% growth in cataloged variants between 2014 and 2020, which outweighed the growth in indicated affected individuals.

These three datasets have reduced the false positive rate of their cataloged variants over time, yet false positive rates currently differ greatly between them. As of 2020, Full ClinVar variants indicate 0.22 affected individuals per 1,000 cataloged pathogenic variants, which is an order of magnitude lower than Select HGMD variants, which indicate 1.9 affected individuals per 1,000 cataloged pathogenic variants (Fig. 4.1G). This, in turn, is an order of magnitude lower than Full HGMD variants, which indicate 25 affected individuals per 1,000 cataloged pathogenic variants.

Reliability of genotypes

To ensure that the inferred pathogenic genotypes we observed in 1KGP were not caused by errors from sequencing or downstream variant and genotype calling, we independently confirmed nearly all Select ClinVar variants, Full ClinVar variants, and Select HGMD variants present in an inferred pathogenic genotype. We re-called a subset of these genotypes using available low-coverage genome sequencing and high-coverage exome sequencing data from 1KGP (see Methods). We found that nearly all annotated pathogenic variants in this subset passed variant quality score recalibration (VQSR) filtering, and most genotypes in the indicated affected individuals had a genotype quality (GQ) larger than 30. For variants that we did not attempt to re-call or for which re-call quality was poor, we confirmed genotypes using high-coverage whole genome sequencing by either Complete Genomics or the New York Genome Center (see Methods). Out of the entire set of 52 genotypes indicated as pathogenic, there were just two for which genotype quality was below 30. One was TAZ:c.383T>C present in Select HGMD variants and found in a hemizygous state in HG03196. The other genotype consisted of a pair of compound heterozygous variants (PRODH:c.1357C>T;c.1322T>C) present in Full ClinVar variants and harbored by NA19372. Overall, we confirmed that 96% of the inferred pathogenic genotypes are high quality and reproducible. This suggests that the over-representation of putatively pathogenic genotypes in 1KGP is unlikely to be explained by errors introduced by sequencing or data processing.

ClinVar variant reclassification

Between 2014 and 2020, 11 variants in the Full ClinVar dataset were part of an inferred pathogenic genotype in at least one 1KGP individual. As of December 2020, 10 of these 11 variants have been reclassified in ClinVar to a non-pathogenic category. 8 variants were reclassified to Conflicting, 1 variant to VUS, and 1 variant to B/LB. One variant remains classified as P with 0 review stars. These variants were present in 7 genes: *OTC* (3), *ASS1* (2), *PRODH* (2), *ACADS* (1), *MMAB* (1), *MMUT* (1), and *SLC22A5* (1). Variants within the same gene tended to be initially contributed by the same submitter. For example, GenMed Metabolism Lab submitted the first interpretation for all three variants we identified in *OTC*, and OMIM first provided both *PRODH* variants. For each variant, we also recorded the submitter that contributed the first non-pathogenic classification but did not identify any patterns.

Among these 11 variants, we noticed a trend in which variants were initially submitted as P or LP when seen in an affected individual, even though there was limited evidence for pathogenicity. As more information because available, such as MAF, later submitters, most using defined criteria, interpreted these variants as VUS, B, or LB. One illustrative case is the variant A135T in *MMAB* (Table 1). Through a semi-automated process, this variant was extracted from a GeneReviews table to a ClinVar record in February 2016 as P and included two articles to support the interpretation[61,62]. According to these articles, researchers found this variant in a heterozygous state in three African ancestry individuals with methylmalonic acidemia (MMA) cblB type. In addition to A135T, each of these individuals also harbored a suspected pathogenic variant, although it was not confirmed to be in trans. Both articles claim the variant was absent from control samples, for which ancestry information was not provided. We now know the MAF of this variant in African ancestry individuals is approximately 1% in 1KGP and gnomAD exomes, corresponding to a disease incidence of 1 in 10,000 assuming complete penetrance. However, MMA cblB type occurs in less than 1 in 50,000 births, and has not been seen at elevated levels in individuals of African ancestry[51]. This variant was observed in a homozygous state in an African ancestry male in 1KGP, who most likely did not have MMA cblB type, which is a neonatal-onset disorder that results in severe disability and sometimes death without treatment. A plausible explanation is that the three affected individuals from the literature happened to carry this putatively benign allele, and due to inadequate information about its frequency this allele was mistakenly associated with MMA. Since the P submission, GeneDx used variant classification criteria to interpret this variant as VUS, citing the relatively high variant frequency as evidence for benignity. Invitae (with criteria) and Natera (without criteria) have interpreted the variant as B.

We examined the ClinVar variant responsible for the single indicated affected individual in 2020 and found this variant could plausibly cause disease. GenMed Metabolism Lab submitted this variant, G50R in *OTC*, an X-linked gene, in 2014 and cited an article in which researchers found this variant in a male with late-onset Ornithine transcarbamylase deficiency (OTCD) but did not provide the age of onset[63]. OTCD is known to have a variable age of onset in a sizeable fraction of cases, and researchers have identified one individual who was 44 years old when disease onset began[42]. Plausibly, this variant may be associated with late onset OTCD and the 1KGP hemizygous South Asian ancestry male (NA21124) has not yet reached the age of onset.

HGMD variant reclassification

In 2014, 16 Select HGMD variants contributed to an inferred pathogenic genotype in at least one 1KGP individual. By December 2020, 8 of these variants were reclassified to DM?, and an additional 3 DM variants were cataloged that contributed to an inferred pathogenic genotype. In

total, we observed 19 Select variants in an inferred pathogenic genotype, which were present in 11 genes: *OTC* (4), *PAH* (3), *ASS1* (2), *CBS* (2), *CPT2* (2), *ACAD8* (1), *ACADS* (1), *ACADVL* (1), *SLC22A5* (1), *SLC25A13* (1), and *TAZ* (1). We did not evaluate the Full HGMD variants in detail, but we do note that of the 32 DM and DM? variants that contributed to an inferred pathogenic genotype in 2014, none were reclassified to a non-disease-causing category by 2020.

HGMD rarely provides explanations for variant reclassification, so it is difficult to directly investigate why certain variants were reclassified. Instead, we examined the evidence for pathogenicity of the 19 Select variants identified in an inferred pathogenic genotype in 2014, 2016, or 2020. For each variant, we reviewed the articles cited by HGMD. According to the cited articles, researchers observed these variants in probands who were diagnosed with an IEM. None of the articles provided evidence for pathogenicity equivalent to the ACMG/AMP guidelines, which is not surprising given that most of the articles were published prior to 2015. Additionally, 12 out of 19 studies (63%) did not show any direct evidence for the functional effect of the variant, such as experimental assays of gene expression or enzymatic activity, and therefore did not conclusively assign pathogenicity to the variant. Assay absence was highly correlated with later reclassification from DM to DM?. Of the 5 variants classified as DM in 2014 for which assays were performed, all remained DM through 2020. Of the 11 variants for which no assay was performed, 8 were reclassified to DM? by 2020. Despite the predictive power of assay presence, the results of the assays were not always conclusive. For example, we found one 1KGP individual was homozygous for the variant c.1108A>G (M370V) in *ACADS*, which was cataloged by HGMD as DM (Table 2). Yet, the original article cited by HGMD indicates that the variant c.1108A>G has a much more mild effect on tetramerization than all other putatively pathogenic variants tested[64]. Similarly, functional assays of the variant c.1105C>T (R369C) in *CBS* in a yeast model indicated no effect on enzyme function in the article cited by HGMD[65] (Table 2). Among the 19 studies cited by HGMD for these Select variants, only three studies directly measure the enzymatic activity of the observed variant[66-68]. One of these studies described the variant c.374C>T (T125M) in *OTC*, which was observed in a male newborn who died at the age of 14 days[66]. A biochemical assay verified the variant OTC enzymatic activity was <1% that of wild type in liver tissue. Surprisingly, we observed that one African ancestry male, NA19117, possesses c.374C>T in his single copy of the X-linked *OTC* gene. Although the genotype in this individual was called with low quality, this same genotype was re-confirmed with high genotype quality (GQ=187) by Complete Genomics.

We observed a single variant (ACADS:c.1108A>G) that led to an inferred pathogenic genotype in 1KGP that was present in both the Select ClinVar and Select HGMD datasets. 6 variants that led to an inferred pathogenic genotype were shared by the Full ClinVar and Select HGMD datasets. A total of 8 variants that led to an inferred pathogenic genotype were shared by the Full ClinVar and Full HGMD datasets. By the end of 2020, 7 of these 8 variants were reclassified to a non-pathogenic category in ClinVar, while in HGMD, 4 of the variants were classified as DM, and 4 were classified as DM?

Considering expanded stand-alone benign guidelines

We next considered how the use of updated BA1 guidelines changed the number of 1KGP individuals who were indicated affected. In accordance with the updated 2018 BA1 guidelines[44], we removed variants from consideration that had a MAF > 5% in any gnomAD exomes continental population. This had no effect on our analysis of Select or Full ClinVar variants (Fig.

4.2A,B). Applying these guidelines to Select HGMD variants led to the removal of 1 variant in 2014 and 2016 (Fig. 4.2C). We found that this reduced the Select HGMD indicated affected individuals by 15 African ancestry individuals in 2014 and 2016, while the 2020 individuals remained at 11. We next applied these guidelines to the Full HGMD variants, which led to the removal of 8 variants in all three years and reduced the number of affected individuals by 75% in 2014 and 62% in 2020 (Fig. 4.2D). Additionally, there was no remaining significant ancestry skew after correcting for multiple tests (see Methods). When we implemented the 2018 BA1 guidelines, the ClinVar and HGMD datasets had similar rates of false positive individuals in 2014, and only recently have their rates diverged (Fig. 4.2E).

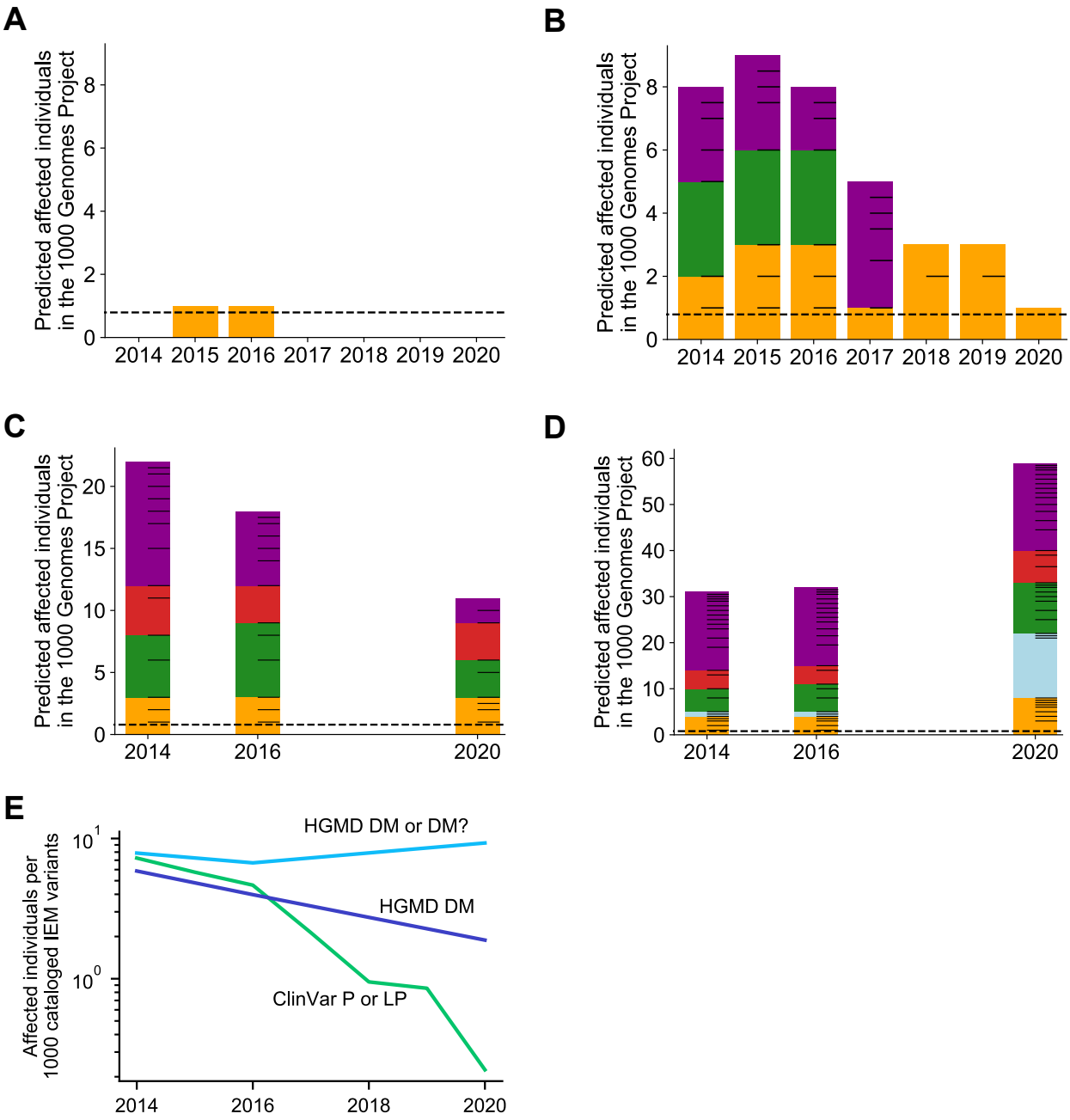

Fig. 4.2 Number of 1KGP individuals indicated affected for screened IEMs with Ghosh et al. guidelines applied. Bar coloring, tick marks, and dashed lines are used as described in Fig. 4.1. The number of 1KGP individuals with a pathogenic genotype for a variant in **A** Select ClinVar variants annotated as pathogenic. **B** Full ClinVar variants annotated as pathogenic. **C** Select HGMD variants. **D** Full HGMD variants. **E** The number of affected individuals relative to the number of variants classified in each variant set.

Comparison of inferred incidence with known incidence of screened IEMs

Next, we sought to characterize the extent of misclassified rare variants that could not be removed by a MAF filter or identified as part of an inferred pathogenic genotype. To do this, we compared the screened IEM incidence inferred from each database with the known incidence of screened IEMs. The aggregate incidence of screened IEMs is estimated to be 1 in 3,200 births[51]. This includes a small number of X-linked IEMs, which are extremely rare, with an estimated aggregate incidence of 1 in 450,000 births. We used these values as baselines to compare with the inferred incidence of screened IEMs. We inferred the screened IEM incidence of each database from the 1KGP allele frequency of annotated pathogenic variants (see Methods), after applying the 2018 BA1 guidelines. Since the inferred incidence of X-linked IEMs is primarily determined by hemizygous males, we consider autosomal and X-linked IEMs separately. For autosomal IEMs, we found that both Full and Select ClinVar variants inferred an incidence greater than the known incidence prior to 2018 (Fig. 4.3A,C). By 2018, the inferred incidence fell below the known incidence, and has remained at 20% of the known incidence for both datasets. For X-linked IEMs, Select ClinVar variants have indicated an incidence of zero since 2014 (Fig. 4.3B). However, Full ClinVar variants have always suggested an incidence orders of magnitude higher than the known incidence, although since 2017 this has been due to just a single variant which primarily is found in East Asian ancestry (Fig. 4.3D). The more comprehensive perspective provided by screened IEM incidence also allows us to observe patterns that were too subtle to be seen in our analysis of indicated affected individuals. For example, we observed that a large fraction of the screened IEM incidence was skewed towards European ancestry from 2015 to 2017 in Select ClinVar variants (Fig. 4.3A). However, due to the extreme rarity of these conditions, it is difficult to precisely infer incidence from 1KGP.

When we considered Select HGMD autosomal variants, we found that the inferred screened IEM incidence has decreased slightly over time, yet in 2020 is triple the known incidence (Fig. 4.4A). The incidence inferred from Full HGMD autosomal variants has increased over time, and in 2020 was 10-fold greater than the known incidence (Fig. 4.4C). As with Full ClinVar variants, the X-linked IEM incidence suggested by Select and Full HGMD variants is orders of magnitude higher than the known incidence (Fig. 4.4B,D). The separation of autosomal and X-linked IEMs suggests that African ancestry skew remains among Full HGMD autosomal variants (Fig. 4.4C), but in our analysis of indicated affected individuals (Fig. 4.2D) this African ancestry skew is largely masked by X-linked variants with high MAF.

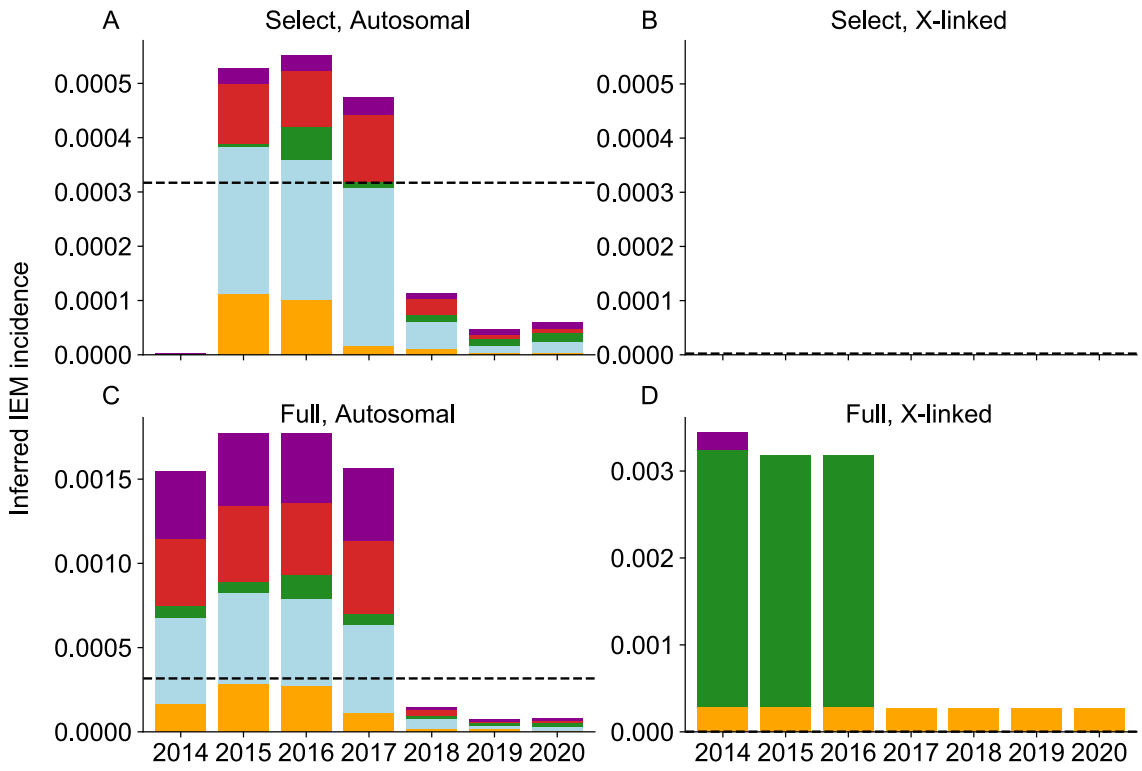

Figure 4.3 Incidence of IEMs inferred by ClinVar variants. Dashed line indicates expected incidence of 1 in 3,200 births. Colors are used as described in Fig. 4.1. The IEM incidence in 1KGP inferred by allele frequency of **A** Select autosomal ClinVar variants annotated as pathogenic. **B** Select X-linked ClinVar variants annotated as pathogenic. **C** Full autosomal ClinVar variants annotated as pathogenic. **D** Full X-linked ClinVar variants annotated as pathogenic.

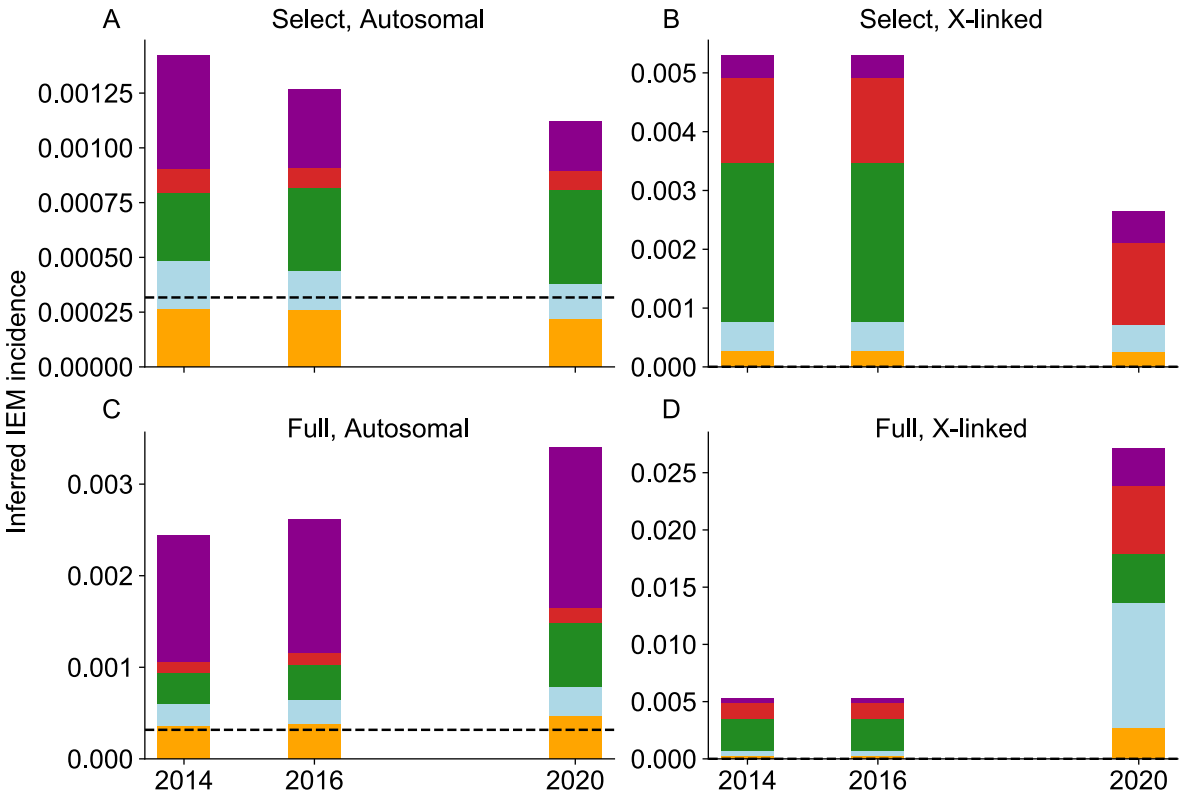

Figure 4.4 Incidence of IEMs inferred by HGMD variants. Dashed line indicates expected incidence of 1 in 3,200 births. Colors are used as described in Fig. 4.1. The IEM incidence in 1KGP inferred by allele frequency of **A** Select autosomal HGMD variants annotated as pathogenic. **B** Select X-linked HGMD variants annotated as pathogenic. **C** Full autosomal HGMD variants annotated as pathogenic. **D** Full X-linked HGMD variants annotated as pathogenic.

Independent validation of major results

We used gnomAD 3.0 genomes to assess the reproducibility of our major findings. We considered gnomAD individuals from the five continental ancestries (African, Latino, East Asian, European, and South Asian; n = 63,269). gnomAD 3.0 does not include any individuals sampled in 1KGP. However, gnomAD does include individuals enrolled in genetic studies. Additionally, gnomAD does not provide individual-level data, so we were unable to identify compound heterozygous variants. These are significant limitations that restrict our confidence in absolute values derived from this analysis. Instead, we focus on robust claims that can be made from trends over time in the relative values we obtained.

For each cataloged variant, we recorded the number of homozygotes and hemizygotes in gnomAD. Overall, our gnomAD analysis replicated all major findings from our 1KGP analysis. Across both ClinVar and HGMD, we found that the proportion of individuals in gnomAD that were indicated affected was almost always less than the proportion affected in 1KGP, but not by less than 50%. One exception was the number of gnomAD individuals affected by Full and Select HGMD variants, which was one third of the size expected based on our 1KGP analysis (Fig. 4.5E,F).

The direction of change in indicated affected individuals for all four datasets over time was nearly always consistent with our 1KGP analysis. We found one notable difference when we considered indicated affected individuals using Select ClinVar variants. In 2015 and 2016, we found an unexpectedly large number of indicated affected individuals (Fig. 4.5C). This can be attributed to a single variant (ACADS:c.511C>T) with a gnomAD MAF >3% and which was P with 1 review star in 2015 and 2016. Due to its modest size, 1KGP did not contain any individuals affected by this variant, although the existence of such individuals was suggested by our incidence analysis in 1KGP, which found elevated European ancestry incidence in Select ClinVar variants from 2015 to 2017 (Fig. 4.3A). The variant, which was annotated as P with 1 star in 2015 and 2016, is more prevalent in European ancestry individuals, resulting in European ancestry individuals significantly ($p < 3.8e\times10^{-10}$) over-represented in 2015 and 2016 Select ClinVar variants, with an odds ratio of 4.0 (95% CI:2.4-6.7). When considering Full ClinVar variants, both East Asian ($p<4\times10^{-6}$) and European ($p<1\times10^{-4}$) ancestry individuals were significantly over-represented from 2014 through 2016, with East Asian individuals having an OR of 4.5 (95% CI:2.6-7.6). These results were obtained by applying the 2015 BA1 guidelines, and they were unaltered when the 2018 BA1 guidelines were applied.

In addition to confirming the African ancestry skew in indicated affected individuals in our 1KGP analysis of HGMD variants, we discovered significant ancestry skew ($p<2\times10^{-6}$) towards East Asian ancestry individuals in Full HGMD variants in 2014 and 2016 (Fig. 4.5F), as well as significant skew towards European ancestry individuals ($p<2\times10^{-5}$) in Select HGMD variants in 2020 (Fig. 4.5E). When 2018 BA1 guidelines were applied, significant skew remained for East Asian and European ancestry individuals. Due to the imbalanced ancestry composition of gnomAD, the described ancestry skew is not obvious from visual inspection of the figures.

With the greater number of individuals in gnomAD relative to 1KGP, we were able to directly compare the false positive rate of Select and Full ClinVar variants. Although there were fewer gnomAD individuals predicted affected by Select ClinVar variants, when considering the inferred

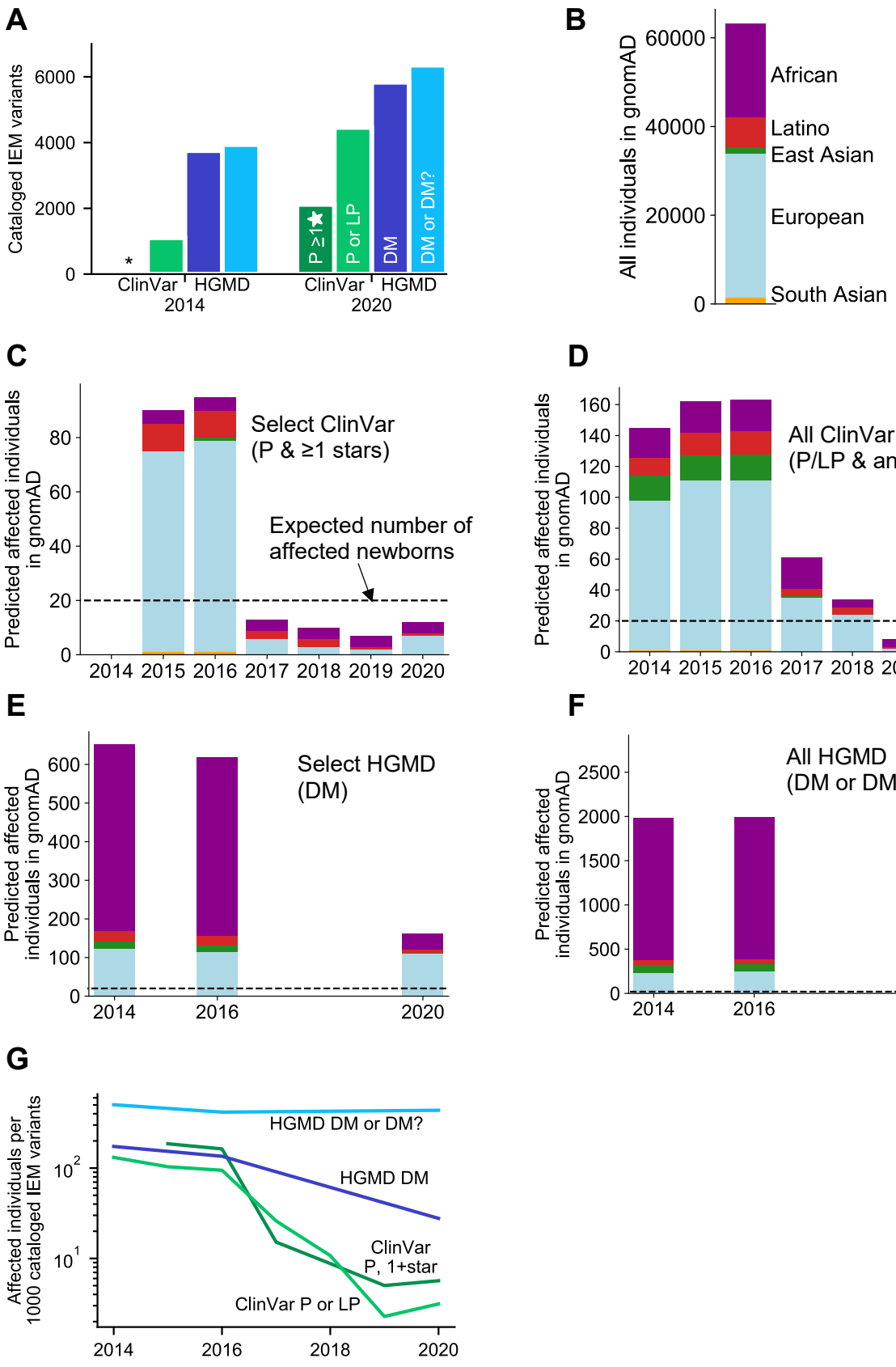
A
Cataloged IEM variants
6000
4000
2000
0
*
P ≥1★
P or LP
DM
DM or DM?
ClinVar
HGMD
2014
ClinVar
HGMD
2020
B
All individuals in gnomAD
60000
40000
20000
0
African
Latino
East Asian
European
South Asian
C
Predicted affected individuals in gnomAD
80
60
40
20
0
Select ClinVar
(P & ≥1 stars)
Expected number of
affected newborns
2014
2015
2016
2017
2018
2019
2020
D
Predicted affected individuals in gnomAD
160
140
120
100
80
60
40
20
0
All ClinVar
(P/LP & any stars)
2014
2015
2016
2017
2018
2019
2020
E
Predicted affected individuals in gnomAD
600
500
400
300
200
100
0
Select HGMD
(DM)
2014
2016
2020
F
Predicted affected individuals in gnomAD
2500
2000
1500
1000
500
0
All HGMD
(DM or DM?)
2014
2016
2020
G
Affected individuals per 1000 cataloged IEM variants
10^2
10^1
HGMD DM or DM?
HGMD DM
ClinVar
P, 1+star
ClinVar P or LP
2014
2016
2018
2020

Figure 4.5 Number of gnomAD individuals indicated affected for screened IEMs by ClinVar or HGMD over time. **A** Number of screened IEM variants present in ClinVar or HGMD in 2014 and 2020 (identical to Fig. 4.1). **B** Ancestry composition of individuals in gnomAD. Due to the imbalanced ancestry composition, ancestry skew is difficult to discern visually in the following panels. **C**-**F** Bars are colored by ancestry as shown in **B**. Dashed black lines indicate the aggregate population incidence of screened IEMs. The number of gnomAD individuals with a pathogenic genotype for a variant in **C** Select ClinVar variants annotated as pathogenic, defined as variants with a P interpretation with at least 1 review star. Variants that also have conflicting interpretations (with VUS or B/LB) with 1 or more review stars are removed. **D** Full ClinVar variants annotated as pathogenic, defined as variants with a P or LP interpretation. Variants that also have conflicting interpretations (with VUS or B/LB) are removed. **E** Select HGMD variants, defined as variants classified as DM. 2014, 2016, and 2020 are shown because they are the years for which we have archived HGMD data. **F** Full HGMD variants, defined as variants classified as DM or DM?. **G** The number of affected individuals relative to the number of variants classified in each variant set. *Data not available because existing review star framework was not in place until 2015.

affected individuals per cataloged variant, we found that there was little difference between Select and Full ClinVar variants.

Comparison of variant reclassification between ClinVar and HGMD

Our analysis of reclassified variants has so far considered only those variants which contributed to an inferred pathogenic genotype in 1KGP individuals. To identify broad trends in variant reclassification in ClinVar and HGMD, we considered all screened IEM variants that were reclassified in ClinVar or HGMD between 2014 and 2020.

Out of 16,857 ClinVar variants, 3,772 (22%) were reclassified between April 2014 and December 2020. Of these reclassified variants, 28% were reclassified 2 or more times. To simplify our analysis, for each variant we considered only the variant's classification when it first entered ClinVar and the variant's classification at the end of 2020. Of the 4,917 P/LP variants in ClinVar between 2014 and 2020, we found 1,655 (34%) were reclassified by the end of 2020 (Fig. 4.6A). 78% of these reclassifications were towards greater evidence for pathogenicity, and the remaining 22% were towards reduced evidence for pathogenicity (8% of all P/LP variants). The most common reclassification towards greater evidence for pathogenicity was from P/LP 1 star to P/LP 2 stars. The most common reclassification towards reduced evidence for pathogenicity was from P/LP 1 star to Conflicting.

HGMD screened IEM variants were reclassified substantially less often than those in ClinVar. Out of 4,777 variants classified as DM or DM? in 2014 or 2016, just 37 (0.8%) were reclassified. 7 of these reclassifications were from DM? to DM, and the remaining 30 were towards reduced evidence for pathogenicity (0.6% of all DM or DM? variants). The most common reclassification towards reduced evidence for pathogenicity was from DM to DM?.

When considering variants reclassified towards reduced evidence for pathogenicity, we found that ClinVar variants were reclassified at a rate 12-fold greater than those in HGMD. We recognize this analysis is impacted by the greater number of available time samples and variant categories in ClinVar compared to HGMD. However, when we repeat this analysis considering

only ClinVar variants at time points for which HGMD data is available, while also collapsing ClinVar pathogenic variants to just 2 categories (see Methods), this result stands.

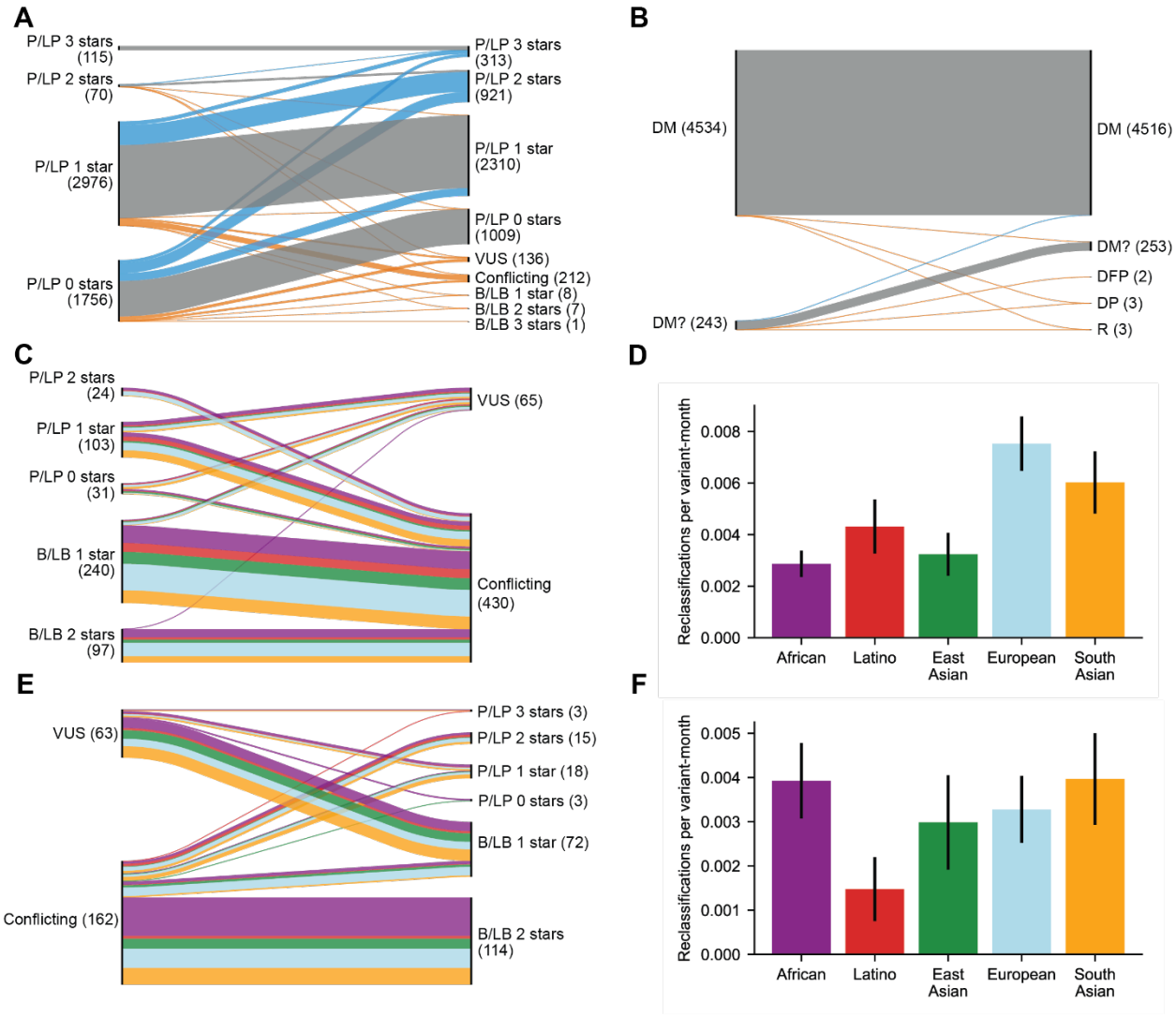

Figure 4.6 Variant reclassification in ClinVar and HGMD **A** Reclassification paths of P/LP ClinVar variants from 2014 (or first submission thereafter) to 2020, visualized in a Sankey plot in which line width represents the number of reclassified variants. Blue lines indicate increasing annotated pathogenicity or review stars, orange lines indicate increasing annotated benignity or reduced confidence of pathogenicity, and gray lines indicate no change. Numbers in parentheses provide variant counts of initial and final classifications for each category. **B** Reclassification paths of DM and DM? HGMD variants from 2014 to 2020. **C** Reclassification paths of ClinVar variants from P/LP or B/LB to VUS or Conflicting. We plot only variants that could be assigned to a principal ancestry. Variant paths are colored by ancestry as in **D**. **D** Rate of reclassification of variants shown in **C** when normalized by the historical ancestry composition of variants in ClinVar. **E** Reclassification paths of ClinVar variants from VUS or Conflicting to P/LP or B/LB. **F** Rate of reclassification of variants shown in **E** when normalized by historical ancestry composition of variants in ClinVar.

Variant reclassification rates by ancestry

In our earlier analysis, we identified ancestry skew in likely misclassified variants. Next, we investigate whether ancestry influences overall reclassification rates of variants in ClinVar. Historically, large-scale exome and genome sequencing projects (from which MAF is often derived) have undersampled non-European individuals[69,70]. Thus, we suspected that non-European individuals may shoulder a larger burden of variants that were initially classified as P or LP due to uncertain MAF and later reclassified to be VUS or Conflicting. At the same time, we recognized that the largest ClinVar submitters are located in countries where a majority of the population has European ancestry. Consequently, variants common in European ancestry individuals may have a greater chance of being interpreted by multiple submitters which could lead to Conflicting interpretations.

To distinguish which of these effects likely dominate in ClinVar, we determined whether variants present in specific ancestries were disproportionately likely to be reclassified. First, for each variant we used gnomAD exomes to identify the continental ancestry group with the highest allele frequency, and we assigned the variant to that ancestry group. gnomAD exome allele frequencies were normalized to avoid bias from sample size differences between ancestries (see Methods). We first considered variants for which the classification was reduced in confidence, which includes P/LP and B/LB variants that were reclassified to VUS or Conflicting. For those variants that could be assigned to an ancestry, we visualized reclassifications using Sankey diagrams in which line width represents the number of reclassified variants, and lines were colored by ancestry (Fig. 4.6C). We observed that European ancestry variants were the largest group in most reclassification paths. However, this analysis did not account for the differences in ancestry composition of the variants submitted to ClinVar. To control for this potential bias, for each ancestry we normalized by both the number of variants assigned to that ancestry and the duration in which they were in ClinVar which we measure in variant-months (see Methods). One variant-month is equivalent to a single variant classified in ClinVar for one month. We normalized only by variants that could have contributed to the reclassification (in this case, P/LP and B/LB). Controlling for the ancestry composition of variants in ClinVar, we found that variants for which European ancestry individuals had the highest MAF were reclassified towards greater uncertainty at a rate of ~0.8% per variant-month (Fig. 4.6D). This was approximately twice the rate of reclassification for variants for which African, East Asian, or Latino ancestries (all $p < 8\times10^{-5}$) had the highest MAF (Fig. 4.6D). This is consistent with our observation that amongst all variants classified in ClinVar, a larger fraction of European ancestry variants were classified as Conflicting. South Asian variants were also found to have elevated reclassification towards greater uncertainty of approximately 0.6% per variant-month, significantly higher than East Asian or African variants (both $p < 2\times10^{-4}$).

We also considered variants for which classification increased in confidence, which includes VUS or Conflicting variants that were reclassified to P/LP or B/LB. After visualizing these reclassifications with Sankey plots, we observed that in many reclassification paths, European variants were not the largest group (Fig. 4.6E), in contrast with reclassification paths towards less confidence. Indeed, when we normalized by the ancestry composition of variants in ClinVar, we found no significant difference between variants most common in African, East Asian, European, or South Asian ancestry, each of which was reclassified at ~0.3% per variant-month (Fig. 4.6F). The exception were variants most common in Latino ancestry, which were reclassified at ~0.1% per variant-month.

Discussion

Variant databases are under continuous development and growth[23,24]. Several studies have attempted to capture this progress at different snapshots in time, although these studies have generally looked at different database elements, making comparisons across time difficult[25-27,71]. Here, we investigated not a single point in time, but evaluated systematically the same disorders over 6 years across two different databases. In both databases, we observed a decrease over time in the number of 1KGP individuals indicated affected by an IEM. Based on the high temporal resolution the ClinVar archives afford, we can see this change was most pronounced in 2016 through 2018 after the establishment of the 2015 ACMG/AMP guidelines and coincident with allele frequency resources such as ExAC. We believe screened IEMs provide an informative lens that reveals broader database trends that may be representative of thousands of rare genetic disorders.

Perhaps our most striking finding is the large difference between the number of affected individuals predicted by HGMD and ClinVar in 2020. However, this difference is not entirely surprising. HGMD states that its curation policy is "to err on the side of inclusion and enter a variant into the database even if its pathological relevance may be questionable" and uses DM? classifications for this purpose as well as frequency flags in its online interface[75]. On the other hand, the clinicians and genetic testing laboratories that contribute to ClinVar are typically concerned with the immediate clinical implications of a variant. While they don't want to pass over variants that could explain proband phenotypes, they are also loath to misinform patients or begin unnecessary interventions that may be irreversible. Thus, ClinVar contributors are invested in maintaining a database of pathogenic variants with minimal false negatives and false positives. An additional factor may be the increasing use of assertion criteria in variants contributed to ClinVar, which compels contributors to delineate the pieces of evidence leading to a classification. In contrast, many journals do not require these pieces of evidence. Therefore, this analysis should not be seen as a duel between two competing databases, but rather a quantitative comparison of the outcomes of two distinct variant cataloging methods. These distinct methods led to the 100-fold difference between the false positive rate of individuals indicated affected by Full ClinVar variants and Full HGMD variants, observed in both 1KGP (Fig. 4.1G) and gnomAD (Fig. 4.5G). While the Full HGMD rate (~25 indicated affected individuals per 1,000 cataloged variants) is still relatively low, our analysis allows us to quantify the difference between the two databases. It is possible that a clinical analysis using HGMD, which includes a greater number of variants than ClinVar, would result in a higher sensitivity analysis, but we are not able to assess false negatives in this study. Understanding this difference between these databases may be valuable to not only clinical researchers, but also to non-domain experts such as computational researchers, who sometimes use HGMD and ClinVar interchangeably to develop variant interpretation methods[76].

Due to founder mutations, individual IEM variants are often enriched in a single ancestry. However, when we consider the total burden of all screened IEMs, continental ancestry groups appear to be affected at similar rates[51]. We found that African ancestry individuals were disproportionately affected in 2014, when HGMD Select variants were considered, but this skew was resolved by 2020. Yet, all of the DM variants causing the ancestry skew in 2014 were reclassified to DM?. Thus, when considering HGMD Full variants, we found that significant African ancestry skew remained. Encouragingly, when we applied the 2018 BA1 guidelines, we observed no significant ancestry skew among Full or Select HGMD variants. This suggests that

much of the observed ancestry skew is due to population-specific common variants. This likely reflects the historical lack of African ancestry samples in large sequencing projects[72,73]. HGMD in particular may be susceptible to these factors, since it catalogs variants directly from publications, including older literature that was written when common variants in African ancestry individuals were poorly characterized. When older studies are given the same credence as recent ones, these disparities are more likely to be perpetuated.

In addition to the 2018 BA1 guidelines, Whiffin et al.[78] have also proposed disorder-specific MAF thresholds which are supported by recent ACMG/AMP guideline specifications[79]. For example, under this system *PAH* variants would have a stand-alone benign MAF threshold of 1.5% assuming a maximum incidence of 1 in 5,000 births. However, we decided not to pursue Whiffin et al. thresholds due to the heterogeneity of our disorders and complications arising from incomplete penetrance in some disorders. Additionally, many screened IEMs are significantly more common in one ancestry group, due to founder effects, which makes it difficult to define thresholds.

Among ClinVar variants that were reclassified, very rarely did the initial submitter change their interpretation, and instead nearly all were reclassified due to conflicting interpretations that largely included assertion criteria. We carefully examined 10 ClinVar variants which previously contributed to an inferred pathogenic genotype but have since been re-classified to a non-pathogenic interpretation. Eight of these variants are currently interpreted as Conflicting. For many variants, this is an accurate descriptor and reflects enduring disagreement among submitters. However, for some variants this may be a byproduct of ClinVar's definition of Conflicting. Specifically, if any P or VUS classification includes assertion criteria (one review star), then regardless of the number of B or LB classifications submitted, the record remains conflicting until the P or VUS submitter changes or retracts their submission. If a submitter is no longer active, then an older submission becomes impossible to change. Although this system has advantages (historical knowledge is not lost), it may also impede the resolution of variants and indicate conflict when there is large consensus. For example, c.323G>T in *ASS1* is currently listed in ClinVar as conflicting, yet it has the following interpretations with at least 1 review star: 1 B, 4 LB, and 1 VUS. The VUS interpretation is from 2017, while the 5 B/LB interpretations are more recent. Although researchers have found that older variant classifications tend to be less accurate, their influence persists[31]. This is even more true for HGMD, which predates ClinVar and thus also contains a large fraction of variants classified without assertion criteria. Given the rapid increase in our ability to determine variant MAF and predict variant pathogenicity, even in the past 5 years, it may be reasonable to require submitters to refresh older interpretations. Under such a system, submitters would need to confirm their interpretations after several years, or the interpretations would be deprecated. This would reduce the influence of 'zombie' interpretations that persist although their submitter is no longer active. Regardless of the exact strategy, methods to confirm the validity of older classifications will be valuable.

Our analysis of ClinVar and HGMD variants revealed a few variants that do not fit the model of screened IEM variants, which typically result in severe, highly-penetrant disorders that begin in infancy or early childhood. For example, we discovered inferred pathogenic genotypes in 1KGP that included c.512C>G in ACAD8 (associated with asymptomatic disease), c.374C>T in *OTC* (our results suggest this variant has incomplete penetrance), and c.148G>A in *OTC* (observed in late onset disease). Asymptomatic IEMs occur when a proband does not have any noticeable

signs of disease, but their metabolites reveal a disease phenotype. These variants are generally classified as disease-causing due to their potential to cause disease. However, as our analysis shows, some individuals will be predicted to have a disease, even though they may never develop symptoms (asymptomatic disease, incomplete penetrance) or symptoms may appear much later in life (late-onset). That we are identifying these variants (as well as variants in diseases known to be asymptomatic such as ACADS and PRODH) implies the databases may in fact be performing better than our analysis suggests. It is possible that some of these variants do cause symptomatic disease in some individuals but disease has not manifested in the 1KGP individuals. These variants may inhabit a gray zone between pathogenic and benign, and they contribute to existing appeals to reconsider the binary paradigm of pathogenic and benign classifications[25,77]. Our work shows that a feature such as optional flags on a ClinVar or HGMD record would be useful. Submitters or curators could then flag entries for various non-standard features for which there is evidence, such as asymptomatic disease, incomplete penetrance, or late-onset in a disease that is typically early onset. This information would then be readily available, without the need to search through supplemental materials of cited publications or detailed explanations provided by submitters. This would be a step towards a classification system that recognizes that variant pathogenicity is multi-faceted, and it would also enable greater interpretability of variant classification data for non-domain experts.

Our gnomAD analysis supported the major findings or our 1KGP analysis. However, we noted a persistent issue in which the number of indicated affected gnomAD individuals was proportionally about half of that expected from our 1KGP results. This is potentially explained by our inability to identify compound heterozygotes. In IEM cohorts, a majority of pathogenic genotypes are caused by compound heterozygous variants[74]. However, in our 1KGP analysis, compound heterozygotes rarely exceeded 20% of indicated affected individuals, possibly due to high-frequency false positives which contributed to a disproportionately large number of homozygotes. Alternatively, the reduction in gnomAD indicated affected individuals may be caused by the imbalanced ancestry composition of gnomAD, specifically the large fraction of European genomes (which had few affected individuals in our 1KGP analysis) compared with the relative paucity of South Asian or East Asian genomes (which contributed to a large fraction of the affected 1KGP individuals). Despite gnomAD's imbalanced ancestry composition, its greater size did allow us to compare the accuracy per variant of Select and Full ClinVar, suggesting that there was little difference in accuracy between the two datasets. Additionally, our gnomAD analysis revealed ancestry skew towards East Asian and European individuals in both ClinVar and HGMD that could not be definitively detected by 1KGP.

This work has several limitations. Among rare diseases, the variants associated with screened IEMs are unusually well-studied thanks to newborn screening programs. Thus, screened IEMs are not necessarily representative of many rare diseases. Furthermore, our primary analysis was limited by the comparably small size of 1KGP relative to the rarity of IEMs. At the same time, 1KGP has several advantages, including its approximately even representation of the 5 major continental ancestries and its open availability of genomes, which allowed us to identify individuals who are compound heterozygous for annotated pathogenic variants and to validate the quality of nearly all analyzed variants. These unique features give 1KGP enduring value. Our analysis was particularly sensitive to putatively misclassified variants on the X chromosome since we considered males who were hemizygous for an annotated pathogenic variant to be affected. This explains the relatively high number of observed *OTC* and *TAZ* variants flagged by our analyses of ClinVar and HGMD, despite the extreme rarity of their associated disorders.

Finally, since few ClinVar submitters provide detailed explanation for their interpretation, and HGMD does not provide detailed explanation for its classifications, for many variants it is difficult to determine with confidence why interpretations changed over time.

We have investigated how the false positive rate of ClinVar and HGMD variants has changed over time. Our results suggest that ClinVar has a lower false positive rate than HGMD due to variant reclassification occurring in the past few years. We noted patterns in variant reclassification, and found that variant interpretation guidelines and diverse allele frequency databases principally contributed to these reclassifications. In agreement with the lower false positive rate of ClinVar variants, we found that annotated pathogenic ClinVar variants are reclassified 12-fold more often than those in HGMD, suggesting that misclassified variants are more readily reclassified in ClinVar than HGMD. We also discovered that variants common in European and South Asian individuals were significantly more likely to be reclassified from P/LP or B/LB to VUS or Conflicting. We conclude that although the allele frequency of variants common in European individuals has been known for longer, due to the increased chance they will be annotated by multiple submitters, they are more often reclassified from a confident category to a less confident category in ClinVar. We anticipate that this work will be a valuable benchmark of the progress that has been made in variant interpretation, of interest to the individuals who maintain these databases, the clinical researchers who use these databases regularly, and the computational researchers who use these databases for training and testing methods.

References

1 Schieppati, A., Henter, J.-I., Daina, E. & Aperia, A. Why rare diseases are an important medical and social issue. *The Lancet* **371**, 2039-2041 (2008).

2 Clark, M. M. *et al.* 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** (2018).

3 Bick, D., Jones, M., Taylor, S. L., Taft, R. J. & Belmont, J. Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases. *J. Med. Genet.* **56**, 783-791 (2019).

4 Frésard, L. & Montgomery, S. B. Diagnosing rare diseases after the exome. *Molecular Case Studies* **4**, a003392 (2018).

5 Schofield, D., Rynehart, L., Shresthra, R., White, S. M. & Stark, Z. Long-term economic impacts of exome sequencing for suspected monogenic disorders: diagnosis, management, and reproductive outcomes. *Genet. Med.* **21**, 2586-2593 (2019).

6 Jayasinghe, K. *et al.* Clinical impact of genomic testing in patients with suspected monogenic kidney disease. *Genet. Med.*, 1-9 (2020).

7 Wise, A. L. *et al.* Genomic medicine for undiagnosed diseases. *The Lancet* **394**, 533-540 (2019).

8 Richards, 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. *Genet. Med.* **17**, 405 (2015).

9 Niehaus, A. *et al.* A survey assessing adoption of the ACMG-AMP guidelines for interpreting sequence variants and identification of areas for continued improvement. *Genet. Med.* **21**, 1699-1701 (2019).

10 Amendola, L. M. *et al.* Variant Classification Concordance using the ACMG-AMP Variant Interpretation Guidelines across Nine Genomic Implementation Research Studies. *The American Journal of Human Genetics* **107**, 932-941 (2020).
11 Amendola, L. M. *et al.* Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. *The American Journal of Human Genetics* **98**, 1067-1076 (2016).
12 Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220 (2013).
13 Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
14 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
15 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
16 Phan, L. *et al. ALFA: Allele Frequency Aggregator*, <www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/> (2020).
17 Fokkema, I. F. *et al.* LOVD v. 2.0: the next generation in gene variant databases. *Hum. Mutat.* **32**, 557-563 (2011).
18 Horaitis, O., Talbot, C. C., Phommarinh, M., Phillips, K. M. & Cotton, R. G. A database of locus-specific databases. *Nat. Genet.* **39**, 425-425 (2007).
19 Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789-D798 (2015).
20 Pagon, R. A. *et al.* GeneTests-GeneClinics: Genetic testing information for a growing audience. *Hum. Mutat.* **19**, 501-509 (2002).
21 Brenner, S. E. Common sense for our genomes. *Nature* **449**, 783-784 (2007).
22 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-D1067 (2018).
23 Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835-D844 (2020).
24 Stenson, P. D. *et al.* The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum. Genet.*, 1-11 (2020).
25 Shah, N. *et al.* Identification of misclassified ClinVar variants via disease population prevalence. *The American Journal of Human Genetics* **102**, 609-619 (2018).
26 Tarailo-Graovac, M., Zhu, J. Y. A., Matthews, A., Van Karnebeek, C. D. & Wasserman, W. W. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. *Genet. Med.* **19**, 1300-1308 (2017).
27 Xue, Y. *et al.* Deleterious-and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *The American Journal of Human Genetics* **91**, 1022-1032 (2012).
28 Cassa, C. A., Tong, M. Y. & Jordan, D. M. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.* **34**, 1216-1220 (2013).
29 Wright, C. F. *et al.* Evaluating variants classified as pathogenic in ClinVar in the DDD Study. *Genet. Med.* **23**, 571-575 (2021).

30 Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* **359**, 1233-1239 (2018).
31 Yang, S. *et al.* Sources of discordance among germ-line variant classifications in ClinVar. *Genet. Med.* **19**, 1118-1126 (2017).
32 Manrai, A. K. *et al.* Genetic misdiagnoses and the potential for health disparities. *New Engl. J. Med.* **375**, 655-665 (2016).
33 Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886-D894 (2019).
34 Ioannidis, N. M. *et al.* REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
35 Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161-1170 (2018).
36 Sharo, A. G., Hu, Z. & Brenner, S. E. StrVCTVRE: A supervised learning method to predict the pathogenicity of human structural variants. *BioRxiv* (2020).
37 Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062-D1067 (2017).
38 Harrison, S. M. & Rehm, H. L. Is 'likely pathogenic'really 90% likely? Reclassification data in ClinVar. *Genome medicine* **11**, 1-4 (2019).
39 Adhikari, A. N. *et al.* The role of exome sequencing in newborn screening for inborn errors of metabolism. *Nat. Med.* **26**, 1392-1397 (2020).
40 Adhikari, A. N. *et al.* Genomic Analysis of Historical Cases with Positive Newborn Screens for Short-Chain Acyl-CoA Dehydrogenase Deficiency Shows That a Validated Second-Tier Biochemical Test Can Replace Future Sequencing. *International journal of neonatal screening* **6**, 41 (2020).
41 Phang, J. M. Disorders of proline and hydroxyproline metabolism. *The metabolic basis of inherited disease* (1995).
42 Finkelstein, J., Hauser, E., Leonard, C. & Brusilow, S. Late-onset ornithine transcarbamylase deficiency in male patients. *The Journal of pediatrics* **117**, 897-902 (1990).
43 Grünert, S. C. Clinical and genetical heterogeneity of late-onset multiple acyl-coenzyme A dehydrogenase deficiency. *Orphanet journal of rare diseases* **9**, 1-8 (2014).
44 Ghosh, R. *et al.* Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum. Mutat.* **39**, 1525-1530 (2018).
45 Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* **48**, D941-D947 (2020).
46 Hall, P. L. *et al.* Postanalytical tools improve performance of newborn screening by tandem mass spectrometry. *Genet. Med.* **16**, 889-895 (2014).
47 Jacquet, H. *et al.* The severe form of type I hyperprolinaemia results from homozygous inactivation of the PRODH gene. *J. Med. Genet.* **40**, e7-e7 (2003).
48 Di Rosa, G., Nicotera, A. G., Lenzo, P., Spanò, M. & Tortorella, G. Long-term neuropsychiatric follow-up in hyperprolinemia type I. *Psychiatric genetics* **24**, 172-175 (2014).
49 Cleynen, I. *et al.* Genetic contributors to risk of schizophrenia in the presence of a 22q11. 2 deletion. *Mol. Psychiatry*, 1-15 (2020).

50 Ghasemvand, F., Omidinia, E., Salehi, Z. & Rahmanzadeh, S. Relationship between polymorphisms in the proline dehydrogenase gene and schizophrenia risk. *Genet Mol Res* **14**, 11681-11691 (2015).
51 Feuchtbaum, L., Carter, J., Dowray, S., Currier, R. J. & Lorey, F. Birth prevalence of disorders detectable through newborn screening by race/ethnicity. *Genetics in medicine* **14**, 937-945 (2012).
52 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
53 Pedersen, B. S. & Quinlan, A. R. cyvcf2: fast, flexible variant analysis with Python. *Bioinformatics* (2017).
54 Team, R. C. R: A language and environment for statistical computing. (2013).
55 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498, doi:http://www.nature.com/ng/journal/v43/n5/abs/ng.806.html#supplementary-information (2011).
56 Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.11-11.10.33, doi:10.1002/0471250953.bi1110s43 (2013).
57 Drmanac, R. *et al.* Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science* **327**, 78-81, doi:10.1126/science.1181498 (2010).
58 Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021).
59 McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biology* **17**, 122 (2016).
60 Lupton, R. C. & Allwood, J. M. Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. *Resources, Conservation and Recycling* **124**, 141-151 (2017).
61 Dobson, C. M. *et al.* Identification of the gene responsible for the cblB complementation group of vitamin B12-dependent methylmalonic aciduria. *Hum. Mol. Genet.* **11**, 3361-3369 (2002).
62 Lerner-Ellis, J. P. *et al.* Mutation and biochemical analysis of patients belonging to the cblB complementation class of vitamin B12-dependent methylmalonic aciduria. *Mol. Genet. Metab.* **87**, 219-225 (2006).
63 Tuchman, M., Morizono, H., Rajagopal, B., Plante, R. & Allewell, N. Identification of 'private'mutations in patients with ornithine transcarbamylase deficiency. *J. Inherited Metab. Dis.* **20**, 525-527 (1997).
64 Pedersen, C. B. *et al.* The ACADS gene variation spectrum in 114 patients with short-chain acyl-CoA dehydrogenase (SCAD) deficiency is dominated by missense variations leading to protein misfolding at the cellular level. *Hum. Genet.* **124**, 43-56 (2008).
65 Kim, C. E. *et al.* Functional modeling of vitamin responsiveness in yeast: a common pyridoxine-responsive cystathionine β-synthase mutation in homocystinuria. *Hum. Mol. Genet.* **6**, 2213-2221 (1997).
66 Gilbert-Dussardier, B. *et al.* Partial duplication [dup. TCAC (178)] and novel point mutations (T125M, G188R, A209V, and H302L) of the ornithine transcarbamylase gene in congenital hyperammonemia. *Hum. Mutat.* **8**, 74-76 (1996).

67 Wongkittichote, P., Tungpradabkul, S., Wattanasirichaigoon, D. & Jensen, L. T. Prediction of the functional effect of novel SLC25A13 variants using a S. cerevisiae model of AGC2 deficiency. *J. Inherited Metab. Dis.* **36**, 821-830 (2013).
68 Zielonka, M. *et al.* Early prediction of phenotypic severity in Citrullinemia Type 1. *Annals of clinical and translational neurology* **6**, 1858-1871 (2019).
69 Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26-31 (2019).
70 Petrovski, S. & Goldstein, D. B. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. *Genome biology* **17**, 1-3 (2016).
71 Xiang, J. *et al.* Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. *Scientific Reports* **10**, 1-5 (2020).
72 Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489-494 (2009).
73 Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584-591 (2019).
74 Blau, N. Genetics of phenylketonuria: then and now. *Hum. Mutat.* **37**, 508-515 (2016).
75 Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136**, 665-677 (2017).
76 Jagadeesh, K. A. *et al.* S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat. Genet.* **51**, 755-763 (2019).
77 Manrai, A. K., Ioannidis, J. P. & Kohane, I. S. Clinical genomics: from pathogenicity claims to quantitative risk estimates. *Jama* **315**, 1233-1234 (2016).
78 Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151-1158 (2017).
79 Harrison, S. M., Biesecker, L. G. & Rehm, H. L. Overview of specifications to the ACMG/AMP variant interpretation guidelines. *Current protocols in human genetics* **103**, e93 (2019).

Chapter 5: Conclusions and Future Directions

Clinicians and researchers face a crucial challenge in the molecular diagnosis of rare diseases. Despite the genetic etiology of these diseases, less than half are able to be resolved by exome or genome sequencing. Structural variants explain a portion of these unresolved cases, as they are challenging both to detect and interpret. To improve the interpretation of structural variants, I developed StrVCTVRE, a method to automatically prioritize exon-affecting deletions and duplications. I anticipate that this method will not only assist clinical researchers to resolve cases but also enable gene-phenotype discoveries. Next, I compared the ability of two recent long-read sequencing methods to detect pathogenic structural variants in a rare disease clinical setting. I found that these recent methods detect more diagnostic and candidate structural variants than can be detected with short-read sequencing. Although their associated structural variant calling algorithms would benefit from further refinement, these long-read sequencing methods and others like them show promise in resolving cases by detecting diagnostic structural variants that may otherwise be missed. Finally, I investigated how rare disease variants have been reclassified over time in the variant curation databases ClinVar and HGMD. I found that the unique methodologies used by each database have resulted in a large difference in their rates of variant reclassification, with ClinVar variants reclassified 12-fold more often than those in HGMD. Consequently, HGMD predicts a much larger number of false positive individuals than ClinVar and with a significant skew towards African ancestry individuals. Overall, my work suggests that our ability to identify variants that cause rare disease is slowly improving. While I am glad to have contributed to this progress during my PhD training, there remains much further work to be done.

Exon-affecting variants made up the majority of the variants in my analysis. This is due to two related reasons: it is difficult to determine whether an intronic or intergenic variant is pathogenic due to our limited understanding of the function of these regions; and the majority of classified variants (which are used to train and test methods) overlap exons. Yet, noncoding variants are the cause of a portion of unresolved rare disease cases. Fortunately, existing limitations should begin to erode as biological data increases. Valuable data will continue to come from ENCODE[1] and similar projects that seek to experimentally identify functional elements in the human genome and epigenome. Additional insights will be gained by aggregating data from the enormous number of human genomes that will be sequenced in the near future, through large-scale projects like the UK Biobank[2]. Finally, diverse functional genomic data such as expression[3] and proteomics data[4] will enable researchers to model cellular networks and pathways that are disrupted by pathogenic variants. By integrating these data, I am optimistic that researchers will be able to further refine and test variant impact predictors that will provide accurate clinical interpretations for noncoding regions of the genome.

An additional promising direction in rare disease diagnosis is the interrogation of RNA. When the tissue of interest can be biopsied, RNA provides a window into disease that is one step closer to the actual phenotype than DNA. Consider, RNA can readily reveal defects in splicing and expression which may be completely opaque in analysis of the corresponding DNA. Additionally, because gene expression is often the result of a finely-tuned network of co-regulation, RNA can reveal network signatures of a genetic variant that would be invisible in DNA. Already RNA analysis shows promise in specific cases[5,6], and I am optimistic that the integration of RNA and DNA analysis will resolve rare disease cases that would otherwise go unresolved.

An alternate path to improve variant detection is to improve the substrate against which we call variants, the human reference genome. There has not been a coordinate-altering update to the human genome since GRCh38 was released in 2013[7] (although the Telomere-to-Telomere Consortium recently released the first apparently complete sequence of a human genome[8]). Recent proposals to improve the reference genome have included constructing pan-genomes[9]. A pan-genome would include all common variation in the human population. These additional sequences would be considerable, given that most individuals have >20,000 structural variants relative to the reference genome. These sequences would likely reduce read misalignments and also enable variant calling in regions which are currently absent from the reference genome. The urgency of this approach has been highlighted by a recent study showing that sequencing of 910 African ancestry individuals reported 300 Mb of novel sequence that is absent from the reference genome[10]. Pan-genomes will likely require novel alignment and calling algorithms. Perhaps the most revolutionary innovation would be to create a graph representation of the genome, rather than a linear string[11]. Until a single pan-genome is created for all humans, a useful intermediate step will be the population-specific pan-genome, which is a pan-genome focused on a single ancestry[12]. I anticipate that these approaches will allow researchers to detect pathogenic variants which are currently missed, improving the diagnostic yield in rare disease cases.

It is likely that sequencing will play a growing role in our own lives and those of the next generation. We are approaching an era in which genome sequencing at birth may become a widespread practice with the potential to revolutionize healthcare. This influx of personal genome data will be accompanied by advances in our ability to interpret these data. Screening newborns for complex diseases such as autism, asthma, or depression and providing prophylactic care will likely be possible. Sufficient advances may even enable pre-implantation screening of embryos for these diseases. Given the potentially widespread effects of these innovations on society, it will be critical to consider the ethical, legal, and social implications of these technologies to improve upon—rather than exacerbate—existing health disparities. A primary consideration when developing methods to improve clinical care is to ensure that the clinical value of sequencing is similar across ancestry groups. However, European ancestry individuals comprise 81% of individuals in genome-wide association studies[13]. Variant databases are similarly biased, with a recent study finding ClinVar missed a large number of hearing impairment variants that primarily affect African ancestry individuals[14]. Given the future importance of polygenic risk scores and cataloged variants, how can researchers ensure that genetic databases reflect human diversity? Projects to sequence underrepresented populations, such as H3 Africa[15], will be increasingly essential. As a greater number of humans are sequenced for clinical research, there will inevitably be concerns regarding how these data can be effectively used to improve healthcare while reducing personal risk. One option to improve data privacy and security is through a federated learning model, in which data are not shared directly but new methods are able to learn on samples across data silos[16]. Given the increasing risk of data leaks and ease of identifying individuals from genetic data, new legislation may also be useful to limit potential risks to research participants[17]. Although these challenges are concerning, I am optimistic that by being honest about current inequalities, researchers will build the determination to tackle and eventually overcome these limitations.[*]

[*] This paragraph was adapted from a published article[18] and originally written by Julia Brown. This work is included with permission from the authors.

Our ability to identify the variants that cause rare diseases is already rapidly growing, and will likely accelerate as new data pours in. Although we are still quite limited in our ability to predict the impacts of most variants, research into rare disease variants have already yielded great advances in both biology and clinical care[19-21]. Considering the track record of advances that resulted from rare disease research, I anticipate that the field will continue to generate innovations that benefit not only those with a rare disease, but all of humanity.

References

1 Moore, J. E. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699-710 (2020).
2 Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature, 1-10 (2021).
3 Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318-1330 (2020).
4 Deutsch, E. W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res., gkw936 (2016).
5 Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science Translational Medicine 9, eaal5209 (2017).
6 Frésard, L. et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. 25, 911-919 (2019).
7 Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849-864 (2017).
8 Nurk, S. et al. The complete sequence of a human genome. bioRxiv, 2021.2005.2026.445798, doi:10.1101/2021.05.26.445798 (2021).
9 Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. Nature Reviews Genetics 21, 243-254 (2020).
10 Sherman, R. M. et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet. 51, 30-35 (2019).
11 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907-915 (2019).
12 Li, Q. et al. Building a Chinese pan-genome of 486 individuals. Communications biology 4, 1-14 (2021).
13 Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. Nature News 538, 161 (2016).
14 Chakchouk, I. et al. Disparities in discovery of pathogenic variants for autosomal recessive non-syndromic hearing impairment by ancestry. Europ. J. Hum. Genet. 27, 1456-1465 (2019).
15 Consortium, H. A. Enabling the genomic revolution in Africa. Science 344, 1346-1348 (2014).
16 Rieke, N. et al. The future of digital health with federated learning. NPJ digital medicine 3, 1-7 (2020).
17 Clayton, E. W., Evans, B. J., Hazel, J. W. & Rothstein, M. A. The law of genetic privacy: applications, implications, and limitations. Journal of Law and the Biosciences 6, 1-36 (2019).

18 McInnes, G. et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. The American Journal of Human Genetics 108, 535-548 (2021).
19 Taylor, A. How studying a disease that affects hundreds of people could save millions of lives. The Conversation (2016).
20 Endo, A. A gift from nature: the birth of the statins. Nat. Med. 14, 1050-1052 (2008).
21 Gillmore, J. D. et al. CRISPR-Cas9 in vivo gene editing for transthyretin amyloidosis. New Engl. J. Med. 385, 493-502 (2021).