Unique Capabilities of Genome Sequencing for Rare Disease Diagnosis

Monica H Wojcik, MD, MPH[1-3]; Gabrielle Lemire, MD[3]; Maha S Zaki, MD, PhD[4]; Mariel Wissman, BS[5-7]; Wathone Win, BS[2]; Sue White, MD[8,9]; Ben Weisburd, BS[3]; Leigh B Waddell, PhD[10,11]; Jeffrey M Verboon, MS[5-7]; Grace E. VanNoy, MS[3]; Ana Töpf, PhD[12];Tiong Yang Tan, MBBS, PhD[13-15]; Volker Straub, MD, PhD[12]; Sarah L Stenton, MD, PhD[2,3]; Hana Snow, BSE[3]; Moriel Singer-Berk, MS[3]; Josh Silver, MS[16,17]; Shirlee Shril, MS[18]; Eleanor G Seaby, MD[3]; Ronen Schneider, MD[18]; Vijay G Sankaran, MD, PhD[7, 19-21]; Alba Sanchis-Juan, PhD[7,22-24]; Kathryn A Russell, BS[3]; Karit Reinson, MD, PhD[25,26]; Gianina Ravenscroft, PhD[27]; Eric A Pierce, MD, PhD[28]; Emily M Place, MS[28]; Sander Pajusalu, MD, PhD[25,26]; Lynn Pais, MS[3]; Katrin, Õunap, MD, PhD[25,26]; Ikeoluwa Osei-Owusu, PhD[3]; Volkan Okur, MD[29]; Kaisa Teele Oja[25,26]; Melanie O'Leary, MS[3]; Emily O'Heir, BS[3]; Chantal Morel, MD[16,30]; Rhett G Marchant, BSc [10,11,31]; Brian E Mangilog, BA[3]; Jill A Madden, PhD[2]; Daniel MacArthur, PhD[3]; Alysia Lovgren, PhD[3]; Jordan P Lerner-Ellis, PhD[32-34]; Jasmine Lin, MS[2]; Nigel Laing, PhD[27]; Friedhelm Hildebrandt, MD[18]; Emily Groopman, MD, PhD[3]; Julia Goodrich, PhD[3]; Joseph G Gleeson, MD[35]; Roula Ghaoui, PhD[36-38]; Casie A Genetti, MS[2]; Hanna T Gazda, MD, PhD[2]; Vijay S. Ganesh[3,39]; Mythily Ganapathy, PhD[40]; Lyndon Gallacher, MGC[41]; Jack Fu, PhD[7,22-24]; Emily Evangelista, MS[3]; Eleina England, MS[3]; Sandra Donkervoort, MS[42]; Stephanie DiTroia, PhD[3]; Sandra T Cooper, PhD[10,11,31]; Wendy K Chung, MD, PhD[18]; John Christodoulou[43]; Katherine R Chao, BS[3]; Liam D Cato, MD[7, 19-21]; Kinga M Bujakowska, PhD[28]; Samantha J Bryen, PhD[10,11,31]; Harrison Brand, PhD[7,22-24]; Carsten Bonnemann, MD[42]; Alan H Beggs, PhD PhD[2]; Samantha M Baxter[3]; Pankaj B Agrawal, MD, MMSC[3,44]; Michael Talkowski, PhD[7,22-24, 45]; Chrissy Austin-Tse, PhD[3]; Heidi L Rehm, PhD[3,22,23]; Anne O'Donnell-Luria, MD, PhD[2,3,22,23].

1. Divisions of Newborn Medicine and Genetics and Genomics, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
2. Manton Center for Orphan Disease Research, Division of Genetics and Genomics, Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA
3. Broad Center for Mendelian Genomics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
4. Clinical Genetics Department, Human Genetics and Genome Research Institute, National Research Centre, Cairo, Egypt
5. Division of Hematology and Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
6. Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA
7. Broad Institute of MIT and Harvard, Cambridge, Massachusetts , USA.
8. Victorian Clinical Genetics Service, Murdoch Children's Research Institute, Parkville, Australia
9. Department of Pediatrics, University of Melbourne, Melbourne, Australia
10. Kids Neuroscience Centre, Kids Research, Children's Hospital at Westmead, Westmead, New South Wales, Australia
11. Discipline of Child and Adolescent Health, Faculty of Medicine and Health, The University of Sydney, Westmead, New South Wales, Australia
12. John Walton Muscular Dystrophy Research Centre, Clinical and Translational Research Institute, Newcastle University and NHS Trust, Newcastle upon Tyne, United Kingdom
13. Victorian Clinical Genetics Services, Melbourne, Australia
14. Murdoch Children's Research Institute, Melbourne, Australia
15. Department of Pediatrics, University of Melbourne, Melbourne, Australia
16. Fred A. Litwin Family Centre in Genetic Medicine, University Health Network, Toronto, Ontario, Canada

17. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
18. Department of Pediatrics, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA
19. Division of Hematology and Oncology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA
20. Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA
21. Harvard Stem Cell Institute, Cambridge, Massachusetts, USA
22. Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
23. Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
24. Department of Neurology, Harvard Medical School, Boston, Massachusetts, USA
25. Department of Clinical Genetics, Genetics and Personalized Medicine Clinic, Tartu University Hospital, Tartu, Estonia
26. Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia
27. Harry Perkins Institute of Medical Research and Centre for Medical Research, University of Western Australia
28. Ocular Genomics Institute, Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, USA
29. Molecular Diagnostics, New York Genome Center, New York, USA
30. Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
31. Functional Neuromics, Children's Medical Research Institute, Westmead, New South Wales, Australia
32. Pathology and Laboratory Medicine, Mount Sinai Hospital, Sinai Health, Toronto, Ontario, Canada
33. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada
34. Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, Sinai Health, Toronto, Ontario, Canada
35. Department of Neurosciences, Rady Children's Institute for Genomic Medicine, University of California, San Diego, La Jolla, California, USA
36. Department of Neurology, Central Adelaide Local Health Network/Royal Adelaide Hospital, Adelaide, SA, Australia
37. Adelaide Medical School, The University of Adelaide, Adelaide, SA, Australia
38. Department of Genetics & Molecular Pathology, SA Pathology, Adelaide, SA, Australia
39. Department of Neurology, Brigham and Women's Hospital, Boston, Massachusetts
40. Department of Pathology & Cell Biology, Columbia University Irving Medical Center, New York, USA
41. Victorian Clinical Genetics Services, Murdoch Children's Research Institute and Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia
42. Neuromuscular and Neurogenetic Disorders of Childhood Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland, USA
43. Brain and Mitochondrial Research Group, Murdoch Children’s Research Institute and Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia
44. Division of Neonatology, Department of Pediatrics, University of Miami School of Medicine and Holtz Children's Hospital, Jackson Health System, Miami, FL
45. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

Corresponding author: Monica H Wojcik, MD, MPH, monica.wojcik@childrens.harvard.edu, 300 Longwood Ave BCH 3036, Boston, MA, 02115; Anne O’Donnell-Luria, anne.odonnell@childrens.harvard.edu;, 300 Longwood Ave, Boston, MA, 02115.

Abstract

Background: Causal variants underlying rare disorders may remain elusive even after expansive gene panels or exome sequencing (ES). Clinicians and researchers may then turn to genome sequencing (GS), though the added value of this technique and its optimal use remain poorly defined. We therefore investigated the advantages of GS within a phenotypically diverse cohort.

Methods: GS was performed for 744 individuals with rare disease who were genetically undiagnosed. Analysis included review of single nucleotide, indel, structural, and mitochondrial variants.

Results: We successfully solved 218/744 (29.3%) cases using GS, with most solves involving established disease genes (157/218, 72.0%). Of all solved cases, 148 (67.9%) had previously had non-diagnostic ES. We systematically evaluated the 218 causal variants for features requiring GS to identify and 61/218 (28.0%) met these criteria, representing 8.2% of the entire cohort. These included small structural variants (13), copy neutral inversions and complex rearrangements (8), tandem repeat expansions (6), deep intronic variants (15), and coding variants that may be more easily found using GS related to uniformity of coverage (19).

Conclusion: We describe the diagnostic yield of GS in a large and diverse cohort, illustrating several types of pathogenic variation eluding ES or other techniques. Our results reveal a higher diagnostic yield of GS, supporting the utility of a genome-first approach, with consideration of GS as a secondary or tertiary test when higher-resolution structural variant analysis is needed or there is a strong clinical suspicion for a condition and prior targeted genetic testing has been negative.

Introduction

Massively-parallel ("next generation") sequencing has revolutionized clinical medicine by uncovering the causal variants underlying rare conditions, particularly via large gene panels or exome sequencing (ES).[1] Even if a clinical diagnosis has been made, identifying the molecular diagnosis – the underlying genomic change responsible for disease – can provide new clinical insight and support individualized therapies[2] and familial testing and counseling. However, many causal variants, including those uniquely amenable to precision therapies,[3] remain elusive even after ES or gene panel sequencing. Thus, a majority of rare disease patients remain molecularly-undiagnosed,[4,5] and clinicians may thus turn to genome sequencing (GS). Although more costly than ES, which only evaluates the ~2% of the genome that is protein-coding, GS has many potential advantages, including the ability to detect variants that ES cannot easily detect, such as certain structural variants (SVs, encompassing a broader range of copy number gains and losses in addition to copy-neutral inversions, retroviral insertions and other, more complex events), tandem repeat expansions (TREs), and deep intronic variants. Moreover, the lack of an exon-capture step in GS leads to improved uniformity of sequence coverage, increasing detection sensitivity for single nucleotide variants (SNVs) and insertions/deletions (indels) in certain coding regions compared to ES.[6-10] However, with the enhanced technical sensitivity of GS comes a higher analytic burden due to the millions of non-coding or structural variants identified. To date, the few publications that have compared the relative yield of GS versus ES have found that higher detection sensitivity of GS is associated with only a modest increase in diagnostic yield.[11-14] Thus, the incremental benefit of GS remains unclear.

Through the Center for Mendelian Genomics (CMG) and the Rare Genomes Project (RGP) at the Broad Institute of MIT and Harvard, we have sequenced and analyzed over 8,000 families with

rare, suspected monogenic disease, employing a variety of analytic techniques to identify pathogenic variants responsible for a variety of phenotypes.[15] We therefore evaluated the diagnostic yield of GS for rare disease diagnosis within this cohort, focusing on features enabling successful diagnosis via GS, particularly where ES or other methods had been unsuccessful.

Methods

The Broad Institute CMG and RGP are research studies aimed at genetic diagnosis and discovery for people with suspected rare Mendelian disorders. The Broad CMG was established in 2016 as part of an initiative funded by the National Institutes of Health to identify novel disease genes underlying Mendelian disorders using ES and GS.[15-17] Families sequenced through the Broad CMG are enrolled by collaborating investigators in research studies under a local institutional review board (IRB)-approved protocol that includes a provision for data sharing. RGP was launched in 2017 as a patient-centered initiative to directly recruit individuals and families with suspected Mendelian disorders throughout the United States for sequencing and analysis via the Broad CMG. Potential participants self-refer through the study website. Both studies have been approved by the Mass General Brigham IRB.

Over a five-year period, from April 2016 to April 2021, 744 families underwent GS via the Broad CMG (354 families) or RGP (390 families) following an unrevealing prior diagnostic evaluation.

Case selection

Individuals or families were selected for GS after case review by at least two members of CMG/RGP team (genetic counselors, clinical geneticists or other relevant subspecialists). Cases approved for GS were thought to have a high likelihood of an underlying genetic disorder in addition to appropriate, non-diagnostic prior testing justifying the need for GS, such as ES or

targeted testing of specific disease genes. The RGP study also includes families who are unable to access genetic testing due to lack of insurance coverage or other barriers to care and, therefore, may have not had any prior genetic testing.

Sequencing and Analysis

GS and data processing were performed by the Genomics Platform at the Broad Institute. Our approaches to data analysis continue to evolve as we develop innovative approaches to derive the maximum diagnostic yield. Additionally, our Broad CMG collaborators use multiple "homegrown" approaches to data analysis based on their experience and the genetic architecture of the various phenotypes. Generally, we analyze our ES/GS data under the assumptions that the affected individual/family has a severe, rare Mendelian condition, and our analysis filters reflect these assumptions (Table S1). Our "first pass" analysis for SNV and indels includes evaluation for both *de novo* dominant or recessive conditions. For most families, we also perform a search with a list of priority genes based on phenotype and incorporating literature reports, Online Mendelian Inheritance in Man (OMIM)[18], and internal communication with researchers/collaborators and relax our filtering strategy when using these lists. Additional sequencing methods and analytic tools are further described in the Supplement.

Variant interpretation

Candidate variants (SNVs, indels and SVs) in known disease-associated genes were classified according to established criteria[19-21] (additional details in the Supplement). We considered a case solved if a pathogenic or likely pathogenic variant was found in a known disease gene that explained the phenotype or if a variant was found in a novel disease gene with moderate/strong supporting evidence for the individual's phenotype by ClinGen criteria[22]; we considered cases

likely solved when a variant was identified in a known disease gene that was classified as a variant of uncertain significance (VUS) by ACMG/AMP criteria but the multidisciplinary CMG team and referring provider, when relevant, considered the variant causative based upon supportive clinical data. Re-analysis of unsolved cases remains ongoing; the solve status reported here is as of May 1, 2023.

Evaluation of diagnoses

We systematically evaluated all solved/likely solved cases to determine variants requiring GS for identification (Figure). These included deep intronic non-coding variants as well as coding variants such as SNVs and indels poorly covered by standard ES platforms. In addition, we evaluated all TREs and SVs to identify those that were missed on prior exome or could not reliably be found using ES data. For example, larger CNVs in well-covered regions may be identified on ES data if dedicated CNV calling pipelines are applied. Because the performance of many CNV callers on ES data decays for CNVs smaller than 3 exons, [23-25] any heterozygous CNV with less than 3 targets was considered to require GS as it would not be reliably detected by ES. We also considered CNVs involving noncoding regions or regions that are challenging to sequence (such as high sequence homology, exons with high GC content) as requiring GS to detect and manually reanalyzed the ES data when available with updated CNV calling to see if we could retrospectively identify the CNV. Comparisons of diagnostic yield across categories were made using the Fisher's exact test or chi square test when appropriate.

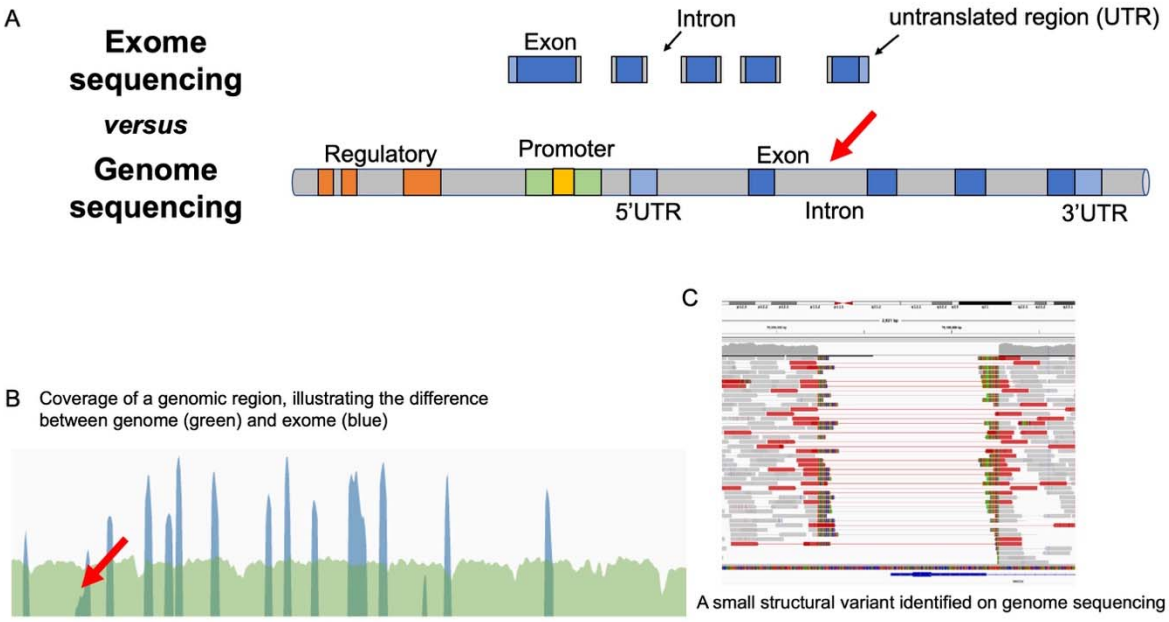

Figure 1. Variants assessed as requiring genome sequencing over exome sequencing to be identified within our cohort. These include: (A) deep intronic variants (arrow) unlikely to be reliably detected by typical ES methodologies (greater than 20 base pairs upstream/downstream from the beginning or end of an exon), (B) indels or SNVs that were missed on prior ES due to poor coverage of the region (arrow), tandem repeat expansions (TREs), and some SVs such as copy-neutral inversions, (C) small (~50-2000 base pair) and/or largely intronic CNVs, or complex events involving more than one type of SV.

Results:

Diagnostic yield

Of the 744 families who underwent GS, we consider 218 (29.3%) solved or likely solved (Table S2). Characteristics of the probands undergoing GS are presented in Table 1. Most solves (157/218, 72.0%) were identified in previously known disease genes and the remainder represented novel disease gene discoveries (61/218, 32.1%), with one case involving both a novel gene and a known gene contributing to a blended phenotype. Our diagnostic yield was lowest in cases sequenced as duos (two affected siblings or a parent/proband pair, 14/77, 18.2%),

compared to proband only (44/215, 20.5%), trios (136/369, 36.9%), or larger family groups (24/83, 28.9%) ($p < 0.001$). Diagnostic yield was lower in RGP-enrolled cases (102/390, 26.1%) compared to CMG collaborator-recruited cases (116/354, 32.8%), though this did not reach statistical significance. Although we allowed for cases to be considered solved by VUS that were clinically-interpreted by our multi-disciplinary team as causal, excluding any solve involving a VUS would result in a diagnostic yield of 130/744 (17.5%).

For the 138/744 (18.5%) cases with VUS, most were in strong candidate novel disease genes (109/138, 80.0%) where ACMG/AMP criteria are not yet appropriate to be applied and the remainder were novel variants in known disease genes (29/138, 21.0%) identified that did not yet have sufficient evidence for a pathogenic classification. A list of candidate disease genes is provided and cases have been submitted to Matchmaker Exchange (Table S3).[26]

The average diagnostic yield varied significantly by phenotype category ($p<0.0001$ by chi-square test), with the highest yields seen in probands with neurodevelopmental conditions or syndromic anomalies (Figure 2). Our solve rate did not significantly differ by imputed ancestry (Table S3) either when considered across multiple groups or when all European were compared to non-European ancestries, though this may be due to the small sample size as the solve rates were lower for African/African American, Admixed American, and East Asian (22%, 15%, 17%) as compared to European (43%).

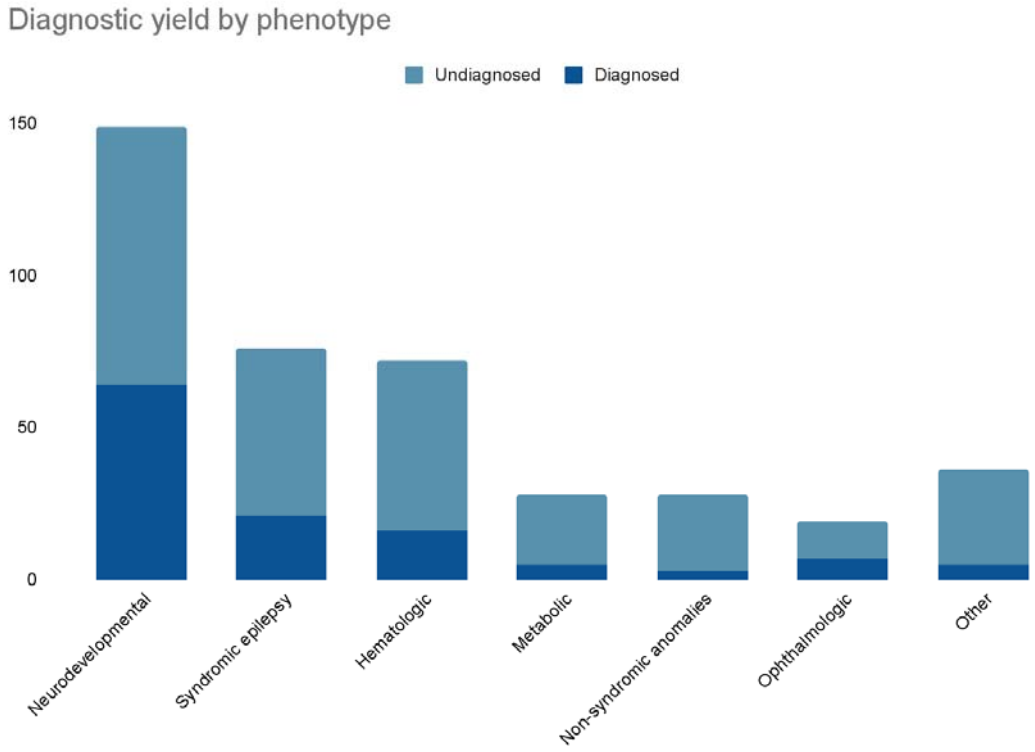

Figure 2. Diagnostic yield by phenotype category. Probands were categorized by most prominent phenotypic features.

Solves requiring GS

We determined that 62 of our 218 solved/likely solved cases required GS to detect the causal variant, comprising 28.3% of the diagnosed cases and 8.3% of the entire cohort (Figure 3). Of these 62 cases, 55 had prior negative ES, and two others had prior targeted sequencing of the causal gene (*ADA*, *DMD*). The remaining five cases were empirically determined to require GS as they involved a deep intronic variant, a complex SV and three TREs unlikely to be identified via ES pipelines.

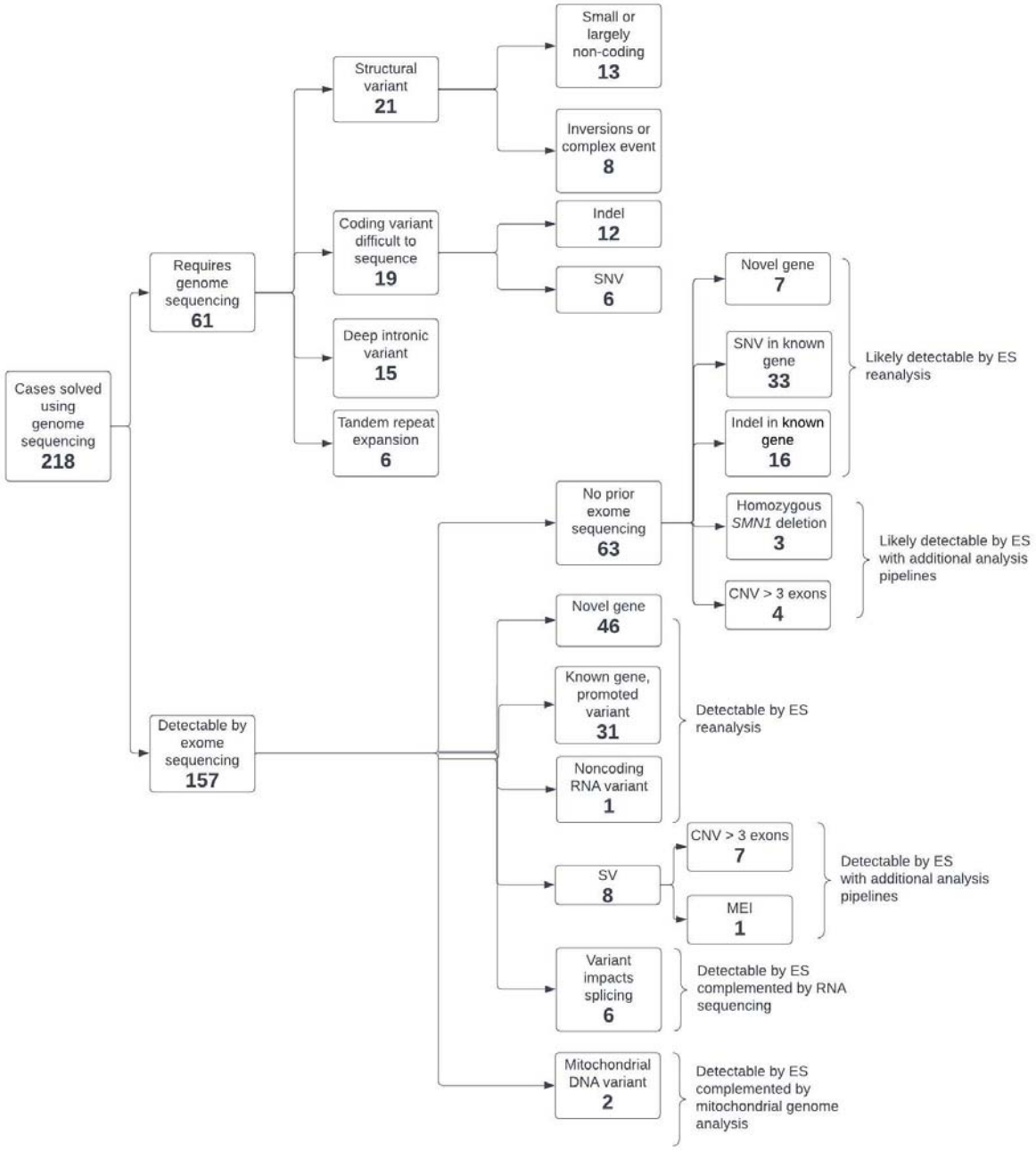

Figure 3. Classification of solved cases. Of all 218 cases solved/likely solved by GS, the types of variants requiring GS (61) to identify are displayed. Twenty-one (34%) were SVs, including 11 deletions (2 non-coding regions), 2 duplications, 4 inversions, 3 complex del/dup events, and one mobile element insertion (MEI) in a non-coding region. Additionally, there were 6 tandem repeat expansions identified. Nineteen cases (31%) requiring GS involved coding variants missed on prior ES due to poor coverage of the region, either resulting in no variant call at that site or a poor-quality variant that was filtered out during the analysis process. Most of these (13/19) were indels (small insertion/deletions usually of 10-15 bp in size) of which 8/13 were deletions and 5/13 were involved indels of different types. Fifteen (25%) of the diagnoses

requiring GS involved deep intronic non-coding variants (complemented by RNA sequencing). Three cases had variants that were missed for more than one reason (SVs in non-coding regions for two and a deep intronic variant in a novel gene in one). Of the cases solved by variants that were detectable by ES, most (93/156, 59.6%) had prior ES that missed the variant, whereas the remaining 63 did not have prior ES but were solved via variants expected to be detectable by current ES methodologies. The three cases of *SMN1* homozygous deletions responsible for spinal muscular atrophy (SMA) were identified in RGP probands (two were concurrently identified via clinical genetic testing, one was an adult who had been clinically diagnosed with SMA of unknown type but never received testing). Two diagnostic mitochondrial DNA (mtDNA) variants were identified by GS and missed by prior ES, although these variants were also identified by mitochondrial genome sequencing.

Diagnostic yield of GS detectable by ES

Most cases solved via GS previously had non-diagnostic ES (148/218, 67.9%) (Figure 4), confirming that these variants were overlooked previously. Overall, 94/148 (63.5%) exome-negative cases later solved by GS could have been found by exome reanalysis, most commonly because the diagnosis involved a recent novel gene-disease discovery (46/94, 48.9%), a variant in a known gene that was re-interpreted as disease causing over time (31/94, 33.0%), or analysis of variants in noncoding disease-associated transcripts (1/94, 1.1%).

Expanding the methodologies for variant calling applied to ES is predicted to uplift diagnosis by 3.2% (24/744) overall for the cohort. If CNV calling (8), mobile element insertion calling (1),[27] and mitochondrial genome variant calling (2) were performed on ES data, an additional 11/94 (11.7%) would be solved. RNA sequencing analysis helped identify or validate the impact of a variant on gene expression and/or splicing for deep intronic variants (6/94, 6.4%). For the 63 families without prior ES that could have been found by exome analysis, seven diagnoses required CNV calling (4) or specialized calling of *SMN1* deletions responsible for SMA (3).[28] Indels were a common source of variants requiring GS, so it remains possible that some of the 16 coding indels identified in individuals without prior ES may be missed by ES. Given that this

cannot be ascertained with the available data, we have not counted them as requiring GS to avoid over-reporting.

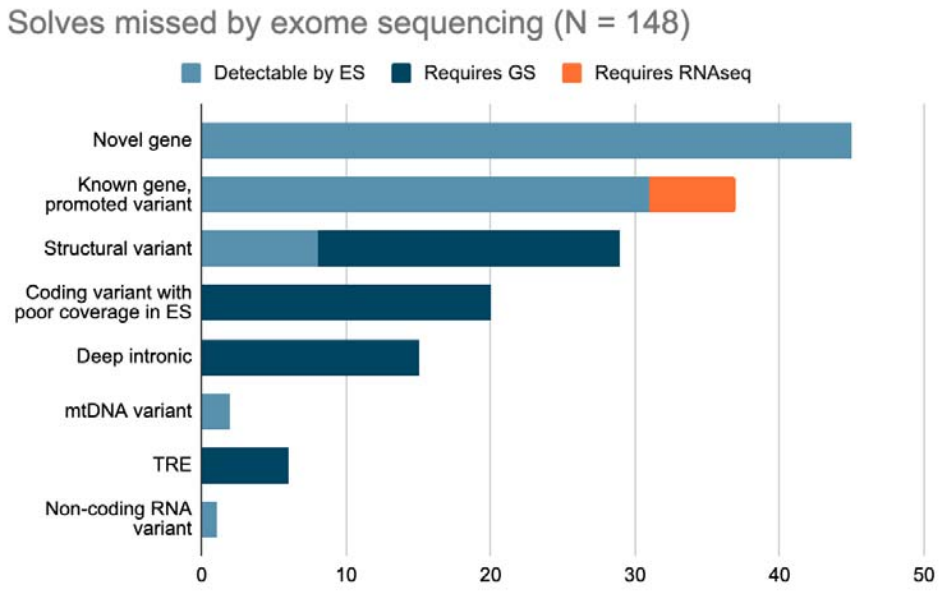

Figure 4. Reasons cases were not solved by prior exome sequencing.

Maximizing the yield of GS

Of the 61 solves/likely solves requiring GS, 54 (89%) were in known disease genes and most were identified by untargeted analysis either due to improved coverage of coding regions (19), or by genome-wide evaluation for rare SVs (21) or known pathogenic TREs (6). Deep intronic variants (15) were typically identified in a more targeted fashion, by evaluating genes associated with the specific phenotype, and the impact of several were validated by RNA sequencing (7/15) or other functional evaluation (2/15). Additionally, eight diagnoses involving SVs benefited from a strong clinical suspicion based upon phenotypic or biochemical data that was used to identify the underlying causal variant(s) that had been missed by standard genetic testing methodologies (*FBN1, ADA, QDPR, NIPBL, RPGRIP1, DMD, RSP19, EDA*). [29,30] In these cases, a clinical

diagnosis suggested a particular molecular genetic diagnosis - often narrowing the differential to a single gene - though the diagnosis eluded either ES or more targeted testing.

Only seven solved cases requiring GS were in novel disease genes, reflecting the challenges in prioritizing and validating variants beyond ES for genes not currently associated with human disease and the power of GS analyzed via research consortia. These diagnoses include: i) a homozygous deep intronic variant in a novel mitochondrial disease gene, *NDUFB10*, ultimately solved via a 'multi-omic' approach;[31] ii) a deep intronic variant in *RPL17*; iii) a homozygous splice-impacting variant in *CYS1* missed on ES due to poor coverage of a GC-rich region; [32] iv) a structural long non-coding RNA variant in *CHASERR*; v) a missense variant in *PNPLA7* that was missed due to poor coverage on ES present in trans with a splice-impacting variant; vi) a single exon deletion in the loss-of-function-constrained *ZFHX3* gene; vii) and a homozygous 16 kb deletion in *WBP4* identified on GS complemented by RNA sequencing.[33] The latter four diagnoses were aided by identification of other affected cases via the Matchmaker Exchange.[34]

Discussion:

We describe a large cohort of individuals with rare disease for whom a plausible diagnosis was found using GS, often after being missed by prior testing approaches. These cases illustrate several types of pathogenic variation that may be missed by ES or other standard genetic testing methodologies, particularly small or copy-neutral SVs, deep intronic variants, and TREs. Importantly, these cases were solved using short read GS, which is currently clinically available. Long read GS, which is anticipated to soon be clinically available, should also be able to identify these variants described here as well as additional findings undetectable by short read GS. For this cohort, in order to achieve the overall solve rate of 29.3% after prior negative testing, 13.3%

of families needed ES reanalysis, 2.3% needed additional methodologies (calling of CNV, mitochondrial genome variants, mobile element insertions, and *SMN1* deletions on ES data), and 8.3% of families truly required sequencing of the genome.

Most of our solved and likely solved cases requiring GS (47/61, 77%) were identified in an untargeted fashion, by systematically applying analytic tools across the entire cohort. As most diagnoses identified in the cohort could have been found using ES, those that truly required GS offer important insight into optimal clinical application of GS for rare disease diagnosis. Our analysis for SVs was particularly high-yield at identifying variants missed by prior testing, as 35 solves involved SVs, most (20/35) of which required GS to identify. The deep intronic variants described in this cohort were identified by a targeted approach, narrowed down to a single gene or list of genes by phenotyping or RNA sequencing. This reflects not only the challenges in variant prioritization related to the multitude of SVs, deep intronic, or non-coding variants identified by GS data, but also suggests that additional diagnoses may be found in the future by continued reanalysis of GS data from unsolved cases using regularly updated candidate gene lists. These diagnoses also demonstrate the value of careful phenotyping in the evaluation of genetic testing results in addition to prioritization of genetic loci for evaluation by GS via transcriptome analysis (RNA sequencing).[35] In particular, a phenotypically-driven or otherwise targeted approach may mitigate some of the analytic burden of GS analysis, as a deep analysis of a limited number of genetic loci is more feasible than a search for SNVs and SVs across the entire genome. We have previously demonstrated the success of this technique in an individual with early-onset Marfan syndrome, in which a structural variant (deletion) was identified using GS that had been missed by multiple prior approaches, including targeted deletion/duplication analysis via multiplex ligation-dependent probe amplification (MLPA) in addition to ES.[30]

Critical to this approach is the ability to accurately phenotype the patient in order to guide the genomic evaluation[36], and, indeed, the diagnostic yield within our cohort was greatly augmented by the expertise of the investigators providing cohorts for collaborative analysis.

Prior evaluations of the diagnostic yield of GS when compared to ES have demonstrated either a similar[13] or only mildly increased yield[11], with few diagnoses identified that would not be detectable by ES alone - usually SVs.[4,12-14,37,38] In particular, one prior report of 108 cases of GS after non-diagnostic ES identified an incremental yield of 7%, but only 3% for cases requiring GS for diagnosis.[11] A more recent study of GS as a first-line modality for rare disease diagnosis via identified an incremental yield of 37.5% in those who had prior ES, though notably only 16 individuals (1.5% of the larger cohort) had ES prior to GS, making it difficult to directly compare to our data given the small numbers[39]. Furthermore, while 14% of all diagnoses were described by the investigators as requiring GS for detection, this included all SVs identified in the study, even though some may be callable on ES data, which is distinct from our approach.[39] Our larger proportion of diagnoses requiring GS for detection (28.4% of the solved cohort) compared to this and prior studies also likely reflects our increased detection of SVs and variants outside of coding regions, as many prior analyses of GS yield have focused primarily on coding variants[5,13,14], as well the case selection in our cohort, with families with a high probability of monogenic disease (often with prescreening by gene panel or prior ES) selected for sequencing via the Broad CMG.

Overall, our results support the use of GS as a first-line test for rare disease diagnosis due to the ability to detect multiple types of disease-causing variation, replacing both ES and chromosomal microarray in a single test, and superior ability to detect all classes of variation (with the exception of somatic mosaic variants, for which the lower mean coverage of GS reduces

sensitivity). However, routinely applying GS after ES or other non-diagnostic testing is still not likely to be the most efficient strategy, reflected in our relatively low incremental yield after non-diagnostic ES. While sequencing, analysis, and storage costs for short read GS and variation in payor reimbursement may present limitations to its routine application for genetic diagnosis, we anticipate these barriers to be overcome in the future, particularly as sequencing costs continue to fall. For people who remain undiagnosed after clinical ES, our findings support a reasonable diagnostic potential from GS, especially in the case where there is a clinical diagnosis directing focused attention in the GS analysis. Transcriptome analysis, which was utilized for a minority of families here, has a complementary role in GS, and its implementation in clinical testing laboratories and is likely to further improve diagnosis rates. Altogether, our findings provide important insight into the present diagnostic utility of GS and the role it may play in the future.

Acknowledgements: We thank the many families who participate in the Rare Genomes Project or via our Broad CMG collaborators research studies for sharing their samples and medical data.

Funding: The Broad Institute of MIT and Harvard Center for Mendelian Genomics (Broad CMG) was funded by the National Human Genome Research Institute (NHGRI), the National Eye Institute, the National Heart, Lung and Blood Institute grant UM1HG008900 and NHGRI grants U01HG011755 and R01HG009141. This publication has also been made possible in part by CZI grant DAF2019-19927, grant DOI https://doi.org/10.37921/236582yuakxy, from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (funder DOI 10.13039/100014989). FH is supported by NIH 5RC-2DK122397. SP is supported by Estonian Research Council grants PUTJD827, MOBTP175, PSG774. KO, KR and KTO are

supported by Estonian Research Council grant PRG471. UDP-Vic receives financial support from the Murdoch Children's Research Institute and the Harbig Foundation. The research conducted at the Murdoch Children's Research Institute (MCRI) was supported by the Victorian Government's Operational Infrastructure Support Program. The Chair in Genomic Medicine awarded to JC is generously supported by The Royal Children's Hospital Foundation. MYO–SEQ was funded by Sanofi Genzyme, Ultragenyx, LGMD2I Research Fund, Samantha J Brazzo Foundation, LGMD2D Foundation, Kurt+Peter Foundation, Muscular Dystrophy UK and Coalition to Cure Calpain 3. CGB is supported by intramural funds by the NIH National Institute of Neurological Disorders and Stroke. JLE was funded by the McLaughlin Centre (grant #MC-2012–13, #MC-2014-11-1 and MC-2017–12) and CIHR- Champions of Genetics: Building the Next Generation Grant (FRN: 135730). MHW is supported by NIH K23 HD102589, NIH R21 HG012397, and by an early career award from the Thrasher Research Fund. KMB and EAP are supported by grants from the National Eye Institute: R01 EY026904 (KMB/EAP), R01 EY012910 (EAP) and P30 EY014104 (MEEI core support), the Foundation Fighting Blindness: EGI-GE-1218-0753-UCSD, (KMB) and the Research to Prevent Blindness International Research Collaborators Award (KMB). ADD is supported by NIH R01 HD081256, MH 115957. The Rare Disease Flagship acknowledges financial support from the Royal Children's Hospital Foundation, the Murdoch Children's Research Institute and the Harbig Foundation. NGL is supported by the Australian National Health and Medical Research Council (NHMRC) Grant APP2002640. STC received support from National Health and Medical Research Council (NHMRC) of Australia Fellowships APP1048816 and APP1136197, NHMRC Project Grant APP1080587 and NHMRC Ideas Grants APP1106084, APP2002640. SJB received a Muscular Dystrophy Association of New South Wales Sue Connor postgraduate training scholarship. RG

was supported by NHMRC grant APP1074954. SLS is supported by a fellowship from the Manton Center for Orphan Disease Research at Boston Children's Hospital. GR is supported by a NHMRC EL2 Fellowship (APP2007769).

Financial disclosures: DGM is a paid advisor to GlaxoSmithKline, Insitro, Variant Bio and Overtone Therapeutics, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Google, Merck, Microsoft, Pfizer, and Sanofi-Genzyme. MHW has consulted for Illumina and Sanofi. HLR. has received support from Illumina and Microsoft to support rare disease gene discovery and diagnosis. AO'DL has consulted for Tome Biosciences and Ono Pharma USA Inc, and is member of the scientific advisory board for Congenica Inc and the Simons Foundation SPARK for Autism study. STC is director of Frontier Genomics Pty Ltd (Australia) and receives no remuneration (salary or consultancy fees) for this role. Frontier Genomics Pty Ltd has no current financial interests that will benefit from publication of this data. WKC is on the Board of Directors of RallyBio and Prime Medicine.

References

1. Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 2013;369(16):1502-11.
2. Kim J, Hu C, Moufawad El Achkar C, et al. Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. N Engl J Med 2019;381(17):1644-1652.
3. Kim J, Woo S, de Gusmao CM, et al. A framework for individualized splice-switching oligonucleotide therapy. Nature 2023;619(7971):828-836.
4. Shashi V, Schoch K, Spillmann R, et al. A comprehensive iterative approach is highly effective in diagnosing individuals who are exome negative. Genet Med 2019;21(1):161-172.

5. Splinter K, Adams DR, Bacino CA, et al. Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. N Engl J Med 2018;379(22):2131-2139.
6. Lindstrand A, Eisfeldt J, Pettersson M, et al. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. Genome Med 2019;11(1):68.
7. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Hum Mutat 2015;36(8):815-22.
8. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A 2015;112(17):5473-8.
9. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? Hum Genet 2016;135(3):359-62.
10. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet 2015;47(7):717-726.
11. Alfares A, Aloraini T, Subaie LA, et al. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. Genet Med 2018;20(11):1328-1333.
12. Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med 2018;20(4):435-443.
13. Kingsmore SF, Cakici JA, Clark MM, et al. A Randomized, Controlled Trial of the Analytic and Diagnostic Performance of Singleton and Trio, Rapid Genome and Exome Sequencing in Ill Infants. Am J Hum Genet 2019;105(4):719-733.
14. van der Sanden B, Schobers G, Corominas Galbany J, et al. The performance of genome sequencing as a first-tier test for neurodevelopmental disorders. Eur J Hum Genet 2023;31(1):81-88.

15. Baxter SM, Posey JE, Lake NJ, et al. Centers for Mendelian Genomics: A decade of facilitating gene discovery. Genet Med 2022;24(4):784-797.
16. Bamshad MJ, Shendure JA, Valle D, et al. The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. Am J Med Genet A 2012;158A(7):1523-5.
17. Posey JE, O'Donnell-Luria AH, Chong JX, et al. Insights into genetics, human biology and disease gleaned from family based genomic studies. Genet Med 2019;21(4):798-812.
18. McKusick VA. Mendelian Inheritance in Man and its online version, OMIM. American Journal of Human Genetics 2007;80(4):588-604.
19. 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.
20. Riggs ER, Andersen EF, Cherry AM, 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 2020;22(2):245-257.
21. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. Hum Mutat 2018;39(11):1517-1524.
22. Strande NT, Riggs ER, Buchanan AH, et al. Evaluating the Clinical Validity of Gene-Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. Am J Hum Genet 2017;100(6):895-906.
23. Hong CS, Singh LN, Mullikin JC, Biesecker LG. Assessing the reproducibility of exome copy number variations predictions. Genome Med 2016;8(1):82.
24. Fu JM, Satterstrom FK, Peng M, et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat Genet 2022;54(9):1320-1331.

25. Babadi M, Fu JM, Lee SK. GATK-gCNV: A Rare Copy Number Variant Discovery Algorithm and Its Application to Exome Sequencing in the UK Biobank. bioRxiv 2022.08.25.504851

26. Boycott KM, Azzariti DR, Hamosh A, Rehm HL. Seven years since the launch of the Matchmaker Exchange: The evolution of genomic matchmaking. Hum Mutat 2022;43(6):659-667.

27. Torene RI, Galens K, Liu S, et al. Mobile element insertion detection in 89,874 clinical exomes. Genet Med 2020;22(5):974-978.

28. Larson JL, Silver AJ, Chan D, Borroto C, Spurrier B, Silver LM. Validation of a high resolution NGS method for detecting spinal muscular atrophy carriers among phase 3 participants in the 1000 Genomes Project. BMC Med Genet 2015;16:100.

29. Lillevali H, Pajusalu S, Wojcik MH, et al. Genome sequencing identifies a homozygous inversion disrupting QDPR as a cause for dihydropteridine reductase deficiency. Mol Genet Genomic Med 2020:e1154.

30. Wojcik MH, Thiele K, Grant CF, et al. Genome Sequencing Identifies the Pathogenic Variant Missed by Prior Testing in an Infant with Marfan Syndrome. J Pediatr 2019; 213:235-240.

31. Helman G, Compton AG, Hock DH, et al. Multiomic analysis elucidates Complex I deficiency caused by a deep intronic variant in NDUFB10. Hum Mutat 2021;42(1):19-24.

32. Yang C, Harafuji N, O'Connor AK, et al. Cystin genetic variants cause autosomal recessive polycystic kidney disease associated with altered Myc expression. Sci Rep 2021;11(1):18274.

33. Engal E, Oja KT, Maroofian R, et al. Biallelic loss of function variants in WBP4 , encoding a spliceosome protein, result in a variable neurodevelopmental delay syndrome. medRxiv 2023:2023.06.19.23291425.

34. Sobreira NLM, Arachchi H, Buske OJ, et al. Matchmaker Exchange. Current protocols in human genetics 2017;95:9.31.15.

35. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Trans Med 2017;9(386):10.1126/scitranslmed.aal5209.
36. Foley AR, Donkervoort S, Bonnemann CG. Next-generation sequencing still needs our generation's clinicians. Neurol Genet 2015;1(2):e13.
37. Palmer EE, Sachdev R, Macintosh R, et al. Diagnostic Yield of Whole Genome Sequencing After Nondiagnostic Exome Sequencing or Gene Panel in Developmental and Epileptic Encephalopathies. Neurology 2021;96(13):e1770-e1782.
38. Lowther C, Valkanas E, Giordano JL, et al. Systematic evaluation of genome sequencing for the assessment of fetal structural anomalies. bioRxiv 2020.08.12.248526
39. Investigators GPP, Smedley D, Smith KR, et al. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. N Engl J Med 2021;385(20):1868-1880.
40. Karczewski KJ, Francioli LC, Tiao G, et al. Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2021;590(7846):E53.

Table 1. Demographics of sequenced probands

	Probands (N, %)
Sex	
Male	406 (54.6%)
Female	338 (45.4%)
Age*	Median, IQR
	10.2 (4 – 36)

Imputed ancestry**	
African/African American	22 (3.0%)
Ashkenazi Jewish	33 (4.4%)
East Asian	12 (1.6%)
European	570 (76.6%)
Admixed American	27 (3.6%)
Middle Eastern	2 (0.3%)
South Asian	13 (1.7%)
Multiple/Unassigned	65 (8.7%)

* Age unknown for 116 participants

**For ancestry imputation, we computed principal components (PCs) on high-quality bi-allelic autosomal SNVs in our rare disease cohorts using the gnomAD v2 method[40]. The sample scores from this projection are input into the gnomAD v2 random forest model and an ancestry assigned to all samples for which the probability of that ancestry is > 90%. All samples that do not meet the 90% threshold are assigned as Multiple/Unassigned.