6-2022

Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes.

Ana S A Cohen
Children's Mercy Hospital

Emily G. Farrow
Children's Mercy Hospital

Ahmed Abdelmoity
Children's Mercy Hospital

Joseph Alaimo
Children's Mercy Hospital

Shivarajan Manickavasagam Amudhavalli
Children's Mercy Hospital

See next page for additional authors

Recommended Citation

Cohen ASA, Farrow EG, Abdelmoity AT, et al. Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. Genet Med. 2022;24(6):1336-1348. doi:10.1016/j.gim.2022.02.007

Creator(s)

Ana S A Cohen, Emily G. Farrow, Ahmed Abdelmoity, Joseph Alaimo, Shivarajan Manickavasagam Amudhavalli, John Anderson, Lalit R. Bansal, Lauren E. Bartik, Primo Baybayan, Bradley Belden, Courtney D. Berrios, Rebecca L. Biswell, Pawel Buczkowicz, Orion Buske, Shreyasee Chakraborty, Warren A. Cheung, Keith A. Coffman, Ashley M. Cooper, Laura A. Cross, Tom Curran, Thuy Tien T. Dang, Mary M. Elfrink, Kendra Engleman, Erin Day Fecske, Cynthia Fieser, Keely M. Fitzgerald, Emily Fleming, Randi N. Gadea, Jennifer L. Gannon, Rose N. Gelineau-Morel, Margaret Gibson, Jeffrey Goldstein, Elin Grundberg, Kelsee Halpin, Brian S. Harvey, Bryce Heese, Wendy Hein, Suzanne M. Herd, Susan Starling Hughes, Mohammed Ilyas, Jill Jacobson, Janda L. Jenkins, Shao Jiang, Jeffrey J. Johnston, Kathryn Keeler, Jonas Korlach, Jennifer Kussman, Christine Lambert, Caitlin E. Lawson, Jean-Baptist LePichon, J Steven Leeder, Vicki C. Little, Daniel A. Louiselle, Michael Lypka, Brittany D. McDonald, Neil Miller, Ann Modrcin, Annapoorna Nair, Shelby H. Neal, Christopher M. Oermann, Donna M. Pacicca, Kailash Pawar, Nyshele L. Posey, Nigel Price, Laura M B Puckett, Julio Quezada, Nikita Raje, William J. Rowell, Eric T. Rush, Venkatesh Sampath, Carol J. Saunders, Caitlin Schwager, Richard M. Schwend, Elizabeth Shaffer, Craig Smail, Sarah E. Soden, Meghan Strenk, Bonnie Sullivan, Brooke Sweeney, Jade B. Tam-Williams, Adam Walter, Holly Welsh, Aaron M. Wenger, Laurel K. Willig, Yun Yan, Scott T. Younger, Dihong Zhou, Tricia N. Zion, Isabelle Thiffault, and Tomi Pastinen

Genetics in Medicine (2022) 24, 1336–1348

ELSEVIER

Genetics in Medicine

www.journals.elsevier.com/genetics-in-medicine

ARTICLE

Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes

ARTICLE INFO

Article history:
Received 15 October 2021
Received in revised form
5 February 2022
Accepted 7 February 2022
Available online 16 March 2022

Keywords:
AI
Genome sequencing
Long read sequencing
NGS
Rare disease

ABSTRACT

Purpose: This study aimed to provide comprehensive diagnostic and candidate analyses in a pediatric rare disease cohort through the Genomic Answers for Kids program.

Methods: Extensive analyses of 960 families with suspected genetic disorders included short-read exome sequencing and short-read genome sequencing (srGS); PacBio HiFi long-read genome sequencing (HiFi-GS); variant calling for single nucleotide variants (SNV), structural variant (SV), and repeat variants; and machine-learning variant prioritization. Structured phenotypes, prioritized variants, and pedigrees were stored in PhenoTips database, with data sharing through controlled access the database of Genotypes and Phenotypes.

Results: Diagnostic rates ranged from 11% in patients with prior negative genetic testing to 34.5% in naive patients. Incorporating SVs from genome sequencing added up to 13% of new diagnoses in previously unsolved cases. HiFi-GS yielded increased discovery rate with >4-fold more rare coding SVs compared with srGS. Variants and genes of unknown significance remain the most common finding (58% of nondiagnostic cases).

Conclusion: Computational prioritization is efficient for diagnostic SNVs. Thorough identification of non-SNVs remains challenging and is partly mitigated using HiFi-GS sequencing. Importantly, community research is supported by sharing real-time data to accelerate gene validation and by providing HiFi variant (SNV/SV) resources from >1000 human alleles to facilitate implementation of new sequencing platforms for rare disease diagnoses.

Introduction

The Children's Mercy Research Institute in Kansas City established a large-scale genomic disease program named "Genomic Answers for Kids" (GA4K) to expand diagnostic capabilities and catalog rare disease genomes and phenotypes within a health care system. Broad recruitment across all pediatric rare diseases resulted in most patients entering the study either with negative or no prior genetic testing. Recent studies have shown >10% rate of new findings upon reanalysis of exome sequencing (ES) or genome sequencing (GS) data in patients with a history of negative genetic testing.[1-4] The predominant factors in identifying new diagnoses were recent publications establishing novel gene–disease associations, often through data-sharing efforts such as GeneMatcher (GM) (upgrade from gene of uncertain significance [GUS]), or expanding the phenotypic spectrum of established disease genes (upgrade from variant

Ana S.A. Cohen and Emily G. Farrow contributed equally.

*Correspondence and requests for materials should be addressed to Isabelle Thiffault, Genomic Medicine Center, Children's Mercy Kansas City, 2401 Gilham Rd, Kansas City, Missouri 64108. *E-mail address:* ithiffault@cmh.edu. or *Tomi Pastinen, Genomic Medicine Center, Children's Mercy Kansas City, 2401 Gilham Rd, Kansas City, Missouri 64108. *E-mail address:* tpastinen@cmh.edu

A full list of authors and affiliations appears at the end of the paper.

doi: https://doi.org/10.1016/j.gim.2022.02.007

of uncertain significance).[1,3,5] The next most helpful strategy to increase diagnostic yield was the incorporation of sequencing data from additional family members, particularly for singletons.[4] Furthermore, given the continued advances in technology and expanded availability of public data, samples sequenced and/or analyzed >3 to 5 years ago may also benefit from resequencing to enhance coverage and/or repipelining to incorporate improved filtering methods and more extensive population data.[3,6]

The variable success in analyses/reanalyses is largely explained by patient ascertainment and testing schemes, although differing variant prioritization strategies are also likely to play a role. Specifically, depending on the relative weight placed on inheritance, variant-effect properties, and the identity/function of the gene harboring the rare variant, the ranking of candidate variants may yield very different results. Multiple machine-learning tools have emerged to balance the variant/locus characteristics in an attempt to systematically extract optimal candidate prioritization.[7] The integration of such tools in rare disease molecular analyses has been shown by several centers primarily for small, selected cohorts.[8-12] The universal feature is the patient's phenotype coded through human phenotype ontology (HPO) terms as a basis for prioritization, followed by the deployment of variable ranking algorithms.[13] However, the utility of incorporating such tools for a systematic first-pass analysis of patient data within a large, unselected, and phenotypically diverse pediatric rare disease diagnostic setting is unknown.

Although variant prioritization strategies continue to improve, the choice of technology in genome-wide sequencing and primary data processing strategy have remained comparatively stable, despite missing some variant types, including structural variants (SVs).[14,15] At our center, the performance of short-read (sr) genome sequencing (srGS) and ES was similar when used in the diagnostic evaluation of suspected pediatric genetic disease on the same Illumina platform.[16] However, alternative platforms have the potential to reduce uncertainty of chemistry-dependent errors and omissions, and scalable alternatives have emerged for short-read polymerase chain reaction (PCR)-free genomes such as DNA NanoBall sequencing.[17] Furthermore, long-read GS (lrGS) has been shown to detect variants missed by short-read sequencing, specifically complex SVs, including inversions and inverted duplications, as well as repeat expansions and variants in difficult-to-map regions.[18] In addition, lrGS also has the potential to resolve phasing of variants in autosomal recessive genes when parental samples are unavailable. Recent technological advances in long-read platforms enable the consideration of lrGS for unsolved rare diseases.[19]

In this study, we leveraged a large-scale pediatric genomic medicine program with real-time return of results to explore automation of variant prioritization and expert clinical interpretation, as well as the retesting of prior negative exomes at a scale that has not been previously reported. The results from the analyses of >1000 patients with rare disease highlight the utility of systematic variant prioritization, identify variants in blind spots associated with current technologies, and underscore the imperative for improved sharing strategies of suggestive results across rare disease programs and cohorts.[20]

Materials and Methods

Detailed methods are described in the Supplemental Materials and Methods online. All analyses were completed on Genome Reference Consortium Human Build 38 (GRCh38).

Cohort

The case cohort described included 1083 affected patients from 960 families, with a total of 2957 sequenced individuals collectively (detailed in Supplemental Tables 1 and 2). Cases included 595 males and 488 females, aged 1 to 55 years (older individuals were typically ascertained as follow-up from an affected family member). Of these patients, 158 (14.6%) were singletons, whereas the remaining 955 had at least 1 additional family member sequenced. Patients were referred from 22 different specialties, with the largest proportion nominated by Clinical Genetics (47.7%) followed by Neurology (22.9%). Given the broad referral pool, we acknowledge the limitations in the ethnic diversity of this population that may reflect systemic health care issues; these will be addressed directly in future studies. A continuum of pediatric conditions is represented, ranging from congenital anomalies to more subtle neurological and neurobehavioral clinical presentations later in childhood. Of the 1083 patients, 125 entered the study with a known genetic diagnosis because the program is building an inclusive rare disease genome resource with solved cases serving to benchmark new methods and processes. Phenotypes were manually extracted from the medical records and primary features recorded in PhenoTips using HPO terminology.[13,21] These structured data were used for automated prioritization tools, whereas expert review used the complete clinical notes for variant prioritization and interpretation. A summary of HPO terms/patient is detailed in Supplemental Table 3.

ES/srGS

Exome libraries were prepared according to the manufacturer's standard protocols using the Illumina TruSeq DNA PCR-Free library preparation kit (Illumina) with 10 cycles of PCR, followed by enrichment with the IDT xGen Exome Research Panel v2, with additional spike-in oligos (Integrated DNA Technologies) to capture the mitochondrial genome and dispersed genomic regions for copy number variation (CNV) detection. PCR-Free genome libraries were prepared according to the manufacturer's standard protocols for Illumina TruSeq DNA PCR-Free library preparation.

MGI sequencing (srGS)

Genome sequencing libraries were constructed using the MGIEasy Universal DNA Library Prep Set (MGI Tech Co, Ltd) according to the manufacturer's standard protocols. srGS was performed on MGI DNBSEQ-G400 (MGI Tech Co, Ltd).

PacBio HiFi long-read GS and analysis

DNA was sheared to a target size of 14 kb using the Diagenode Megaruptor 3 (Diagenode). Single molecule, real-time (SMRT) bell libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (100-938-900, Pacific Biosciences) following the manufacturer's standard protocol (101-693-800) with modifications described in the Supplemental Materials and Methods. Libraries were sequenced on the Sequel IIe Systems using the Sequel II Binding Kit 2.0 (101-842-900, Pacific Biosciences) or 2.2 (102-089-000, Pacific Biosciences) and Sequel II Sequencing Kit 2.0 (101-820-200, Pacific Biosciences) with 30 hour movies/SMRT cell. In total, 175 samples were sequenced to a target of >25× coverage; 297 samples were sequenced on 1 SMRT Cell (average: 10× coverage).

Read mapping, variant calling, and genome assembly were performed using a Snakemake workflow. HiFi reads were mapped using pbmm2 v1.4.0, and SVs were called using pbsv 2.4.0. Single nucleotide variants (SNVs) were called using DeepVariant v1.0 following DeepVariant best practices for PacBio reads.[22] De novo assembly was performed using hifiasm v0.9-r289 using default parameters.[23]

SV call sets were compared using svpack match, which considers 2 SV calls to match when the variants are of the same type (considering insertion and duplication to be the same), are nearby (start position difference ≤100 base pairs [bp]), and are of similar size (size difference ≤100 bp). To systematically evaluate expansions at known pathogenic tandem repeat loci, tandem genotypes were used to count the length of the tandem repeats in HiFi reads for each sample.[24] Because long (GA)-rich repeats have been noted to have lower coverage in HiFi reads, a complementary system was setup to identify haplotypes with coverage dropouts at the known pathogenic tandem repeat loci.[25] At each locus, the number of reads that span the repeat region were counted per haplotype (on the basis of a Whatshap-haplotagged BAM from phased SNVs) (unpublished—Martin et al, WhatsHap: fast and accurate read-based phasing. bioRxiv. 2016. https://doi.org/10.1101/085050). A coverage dropout was identified as a locus with fewer than 2 spanning reads in a haplotype.

Joint calling of SV and small variants was also completed for HiFi long-read genome sequencing (HiFi-GS). A multisample SV call set was produced by merging single-sample pbsv call sets with JASMINE v1.1.4 (unpublished—Kirsche et al. Population-scale SV comparison and analysis. BioRxiv. 2021. https://doi.org/10.1101/2021.05.27.445886). A multisample small variant call set was produced by running GLnexus v1.2.7 on all single-sample DeepVariant genomic variant call format files using glnexus_cli –config DeepVariant_unfiltered and converting the resulting binary variant call format to variant call format using bcftools view v1.10.[26]

Analyses and variant prioritization pipeline

Figure 1 depicts an overview of the sequence processing, variant calling, and interpretation pipeline. Reanalysis was carried out using ES/srGS data in parallel. Exomiser v12.1 (Sanger Institute) (data version 2102) and AMELIE v3.1.0 were applied for variant prioritization, and the top ranked variants were manually reviewed and flagged for expert interpretation.[27,28] Variant prioritization was restricted, impacting common pathogenic variants (Supplemental Table 4). An additional sequencing platform using srGS was tested in a subset of trios (MGI), whereas HiFi-GS (PacBio) was predominantly deployed for cases without diagnosis after srGS. Finally, an early phase of the study employed 10x Linked-Read GS, predominantly in singleton cases (Supplemental Materials and Methods). Supplemental Table 5 summarizes the different types of data generated for the cohort.

Annotation of SVs for disease relevance used both frequency (minor allele frequency of <1%) in a local sequence modality specific SV warehouse and focused on overlap with OMIM morbid genes, followed by manual curation to interpret the validity of candidate SV calls as well as relevance in context of the phenotype/known transmission of disease at locus.

Clinical validation of research results

Variants identified through research sequencing were reviewed in accordance with American College of Medical Genetics and Genomics/Association for Molecular Pathology criteria; pathogenic and likely pathogenic variants related to the disease phenotype were confirmed in the Children's Mercy Clinical Laboratory Improvement Amendments–certified laboratory through the best applicable validated methods and reported clinically in real-time for incorporation into clinical management.[29]

Results

Machine-assisted interpretation

A combination of 2 publicly available tools was implemented to aid with variant prioritization: Exomiser (E) and AMELIE (A).[27,28] Both tools (E/A) rely on structured

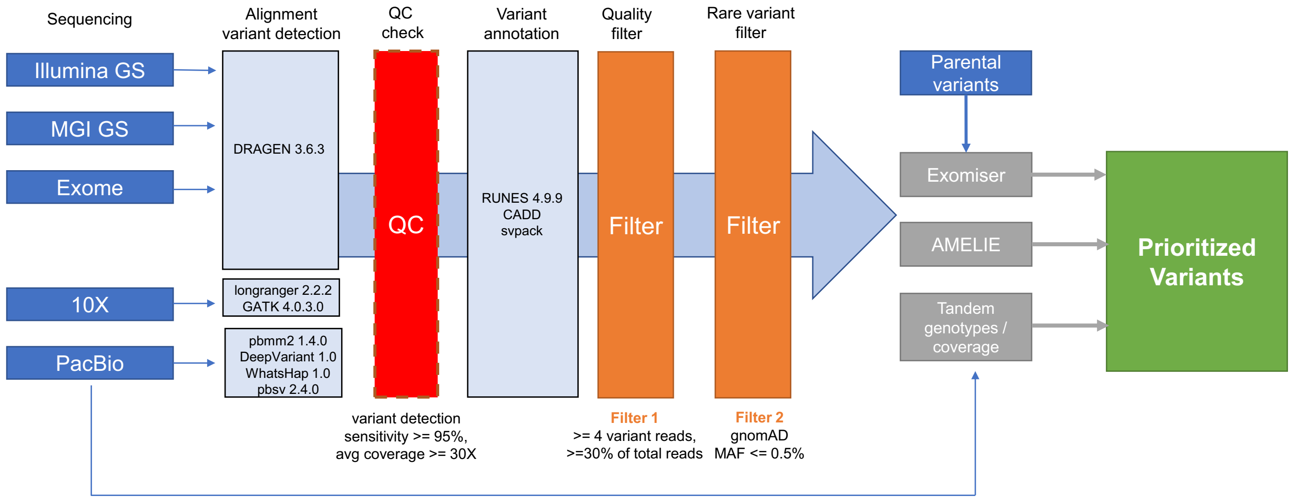

Figure 1 **Genomic Answers for Kids pipeline.** Overview of sequencing, variant calling, and variant prioritization pipeline. Sequencing included exome sequencing and GS through multiple technologies (Illumina, MGI, and 10x for short-reads, and PacBio for long-reads). Standard QC and filtering were applied. Variant prioritization relied on inheritance pattern and AI tools (Exomiser/AMELIE) and tandem genotypes. AI, artificial intelligence; CADD, combined annotation dependent depletion; GATK, genomic analysis toolkit; gnomAD, Genome Aggregation Database; GS, genome sequencing; MAF, minor allele frequency; QC, quality control.

phenotyping (with HPO terms) but apply algorithms that explore different features of variants/genes (Supplemental Materials and Methods). Therefore, we hypothesized that the combination would improve speed and accuracy of analysis. To test the efficacy of these tools, we reviewed the combined top 50 ranked E/A candidate variant list for cases with known molecular diagnoses at study entry (n = 125), with a knowledge of the phenotype of each proband but blinded to the original genetic results. Of these, 88 had diagnostic SNVs serving as a positive control set (other known diagnoses included aneuploidies, microdeletions/duplications, repeat expansions, or special cases such as *SMN1*/2 variants not in the scope of exome or genome interpretation provided in this article and are described in Figure 2A as other mechanism). The causative variant was ranked within this top 50 by E/A in 84 (95.5%) of the positive control cases. Diagnostic variants were sometimes outranked by single pathogenic variants (carrier status) or 2 variants in *cis* in autosomal recessive genes. Absolute E/A score distribution is shown in Figure 2B. Of the 4 cases for which the diagnostic variant was not ranked, 3 had deep intronic pathogenic variants, a recognized limitation of E/A prioritization; therefore, only 1 diagnostic coding variant was missed in this subset of cases.

On expanding the E/A strategy to the entire data set and comparing with expert review, in which variants were prioritized on the basis of multiple criteria (zygosity, segregation, population frequency, gene function, etc.) using our custom software RUNES (Supplemental Materials and Methods), variant prioritization was found to be concordant in approximately 49.8% of 1083 cases (Figure 2A). This means that the top variants manually selected from the combined E/A files for further review were also considered to be the best candidate variants when expert analysts reviewed the full clinical notes and unrestricted variant lists, and yet this variant selection was achieved in a fraction of the time.[30] No strong E/A candidates were identified in approximately 8.4% of cases that were positive for a variant that would not have been annotated by these tools (such as copy number variants, deep intronic variants, SVs, repeat expansions, etc.). Moreover, approximately 30.6% of cases were deemed negative by both expert analysis and combined E/A ranking, giving us an overall consistency in analysis outcome of approximately 88.7% (Figure 2A). Importantly, in approximately 3.4% of cases (n = 37), these tools pointed us toward new candidates who may not have otherwise been considered.

Diagnostic yields stratified by earlier testing history

Of the 958 patients (88.5%; 958/1083) who entered the study without a previous diagnosis, the largest group (584/958) consisted of patients with negative genetic testing history, either through ES, srGS, or panel testing. New ES and srGS/lrGS with (re)analysis yielded definitive diagnoses for 64 of 584 cases (11%). A smaller group of patients, referred to the research study and to clinical ES in parallel, achieved a diagnostic rate of 34.5%, which was 71 of 206 cases, and among the patients that had no clinical genetic testing approved/ordered, the diagnostic rate was 20.2%, which was 34 of 168 cases. The latter group represented patients whose physicians did not order testing and/or those for whom testing was ordered but denied by the insurance. Various modes of reinterpretation success are exemplified in

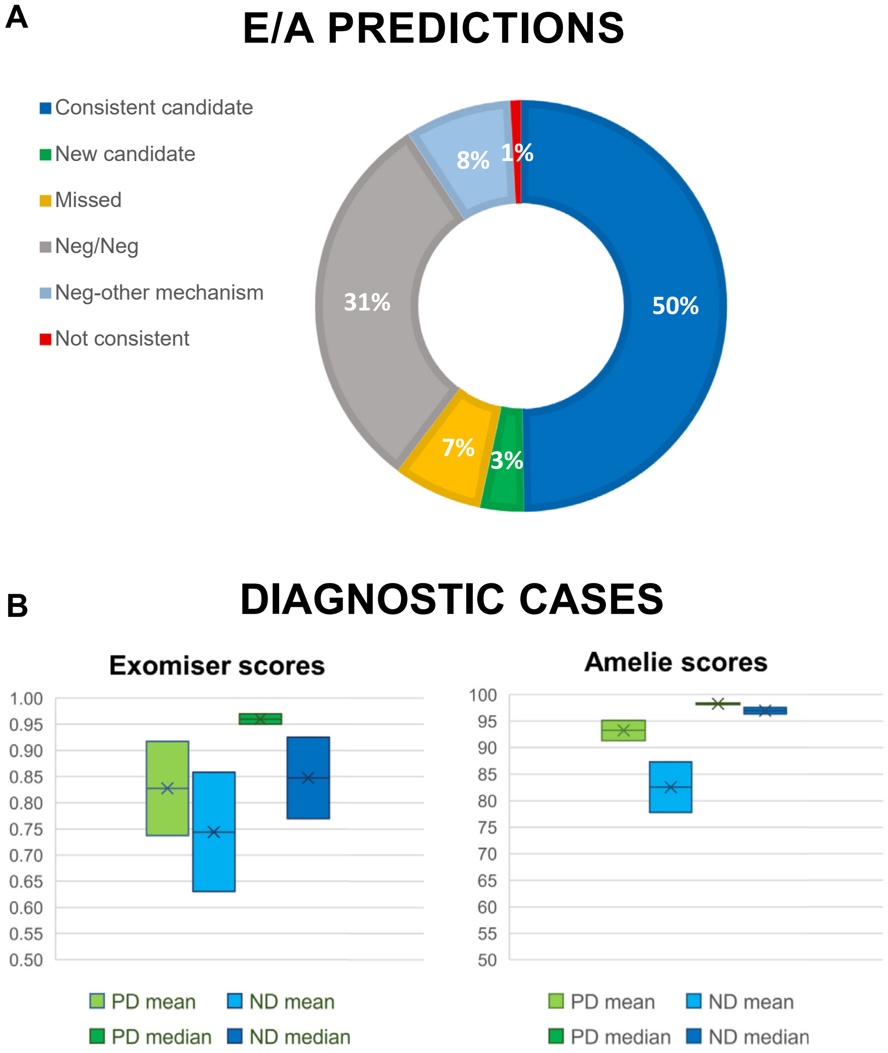

Figure 2 **Variant prioritization tools showed great concordance with expert analysis. A**. Distribution of E/A predictions for all 1083 patients. Prioritization was deemed concordant when the main candidate variant was concordant with expert review ("consistent candidate"), Neg by both E/A and expert review ("Neg-Neg"), or when the causative variant was not an single nucleotide variant and therefore not expected to be ranked by E/A ("Neg-other mechanism"), totaling to almost 89%. Prioritization was deemed nonconcordant when a different candidate variant was highly ranked ("Not consistent") or when the top candidate was not ranked/very low ranked ("Missed"), totaling to about 8%. Finally, approximately 3% had a new strong candidate variant prioritized by E/A that was missed by expert review. **B**. The distribution of E/A scores is shown for cases with known diagnosis at enrollment ("PD") and new diagnosis ("ND"). Exomiser scores range from 0 to 1, with 1 being the highest/best match. AMELIE scores range from 0 to 100, with 100 being the highest/best match. Median is shown to indicate the shift in mean due to a minority of missed rankings (when diagnostic variant was not ranked the lowest score was used). E/A, Exomiser/AMELIE; ND, new diagnosis; Neg, negative; PD, prior diagnosis.

Table 1 (and shown in Supplemental Figures 1-8). We note that 9 of 64 previously tested cases were diagnosed by analyses of GS when ES analyses were negative, suggesting that among cohorts of ES-tested patients the contribution of GS can be >10% of achievable diagnoses. Among previously untested patients (no testing or testing initiated in parallel), GS was required to solve pathogenic variation not detected by ES in 7 of 105 (6.7%) patients because of intronic variation, small deletions difficult to detect with ES, repeat expansion disorders, and disease associated noncoding RNAs not covered in the exome capture.

Effect of GS platforms

GS contributed 6% to 14% of diagnoses (see previous sections). The different platforms assessed in our study exhibited distinct characteristics that can contribute to individual variant types and overall potential for augmentation of ES. We examined 3 srGS platforms: 10x Linked-Read sequencing (10x Genomics, n = 542/587 patients), DNA NanoBall sequencing (MGI, n = 74/180 patients), and PCR-free srGS (Illumina, n = 683/1660 patients), along with a subset of samples assessed by HiFi-GS (PacBio,

Table 1 Example cases for which diagnosis was initially missed and subsequently solved through research analysis

Case	Phenotype	Previous Clinical Testing	Previous Result	Research Test/ Analysis	Diagnostic Finding (GRCh38)	Inheritance	Barrier Overcome by Research Methods
New diagnosis upon reanalysis							
239/240 (Supplemental Figure 1)	Lipodystrophy	Microarray, ES	Neg	10x linked-read GS, srGS	*MFN2* (NM_014874.3): c.2119C>T, (p.Arg707Trp) (homozygous)	AR	Reanalysis revealed atypical disease presentation
272	Narrow chest, small stature, macrocephaly, tall forehead, high palate	Skeletal ciliopathies panel, ES	*HUWEI*	ES, srGS, scRNA	*HUWEI* (NM_031407.5): c.647C>T, (p.Thr216Ile); *MAP3K7* (NM_145331.2): c.745C>T, (p.Pro249Ser)	De novo	Research uncovered second diagnosis (*MAP3K7*), which was not reported by ES in commercial laboratory (reason unknown)
453 (Supplemental Figure 2)	Congenital myotonic dystrophy	*CLCN1*, *DMPK*, *SCN4A* seq, and *DMPK* expansion	Neg	ES, 10x linked-read GS, srGS	*SCN4A* (NM_000334.4): c.4342C>T, (p.Arg1448Cys)	AD (nk)	Not reported by commercial clinical laboratory because of low coverage cutoff
953/954	Precocious puberty, epilepsy, DD	Brain malformation panel, ES	Neg	ES, srGS, GBS	*PTEN* (NM_000314.4): c.269T>C, (p.Phe90Ser)	AD (pat)	Not reported by commercial clinical laboratory because of atypical phenotype
Not detected in previous testing: technology and/or analysis limitations							
110/111 (Figure 3D-F)	Septo-optic dysplasia, hypotonia, strabismus, tremor, DD	Microarray, *HESX1* seq del/dup, *ALSM1* seq, neuro-muscular panel	Neg	ES, srGS, HiFi-GS, GBS	*AARS2* (NM_020745.4): c.595C>T (p.Arg199Cys); 6p21.1(44306618_44310699)x1	AR	Deletion of exons 5 to 7 was difficult to detect; *AARS2*-related disease reported after clinical testing was completed
129 (Supplemental Figure 3A-C)	Profound congenital hypotonia, motor deficits, cerebral visual impairment	Chromosomes, microarray, *DMPK* expansion, neuromuscular panel	Neg	ES, srGS (blood and muscle)	*TBCK* (ENST00000394708.2): c.1039C>T, (p.Arg347Ter)/ c.2060-6793_2235+426del, (p.Glu687Valfs*8)	AR	Single exon deletion in setting of large intronic regions difficult to detect using ES
189-190, 192-193 (Figure 3A-C)	Global DD, dystonia	Microarray, exon array, ES	Neg	ES, srGS, GBS, HiFi-GS	*STARD7*: triplet expansion	AD (pat)	Novel expansion disorder (solved by lrGS)
302 (Supplemental Figure 4)	Autoimmune hypothyroidism, autoimmune neutropenia, immunodeficiency (unknown type, low B-cells)	Microarray, ES	*SPECC1L*	ES, srGS, scRNA	*SPECC1L:* (NM_015330.6): c.1900C>T, (p.Arg634Ter) & *RNU4ATAC* (NR_023343.1): n.37G>A/n.8C>T	AD (pat) and AR	Research uncovered second diagnosis, missed in clinical ES owing to no coverage (noncoding RNA not covered in most ES)

(continued)

Table 1 Continued

Case	Phenotype	Previous Clinical Testing	Previous Result	Research Test/ Analysis	Diagnostic Finding (GRCh38)	Inheritance	Barrier Overcome by Research Methods
305 (Supplemental Figure 3D-E)	Generalized hypotonia, global DD, infantile spasms	Microarray, ES	VUS	10x linked-read GS, srGS	*TBCK* (NM_001163435.3): c.2060-6793_2235+426del, (p.Glu687Valfs*8) homozygous	AR	Single exon deletion in setting of large intronic regions difficult to detect with ES
397/398 (Supplemental Figure 5)	Becker muscular dystrophy	Microarray, ES	Neg	RNAseq (external collaboration)	*DMD* (NM_004006.3): c.6290+3076A>G, (p.Thr3055Serfs*1)	XL (mat)	Deep intronic variant, required functional RNAseq on muscle biopsy to identify the creation of pseudoexon
451 (Supplemetal Figure 6)	Multiple congenital anomalies (including severe heart malformations), slow growth, DD	Microarray	1.73 Mb dup 1q21.1q21.2	ES, 10xlinked-read GS, srGS	*GATA4* (NM_002052.3): c.886G>A, (p.Gly296Ser)	AD (mat)	Research uncovered a second unexpected diagnosis by automated variant prioritization, which was clinically relevant
678 (Supplemental Figure 7)	Lissencephaly	Lissencephaly panel	Neg	10x linked-read, HiFi-GS	*CEP85L* (NM_001042475.2): c.3G>T, (p.Met1?)	AD (nk)	Novel gene not included in panel testing (and poor coverage of exon 1 in 10x GS)
791 (Supplemental Figure 8)	Hypotonia, persistent global DD, epilepsy	Microarray, ES	AOH region 6q15; *HEXB* carrier status	srGS	*CACNA1A* (NM_001127221.1) 19p13.13(13332562-13336361)x1	AD (not mat)	Deletion of exons 7 to 9; CNV analysis of ES analysis not completed clinically
799	Global DD, language delays, hypotonia	None	n/a	ES, srGS	*SHANK3* (ENST00000262795.3) 22q13.33(50690814-50780545)x1	AD (not mat)	Intronic breakpoints detected using GS, CNV analysis not completed using ES

AD, autosomal dominant; *AR*, autosomal recessive; *CNV*, copy number variation; *DD*, developmental delay; *del,* deletion; *dup*, duplication; *ES*, exome sequencing; *GBS*, genome bisulfite sequencing; *GS*, genome sequencing; *HiFi-GS,* HiFi long-read genome sequencing; *lr*, long read; *mat*, maternally inherited; *n/a*, not applicable/not available; *Neg*, negative; *nk*, inheritance not known; *pat*, paternally inherited; *scRNA*, single-cell RNA expression analysis; *seq*, sequencing; *sr*, short read; *VUS*, variant of uncertain significance; *XL*, X-linked.

$n = 274/472$ patients; Supplemental Materials and Methods and Supplemental Table 6). The 10x Linked-Read sequencing exhibited inconsistent coverage across the genome, which resulted in suboptimal variant sensitivity (97.8% mean sensitivity), and was discontinued in favor of other GS platforms that performed similarly (>98.3% sensitivity, >98.8% specificity) against Infinium Global Screening microarray genotypes (Supplemental Figure 9). Considering the moderate increase in diagnostic yield using srGS, lrGS using PacBio HiFi reads was systematically deployed for negative trios, allowing for a thorough comparison of HiFi-GS with srGS with a particular focus on the potential for rare disease variant discovery while controlling for false positives using parental samples. Direct comparison of overall SNV calls and SV calls indicated an approximately 5% increase in SNV called from high coverage lrGS (25× HiFi-GS) vs srGS (35× Illumina GS), with a much more dramatic effect on SV detection with nearly double the discovery rate with lrGS (Supplemental Table 6).

To gauge the effect on potential rare disease SNV alleles, we compared a subset of probands ($n = 102$) using both srGS and HiFi-GS, focusing on rare coding variants. On average, there were 439 coding variants per proband genome, of which 14% were unique to HiFi-GS, in contrast to 6% unique rare coding variants in srGS. Of these variants, transmission (variant detected in parent) supported nearly all (98%) variants observed through both srGS and HiFi-GS, whereas 40% of HiFi-GS–specific variants appeared transmitted and 20% of srGS-specific variants showed evidence of transmission. Extrapolating true positive rates per genome and per technology on the basis of transmission suggested that on average lrGS exclusively detects 31 coding variants and srGS detects 6 coding variants per genome (Supplemental Table 7). More striking differences were observed for family-transmitted rare SVs (minor allele frequency < 1%) generated at our center in either srGS or HiFi-GS data and not seen in publicly available reference data, including database of genomic variants for srGS, Human Pangenome Reference Consortium HiFi-GS, or variants published from ONT-lrGS by Decode for lrGS.[18,31,32] On average, 70 transmitted rare SVs are observed in srGS data and >300 in HiFi-GS data: a greater than 4-fold difference. The discovery advantage for HiFi-GS also applies for transmitted rare coding SVs (Table 2). Similar to earlier reports, the rate of de novo SVs is low and only 2 (noncoding) examples were found in the manual curation of 8 high coverage HiFi-GS trios (Supplemental Table 8).[33]

Enabling rare disease allele discovery by HiFi-GS

A tangible consequence of higher discovery rate of variant detection using HiFi-GS was the detection of 4,369,149 recurrent (observed in at least 2 unrelated individuals) SNVs not reported in Genome Aggregation Database, as well as 115,595 recurrent SVs detected in our aggregated HiFi-GS resource (30,707 not seen in any previously published data sets).[34] These findings serve as a reminder that publicly available data sets remain highly incomplete. To enable new rare disease discovery efforts using HiFi-GS, we are sharing these recurrent variants and their frequencies derived from >1000 alleles of HiFi-GS data (https://github.com/ChildrensMercyResearchInstitute/GA4K). As anticipated, the recurrent variants detected in HiFi-GS were biased to regions with poor srGS resolution (eg, segmental duplications and satellite repeats), but recurrent SVs not in database of genomic variants were widely dispersed across genic regions, and >800 OMIM loci also showed higher than GENCODE average rate of HiFi-GS–specific SNVs (Supplemental Tables 9 and 10).

The current diagnostic evaluation for rare disease relies on a multitude of genome-wide tests (ES, GS, microarray, chromosomes) as well as specialized directed tests (for repeat expansions, methylation defects, etc.). We explored the potential of HiFi-GS to consolidate some of this testing and therefore reduce costs in the diagnostic odyssey for each proband. Developing a toolkit for HiFi-GS in rare disease included the accommodation of specific queries for known repeat expansion loci (Supplemental Table 19). Among our cohort, in which each sample had a minimum of 8× HiFi-GS coverage across 51 loci, we identified 3 pathogenic events (1 *FMR1*, not shown, and 2 *STARD7* expansions, shown in Figure 3A-C). In addition, although not specifically explored, there are known disease genes among the loci with an excess of non–Genome Aggregation Database variation (see previous sections) such as *OTOA* and *STRC*, which are challenging to test owing to known pseudogenes/duplications (Supplemental

Table 2 Structural variation

	Average Proband Counts—Illumina/MGI srGS (49 Trios), >30× Coverage							Average Proband Counts—PacBio HiFi-GS (81 Trios) >25× (Proband) >10× (Parents)						
Structural Variation Group	TOT	BND	CNV	DEL	DUP	INS	INV	TOT	BND	CNV	DEL	DUP	INS	INV
All	11,036	2046	n/a	4299	393	4299	n/a	22,013	52	5	9104	412	12,354	86
Rare	260	98	n/a	43	10	108	n/a	398	4	1	160	15	217	2
Family validated	9127	1537	n/a	3876	339	3375	n/a	21,114	45	5	8768	390	11,824	81
Rare family-validated	69	24	n/a	19	4	22	n/a	332	3	1	136	12	179	2
Rare family-validated coding	20	6	n/a	6	1	7	n/a	119	1	0	46	4	67	1

BND, break-end; *CNV*, copy number variation; *DEL*, deletion; *DUP*, duplication; *HiFi-GS*, HiFi long-read genome sequencing; *INS*, insertion; *INV*, inversion; *n/a*, not available/applicable; *srGS*, short-read genome sequencing ; *TOT*, total.

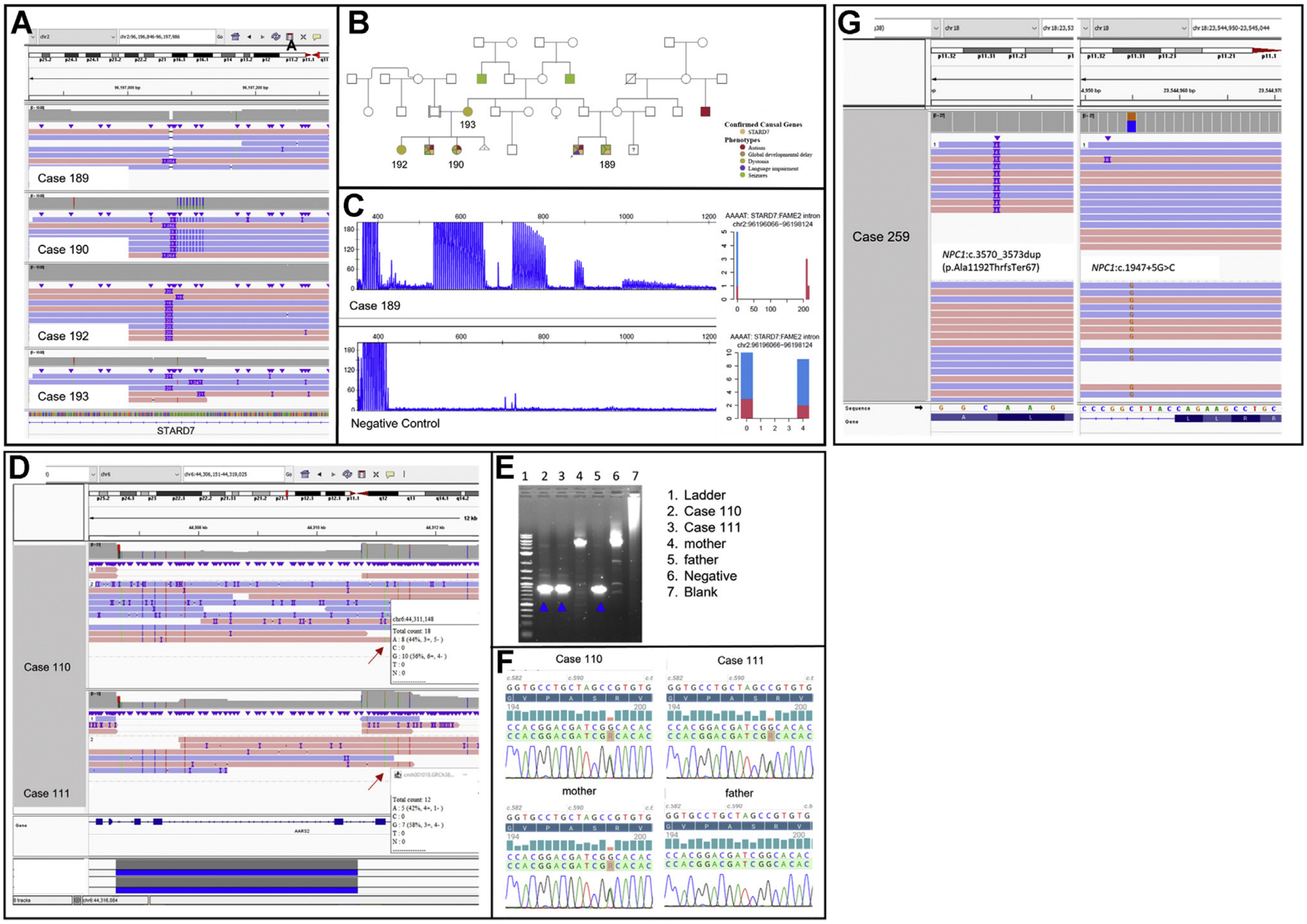

Figure 3 Examples of cases solved by HiFi long-read genome sequencing (HiFi-GS). Long-read genome sequencing addresses challenges in short-read genome sequencing as exemplified by 3 cases. **A**. HiFi-GS identified a novel pentamer expansion in *STARD7*, previously associated with familial adult myoclonic epilepsy, 2 in an extended family. **B**. Pedigree of family with *STARD7* disease, case 193 had adult-onset dystonia, whereas cases 189, 190, and 192 had childhood onset of disease, consistent with anticipation. **C**. Repeat primed-polymerase chain reaction (PCR) confirmed the expansion detected in the HiFi-GS in case 189, which was also detected by the tandem genotyping tool. The negative control had a normal repeat pattern. **D**. Affected siblings cases 110 and 111 were found to be compound heterozygous for 2 pathogenic variants in *AARS2*: NM_020745.4: maternally inherited, c.595C>T (p.Arg199Cys) and a paternally inherited deletion, chr6:44306625-44310745 encompassing exons 5 to 7 of *AARS2*. **E**. Clinical confirmation of the deletion using long-read PCR detected the deletion (arrow) and normal allele in both siblings and unaffected father. **F**. Clinical Sanger confirmation of the maternally inherited c.595C>T (p.Arg199Cys) variant. **G**. Case 259 was clinically diagnosed with Niemann-Pick disease, but parents were unavailable for phasing. HiFi-GS confirmed the pathogenic variants were in *trans*, consistent with autosomal recessive disease. *NPC1*: c.3570_3573dupACTT (p.Ala1192Thrfs*67) (left)/ c.1947+5G>C (right).

Table 9). We note that the current alignment/variant calling pipeline for HiFi-GS also generates phased haplotypes that allow detection of compound heterozygotes even in the case of singletons (Figure 3D-G), with an average phase block of 400 kb (Supplemental Figure 10). Finally, the combination of SV calls and personal assemblies allowed the identification of HiFi-GS signatures for large CNVs clinically detected using microarray (Supplemental Table 12). Furthermore, the implementation of personal assembly data can add bp-level resolution for complex rearrangements interpreted as balanced using cytogenetic assays owing to resolution limitations (Supplemental Figure 11).

New candidate genes after reanalysis across all data and variant prioritization

The joint sequencing results and automated prioritization were reviewed by an expert analysist (genetic counselor or clinical laboratory director) to identify a large fraction of patients (58%) with potential new disease genes. Compelling candidates were systematically submitted to GM.[5] At the time of manuscript submission, 152 candidate genes were active in GM, 12 of which were identified in >1 unrelated family and 6 of which were recently published or were close to publication and therefore in transition from

GUS to diagnostic. More than 36% of submitted GUS had >10 hits in GM, suggesting that they are strong candidates as we had previously reported.[30] Collectively, this underscores the imperative for data sharing and collaboration in rare disease research and diagnosis.

Individual data sharing to enhance variant and gene discovery

Uniform research consents permit sharing of sequences and structured phenotypic data with other rare disease investigators to enhance gene matching beyond the variation submitted to GM. Raw data submitted to database of Genotypes and Phenotypes (phs002206.v2.p1) will allow for joint calling with other available rare disease data sets. Access to processed data for rare variants, de-identified pedigrees, and coded phenotypes will be available to registered users through a cloud-hosted PhenoTips web user interface: https://phenotips-ga4k.cmh.edu (access inquiries for investigators GA4k@cmh.edu) (Supplemental Figure 12).[23] This web user interface provides a simple interface for users to review participant data, identify cohorts of participants on the basis of phenotypic or genotypic attributes, and review rare variants in the context of a specific phenotype. Furthermore, this interface will continue to be dynamically synchronized with the GA4K program, and it already included >1000 additional cases in various stages of ongoing analyses at the time of manuscript submission for a total of 5922 individuals across 2537 families and processed variants for 2069 patients.

Discussion

We developed a comprehensive rare disease phenotype–genotype data repository across a large pediatric health care system in the GA4K program. Full access is provided to enable medical genomic testing, complete annotation for reanalysis, and use by contemporary research genomic tools. Using multiple sequencing methods and analytic approaches, the first 1083 patients evaluated serve as a roadmap to improve rare disease diagnostics and as a catalog of case data for utility in biomedical discovery.

We combined publicly available machine-learning approaches E/A for variant and disease gene prioritization at scale in which the nominated candidate variant was ranked by E/A in 539 (49.8%) cases, supporting the use of machine-learning tools as a first-pass, resource-saving analysis. Primarily retrospective studies suggested higher rates of relevant results, and we replicated similar success to these studies in our observed concordance among previously diagnosed cases.[7,27,35] Importantly, the vast majority of our patients were undiagnosed when entering the study. This allowed us to establish the utility of computationally assisted interpretation among prospective patients with diverse rare disease on a scale far beyond any previously assessed rare disease cohort.[9,35] We also showcase patients having a strong candidate or diagnostic variant identified through machine-learning ranking (subsequently confirmed through expert review) that may not have otherwise been prioritized for further investigation owing to combined supporting data being pulled by artificial intelligence from multiple sources and not easily digested by manual analysis in a timely manner, as expected in a clinical setting (ie, not an obvious candidate that would arise from easily checked metrics such as gene constraint and protein function). This supports the utility of the approach not only for diagnostic evaluation but also as a systematic source for generating hypotheses on disease gene discovery. Importantly, prioritization is still biased given that it will inevitably rank genes that have more linked resources (be it clinical, functional, or otherwise) higher than poorly characterized genes, and therefore, genetic prioritization independent of literature mining remains important for gene discovery.[36]

We showed that there is diagnostic utility in ES reanalyses and/or repeat ES to improve coverage; however, >10% diagnoses that we made in previously negative ES cases were solved with elevation to GS, which, unlike most ES analyses, included systematic CNV calling. As expected, the utility of GS was lower in previously unassessed cases; however, even in this group 1 in 20 diagnoses required GS. Similar to previously unsolved cases, GS contributed primarily to the detection of SVs. Given the known benefits of HiFi-GS in SV detection, we pursued HiFi-GS in unsolved rare diseases beyond earlier demonstration studies as routine streamlining of trios.[18,19] Early results from HiFi-GS showed the expected improvement in detection rates for SVs but also provided first glimpses of diagnostic variation currently only achievable through HiFi-GS, such as the discovery of novel repeat expansions (including repeat size and sequence composition), the solving of CNV breakpoints and orientation/localization, and the resolution of phase in the absence of parental samples. The potential for having full genome analyses by HiFi-GS was explored in this study as proof of concept; further work will elaborate underexplored areas of HiFi-GS utility, such as personal assemblies, haplotype-phasing, and directed work on duplicated gene regions. In the meantime, our HiFi-GS variant catalogs extending across hundreds of individuals provide the first building blocks for using alternative GS methods in clinical settings and particularly for unsolved diseases.

Finally, most unsolved cases in our cohort do have candidate genes and variants but lack sufficient evidence to assign pathogenicity owing to a lack of replication (also known as the “n of 1” problem), with hundreds of genes and variants currently followed through GM. Greater data sharing is paramount for enhancing benefits to participants and advancing scientific progress, along with maximizing the utility of genomic data.[37] Unfortunately, hesitancy toward extensive data sharing persists among investigators because of reasons that include the arduous processes required for data sharing, concerns about participant privacy, and fear of loss of priority in data publication.[37,38] Our study follows regulations and considers recommendations for responsible sharing of pediatric genomic data to support the benefits of data sharing to research participants and patients while protecting privacy.[37]

Data Availability

Processed data for rare variants, de-identified pedigrees, and coded phenotypes are available to registered users through a cloud-hosted PhenoTips web user interface (https://phenotips-ga4k.cmh.edu/). Access inquiries for investigators should be directed to GA4k@cmh.edu (including key to correlate study numbers used in this manuscript). Recurrent variants and their frequencies derived from >1000 alleles of HiFi-GS data are available at: https://github.com/ChildrensMercyResearchInstitute/GA4K

Acknowledgments

We would like to thank the families for participating in our study. This work was made possible by the generous gifts to Children's Mercy Research Institute and Genomic Answers for Kids program at Children's Mercy Kansas City.

Author Information

Conceptualization: T. Curran, T. Pastinen.; Formal Analysis: A.S.A. Cohen, E.G. Farrow, J.T. Alaimo, C.J. Saunders, I. Thiffault, T. Pastinen; Funding Acquisition: T. Curran, T. Pastinen; Investigation: A.S.A. Cohen, E.G. Farrow, A.T. Abdelmoity, J.T. Alaimo, S.M. Amudhavalli, J.T. Anderson, L. Bansal, L. Bartik, B. Belden, C.D. Berrios, R.L. Biswell, W.A. Cheung, K.A. Coffman, A.M. Cooper, L.A. Cross, T. Curran, T.T.T. Dang, M.M. Elfrink, K.L. Engleman, E.D. Fecske, C. Fieser, K. Fitzgerald, E.A. Fleming, R.N. Gadea, J.L. Gannon, R.N. Gelineau-Morel, M. Gibson, J. Goldstein, E. Grundberg, K. Halpin, B.S. Harvey, B.A. Hesse, W. Hein, S.M. Herd, S.S. Hughes, M. Ilyas J. Jacobson, J.L. Jenkins, S. Jiang, J.J. Johnston, K. Keeler, J. Kussman, C. Lawsont, J.-B. Le Pichon, J.S. Leeder, V.C. Little, D.A. Louiselle, M. Lypka, B.D. McDonald, N. Miller, A. Modrcin, A. Nair, S.H. Neal, C.M. Oermann, D.M. Pacicca, K. Pawar, N.L. Posey, N. Price, L.M.B. Puckett, J.F. Quezada, N. Raje, E.T. Rrush, V. Sampath, C.J. Saunders, C. Schwager, R.M. Schwend, E. Shaffer, C. Smail, S. Soden, M.E. Strenk, B.R. Sullivan, B.R. Sweeney, J.B. Tam-Williams, A.M. Walter, H. Welsh, L.K. Willig Y. Yan, S.T. Younger, D. Zhou, T.N. Zion, I. Thiffault, T. Pastinen; Methodology: A.S.A. Cohen, E.G. Farrow, P. Baybayan, B. Belden, C.D. Berrios, S. Chakraborty, W.A. Cheung, T. Curran, M.M. Elfrink, M. Gibson, E. Grundberg, S.M. Herd, J.J. Johnston, J. Korlach, C. Lambert, S. Leeder, B.D. McDonald, N. Miller, A. Nair, S.H. Neal, N.L. Posey, L.M.B. Puckett, W.J. Rowell, C. Saunders, A.M. Walter, A.M. Wenger, S.T. Younger, T.N. Zion, I. Thiffault, T. Pastinen; Software: P. Buczkowicz, O. Buske; Writing-original draft: A.S.A. Cohen, E.G. Farrow, I. Thiffault, T. Pastinen; Writing-review and editing: A.S.A. Cohen, E.G. Farrow, A.T. Abdelmoity, J.T. Alaimo, S.M. Amudhavalli, J.T. Anderson, L. Bansal, L. Bartik, P. Baybayan, B. Belden, C.D. Berrios, R.L. Biswell, P. Buczkowica, O. Buske, S. Chakraborty, W.A. Cheung, K.A. Coffman, A.M. Copper, L.A. Cross, T. Curran, T.T.T. Dang, M.M. Elfrink, K.L. Engleman, E.D. Fecske, C. Fraiser, K. Fitzgerald, E.A. Fleming, R.N. Gadea, J.L. Gannon, R.N. Gelineau-Morel, M. Gibson, J. Goldstein, E. Grundberg, K. Halpin, B.S. Harvey, B.A. Heese, W. Hein, S.M. Herd, S.S. Hughes, M. Ilyas, J. Jacobson, J.L. Jenkins, S. Jiang, J.J. Johnston, K. Keeler, J. Korlach, J. Kuussman, C. Lambert, C. Lawson, J.-B. Le Pichon, J.S. Leeder, V.C. Little, D.A. Louiselle, M. Lypka, B.D. McDonald, N. Miller, A. Modrcin, A. Nair, S.H. Neal, C.M. Oermann, D.M. Pacicca, K. Pawar, N.L. Posey, N. Price, L.M.B. Puckett, J.F. Quezada, N. Raje, W.J. Rowell, E.T. Rush, V. Sampath, C.J. Saunders, C. Schwager, R.M. Schwend, E. Shaffer, C. Smail, S. Soden, M.E. Strenk, B.R. Sullivan, B.R. Sweeney, J.B. Tam-Williams, A.M. Walter, H. Welsh, A.M. Wenger, L.K. Willig, Y. Yan, S.T. Younger, D. Zhou, T.N. Zion, I. Thiffault, T. Pastinen.

Ethics Declaration

All studies were approved by the Children's Mercy Institutional Review Board (study # 11120514). Informed written consent was obtained from all participants before study inclusion.

Conflict of Interest

P. Baybayan, S. Chakraborty, J. Korlach, C. Lambert, W.J. Rowell, and A.M. Wenger are employees and shareholders of Pacific Biosciences. P. Buczkowicz and O. Buske are employees of PhenoTips. N. Miller became an employee of Bionano Genomics after contribution to the work described in this manuscript. All other authors declare no conflicts of interest.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gim.2022.02.007) contains supplementary material, which is available to authorized users.

Authors

Ana S.A. Cohen[1,2,3], Emily G. Farrow[1,3,4], Ahmed T. Abdelmoity[4], Joseph T. Alaimo[2,3], Shivarajan M. Amudhavalli[3,5], John T. Anderson[6], Lalit Bansal[4], Lauren Bartik[3,5], Primo Baybayan[7], Bradley Belden[1], Courtney D. Berrios[1], Rebecca L. Biswell[1], Pawel Buczkowicz[8], Orion Buske[8], Shreyasee Chakraborty[7], Warren A. Cheung[1], Keith A. Coffman[4], Ashley M. Cooper[4], Laura A. Cross[5], Tom Curran[9], Thuy Tien T. Dang[4], Mary M. Elfrink[1], Kendra L. Engleman[5], Erin D. Fecske[4], Cynthia Fieser[4], Keely Fitzgerald[4], Emily A. Fleming[5], Randi N. Gadea[5], Jennifer L. Gannon[5], Rose N. Gelineau-Morel[3,4], Margaret Gibson[1], Jeffrey Goldstein[4], Elin Grundberg[1], Kelsee Halpin[3,4], Brian S. Harvey[6], Bryce A. Heese[5], Wendy Hein[4], Suzanne M. Herd[1], Susan S. Hughes[5], Mohammed Ilyas[3,4], Jill Jacobson[3,4], Janda L. Jenkins[5], Shao Jiang[10], Jeffrey J. Johnston[1], Kathryn Keeler[6], Jonas Korlach[7], Jennifer Kussmann[5], Christine Lambert[7], Caitlin Lawson[5], Jean-Baptiste Le Pichon[4], James Steven Leeder[1], Vicki C. Little[4], Daniel A. Louiselle[1], Michael Lypka[10], Brittany D. McDonald[1], Neil Miller[1,3,11], Ann Modrcin[4], Annapoorna Nair[1], Shelby H. Neal[1], Christopher M. Oermann[4], Donna M. Pacicca[6], Kailash Pawar[4], Nyshele L. Posey[1], Nigel Price[6], Laura M.B. Puckett[1], Julio F. Quezada[3,4], Nikita Raje[3,12], William J. Rowell[7], Eric T. Rush[3,5,13], Venkatesh Sampath[14], Carol J. Saunders[1,2,3], Caitlin Schwager[5], Richard M. Schwend[6], Elizabeth Shaffer[4], Craig Smail[1], Sarah Soden[4], Meghan E. Strenk[5], Bonnie R. Sullivan[5], Brooke R. Sweeney[3,4], Jade B. Tam-Williams[4], Adam M. Walter[1], Holly Welsh[5], Aaron M. Wenger[7], Laurel K. Willig[4], Yun Yan[3,4], Scott T. Younger[1], Dihong Zhou[5], Tricia N. Zion[1,3,4,5], Isabelle Thiffault[1,2,3,*], Tomi Pastinen[1,3,9,*]

Affiliations

[1]Genomic Medicine Center, Children’s Mercy Kansas City, Kansas City, MO; [2]Department of Pathology and Laboratory Medicine, Children’s Mercy Kansas City, Kansas City, MO; [3]UKMC School of Medicine, University of Missouri Kansas City, Kansas City, MO; [4]Department of Pediatrics, Children’s Mercy Kansas City, Kansas City, MO; [5]Division of Genetics, Children’s Mercy Kansas City, Kansas City, MO; [6]Department of Orthopaedic Surgery, Children’s Mercy Kansas City, Kansas City, MO; [7]Pacific Biosciences of California, Inc, Menlo Park, CA; [8]PhenoTips, Toronto, Canada; [9]Children’s Mercy Research Institute, Kansas City, MO; [10]Bionano Genomics, Inc, San Diego, CA; [11]Division of Allergy Immunology Pulmonary and Sleep Medicine, Children’s Mercy Kansas City, Kansas City, MO; [12]Division of Neonatology, Children’s Mercy Kansas City, Kansas City, MO; [13]Department of Internal Medicine, University of Kansas School of Medicine, Kansas City, MO; [14]Division of Neonatology, Children's Mercy Hospital Kansas City, Kansas City, MO

References

1. Bruel AL, Nambot S, Quéré V, et al. Increased diagnostic and new genes identification outcome using research reanalysis of singleton exome sequencing. *Eur J Hum Genet*. 2019;27(10):1519–1531. http://doi.org/10.1038/s41431-019-0442-1.
2. Costain G, Jobling R, Walker S, et al. Periodic reanalysis of whole-genome sequencing data enhances the diagnostic advantage over standard clinical genetic testing. *Eur J Hum Genet*. 2018;26(5):740–744. http://doi.org/10.1038/s41431-018-0114-6.
3. Liu P, Meng L, Normand EA, et al. Reanalysis of clinical exome sequencing data. *N Engl J Med*. 2019;380(25):2478–2480. http://doi.org/10.1056/NEJMc1812033.
4. Tan NB, Stapleton R, Stark Z, et al. Evaluating systematic reanalysis of clinical genomic data in rare disease from single center experience and literature review. *Mol Genet Genomic Med*. 2020;8(11):e1508. http://doi.org/10.1002/mgg3.1508.
5. Sobreira N, Schiettecatte F, Valle D, Hamosh A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat*. 2015;36(10):928–930. http://doi.org/10.1002/humu.22844.
6. Schmitz-Abe K, Li Q, Rosen SM, et al. Unique bioinformatic approach and comprehensive reanalysis improve diagnostic yield of clinical exomes. *Eur J Hum Genet*. 2019;27(9):1398–1405. http://doi.org/10.1038/s41431-019-0401-x.
7. Cipriani V, Pontikos N, Arno G, et al. An improved phenotype-driven tool for rare Mendelian variant prioritization: benchmarking Exomiser on real patient whole-exome data. *Genes (Basel)*. 2020;11(4):460. http://doi.org/10.3390/genes11040460.
8. Krämer A, Shah S, Rebres RA, Tang S, Richards DR. Leveraging network analytics to infer patient syndrome and identify causal genes in rare disease cases. *BMC Genomics*. 2017;18(Suppl 5):551. http://doi.org/10.1186/s12864-017-3910-4.
9. Ji J, Shen L, Bootwalla M, et al. A semiautomated whole-exome sequencing workflow leads to increased diagnostic yield and identification of novel candidate variants. *Cold Spring Harb Mol Case Stud*. 2019;5(2):a003756. http://doi.org/10.1101/mcs.a003756.
10. Wu C, Devkota B, Evans P, et al. Rapid and accurate interpretation of clinical exomes using Phenoxome: a computational phenotype-driven approach. *Eur J Hum Genet*. 2019;27(4):612–620. http://doi.org/10.1038/s41431-018-0328-7.
11. Robinson PN, Ravanmehr V, Jacobsen JOB, et al. Interpretable clinical genomics with a likelihood ratio paradigm. *Am J Hum Genet*. 2020;107(3):403–417. http://doi.org/10.1016/j.ajhg.2020.06.021.
12. Zhao M, Havrilla JM, Fang L, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform*. 2020;2(2):lqaa032. http://doi.org/10.1093/nargab/lqaa032.
13. Köhler S, Gargano M, Matentzoglu N, et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res*. 2021;49(D1):D1207–D1217. http://doi.org/10.1093/nar/gkaa1043.
14. Kobren SN, Baldridge D, Velinder M, et al. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. *Genet Med*. 2021;23(6):1075–1085. http://doi.org/10.1038/s41436-020-01084-8.
15. Stranneheim H, Lagerstedt-Robinson K, Magnusson M, et al. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med*. 2021;13(1):40. http://doi.org/10.1186/s13073-021-00855-5.
16. Thiffault I, Farrow E, Zellmer L, et al. Clinical genome sequencing in an unbiased pediatric cohort. *Genet Med*. 2019;21(2):303–310. http://doi.org/10.1038/s41436-018-0075-8.

17. Li Q, Zhao X, Zhang W, et al. Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. *BMC Genomics*. 2019;20(1):215. http://doi.org/10.1186/s12864-019-5569-5.
18. Ebert P, Audano PA, Zhu Q, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372(6537):eabf7117. http://doi.org/10.1126/science.abf7117.
19. Sabatella M, Mantere T, Waanders E, et al. Optical genome mapping identifies a germline retrotransposon insertion in SMARCB1 in two siblings with atypical teratoid rhabdoid tumors. *J Pathol*. 2021;255(2):202–211. http://doi.org/10.1002/path.5755.
20. Boycott KM, Dyment DA, Innes AM. Unsolved recognizable patterns of human malformation: challenges and opportunities. *Am J Med Genet C Semin Med Genet*. 2018;178(4):382–386. http://doi.org/10.1002/ajmg.c.31665.
21. Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat*. 2013;34(8):1057–1065. http://doi.org/10.1002/humu.22347.
22. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983–987. http://doi.org/10.1038/nbt.4235.
23. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18(2):170–175. http://doi.org/10.1038/s41592-020-01056-5.
24. Mitsuhashi S, Frith MC, Mizuguchi T, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol*. 2019;20(1):58. http://doi.org/10.1186/s13059-019-1667-6.
25. Nurk S, Walenz BP, Rhie A, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res*. 2020;30(9):1291–1305. http://doi.org/10.1101/gr.263566.120.
26. Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and Glnexus. *Bioinformatics*. 2021;36(24):5582–5589. http://doi.org/10.1093/bioinformatics/btaa1081.
27. Birgmeier J, Haeussler M, Deisseroth CA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci Transl Med*. 2020;12(544):eaau9113. http://doi.org/10.1126/scitranslmed.aau9113.
28. Smedley D, Jacobsen JO, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015;10(12):2004–2015. http://doi.org/10.1038/nprot.2015.124.
29. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405–424. http://doi.org/10.1038/gim.2015.30.
30. Thiffault I, Cadieux-Dion M, Farrow E, et al. On the verge of diagnosis: detection, reporting, and investigation of de novo variants in novel genes identified by clinical sequencing. *Hum Mutat*. 2018;39(11):1505–1516. http://doi.org/10.1002/humu.23646.
31. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986–D992. http://doi.org/10.1093/nar/gkt958.
32. Beyter D, Ingimundardottir H, Oddsson A, 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*. 2021;53(6):779–786. http://doi.org/10.1038/s41588-021-00865-4.
33. Kloosterman WP, Francioli LC, Hormozdiari F, et al. Characteristics of de novo structural changes in the human genome. *Genome Res*. 2015;25(6):792–801. http://doi.org/10.1101/gr.185041.114.
34. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–443. Published correction appears in *Nature*. 2021;590(7846):E53. Published correction appears in *Nature*. 2021;597(7874):E3-E4. https://doi.org/10.1038/s41586-020-2308-7.
35. Bone WP, Washington NL, Buske OJ, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet Med*. 2016;18(6):608–617. http://doi.org/10.1038/gim.2015.137.
36. Cornish AJ, David A, Sternberg MJE. PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics*. 2018;34(12):2087–2095. http://doi.org/10.1093/bioinformatics/bty028.
37. Rahimzadeh V, Knoppers BM, Bartlett G. Ethical, legal, and social issues (ELSI) of responsible data sharing involving children in genomics: a systematic literature review of reasons. *AJOB Empir Bioeth*. 2020;11(4):233–245. http://doi.org/10.1080/23294515.2020.1818875.
38. Nick HP, Kehoe K, Gammon A, Contreras JL, Kaphingst KA. Researcher knowledge, attitudes, and communication practices for genomic data sharing. *J Empir Res Hum Res Ethics*. 2021;16(1-2):125–137. http://doi.org/10.1177/1556264620969301.