Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders

Susan M Hiatt, James MJ Lawlor, Lori H Handley, et al.

Genome Res. published online September 19, 2024
Access the most recent version at doi:10.1101/gr.279227.124

Long-read genome sequencing and variant reanalysis increase diagnostic yield in neurodevelopmental disorders

Susan M. Hiatt[1*], James M.J. Lawlor[1], Lori H. Handley[1], Donald R. Latner[1], Zachary T. Bonnstetter[1], Candice R. Finnila[1], Michelle L. Thompson[1], Lori Beth Boston[1], Melissa Williams[1], Ivan Rodriguez Nunez[1], Jerry Jenkins[1], Whitley V. Kelley[1], E. Martina Bebin[2], Michael A. Lopez[2,3,4], Anna C. E. Hurst[4], Bruce R. Korf[4], Jeremy Schmutz[1], Jane Grimwood[1], Gregory M. Cooper[1*]

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806, USA

[2]Department of Neurology, University of Alabama at Birmingham, Birmingham, AL, 35924, USA

[3]Department of Pediatrics, University of Alabama at Birmingham, Birmingham, AL, 35924, USA

[4]Department of Genetics, University of Alabama at Birmingham, Birmingham, AL, 35924, USA

*shiatt@hudsonalpha.org, gcooper@hudsonalpha.org, 601 Genome Way, Huntsville, AL, 35806, USA, 256-327-9490

Running Title:

Increased diagnostic yield with long-reads

Key Words:

LRS Special Issue, Long read sequencing, Clinical sequencing, neurodevelopmental disorder, structural variation, repeat expansion

ABSTRACT

Variant detection from long-read genome sequencing (lrGS) has proven to be more accurate and comprehensive than variant detection from short-read genome sequencing (srGS). However, the rate at which lrGS can increase molecular diagnostic yield for rare disease is not yet precisely characterized. We performed lrGS using Pacific Biosciences “HiFi” technology on 96 short-read-negative probands with rare diseases that were suspected to be genetic. We generated hg38-aligned variants and *de novo* phased genome assemblies, and subsequently annotated, filtered, and curated variants using clinical standards. New disease-relevant or potentially relevant genetic findings were identified in 16/96 (16.7%) probands, nine of which (8/96, ~9.4%) harbored pathogenic or likely pathogenic variants. Nine probands (~9.4%) had variants that were accurately called in both srGS and lrGS and represent changes to clinical interpretation, mostly from recently published gene-disease associations. Seven cases included variants that were only correctly interpreted in lrGS, including copy-number variants, an inversion, a mobile element insertion, two low-complexity repeat expansions, and a 1 bp deletion. While evidence for each of these variants is, in retrospect, visible in srGS, they were either not called within srGS data, were represented by calls with incorrect sizes or structures, or failed quality-control and filtration. Thus, while reanalysis of older srGS data clearly increases diagnostic yield, we find that lrGS allows for substantial additional yield (7/96, 7.3%) beyond srGS. We anticipate that as lrGS analysis improves, and as lrGS datasets grow allowing for better variant frequency annotation, the additional lrGS-only rare disease yield will grow over time.

INTRODUCTION

Although genome and exome sequencing (GS/ES) are increasingly used to identify molecular diagnoses for rare diseases, reported diagnostic rates range from 20-60% (Srivastava et al. 2019; Baxter et al. 2022), indicating that many conditions suspected to be genetic remain refractory to genomic testing. While some tested individuals may have phenotypes resulting from polygenic and/or environmental risk factors (e.g., Niemi et al. 2018), a subset of undiagnosed cases likely result from genetic factors that we are as-yet unable to identify. It is well-known that short-read genome sequencing (srGS) has poor sensitivity to many types of variants, especially structural variants (SVs) and variants affecting repetitive sequences (Wenger et al. 2019; Sanghvi et al. 2018; Mahmoud et al. 2024). Long-read genome sequencing (lrGS), in contrast, has been shown to greatly improve sensitivity to many of the variants missed by srGS (Logsdon et al. 2020), in addition to facilitating *de novo* assemblies to allow for more effective evaluation of structural variation (Cheng et al. 2021). Accordingly, lrGS has great potential to improve rare disease diagnostic testing and has been applied to several rare disease cohorts (Cohen et al. 2022; Miller et al. 2021; Hiatt et al. 2021).

In addition to changes in sequencing technology, the scope of knowledge about genes and our ability to annotate genetic variants has steadily increased. As such, systematic reanalysis of GS/ES data can also lead to the discovery of previously overlooked clinically relevant variants. Diagnostic yield increases from reanalysis have been reported to range from 4-31% depending on a variety of factors, most notably time since the previous analysis (Hiatt et al. 2018; Liu et al. 2019; Schobers et al. 2022; Hartley et al. 2023). While a variety of factors contribute to reanalysis discoveries, they often result from the discovery of new disease genes, which contributes to 42-75% of reanalysis findings (Hiatt et al. 2018; Liu et al. 2019; Schobers et al. 2022; Hartley et al. 2023). This reflects the rapid pace of discovery of new disease genes in

the rare disease research community, which has been facilitated by data sharing via the MatchMaker Exchange and GeneMatcher (Philippakis et al. 2015; Sobreira et al. 2017).

Here we discuss findings from lrGS on a cohort of 96 short-read-negative cases, drawn from several studies focused on rare, suspected genetic diseases, especially early-onset neurodevelopmental disorders. We describe 19 relevant or potentially clinically relevant variants not previously evaluated or considered in 16 cases. We show that both more comprehensive variant detection from lrGS and variant reanalysis contribute to these discoveries. However, lrGS clearly provides unique advantages, and these advantages are likely to grow in the future, to maximize the rate of discovery of highly penetrant variation in any given individual with rare disease.

RESULTS

We selected individuals with rare diseases who had undergone short-read exome sequencing (srES; n=2) or srGS (n=94) in previous research studies yet had no pathogenic or likely pathogenic variants (P/LP) nor variants of uncertain significance (VUS) identified (Bowling et al. 2017; East et al. 2021; Bowling et al. 2022). Most of our cohort consisted of children (89% were <18 years of age at time of enrollment) with a neurodevelopmental disorder (NDD, 70%), multiple congenital anomalies (MCA, 22%), or a suspected genetic myopathy (8%). Probands consisted of 66% males (63/96); genetically inferred ancestries for probands revealed 72% European (69/96), 21% African/African American (20/96), 3% Admixed American (3/96), 1% Southeast Asian (1/96) and 3% unspecified ancestry admixture (Table 1). For these 96 cases, we performed lrGS using Pacific Biosciences “HiFi” sequencing to a median depth of 27X (Table 1, Supplemental Table S1). For a subset (10/96), we also performed lrGS on parents (median parental HiFi depth of 22X, Supplemental Table S2). We also generated *de novo* assemblies for each proband using hifiasm (Cheng et al. 2021), with parental srGS used for k-

mer-based binning and phasing when available. While the assemblies were not used for structural variant calling in this study, they have proven useful in visualizing and evaluating complex structural variation (Hiatt et al. 2021). The median N50 for assembled proband contigs was 29.05 Mb (Table 1).

HiFi reads were aligned to hg38, and variant calling was performed using DeepVariant (Poplin et al. 2018) and pbsv (https://github.com/PacificBiosciences/pbsv, see Methods). A median of 4.4 million SNVs and 970,031 indels were called in each proband (Table 1). We detected a median of 55,586 SVs of varying classes across the 96 probands using pbsv, of which a median of 25,218 are greater than 50bp in length (Table 1, Supplemental Table S3). These counts do not include any variant-quality filtration metrics and are expected to include an undetermined fraction of false positive calls; only variant calls of potential disease interest were evaluated for read support and quality (see Methods). Given that SVs are commonly arbitrarily defined as >50bp in length, our counts generally are consistent with previous studies that have detected a range of 22,000 to 33,000 SVs per genome (Pauper et al. 2021; Cohen et al. 2022; Groza et al. 2024).

Findings from lrGS

lrGS SNVs/indels were annotated with features such as gene overlaps, coding consequences, computational impact scores, and allele frequencies. They were then filtered and analyzed using in-house software that is also used for srGS data (Hiatt et al. 2021). Rare SVs (see Methods) were assessed by visualization of reads in IGV and prioritization and analysis using SvAnna (Danis et al. 2022). All variants of interest were subject to curation using American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP) and ClinGen criteria to identify potentially clinically relevant variation (Richards et al. 2015; Riggs et al. 2020). We ultimately identified 19 potentially “clinically relevant” variants, defined here as being pathogenic, likely pathogenic, or variants of uncertain significance

(P/LP/VUS), in 16 of the 96 cases (Table 2). Seven of these have a case-level classification of Definitive Diagnostic or Likely Diagnostic, which we define as P/LP variants that likely fully explain the reason for testing (Bowling et al. 2022). The remaining nine cases have Uncertain case-level classifications, either due to the variants being VUSs or being P/LP variants in genes whose associated phenotypes do not closely match the observed phenotype. Findings in seven probands exemplify the unique benefits of lrGS and are highlighted below.

lrGS-informed SVs

lrGS uncovered a *de novo*, 4 Mb, copy-neutral, paracentric inversion on Chromosome 3 (NC_000003.12:g.110477273_114639202_inv) in Proband 1 (Figure 1). This inversion spans about 35 protein-coding genes, and one breakpoint of this inversion lies within an intron of *ZBTB20* (MIM: 606025), between exons 6 and 7 in the 5' UTR of NM_001348800.3. The event breakpoints are visible in both long and short reads for this proband, but not in short reads for either parent, suggesting it arose *de novo*. Phasing in long-read data indicates the inversion is on the proband's paternal allele (Figure 1A). While evidence of the breakpoints are visible in the proband's srGS data (Figure 1A, 1B) and the breakpoints are called by manta (v1.6.0, Chen et al. 2016), the event only survived filtration and curation in lrGS (see discussion and Supplemental Figure S1, which compares inversion and breakend calls between srGS and lrGS for this proband). The variant is private to the proband and is predicted to disrupt *ZBTB20*. While the breakpoint lies between two UTR exons, it moves and flips the transcript's promoter and the first six exons away from the remaining exons (Figure 1C). This inversion is easily visualized in alignment of the proband's paternal contig vs. the reference genome in this region (Figure 1D). Loss-of-function (LOF) variation in *ZBTB20* is associated with Primrose Syndrome (MIM: 259050) and 3q13.31 Microdeletion Syndrome (Juven et al. 2020). Proband 1's reported features include moderate intellectual disability (ID), delayed speech and language development, muscular hypotonia, strabismus, and hypoplastic corpus callosum. She is also

non-ambulatory. Several of these features overlap Primrose Syndrome. We have classified this variant as likely pathogenic and the case-level designation is Likely Diagnostic (Table 2).

Proband 2 was originally enrolled for sequencing at the age of ~20 years, and presented with spasticity, ataxia, and leukodystrophy. srGS was negative, but lrGS identified two structural variants (SVs) identified in *trans* in *ALS2* (MIM: 606352). These include a maternally-inherited 4.65 kb deletion (Chr 2:201720435-201725085_del) that removes exons 21-23 of NM_020919.4, and a paternally-inherited ~1.6 Mb deletion (Chr 2:200115181-201739349_del) that spans several genes, including the 3' end of *ALS2* (deletion of exons 12-34 of NM_020919.4, Figure 2A, Table 2, Supplemental Figures S2-S5). These deletions are easily visualized in alignment of the proband's maternal (Figure 2B) and paternal (Figure 2C) contigs vs. the reference genome in this region. While these variants were called in short-read data, the smaller deletion was called as a heterozygous deletion in the mother and as a homozygous deletion in the proband, obfuscating the nature of the variation and raising quality-control concerns (e.g., nested and conflicting copy-number states and Mendelian inconsistency) that prevented effective curation (Supplemental Figures S6, S7). Given the results of the lrGS, the srGS variant calls logically resulted from the small maternal deletion intersecting with the larger, overlapping paternal deletion. This case highlights the difficulties in identification and analysis of overlapping SVs of unknown phase. Variation in *ALS2* is associated with several AR conditions (Juvenile amyotrophic lateral sclerosis 2, MIM: 205100; Juvenile Primary lateral sclerosis, MIM: 606353; and infantile onset ascending Spastic paralysis, MIM: 607225), each of which have features that overlap the proband's presentation. These variants are classified as P/LP and given the degree of overlap with expected phenotypes the case-level designation is Definitive Diagnostic.

Proband 3 is a male, enrolled in our research study at age ~50 years, and has a strong X-linked family history of ID (Figure 3A). srES and srGS were both negative, although in srES, two neighboring SNVs of uncertain significance were called, and manually curated, in the 3'

UTR of *HCFC1* (MIM: 300019). While srGS resulted in no calls in this region, visualization of reads in IGV suggested an insertion of unknown length and consequence (Figure 3C). Variation (mostly missense and proposed regulatory variation) in *HCFC1* has been associated with X-linked recessive Methylmalonic aciduria and homocysteinemia, cblX type (MIM: 309541). However, most affected individuals present with severely delayed psychomotor development, seizures, and methylmalonic aciduria. Proband 3's family reported neither of the latter two features. Given the uncertainties in the identity, structure, and consequence of the potential variants seen in srES and srGS, and the lack of clear phenotypic relevance for the gene, this region was curated but not originally considered to be a strong candidate for clinical relevance. HiFi sequencing identified a 4,902 bp mobile element insertion (MEI) in the 3' UTR of *HCFC1* (Chr X:153948602_ins4902), consisting of both SVA and L1 sequence (Figure 3B). This variant was likely inherited from a heterozygous carrier mother, as indicated by srGS reads at the breakpoints (Figure 3C). While this insertion does not affect protein-coding sequence, it is predicted to increase the length of the 3' UTR from 1,791 nt to 6,693 nt. We subsequently performed 3'-end RNA-seq on blood samples from both the proband and his father, generating ~8 million reads for each sample (see Methods). *HCFC1* shows the greatest expression decrease in the proband relative to his father, an ~8.8-fold reduction, among all genes with at least 10 counts in each sample (Supplemental Figure S8). While these results are consistent with the hypothesis that the insertion has a large effect on *HCFC1* expression and potential activity, they are not definitive. Further, expression or segregation analyses in additional family members could not be assessed. Given the uncertainty of the molecular consequence, the differences between observed phenotypic features and those reported to associate with *HCFC1*, and the lack of additional segregation data, we classified this as a VUS, with a case-level designation of Uncertain.

Proband 4 has a complex *de novo* structural variant affecting 16p13.2 (NC_000016.9:g.(8742452_9220783)dup_ins[(8742452_8879961)_(9000190_9220783)], Table

2, Supplemental Figures S9-S18). Proband 4 first had trio srES and no variants were returned, and this SV was not called. Trio lrGS identified the breakpoints, *de novo* status, phase, and probable order, orientation, and copy number of segments in this SV. However, there remains some uncertainty about the SV, as neither manual curation nor the proband's *de novo* assembly could definitively resolve the full, exact structure of the locus. Two possible structures are shown in Supplemental Figure S9. We believe that this is likely a limitation due to the length of the reads, but it may also be influenced by low coverage in this proband (~17x), lack of parental srGS data for hifiasm input, and a small number of informative variants near the breakpoints to facilitate phasing.

Overlapping duplications in this region have been reported in gnomAD but are rare (Lek et al. 2016). One individual in Decipher (Patient: 251349) has also been reported with a very similar duplication of uncertain consequence (Deciphering Developmental Disorders 2015). This region spans seven genes, three of which are associated with disease: *ABAT*, *PMM2*, and *USP7*. The first (*ABAT*) is intersected by a duplication breakpoint in proband 4, but all other breakpoints lie within intergenic regions (Supplemental Figure S9, S10). *USP7* is the only gene associated with autosomal dominant disease (Hao-Fountain Syndrome, MIM:616863), but this gene is expected to remain at a copy number of two in this proband whereas LOF is generally the mechanism associated with disease (Hao et al. 2015). Some general features of Hao-Fountain Syndrome overlap this proband, but it is not a strong phenotypic fit. The proband is reported to have moderate ID, seizures, microcephaly, and facial dysmorphisms. Based on the uncertain global structure and molecular consequence of the SV in this proband, in addition to the clinical significance of variation in this region, we have classified this variant as a VUS, with a case-level designation of Uncertain. Note that this variant is pending orthogonal validation. Proband 4 was one of the two individuals who only previously had srES rather than srGS, and this duplication is not easily visualized in srES data (Supplemental Figure S19).

lrGS-informed repeat expansions

Improved variant calling in repeat regions is also a benefit of lrGS (Nurk et al. 2022). In addition to analysis of SNVs and SVs in our standard pipeline, we assessed variant calls in 66 tandem repeat expansion (TRE) regions, including both known disease-associated and candidate disease-associated loci (Supplemental Table S4). We intersected TRE regions of interest with pbsv-called variants in each individual and compared these calls to known pathogenic expansion sizes from the literature. We identified several large heterozygous insertions in repeat regions that, while longer than the expected pathogenicity threshold, were predicted after manual curation to be benign based on their sequence content (Nakamura et al. 2020, Supplemental Figure S20). In two probands, we identified large heterozygous insertions in *RFC1* (MIM: 102579), one of which is benign based on sequence content and one of which is expected to be a pathogenic insertion. However, *RFC1*-associated disease (CANVAS, MIM: 614575) is caused by biallelic expansions, which we did not observe (Supplemental Figure S20), suggesting the latter proband is merely a heterozygous carrier.

In Proband 5, we observed a *de novo* 18-bp alanine tract expansion in *PHOX2B* (MIM: 603851, NM_003924.4:c.741_758dup, p.(Ala255_Ala260dup), Table 2, Supplemental Figure S21), associated with Central Hypoventilation Syndrome, with or without Hirschsprung Disease (MIM: 209880). This disorder was clinically suspected, but variation in *PHOX2B* was missed by initial clinical genetic testing and srGS, which was performed on a PCR+ srGS library. As ExpansionHunter is intended to run on PCR-free srGS data, it was not run on this sample (Dolzhenko et al. 2019). This variant is reported as pathogenic in the literature (Amiel et al. 2003), is a strong match to the proband's observed symptoms, and was confirmed independently by additional clinical testing. We have classified this variant as pathogenic, with a case-level classification of Definitive Diagnostic.

In proband 6, we identified a 270 bp insertion in *AFF3* (MIM:601464, NM_001386135.1:c.-64-281_-64-280insGGC[90], Table 2, Supplemental Figure S22). While missense variants in *AFF3* have been associated with KINSSHIP Syndrome (MIM:619297), an expansion of a CGG-repeat in the promoter of this gene and subsequent hypermethylation of the promoter has recently been reported to be associated with NDDs (Jadhav et al. 2023). Most probands in our cohort (77/96) had both alleles matching hg38 in this region (Supplemental Figure S23). Among the 19 probands harboring non-reference alleles, proband 6 had a 270 bp insertion, and the remaining 18 probands had insertions ranging from 24-84 bp in length (8-28 triplet repeats). Similarly, when comparing to a larger database of in-house HiFi genomes (n=266, “set 2”, see Methods), only 53/266 individuals harbor a non-reference allele; the longest insertions outside of Proband 6 are 105 bp (35 repeats) and 93 bp (31 repeats), each in different individuals, while the median non-reference insertion is 36 bp (12 repeats). Jadhav *et al.* suggest that normal variation ranges to up to ~38 repeat units, with >= 61 repeats being a likely pathogenic threshold. Thus, the 270 bp insertion (90 repeat units) in Proband 6 is well above the normal range and pathogenicity threshold reported by Jadhav et al. and more than twice as long as the second longest insertion in our sample of 266 individuals, which is in the normal range reported by Jadhav et al. (Supplemental Table S4, Supplemental Figure S23). Proband 6 was sequenced as a neonate and presented with intrauterine growth restriction (IUGR) and hypoplastic left heart (HLH); Jadhav et al. reported a wide range of symptoms, including intellectual disability/global developmental delays, seizures, behavioral disturbances, and generalized hypotonia, but the specificity of phenotypic overlap with proband 6 is unclear and neither supports nor refutes pathogenicity. While these results are consistent with pathogenicity of the insertion in proband 6, there remains a need to further replicate and confirm the spectrum of normal and pathogenic variation in *AFF3* repeat lengths. Given this uncertainty and the uncertainty regarding the proband’s cognitive development, we have classified this TRE

as a VUS, and the case-level classification is Uncertain. Also note that this variant is pending orthogonal validation.

lrGS-informed SNV/indels

A *de novo SHANK3* single-base deletion, predicted to lead to a frameshift (NM_033517.1:c.3161delT, p.(Leu1054Argfs*10)), was identified by lrGS in proband 7. While two reads in the srGS data support this deletion, the variant was not called by our srGS variant calling pipeline (Supplemental Figure S24). LOF variation in *SHANK3* (MIM: 606230) is associated with Phelan-McDermid syndrome (MIM: 606232). Features of this syndrome are consistent with the proband's features, and we classified this variant as pathogenic (case-level Definitive Diagnostic). Coverage of this region in short read data does not indicate a systematic coverage deficiency, as mean coverage within 50 bp of this variant in the srGS data for the cohort is 27.9x (n=94), while it is 15.1x for proband 7. Further, this gene is not present in the list of medically relevant genes that tend to be poorly covered by srGS (Wagner et al. 2022). These observations suggest the no-call may have resulted from a stochastic loss of alternative allele reads in the srGS data in this proband. Nevertheless, lrGS has been shown to provide better overall sensitivity and specificity to SNVs and indels in Genome-In-A-Bottle (GIAB) gold-standard datasets compared to srGS (Logsdon et al. 2020; Hiatt et al. 2021) and thus detection failures to variants such as this *SHANK3* event are more likely in srGS data in general.

Reinterpretation of SNVs

The remaining 9 cases had relevant variation identified following lrGS that were equally well detected through reanalysis of existing srGS variant data (Table 2, Supplemental Case Reports). In four cases (Probands 8-11, *HNRNPU, CSNK2B, GNB2, MCF2*) we identified variation in genes that had additional published support for association of the gene or the

variant with disease since the time of the most recent analysis. In another four cases (Probands 12-15, *NOTCH3, AFF4, KCNT2, KIF21A, NRXN1*) we identified variation in established disease genes that conflicted with the published data regarding molecular mechanisms or expected mode of inheritance. Lastly, we identified a variant of interest in *SCN1A* that resulted from a targeted analysis of candidate “poison exon” variants (Proband 16, Felker et al. 2023). We note that in 7 of these 9 cases, variants were identified in an unaffected or mildly affected parent, which was somewhat unexpected in these cases due to suspicion of high penetrance (also see Supplemental Case Reports).

Cohort Analysis of SVs

Individual SV case analyses were performed on a rolling basis and variants were filtered and prioritized using the HiFi variant data generated up to that point in time (see Methods, Supplemental Figure S25). However, we also sought to characterize how filtering SVs by frequency could reduce manual curation burden for future analyses by considering five allele frequency resources (Supplemental Figure S26). First, we created a set of “cohort” SVs by performing SV call merging across all 96 probands using Jasmine (Kirsche et al. 2023) and generating allele counts from the merged set (set 1). We then used a second Jasmine merge step to match cohort SVs with SV frequencies from: an in-house set of 266 HiFi genomes including all cohort probands and parents, samples from other internal projects, and public HiFi data (set 2, see Methods); gnomAD v4 SV frequencies from 63,046 short read samples (set 3, Collins et al. 2020); Human Genome Structural Variant Consortium phase 2 (HGSVC2) assembly-based calls from 18 HiFi samples (set 4, Ebert et al. 2021, https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdel_alt.vcf.gz); and a PacBio-provided set of pbsv calls from 103 HiFi samples from the Human Pangenome Reference Consortium (HPRC) and Genome in a Bottle (GIAB) consortia (set 5, Liao et al. 2023; Zook et al. 2019;

https://github.com/PacificBiosciences/svpack/raw/main/resources/HPRC_GIAB.GRCh38.pbsv.vcf.gz). We note that these data are partially redundant with one another (e.g., set 1 is a partial subset of set 2 and some individual samples such as NA12878 are shared between sets 2-5). We also note that gnomAD allele frequencies should be cautiously applied in the context of lrGS SV annotation, as the srGS artifacts inherent in gnomAD variant calls may inappropriately mask legitimate lrGS variants of interest, particularly given that gnomAD sample size is orders of magnitude larger than the other datasets included in this analysis. However, our goal in this analysis was to maximize filtering impact by including calls from as many datasets and discovery methods as possible. Jasmine merging parameters were heuristically tuned until further increased stringency demonstrated an inflection point in the unique variant count (see Methods, Supplemental Figure S26).

When filtering SVs using these SV frequency datasets, we found that probands had a median of 1,721 “rare” SVs, defined as having an allele frequency <1% in each of the public SV databases (sets 3-5) and an allele count <4 in each internal cohort (sets 1 or 2). A small subset of these, a median of 87 SVs per proband, overlapped a genomic position within 50 bp of a RefSeq exon (Supplemental Table S3). We also filtered to identify “private” SVs as those absent from sets 3-5, having a set 1 allele count of 1, and having a set 2 allele count of <= 2 (allowing for uniparental inheritance). This filtering results in a median of 733 SVs per proband, only 40 of which overlap a position 50 bp of a RefSeq exon (Supplemental Table S3). This analysis includes all pbsv calls, and does not include quality, length, read-depth, or other filters that might further reduce the numbers of variants needing curation.

DISCUSSION

There remains a substantial fraction of rare disease suspected to have genetic causes that is refractory to genomic testing, a finding that has been repeatedly shown across many clinical and research projects (e.g., Srivastava et al. 2019; Baxter et al. 2022). Several non-mutually exclusive hypotheses exist to explain these observations. Environmental risk factors, such as teratogenic exposure or infectious disease, may be relevant to some phenotypes. Multigenic contributors are also likely to explain at least some cases. For example, a small but appreciable fraction of “double diagnoses”, in which an affected individual is observed to harbor two distinct diseases resulting from highly penetrant variation in two distinct genes, has been observed in clinical genomic studies (e.g. Posey et al. 2017). Notably, such discoveries are typically only made when the variation in both genes is independently amenable to a pathogenicity determination (i.e., would be P/LP regardless of P/LP variation in the other gene), and it is possible if not probable that at least some conditions result from combinations of variants in different genes that are not pathogenic in isolation (Papadimitriou et al. 2019). At the further end of this spectrum is a polygenic accumulation of many risk-factor alleles, which is known to be relevant to many common, complex diseases and which may contribute to some rare conditions, as has been suggested for at least a subset of NDDs (Niemi et al. 2018).

We find it likely that a substantial fraction of unexplained rare disease arises from highly penetrant variation that we have not yet been able to precisely identify or confidently interpret. The results from this study are consistent with that hypothesis, with over 15% of probands with previously negative testing being now found to harbor a relevant or potentially relevant genetic variant. Further, these observations are consistent with the general picture of rare disease testing in recent years. With the advent of exome capture and sequencing ~15 years ago (Ng et al. 2010, 2009) and subsequent improvements in cost and efficiency, genome-wide detection of highly penetrant variation has greatly accelerated in recent years (Bamshad et al. 2019; Baxter et al. 2022; Hamosh et al. 2022; Boycott et al. 2022). Long-read genome sequencing represents

379 the next phase of that acceleration by facilitating a substantial increase in variant
380 comprehensiveness and accuracy.

381 We note that the benefits of lrGS for detection of highly penetrant variation derive not
382 just from improved variant sensitivity, but specificity gains as well. Indeed, effective analysis
383 requires not just the ability to detect a variant but to do so accurately and with sufficient
384 specificity for routine and scalable use. For example, effectively all the variation we describe in
385 this study as being newly found within lrGS is, in fact, visible in srGS data, at least at the
386 breakpoint levels. However, being retrospectively visible, once the location and structure of a
387 variant is known to exist, is a much lower bar than the ability to prospectively detect, define,
388 filter, and curate such variation. For example, we describe here a 4.9 kb insertion of SVA and L1
389 sequence into the 3' UTR of *HCFC1*. In retrospective analysis, the Mobile-Element Locator Tool
390 (MELT, Gardner et al. 2017) detected a 1.2 kb SVA mobile element at this location in srGS.
391 However, this call is incorrect with respect to size and sequence composition and has only one
392 read flagged as supporting the right breakpoint. As another example, the breakpoints of the
393 inversion observed to overlap *ZBTB20* were correctly called as breakends by manta (Chen et al.
394 2016). However, each of these were retrospective analyses since MELT calls and breakend
395 calls produced by manta are too numerous for manual curation and thus not used as part of our
396 standard variant calling pipelines (Hiatt et al. 2021,Supplemental Figure S1). Related to this,
397 srGS CNV/SV callers are known to produce many false calls and require strict filtering to reduce
398 calls to a reasonable number for curation. For example, among the 94 samples that had srGS
399 in this study, the raw output from the four different CNV/SV callers that we currently run (see
400 Methods) produces a mean of 9,944 total calls per sample, of which only ~50 are subject to
401 manual curation.

402 Our results are also consistent with other studies of the benefits of lrGS for discovering
403 genetic contributors to disease. In 2021, for example, we showed in a small pilot project that two

of six previously srGS-negative probands harbored clinically relevant variation uniquely interpretable by lrGS (Hiatt et al. 2021). Since that time, several other studies have also used lrGS for molecular diagnosis of rare disease. For example, Cohen and colleagues showed increased yield of GS (both lrGS and srGS) in exome-negative cases (Cohen et al. 2022). Unique discoveries for lrGS included detection of novel repeat expansions of *STARD7* and a compound heterozygous SNV/deletion that was easier to detect in lrGS. However, the increase in diagnostic yield (~13%) was mainly from variants interpretable in either lrGS or srGS. Other studies have also shown the diagnostic value of lrGS, especially in small, well-phenotyped cohorts or families (Sakamoto et al. 2024; Kilich et al. 2024; Audet et al. 2023; Del Gobbo et al. 2023; Fukuda et al. 2023; Miller et al. 2021). Our results are similar to these studies at a high-level. However, some important differences are worth noting. One difference is that we have shown the value of lrGS in singletons. While srGS data was available for most of the probands' parents (79/96, ~82%), only 10/96 probands had parent lrGS data. Our lrGS variant filtering strategies were sufficient to allow curation of proband variants without parental lrGS data. However, once flagged for interest, inheritance data could often be assessed by looking at parental srGS reads. Based on this experience, prioritization of proband lrGS with targeted validation in parents is an effective way to increase diagnostic yield with lrGS while reducing lrGS sequencing needs.

We have also provided a direct, systematic comparison of lrGS to contemporaneously analyzed srGS in previously negative cases. Our results thus allow for the separation and description of clinically relevant variants that are "new" by virtue of benefitting from the unique advantages of lrGS versus those that are "new" by virtue of reanalysis, and which could be detected and accurately interpreted via lrGS or srGS-reanalysis. In that context, we note that the benefits of reanalysis of older srGS data remain considerable. In the results described here, the "new" discoveries in 9 of 16 cases in lrGS were called correctly and interpretable within srGS data. These observations reflect the rapid pace of gene discovery in rare disease (Baxter

et al. 2022; Boycott et al. 2022; Hamosh et al. 2022), and are consistent with other studies. For example, we previously showed that the probability of a negative srGS dataset harboring a clinically relevant variant increased from 1% within one year of a previous analysis to ~22% if more than three years have passed since a previous analysis (Hiatt et al. 2018). Several other studies found similar results, with many reanalysis findings being due to recent publications of new gene-disease associations (Liu et al. 2019; Schobers et al. 2022; Hartley et al. 2023).

One possible optimal path to maximizing overall yield in previously srGS-negative individuals is to include srGS-reanalysis prior to lrGS. However, this reflects a cost/benefit ratio that depends on the cost of the analysis step in relation to the costs of sequencing. While lrGS costs currently remain higher than srGS, as lrGS costs decrease there may reach a point where the cost of variant analysis alone (which is non-trivial and requires both software and compute resources as well as manual curation) is substantial relative to sequencing costs per se and which might favor a process of simply performing lrGS. Further, evaluation of a study design or testing process needs to consider sequencing as only one of several parts of the overall process (e.g., participant recruitment and phenotyping). Again, especially as lrGS costs decline, these factors are likely to increasingly favor use of lrGS.

One caveat to this study is that the results were generated over a period of time with considerable change in lrGS protocols. For example, most probands in this study were sequenced on Sequel IIE machines (n=64) before the Revio (n=32) became available. Given the costs of data production, the 64 Sequel IIE samples were covered at lower-depth (median coverage 24.65X) than the 32 Revio-sequenced samples (median coverage 30.09X, Supplemental Figure S27), which may have reduced our sensitivity to variants in the earlier samples. Further, methylation data were not available in the early period of this study, although methylation calls are now being generated and are available for the most recent 44 probands, which may also impact diagnostic yield. For example, evaluation of the *AFF3* expansion in Proband 6 (Supplemental Figure S22, S23) would benefit from an assessment of methylation

levels at this locus, as hypermethylation, in addition to insertion length, is likely associated with disease risk (Jadhav et al. 2023). Additionally, some probands may harbor methylation alterations that are clinically relevant even in the absence of a pathogenic genetic variant (Aref-Eshghi et al. 2021).

In addition to changes in sequencing, informatic changes have also been considerable over the course of the data generation for this study. While we present and describe a uniform set of variant-calls, assemblies, and annotations (see Methods), analysis of individual samples took place simultaneously with the optimization of variant-calling and annotation pipelines. One particularly relevant change is variant-frequency annotations. While the key strength of lrGS is its ability to see variants that are invisible or poorly detected in srGS, the ability to discriminate genuine highly penetrant variation from the background of benign alleles depends on the ability to annotate and remove alleles that are common in the general population. As the main allele frequency resources are built from srGS data (e.g., Lek et al. 2016), we have limited ability to filter away likely benign alleles among the variants uniquely detected by lrGS. This ability, however, improved as more lrGS datasets were produced over the course of this study (Ebert et al. 2021; Nurk et al. 2022). Projects like the COLORs consortium (https://colorsdb.org/, currently with lrGS data from almost 1400 individuals) are likely to improve frequency annotation in the future and continue to improve variant curation efficacy. Accumulating lrGS data from as many samples and studies as possible is critical for the long-term maximization of lrGS benefits. Further, assembly-based SV calling is an exciting opportunity that will likely improve SV detection and prioritization (Groza et al. 2024).

In sum, rare disease genetics continues to be a rapidly advancing field. With data-sharing (Sobreira et al. 2015; Philippakis et al. 2015; Muenzen et al. 2022) and technology improvement (Wenger et al. 2019), the overall diagnostic yield for individuals with rare disease is increasing at a considerable pace each year. In that context, lrGS has clear benefits over srGS, providing substantial gains to variant specificity and sensitivity, especially for complex and

repeat-associated variants (Schuy et al. 2022). We anticipate that the degree of improvement will widen over time, as sequencing and analysis pipelines mature and as lrGS datasets grow.

METHODS

Short-read sequencing and variant calling

Probands, their parents, and, when appropriate, affected siblings, were enrolled in one of four research studies aimed at identifying genetic causes of rare disease (Bowling et al. 2017; East et al. 2021; Bowling et al. 2022 and Pediatric Genomics (PGEN), unpublished). These studies were monitored by Western IRB and UAB IRB (WIRB 0071, UAB IRB protocols 170303004, 300000328, and 130201001). Short read exome (srES) or short-read genome sequencing (srGS) was performed as described (Bowling et al. 2022; East et al. 2021; Hiatt et al. 2021; Bowling et al. 2017). Briefly, whole blood genomic DNA was isolated using the QIAsymphony (Qiagen), and sequencing libraries were constructed by the HudsonAlpha Genomic Services Lab or the Clinical Services Laboratory, LLC, using a standard protocol that generally included PCR amplification (86/96). Genomes were sequenced at an approximate mean depth of 30X, with at least 80% of base positions reaching 20X coverage. Exomes were sequenced to a mean depth of 71X. For short read reanalysis, srES and srGS reads were aligned to hg38 and small variants and CNVs were called with a uniform pipeline.SNVs/indels and CNVs were curated using an in-house software tool, as previously described (Hiatt et al. 2021). Expected sample relatedness was confirmed with Somalier (v. 0.2.10, (Pedersen et al. 2020) and predicted major genetic ancestries were calculated with peddy (v. 0.4.1, Pedersen and Quinlan 2017). srGS data for participants who consented to controlled-access sharing in NIH-funded studies are available via dbGaP/AnVIL (CSER1: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001089.v4.p1, dbGaP accession phs001089; SouthSeq: https://anvilproject.org/data/studies/phs002307, dbGaP accession phs002307). Supplemental Table S6 provides the applicable srGS sample ID and dbGaP accession of all

individuals who consented to sharing, and the dbGaP/AnVIL metadata may be used to locate relevant parent and familial srGS data.

Long-Read sequencing, variant calling, analysis and de novo assemblies

Long-read sequencing was performed using HiFi (CCS) mode on either a PacBio Sequel II or Revio instrument (Pacific Biosciences of California, Inc.). Libraries were constructed using a SMRTbell Template Prep Kit (V1.0, 2.0 or 3.0) and tightly sized on a SageELF or BluePippin instrument (Sage Science, Beverly, MA, USA). Sequencing was performed using a 2 hour pre-extension with either 24 or 30 hour movie times. The resulting raw data was processed using either the CCS3.4 or CCS4 algorithm, as the latter was released during the course of the study. Comparison of the number of high-quality indel events in a read versus the number of passes confirmed that these algorithms produced comparable results. Probands were sequenced on 2-3 Sequel II or one Revio SMRT cell. Top off sequencing was performed if the sequencing did not meet the desired coverage (>20X). This resulted in an average estimated HiFi read depth of 26.1X (range 16.7-41) of raw, unaligned sequence data for probands. For 10 families, parents were also sequenced on 2-3 Sequel II SMRT cells, with an average estimated depth of 21.6x (range 14-28) of raw, unaligned sequence data for parents. Aligned sequencing metrics are shown in Supplemental Tables S1 and S2. HiFi reads were aligned to the hg38 no-alt analysis set (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz) using pbmm2 (v. 1.10.0, https://github.com/PacificBiosciences/pbmm2). SNVs and small indels, were called with DeepVariant (v. 1.5.0) and used to haplotag the aligned reads with whatshap (v. 1.7)(Martin et al.). Structural variants were called using pbsv v2.9.0 (https://github.com/PacificBiosciences/pbsv), which discovers SVs based on read alignment and split reads. Read-depth-based SV discovery was not performed. For the 10 cases with parent

lrGS, candidate de novo SVs required a proband genotype of 0/1 and parent genotypes of 0/0, with ≥ 6 alternate reads in the proband along with 0 alternate reads and ≥5 reference reads in the parents.

De novo assemblies were generated for all probands using hifiasm (v. 0.15.2 (r334 or v.0.16.1-r375, Cheng et al. 2021). Hifiasm was used to create two assemblies. First, the default parameters were used, followed by two rounds of Racon (v1.4.10) polishing of contigs (Vaser et al. 2017). For cases with parent GS data, trio-binned assemblies were built using k-mers (srGS). The k-mers were generated using yak (v0.1) using the suggested parameters for running a hifiasm trio assembly (k-mer size=31 and Bloom filter size of 2**37). Maternal and paternal contigs went through two rounds of Racon (v1.4.10) polishing. Individual parent assemblies were also built with hifiasm (v0.15.2) using default parameters. The resulting contigs went through two rounds of Racon (v1.4.10) polishing.

Coordinates of breakpoints were defined by a combination of assembly-assembly alignments using minimap2 (Li 2018) followed by use of `bedtools bamToBed` (Quinlan and Hall 2010), visual inspection of CCS read alignments, and BLAT. Dot plots illustrating sequence differences were created using Gepard (Krumsiek et al. 2007).

Structural Variant Merging and Filtering for Case Analysis

Overall structural variant counts in Supplemental Table S1 were generated from the proband pbsv VCFs using BCFtools (v1.15.1)(Danecek et al. 2021) and awk (e.g., `bcftools filter -i 'SVTYPE=="DEL" <input.vcf> | {a[i++]=$1; sum+=$1} END{asort(a); min=a[1]; max=a[i]; if(i%2==1) median=a[int(i/2)+1]; else median=(a[i/2]+a[i/2+1])/2; mean=sum/i; print mean, median, min, max;`).

Counts of SVs > 50 bp in length in Table 1 were created by addition of the BCFtools filters `-e 'SVLEN>-50 && SVLEN<50'`. No quality filtration or read support filtration was used.

An internal allele frequency catalog was constructed and periodically updated using all available internally sequenced HiFi pbsv variant calls and pbsv calls generated from public HiFi sequencing data (n=266 at the end of this analysis, "set 2"). Samples included: a majority of participants from this cohort and our previous NDD pilot cohort (Hiatt et al. 2021, 120/266); samples from HudsonAlpha non-NDD projects (35/266); HudsonAlpha-sequenced data from HG001, HG003, HG004, HG006, and HG007 (5/266); HG00514, HG00731, HG00732, NA19240 from HGSVC2 (Ebert et al. 2021) (4/266); CHM13 (1/266) (Nurk et al. 2022), (https://www.ncbi.nlm.nih.gov/sra/?term=SRX789768*+CHM13)); and the HPRC Year 1 and HPRC_PLUS releases (101/266)(Liao et al. 2023). pbsv calls were merged naively using `bcftools merge` (v. 1.15.1), i.e., merging only variants that were identical in terms of location, reference sequence, and alternate sequence. The internal pbsv allele frequency catalog included affected probands as well as parent/child trios. Supplemental Figures 25 and 26 detail the pipeline for the creation and use of this catalog.

Structural Variant Annotation and Curation

For individual case structural variant analysis, a frequency-filtered subset of the proband's pbsv calls was generated using `bcftools annotate` and `bcftools filter`, requiring that calls overlap a genomic position within 50bp of a RefSeq exon and have an allele count of < 4 in the HudsonAlpha internal allele frequency catalog (set 2, described above). Exon regions were defined as RefSeq exons +/- 50bp (calculated from `bedtools slop`) and were used to restrict output from BCFtools to only calls overlapping those regions (the `-R` command line option). For SVs that span multiple bases of the reference genome (duplications, deletions, and inversions), this filter requires only at least one bp of overlap between the SV span and the target regions.

For SVs that exist at a single base of the reference genome (insertions and breakends), the position must be within the target regions. Each call was visualized using a custom pipeline to automatically generate IGV screenshots (see Supplemental Code for implementation, https://github.com/HudsonAlpha/igv-grapher) (Robinson et al. 2011). Additionally, these filtered pbsv variants were annotated, prioritized, and visualized with SvAnna (v1.0.4, annotations v.2204 or v.2304, Danis et al. 2022) based on manually curated HPO terms for each case. Supplemental Figure S25 details the SV case analysis pipeline from variant calling through curation.

Variant interpretation and orthogonal confirmation

Variant interpretation was performed using ACMG and ClinGen (Richards et al. 2015; Riggs et al. 2020). Variants of interest were either clinically confirmed by the HudsonAlpha Clinical Services Lab, confirmed within a research lab, and/or were supported by short-read data, except where noted in the text (Table 2).

Structural Variant Merging and Filtering for Cohort-level Analysis

In order to provide a cohort-level descriptive analysis of SVs and assess the filtering efficacy of combining multiple allele frequency resources, a more robustly merged cohort catalog (n=96) was constructed using Jasmine. Jasmine allows for merging of similar structural variants that may have non-identical representation in terms of genomic position or variant sequence via spanning a structural variant proximity graph. First, heuristic trials were conducted to determine a set of stringent Jasmine merge options that would still effectively reduce the resulting count of distinct SVs. The centroid merging strategy was chosen instead of the Jasmine default to lessen the risk of over-merging two dissimilar SVs connected through a third SV of intermediate position and length. In a single trial of Jasmine merging strategies, we observed that using the centroid merging strategy increased the unique SV count by ~8,000 while using the most-

stringent clique merging increased the unique SV count by ~99,000. Supplemental Figure S28 shows the effects of the minimum overlap percentage threshold on the unique deletion count and the minimum sequence identity threshold on unique insertion count. Jasmine was ultimately run with options `--centroid_merging --min_overlap=0.65 --min_sequence_id=0.75 --output_genotypes` to create the merged catalog. This cohort catalog was then annotated with five sets of structural variant frequency annotation. Cohort allele frequencies were generated from the merged set with `bcftools +fill-tags` (set 1, n=96). A second Jasmine merge was used to combine the cohort catalog allele count with additional allele frequency resources: the HudsonAlpha internal pbsv callset (described above) (set 2, n=266); gnomAD structural variants (Collins et al. 2020, v4.0, https://gnomad.broadinstitute.org/news/2023-11-v4-structural-variants) (set 3, n=63,046); HGSVC2 structural variants (Ebert et al. 2021) https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v2.0/integrated_callset/variants_freeze4_sv_insdel_alt.vcf.gz) (set 4, n=18); pbsv calls for individuals from the Human Pangenome Reference Consortium and Genome in a Bottle (from PacBio, Liao et al. 2023; Zook et al. 2019, https://github.com/PacificBiosciences/svpack/raw/main/resources/HPRC_GIAB.GRCh38.pbsv.vcf.gz) (set 5, n=103). The second Jasmine merge was run with the prior settings but without the `--output_genotypes` option and records without genotypes were discarded. A custom Python script was used to transfer allele frequency annotations from the annotation source VCFs to the merged VCF based on unique variant identifiers (which were created for each allele frequency resource as needed) and the IDLIST record from Jasmine (see Supplemental Code, https://github.com/HudsonAlpha/jasmine_sv_annotations). The annotated catalog was split into individual VCFs with BCFtools (`bcftools view -I -s <id>`) which were filtered with BCFtools to generate the counts in Supplemental Table S2. The “rare” filter described in

Supplemental Table S3 was defined as excluding variants with any of the following: within-cohort allele count > 3, in-house allele count > 3, gnomAD maximum population allele frequency (POPMAX_AF) >= 1%, HPRC_GIAB allele frequency >=1%, or HGSVC2 allele frequency >= 1%. The "proband-exclusive" filter described in Supplemental Table S3 was defined as excluding variants with any of the following: within-cohort allele count > 1, in-house allele count > 2, gnomAD allele count > 0, HPRC_GIAB allele count > 0, or HGSVC2 allele count > 0. The in-house allele count cutoff was set at 2 in order to account for the fact that some parental samples were included in the in-house frequency database. Exon regions were defined as RefSeq exons +/- 50bp (described above). Variant filtering and counting was performed with BCFtools and command-line tools, e.g., `bcftools view -R refseq_exons_plus50bp.bed.gz -e 'AC>3 | inhouse_pbsv_AC >3 | gnomad4_POPMAX_AF >= 0.01 | hgsvc2_AF >= 0.01 | hprc_giab_pbsv_AF >= 0.01' -H <input_file> | wc -l` for the rare filter and `bcftools view -R refseq_exons_plus50bp.bed.gz -e 'AC>1 | inhouse_pbsv_AC>2 | gnomad4_AC>0 | hgsvc2_AC>0 | hprc_giab_pbsv_AC>0' -H <input_file> | wc -l` for the exclusive filter. No filtering based on SV length was used; counts include insertions and deletions less than 50bp in length that were called by pbsv. Supplemental Figure 26 details the cohort SV merging and analysis pipeline.

Sequencing Metrics

Sequencing metrics were generated from the aligned BAMs using the Sentieon implementations of Picard sequencing metrics (Kendig et al. 2019). Sentieon algorithms QualityYield, AlignmentStat were run with default settings. Sentieon `WgsMetricsAlgo` was run with settings `--include_unpaired true --min_map_qual 0 --min_base_qual 0 --coverage_cap 5000`. Further coverage metrics were generated with cramino using the `--`

659 `phased option` (De Coster and Rademakers 2023). Supplemental Table S5 maps each tool
660 or command and output field name to the corresponding entry in Supplemental Tables S1 and
661 S2. SNV and indel counts and ratios were calculated with rtg vcfstats
662 (https://github.com/RealTimeGenomics/rtg-tools). Summary statistics and graphs were
663 calculated with R (v4.3.1), RStudio (v2023.9.1 build 494), and ggplot2 (v3.4.3)(Wickham 2011;
664 Team 2020; R Core Team 2023). Coverage across regions of interest, such as *SHANK3,* was
665 calculated using `samtools bedcov` with default parameters.
666
667 *Repeat region detection and analysis*
668 We curated a BED file of disease-associated low-complexity repeat regions in 66 genes from
669 previous studies (Hiatt et al. 2021; Cohen et al. 2022 and references therein). Variant calls from
670 pbsv were extracted from these regions +/- 30bp (Supplemental Table S4). Reads were also
671 visualized using the Integrated Genomics Viewer (IGV). Coverage across low complexity repeat
672 regions was calculated using `samtools bedcov` with default parameters. A coverage of at
673 least 8x across the low-complexity region was required for inclusion in Supplemental Table S4
674 and Supplemental Figure S23. We also used TRGT (Dolzhenko et al. 2024) and companion tool
675 TRVZ for visualization of a subset of calls including those for display in Figure S3. TRGT was
676 fed an hg38 reference genome FASTA, BED catalog of tandem repeats, and a sample's BAM to
677 generate a VCF containing genotypes for each tandem repeat from the provided catalog in the
678 given sample and a BAM containing only reads that span the repeat sequences. Output VCFs
679 and BAMs were sorted, indexed, and fed into TRVZ with the same hg38 reference FASTA and
680 BED catalog of tandem repeats to generate pileup plots for any desired variants.
681
682 *3' mRNA-seq*

Total RNA was isolated from blood samples in PAXgene RNA tubes (PreAnalytiX #762165) according to the manufacturer's instructions and stored short-term at -20°C. RNA was isolated using the PAX gene Blood RNA Kit (Qiagen #762164) according to the manufacturer's instructions. Isolated RNA was quantified by the Qubit RNA HS Assay Kit (Thermo Q32855). 425 ng of RNA was used as input for the QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina and UMI Second Strand Synthesis Module for QuantSeq FWD (Illumina, Read 1) from Lexogen (015.96 and 081.96, respectively). Libraries were quantified using the Qubit DNA HS Assay Kit (Thermo Q32854) and visualized with the Bioanalyzer High Sensitivity DNA Analysis kit (Agilent 5067–4626) and 2100 Bioanalyzer Instrument (Agilent). Sequencing was carried out using Illumina NextSeq 75 bp single-end. UMIs were first extracted from the reads with UMI-tools extract with regex. Reads were then trimmed with bbduk (https://sourceforge.net/projects/bbmap/) and aligned to hg38-GENCODEv42 using STAR (Dobin et al. 2013) with the Lexogen recommended parameters for QuantSeq. Bams were deduplicated by UMI and mapping coordinates using UMI-tools dedup(Smith et al. 2017). Count tables were generated using htseq-count with the intersection-nonempty method (Anders et al. 2015).

DATA ACCESS

82 out of 96 families (85%) consented to controlled-access genomic data sharing, including all probands described in detail in the manuscript and listed in Table 2. Consented samples include 100 lrGS genomes (82 probands and 18 parents, comprising 9 full trios and 73 singletons). For participants who consented to controlled-access sharing, the lrGS data generated in this study will be available through AnVIL (https://anvilproject.org/data/studies) under accession number phs003537. Researchers may apply for access via dbGaP (https://www.ncbi.nlm.nih.gov/gap/). Supplemental Table S6 describes the samples and de-identified file identifiers and, when

applicable, their corresponding short-read dbGaP identifiers and other metadata. Custom scripts for automated IGV plot generation and Jasmine structural variant merging and annotation are available in Supplemental Code and on GitHub at https://github.com/HudsonAlpha/igv-grapher https://github.com/HudsonAlpha/igv-grapher.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank the families who have contributed to our studies and our collaborating physicians and clinical staff for recruitment and enrollment for these research studies. Some reagents were provided by PacBio as part of an early-access testing program. The CSER1 project was supported by a grant from the US National Human Genome Research Institute (NHGRI; UM1HG007301). The SouthSeq project (U01HG007301) was supported by the Clinical Sequencing Evidence-Generating Research (CSER2) consortium, which was funded by the National Human Genome Research Institute with co-funding from the National Institute on Minority Health and Health Disparities and the National Cancer Institute. The Alabama Genomic Health Initiative is an Alabama-State earmarked project (F170303004) through the University of Alabama in Birmingham. The PGEN cohort was funded by the Alabama Pediatric Genomics Initiative. lrGS of some samples was supported by a Research Grant from the Muscular Dystrophy Association (MDA 963255). SMH, JMJL, LHH, DRL, ZTB, CRF, MLT, LBB, MW, IRN, and JJ generated, analyzed, and/or interpreted data. JMJL and CRF coordinated data sharing to AnVIL. WVK, EMB, MAL, ACEH, and BRK recruited and enrolled patients and provided clinical assessment. SMH, JMJL, JS, JG, and GMC designed the study and oversaw

733 interpretation of data. SMH, JMJL and GMC drafted and revised the manuscript. All authors
734 read and approved the final version of the manuscript.
735
736

REFERENCES

Amiel J, Laudier B, Attié-Bitach T, Trang H, De Pontual L, Gener B, Trochet D, Etchevers H, Ray P, Simonneau M, et al. 2003. Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome. *Nat Genet* **33**: 459–461. https://pubmed.ncbi.nlm.nih.gov/12640453/ (Accessed February 5, 2024).

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169. https://dx.doi.org/10.1093/bioinformatics/btu638 (Accessed August 1, 2024).

Aref-Eshghi E, Kerkhof J, Pedro VP, France G DI, Barat-Houari M, Ruiz-Pallares N, Andrau JC, Lacombe D, Van-Gils J, Fergelot P, et al. 2021. Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders. *Am J Hum Genet* **108**: 1161–1163. https://pubmed.ncbi.nlm.nih.gov/34087165/ (Accessed February 27, 2024).

Audet S, Triassi V, Gelinas M, Legault-Cadieux N, Ferraro V, Duquette A, Tetreault M. 2023. Integration of multi-omics technologies for molecular diagnosis in ataxia patients. *Front Genet* **14**: 1304711. http://www.ncbi.nlm.nih.gov/pubmed/38239855 (Accessed February 13, 2024).

Bamshad MJ, Nickerson DA, Chong JX. 2019. Mendelian Gene Discovery: Fast and Furious with No End in Sight. *Am J Hum Genet* **105**: 448–455. https://pubmed.ncbi.nlm.nih.gov/31491408/ (Accessed July 19, 2022).

Baxter SM, Posey JE, Lake NJ, Sobreira N, Chong JX, Buyske S, Blue EE, Chadwick LH, Coban-Akdemir ZH, Doheny KF, et al. 2022. Centers for Mendelian Genomics: A decade of facilitating gene discovery. *Genet Med* **24**: 784–797. https://pubmed.ncbi.nlm.nih.gov/35148959/ (Accessed April 16, 2023).

Bowling KM, Thompson ML, Amaral MD, Finnila CR, Hiatt SM, Engel KL, Cochran JN, Brothers KB, East KM, Gray DE, et al. 2017. Genomic diagnosis for children with intellectual disability and/or developmental delay. *Genome Med* **9**: 43. http://www.ncbi.nlm.nih.gov/pubmed/28554332.

Bowling KM, Thompson ML, Finnila CR, Hiatt SM, Latner DR, Amaral MD, Lawlor JMJ, East KM, Cochran ME, Greve V, et al. 2022. Genome sequencing as a first-line diagnostic test for hospitalized infants. *Genet Med* **24**: 851–861. https://pubmed.ncbi.nlm.nih.gov/34930662/ (Accessed May 29, 2023).

Boycott KM, Azzariti DR, Hamosh A, Rehm HL. 2022. Seven years since the launch of the Matchmaker Exchange: The evolution of genomic matchmaking. *Hum Mutat* **43**: 659–667. https://pubmed.ncbi.nlm.nih.gov/35537081/ (Accessed July 19, 2022).

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220–2. https://github.com/Illumina/manta. (Accessed June 22, 2020).

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**: 170–175. https://pubmed.ncbi.nlm.nih.gov/33526886/ (Accessed February 6, 2024).

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B, et al. 2022. Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genetics in Medicine* **24**: 1336–1348. https://pubmed.ncbi.nlm.nih.gov/35305867/ (Accessed August 24, 2022).

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera A V., Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451. https://pubmed.ncbi.nlm.nih.gov/32461652/ (Accessed February 15, 2024).

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: 1–4. https://dx.doi.org/10.1093/gigascience/giab008 (Accessed August 1, 2024).

Danis D, Jacobsen JOB, Balachandran P, Zhu Q, Yilmaz F, Reese J, Haimel M, Lyon GJ, Helbig I, Mungall CJ, et al. 2022. SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing. *Genome Med* **14**. https://pubmed.ncbi.nlm.nih.gov/35484572/ (Accessed February 12, 2024).

De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**. https://dx.doi.org/10.1093/bioinformatics/btad311 (Accessed August 1, 2024).

Deciphering Developmental Disorders S. 2015. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**: 223–228. https://www.ncbi.nlm.nih.gov/pubmed/25533962.

Del Gobbo GF, Wang X, Couse M, Mackay L, Goldsmith C, Marshall AE, Liang Y, Lambert C, Zhang S, Dhillon H, et al. 2023. Long-read genome sequencing reveals a novel intronic retroelement insertion in NR5A1 associated with 46,XY differences of sexual development.

Am J Med Genet A. http://www.ncbi.nlm.nih.gov/pubmed/38131126 (Accessed February 13, 2024).

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15. /pmc/articles/PMC3530905/ (Accessed August 1, 2024).

Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, Emig-Agius D, Gross A, Narzisi G, Bowman B, et al. 2019. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**: 4754–4756. https://pubmed.ncbi.nlm.nih.gov/31134279/ (Accessed June 25, 2024).

Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and visualization of tandem repeats at genome scale. *Nature Biotechnology 2024* 1–9. https://www.nature.com/articles/s41587-023-02057-3 (Accessed February 15, 2024).

East KM, Kelley W V., Cannon A, Cochran ME, Moss IP, May T, Nakano-Okuno M, Sodeke SO, Edberg JC, Cimino JJ, et al. 2021. A state-based approach to genomics for rare disease and population screening. *Genet Med* **23**: 777–781. https://pubmed.ncbi.nlm.nih.gov/33244164/ (Accessed July 26, 2022).

Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science (1979)* **372**.

Felker SA, Lawlor JMJ, Hiatt SM, Thompson ML, Latner DR, Finnila CR, Bowling KM, Bonnstetter ZT, Bonini KE, Kelly NR, et al. 2023. Poison exon annotations improve the yield of clinically relevant variants in genomic diagnostic testing. *Genet Med* 100884. https://pubmed.ncbi.nlm.nih.gov/37161864/ (Accessed May 29, 2023).

Fukuda H, Mizuguchi T, Doi H, Kameyama S, Kunii M, Joki H, Takahashi T, Komiya H, Sasaki M, Miyaji Y, et al. 2023. Long-read sequencing revealing intragenic deletions in exome-negative spastic paraplegias. *J Hum Genet* **68**: 689–697.

Gardner EJ, Lam VK, Harris DN, Chuang NT, Scott EC, Stephen Pittard W, Mills RE, Devine SE. 2017. The mobile element locator tool (MELT): Population-scale mobile element discovery and biology. *Genome Res* **27**: 1916–1929.

Groza C, Schwendinger-Schreck C, Cheung WA, Farrow EG, Thiffault I, Lake J, Rizzo WB, Evrony G, Curran T, Bourque G, et al. 2024. Pangenome graphs improve the analysis of structural variants in rare genetic diseases. *Nat Commun* **15**: 657. https://pubmed.ncbi.nlm.nih.gov/38253606/ (Accessed February 13, 2024).

Hamosh A, Wohler E, Martin R, Griffith S, Rodrigues E da S, Antonescu C, Doheny KF, Valle D, Sobreira N. 2022. The impact of GeneMatcher on international data sharing and collaboration. *Hum Mutat* **43**: 668–673. https://pubmed.ncbi.nlm.nih.gov/35170833/ (Accessed July 19, 2022).

Hao YH, Fountain MD, Fon Tacer K, Xia F, Bi W, Kang SHL, Patel A, Rosenfeld JA, Le Caignec C, Isidor B, et al. 2015. USP7 Acts as a Molecular Rheostat to Promote WASH-Dependent Endosomal Protein Recycling and Is Mutated in a Human Neurodevelopmental Disorder. *Mol Cell* **59**: 956–969. https://pubmed.ncbi.nlm.nih.gov/26365382/ (Accessed February 11, 2024).

Hartley T, Soubry É, Acker M, Osmond M, Couse M, Gillespie MK, Ito Y, Marshall AE, Lemire G, Huang L, et al. 2023. Bridging clinical care and research in Ontario, Canada: Maximizing diagnoses from reanalysis of clinical exome sequencing data. *Clin Genet* **103**: 288–300.

Hiatt SM, Amaral MD, Bowling KM, Finnila CR, Thompson ML, Gray DE, Lawlor JMJ, Cochran JN, Bebin EM, Brothers KB, et al. 2018. Systematic reanalysis of genomic data improves quality of variant interpretation. *Clin Genet* **94**.

Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Partridge EC, Boston LB, Williams M, Plott CB, Jenkins J, et al. 2021. Long-read genome sequencing for the molecular diagnosis of neurodevelopmental disorders. *HGG Adv* **2**: 100023. http://www.ncbi.nlm.nih.gov/pubmed/33937879 (Accessed June 9, 2021).

Jadhav B, Garg P, van Vugt JJ, Ibanez K, Gagliardi D, Lee W, Shadrina M, Mokveld T, Dolzhenko E, Martin-Trujillo A, et al. 2023. A phenome-wide association study of methylated GC-rich repeats identifies a GCC repeat expansion in AFF3 as a significant cause of intellectual disability. *medRxiv*. https://pubmed.ncbi.nlm.nih.gov/37205357/ (Accessed February 6, 2024).

Juven A, Nambot S, Piton A, Jean-Marçais N, Masurel A, Callier P, Marle N, Mosca-Boidron AL, Kuentz P, Philippe C, et al. 2020. Primrose syndrome: a phenotypic comparison of patients with a ZBTB20 missense variant versus a 3q13.31 microdeletion including ZBTB20. *Eur J Hum Genet* **28**: 1044–1055. https://pubmed.ncbi.nlm.nih.gov/32071410/ (Accessed February 15, 2024).

Kendig KI, Baheti S, Bockol MA, Drucker TM, Hart SN, Heldenbrand JR, Hernaez M, Hudson ME, Kalmbach MT, Klee EW, et al. 2019. Sentieon DNASeq Variant Calling Workflow Demonstrates Strong Computational Performance and Accuracy. *Front Genet* **10**. https://pubmed.ncbi.nlm.nih.gov/31481971/ (Accessed February 15, 2024).

Kilich G, Hassey K, Behrens EM, Falk M, Vanderver A, Rader DJ, Cahill PJ, Raper A, Zhang Z, Westerfer D, et al. 2024. Kagami Ogata syndrome: a small deletion refines critical region

for imprinting. *NPJ Genom Med* **9**: 5. http://www.ncbi.nlm.nih.gov/pubmed/38212313 (Accessed February 13, 2024).

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC. 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20**: 408–417. https://pubmed.ncbi.nlm.nih.gov/36658279/ (Accessed February 15, 2024).

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. **23**: 1026–1028. http://mips.gsf.de/services/analysis/gepard (Accessed May 18, 2020).

Lek M, Karczewski KJ, Minikel E V, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291. https://www.ncbi.nlm.nih.gov/pubmed/27535533.

Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. https://github.com/ruanjue/smartdenovo; (Accessed September 7, 2020).

Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**.

Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, Rosenfeld J, Magoulas PL, Braxton A, Ward P, et al. 2019. Reanalysis of Clinical Exome Sequencing Data. *New England Journal of Medicine* **380**: 2478–2480.

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21**: 597–614. https://www.nature.com/articles/s41576-020-0236-x (Accessed November 4, 2020).

Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, Handsaker RE, Hall S, Pionzio A, Schatz MC, et al. 2024. Utility of long-read sequencing for All of Us. *Nat Commun* **15**: 837. https://pubmed.ncbi.nlm.nih.gov/38281971/ (Accessed February 5, 2024).

Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schönhuth A, Marschall T. WhatsHap: fast and accurate read-based phasing. https://doi.org/10.1101/085050 (Accessed August 1, 2024).

Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, Lewis AP, Fuerte EPA, Paschal CR, Walsh T, et al. 2021. Targeted long-read sequencing identifies missing disease-causing variation. *Am J Hum Genet* **108**: 1436–1449. https://pubmed.ncbi.nlm.nih.gov/34216551/ (Accessed February 13, 2024).

952 Muenzen KD, Amendola LM, Kauffman TL, Mittendorf KF, Bensen JT, Chen F, Green R, Powell
953 BC, Kvale M, Angelo F, et al. 2022. Lessons learned and recommendations for data
954 coordination in collaborative research: The CSER consortium experience. *Human Genetics*
955 *and Genomics Advances* **3**: 100120.
956 http://www.cell.com/article/S2666247722000367/fulltext (Accessed August 1, 2024).
957
958 Nakamura H, Doi H, Mitsuhashi S, Miyatake S, Katoh K, Frith MC, Asano T, Kudo Y, Ikeda T,
959 Kubota S, et al. 2020. Long-read sequencing identifies the pathogenic nucleotide repeat
960 expansion in RFC1 in a Japanese case of CANVAS. *J Hum Genet* **65**: 475–480.
961 https://www.nature.com/articles/s10038-020-0733-y (Accessed June 27, 2021).
962
963 Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs
964 EW, Nickerson DA, et al. 2010. Exome sequencing identifies the cause of a mendelian
965 disorder. *Nat Genet* **42**: 30–35. https://pubmed.ncbi.nlm.nih.gov/19915526/ (Accessed
966 February 27, 2024).
967
968 Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M,
969 Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel
970 sequencing of 12 human exomes. *Nature* **461**: 272–276.
971 https://pubmed.ncbi.nlm.nih.gov/19684571/ (Accessed February 13, 2024).
972
973 Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, McAloney K, McRae J,
974 Radford EJ, Yu S, et al. 2018. Common genetic variants contribute to risk of rare severe
975 neurodevelopmental disorders. *Nature* **562**: 268–271.
976 https://pubmed.ncbi.nlm.nih.gov/30258228/ (Accessed February 11, 2024).
977
978 Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N,
979 Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science*
980 **376**: 44–53. https://pubmed.ncbi.nlm.nih.gov/35357919/ (Accessed January 28, 2024).
981
982 Papadimitriou S, Gazzo A, Versbraegen N, Nachtegael C, Aerts J, Moreau Y, Van Dooren S,
983 Nowé A, Smits G, Lenaerts T. 2019. Predicting disease-causing variant combinations. *Proc*
984 *Natl Acad Sci U S A* **116**: 11878–11887. https://pubmed.ncbi.nlm.nih.gov/31127050/
985 (Accessed February 14, 2024).
986
987 Pauper M, Kucuk E, Wenger AM, Chakraborty S, Baybayan P, Kwint M, van der Sanden B,
988 Nelen MR, Derks R, Brunner HG, et al. 2021. Long-read trio sequencing of individuals with
989 unsolved intellectual disability. *European Journal of Human Genetics* **29**: 637–648.
990
991 Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, Bronner MP, Underhill
992 HR, Quinlan AR. 2020. Somalier: rapid relatedness estimation for cancer and germline
993 studies using efficient genome sketches. *Genome Med* **12**.
994 https://pubmed.ncbi.nlm.nih.gov/32664994/ (Accessed February 15, 2024).
995

Pedersen BS, Quinlan AR. 2017. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am J Hum Genet* **100**: 406–413. https://pubmed.ncbi.nlm.nih.gov/28190455/ (Accessed February 15, 2024).

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske OJ, Carey K, Doll C, et al. 2015. The Matchmaker Exchange: A Platform for Rare Disease Gene Discovery. *Hum Mutat* **36**: 915–921. http://doi.wiley.com/10.1002/humu.22858 (Accessed November 25, 2018).

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal snp and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**: 983.

Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, Walkiewicz M, Bi W, Xiao R, Ding Y, et al. 2017. Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N Engl J Med* **376**: 21–31. https://pubmed.ncbi.nlm.nih.gov/27959697/ (Accessed February 15, 2024).

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. https://pubmed.ncbi.nlm.nih.gov/20110278/ (Accessed August 1, 2024).

R Core Team. 2023. R Core Team 2023 R: A language and environment for statistical computing. R foundation for statistical computing. https://www.R-project.org/. *R Foundation for Statistical Computing*.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**: 405–424. https://www.ncbi.nlm.nih.gov/pubmed/25741868.

Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, Raca G, Ritter DI, South ST, Thorland EC, et al. 2020. 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). *Genetics in Medicine* **22**: 245–257. https://pubmed.ncbi.nlm.nih.gov/31690835/ (Accessed September 8, 2020).

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26. https://pubmed.ncbi.nlm.nih.gov/21221095/ (Accessed February 15, 2024).

Sakamoto M, Kurosawa K, Tanoue K, Iwama K, Ishida F, Watanabe Y, Okamoto N, Tsuchida N, Uchiyama Y, Koshimizu E, et al. 2024. A heterozygous germline deletion within USP8 causes severe neurodevelopmental delay with multiorgan abnormalities. *J Hum Genet* **69**: 85–90. http://www.ncbi.nlm.nih.gov/pubmed/38030753 (Accessed February 13, 2024).

Sanghvi R V., Buhay CJ, Powell BC, Tsai EA, Dorschner MO, Hong CS, Lebo MS, Sasson A, Hanna DS, McGee S, et al. 2018. Characterizing reduced coverage regions through comparison of exome and genome sequencing data across 10 centers. *Genet Med* **20**: 855–866. https://pubmed.ncbi.nlm.nih.gov/29144510/ (Accessed February 5, 2024).

Schobers G, Schieving JH, Yntema HG, Pennings M, Pfundt R, Derks R, Hofste T, de Wijs I, Wieskamp N, van den Heuvel S, et al. 2022. Reanalysis of exome negative patients with rare disease: a pragmatic workflow for diagnostic applications. *Genome Med* **14**.

Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. 2022. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends in Genetics* **0**. http://www.cell.com/article/S0168952522001457/fulltext (Accessed July 31, 2022).

Smith T, Heger A, Sudbery I. 2017. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**: gr.209601.116. https://genome.cshlp.org/content/early/2017/01/18/gr.209601.116 (Accessed August 1, 2024).

Sobreira N, Schiettecatte F, Valle D, Hamosh A. 2015. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat* **36**: 928–30. http://doi.wiley.com/10.1002/humu.22844 (Accessed November 25, 2018).

Sobreira NLM, Arachchi H, Buske OJ, Chong JX, Hutton B, Foreman J, Schiettecatte F, Groza T, Jacobsen JOB, Haendel MA, et al. 2017. Matchmaker Exchange. *Curr Protoc Hum Genet* **95**: 9.31.1-9.31.15. http://doi.wiley.com/10.1002/cphg.50 (Accessed November 25, 2018).

Srivastava S, Love-Nichols JA, Dies KA, Ledbetter DH, Martin CL, Chung WK, Firth H V., Frazier T, Hansen RL, Prock L, et al. 2019. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med* **21**: 2413–2421. https://pubmed.ncbi.nlm.nih.gov/31182824/ (Accessed December 6, 2022).

Team Rs. 2020. RStudio: Integrated Development for R. *RStudio, Inc, Boston, MA*.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**: 737–746. https://pubmed.ncbi.nlm.nih.gov/28100585/ (Accessed August 1, 2024).

Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, Hwang YC, Gupta R, Wenger AM, Rowell WJ, et al. 2022. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**: 672–680. https://pubmed.ncbi.nlm.nih.gov/35132260/ (Accessed February 15, 2024).

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155–1162.

Wickham H. 2011. ggplot2. *Wiley Interdiscip Rev Comput Stat* **3**.

Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al. 2019. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**: 561–566.

FIGURE LEGENDS

Figure 1. A *de novo*, 4 Mb paracentric inversion in proband 1, affecting *ZBTB20*. A, B. Visualization of a subset of proband and parent reads in IGV at the 5' (A, Chr 3:110,477,108-110,477,357) and 3' (B, Chr 3:114,639,120-114,639,284) breakpoints (black arrowheads) indicate a *de novo* event. Reads in A show the SV is present on the proband's paternal allele. C. Schematic of the proband's inversion. Section B-C is inverted, and junctions 1 and 2 (jct1, jct2) are displayed. The inversion disrupts *ZBTB20*, with a promoter (P) and exons 1-6 of NM_001348800.3 (yellow boxes) remaining at the distal end of the chromosome near jct2/D and exons 7-12 of NM_001348800.3 (green boxes) moving upstream to the breakpoint at the more proximal end of the chromosome (near jct1/A). Sequence level resolution of the inversion is also shown. D. Alignment of the proband's assembled paternal contig versus the reference genome supports the inversion.

Figure 2. Two *ALS2* deletions in trans in Proband 2. A. Visualization of proband and parent reads in IGV (Chr 2:201,719,400-201,741,000) indicate two overlapping deletions in *ALS2*; a smaller maternal deletion (pink bar) and a larger paternal deletion (blue bar/arrow). Note that the three breakpoints shown here overlap *Alu* sequence in the RepeatMasker track. Alignment

of the proband's assembled maternal (B) or paternal contig (C) versus the reference genome support the two deletions (red dashed lines).

Figure 3. Proband 3 has a 4 kb insertion in the 3' UTR of HCFC1. A. The proband's family has a history of X-linked intellectual disability, as the proband (arrow) and two other male relatives (gray squares) are affected. B. Model of the relative length of the insertion in the 3' UTR NM_005334.3. Only exons 24-26 are shown. C. The hemizygous insertion is likely inherited from a heterozygous carrier mother, as indicated by srGS reads. Note that some reads contain unaligned ends (multicolored bases) while some span the entire insertion (purple line at the 5' end of the TSD); all the proband's reads support the insertion. A target site duplication (TSD) is a hallmark of an L1-mediated insertion.

Table 1. Demographics, and sequencing and assembly metrics for the cohort

Sex	Male	63 (66%)		
	Female	33 (34%)		
Predicted Major Genetic Ancestry	European (EUR)	71.88%		
	African (AFR)	20.83%		
	Admixed American (AMR)	3.13%		
	Southeast Asian (SAS)	1.04%		
	Unspecified Admixed (UNKNOWN)	3.13%		
		Median	**Min**	**Max**
Sequencing Metrics	Sequenced Bases (Gb)	80.9	53.5	131.5
	Mean Read Length (Sequenced)	16,771	11,295	21,412
	Median Coverage (X)	27	18	44
	Percent Covered at 10x	97.80%	95.10%	99.00%
	Percent Covered at 20x	84.90%	36.10%	97.50%
Proband Assembly N50 (Mb)	All contigs	29.05	2.10	71.30
	Maternal contigs	29.30	2.10	71.30
	Paternal contigs	28.70	2.20	67.90
	hap1 contigs	31.70	18.20	42.90
	hap2 contigs	27.85	14.40	40.80
Small Variant Metrics (DeepVariant)	SNVs	4,404,564	3,997,350	5,299,670
	Total indels	970,031	926,286	1,153,616
Structural Variant Metrics (pbsv)	Total Structural Variants	55,586	53,834	65,120
	Deletion	24,757	23,799	29,531
	Duplication	3,017	2,820	3,649
	Insertion	27,594	26,441	32,437
	Inversion	121	99	155
	Breakend	193	124	312
	Structural Variants >50bp	25,218	23,994	29,159
	Deletion	10,303	9,804	12,306
	Duplication	641	536	742
	Insertion	14,012	13,344	16,159
	Inversion	121	99	155
	Breakend	193	124	312

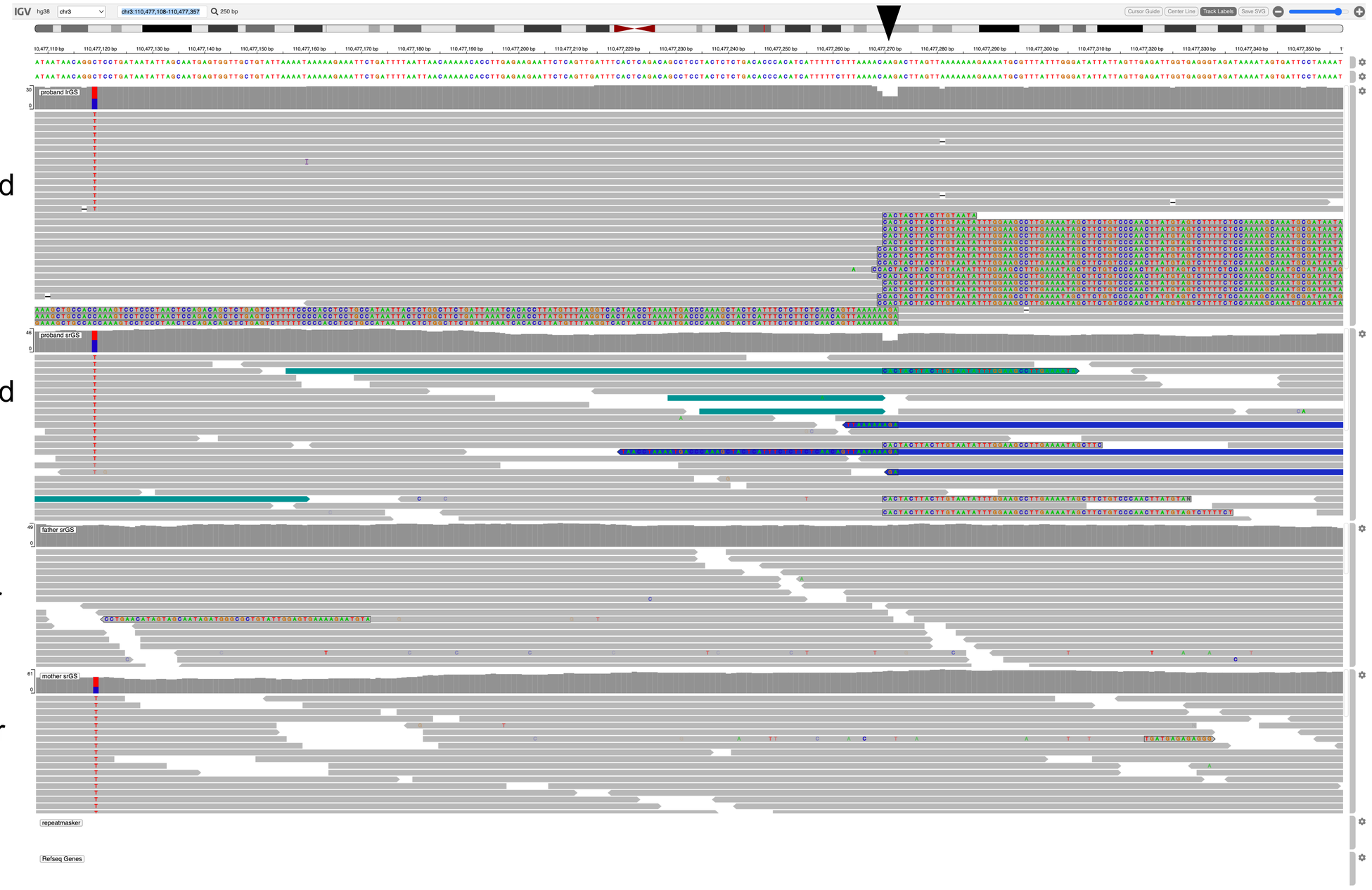
Table 2. Variants identified by long-read sequencing.

Proband ID	Gene(s) Affected	HGVS Nomenclature	Inheritance	Variant Classification	Case-Level Classification	ACMG/ClinGen Evidence Codes	SV, SNV or TRE	Step of srGS loss*	Orthogonal Validation
1	*ZBTB20*	NC_000003.12:g.110477273_114639202_inv	de novo	LP	Likely Diagnostic	2C(+1.00), 5A (+0.15)	SV	Filtering/QC	Yes, Research Sanger of breakpoints
2	*ALS2*	NC_000002.12:g.201720435-201725085_del; NC_000002.12:g.200115181-201739349_del	biparental	LP; P	Definitive Diagnostic	2E (+0.90);2D-4 (+0.90), 3B (+0.45)	SVx2	Filtering/QC	Pending, supported by srGS data
3	*HCFC1*	NC_000023.11:g.153948602_ins4902, NM_005334.3:c.*745_ins4902	maternal (X-linked)	VUS	Uncertain	NA	SV	Variant Calling	Yes, Research PCR amplification of breakpoints.
4	*ABAT, PMM2, USP7,* etc.	NC_000016.9:g.(8742452_9220783)dup_ins[(8742452_8879961)_(9000190_9220783)]	de novo	VUS	Uncertain	2K (+0.30)	SV	Variant Calling	Pending
5	*PHOX2B*	NM_003924.4:c.741_758dup, p.(Ala255_Ala260dup)	unknown[#]	P	Definitive Diagnostic	PS4_M, PM1_Strong, PM2_Moderate, PM6_Moderate[#]	TRE	Variant Calling	Yes, Clinical Testing*
6	*AFF3*	NM_001386135.1:c.-64-281_-64-280insGGC[90]	unknown	VUS	Uncertain	NA	TRE	Variant Calling	Pending
7	*SHANK3*	NM_033517.1:c.3161delT, p.(Lys1054Argfs*10)	de novo	P	Definitive Diagnostic	PVS1_VeryStrong, PS2_Strong, PM2_Moderate	SNV	Variant Calling	Yes, Clinical Sanger
8	*HNRNPU*	NM_031844.3:c.660_661dupAGGCGGCGGA, p.(Gly221ArgfsTer25)	de novo	P	Definitive Diagnostic	PVS1_VeryStrong, PS2_Strong, PM2_Moderate	SNV	Curation (Gene-Disease Association)	Yes, Clinical Sanger
9	*CSNK2B*	NM_001320.6:c.202C>T, p.(Gln68Ter)	paternal	LP	Likely Diagnostic	PVS1_VeryStrong, PM2_Moderate	SNV	Curation (Gene-Disease Association)	Yes, Clinical Sanger
10	*GNB2*	NM_005273.4:c.217G>A, p.(Ala73Thr)	maternal	LP	Uncertain	PS4_Moderate, PP2_Supporting, PP3_Supporting	SNV	Curation (Gene-Disease Association)	Yes, Clinical Sanger
11	*MCF2*	NM_005369.5:c.2234G>T, p.(Gly745Val)	maternal (X-linked)	VUS	Uncertain	PM2	SNV	Curation (Gene-Disease Association)	Yes, Clinical Sanger
12	*NOTCH3*	NM_000435.3c.6409_6410delCT, p.(Leu2137GlyfsTer104)	paternal	LP	Likely Diagnostic	PVS1_Strong, PM2_Moderate	SNV	Curation (unexpected mechanism)	Yes, Clinical Sanger
13	*AFF4*	NM_014423.4, c.879delA, p.(His294IlefsTer5)	paternal	VUS	Uncertain	PM2	SNV	Curation (unexpected mechanism)	Yes, Clinical Sanger
14	*KCNT2, KIF21A*	NC_000001.11:g.196329420-196344697_DUP, NM_017641.3:c.847C>T, p.(Arg283Cys); NM_017641.3:c.706C>T, p.(Gln236Ter)	paternal/ biparental	VUS, VUS;LP	Uncertain	2I (+0.45); PM2, PP3; PVS1, PM2	SV; SNV(x2)	Curation (unexpected mechanism)	SV Pending, supported by srGS data; SNVs-Clinical Sanger
15	*NRXN1*	NC_000002.12:49922063_49928691del	de novo	VUS	Uncertain	2E (+0.30), 4C (+0.15), 4M (+0.30)	SV	Curation (unexpected mechanism)	Pending, supported by srGS data
16	*SCN1A*	NM_001165963.4:c.4003-603T>C	paternal	VUS	Uncertain	PM2_Moderate	SNV	Curation (noncoding variation)	Yes, Clinical Sanger

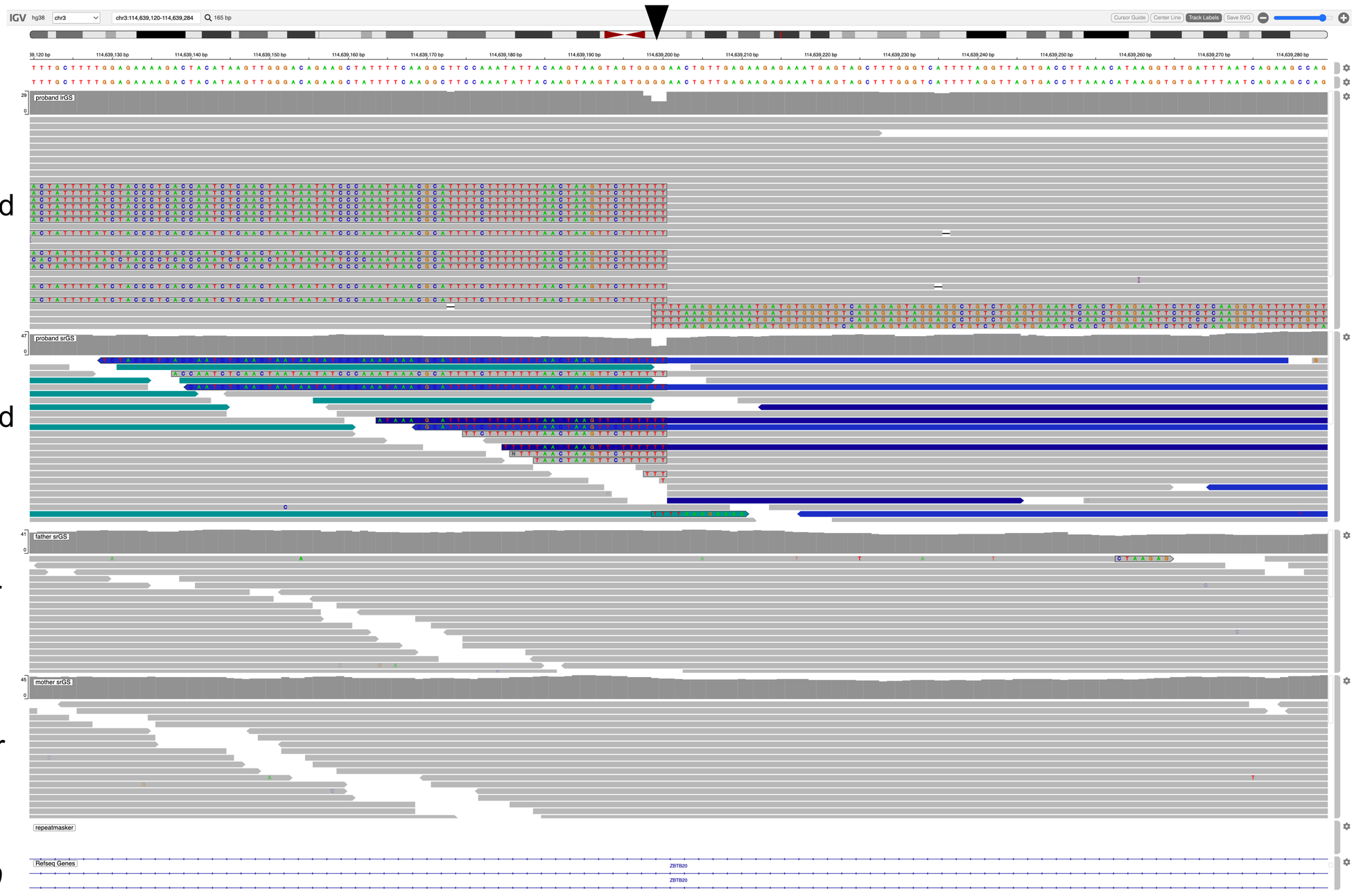
SV, Structural variant; SNV, single nucleotide variant; TRE, tandem repeat expansion. P, Pathogenic; LP, Likely Pathogenic; VUS, Variant of Uncertain Significance; NA, not applicable. *Filtering/QC, our filtering or prioritization strategy did not accurately present this for curation (or did not at all); Variant Calling, the variant was not called or was called incorrectly/inaccurately; Curation, manual curation did not result in flagging of the variant(s). [#]Independent clinical testing indicated that the PHOX2B expansion was de novo; we only sequenced the proband in our research study.

Figure 1

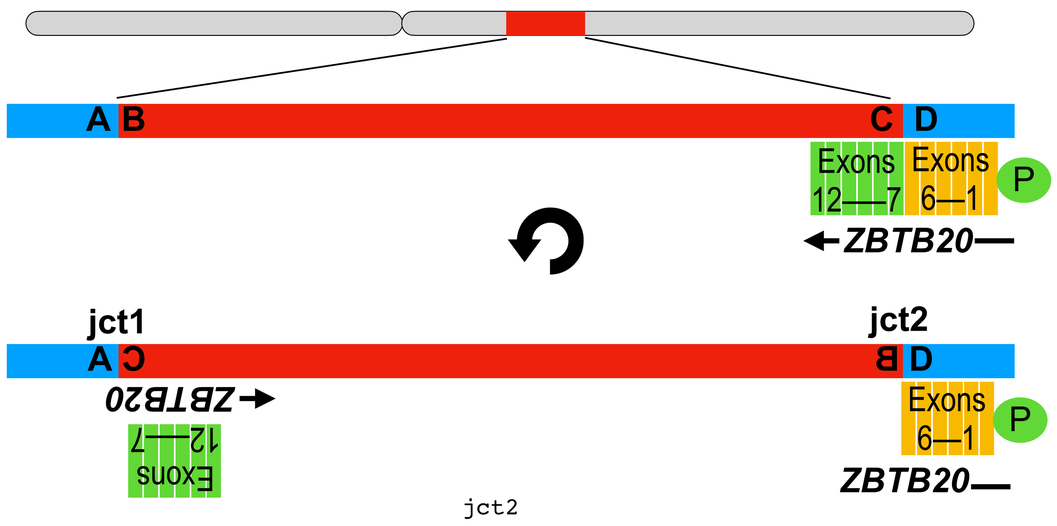
A



B



C



jct1

ACATCATTTTTCTTTAAAACAAGACTTAGTTAAAAAAAGAAAA

ACATCATTTTTCTTTAAAACCACTACTTACTTGTAATATTTGG

TTCTCTTCTCAACAGTTCCCCACTACTTACTTGTAATATTTGG

jct2

TAAACGCATTTTCTTTTTTTAACTAAGTCTTGTTTTAAAGAAAAATGATGT

TAAACGCATTTTCTTTTTTTAACTAAGTTCTTTTTTTAACTGTTGAGAAGAG

TTCAAGGCTTCCAAATATTACAAGTAAGTAGTGGGGAACTGTTGAGAAGAG

D

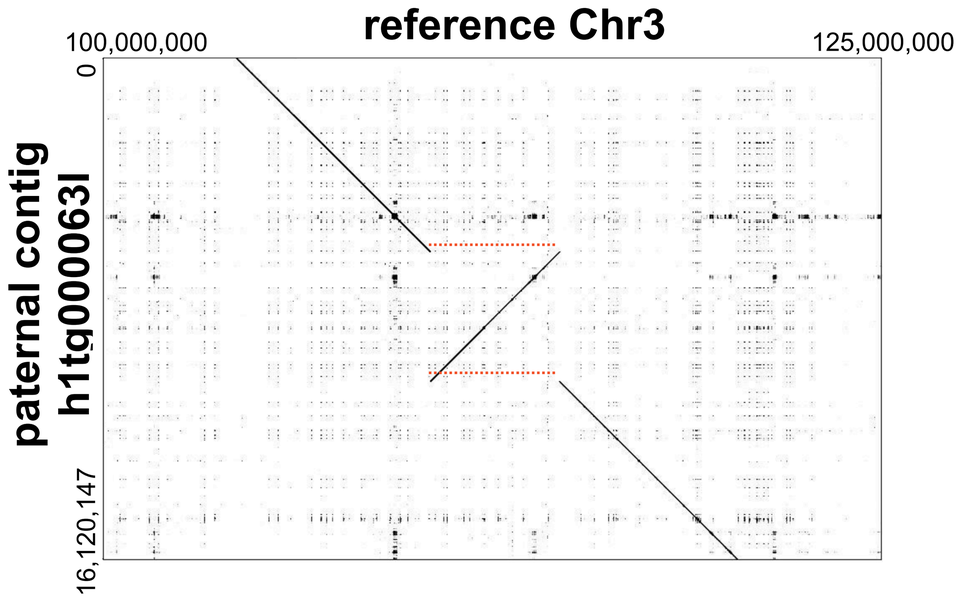
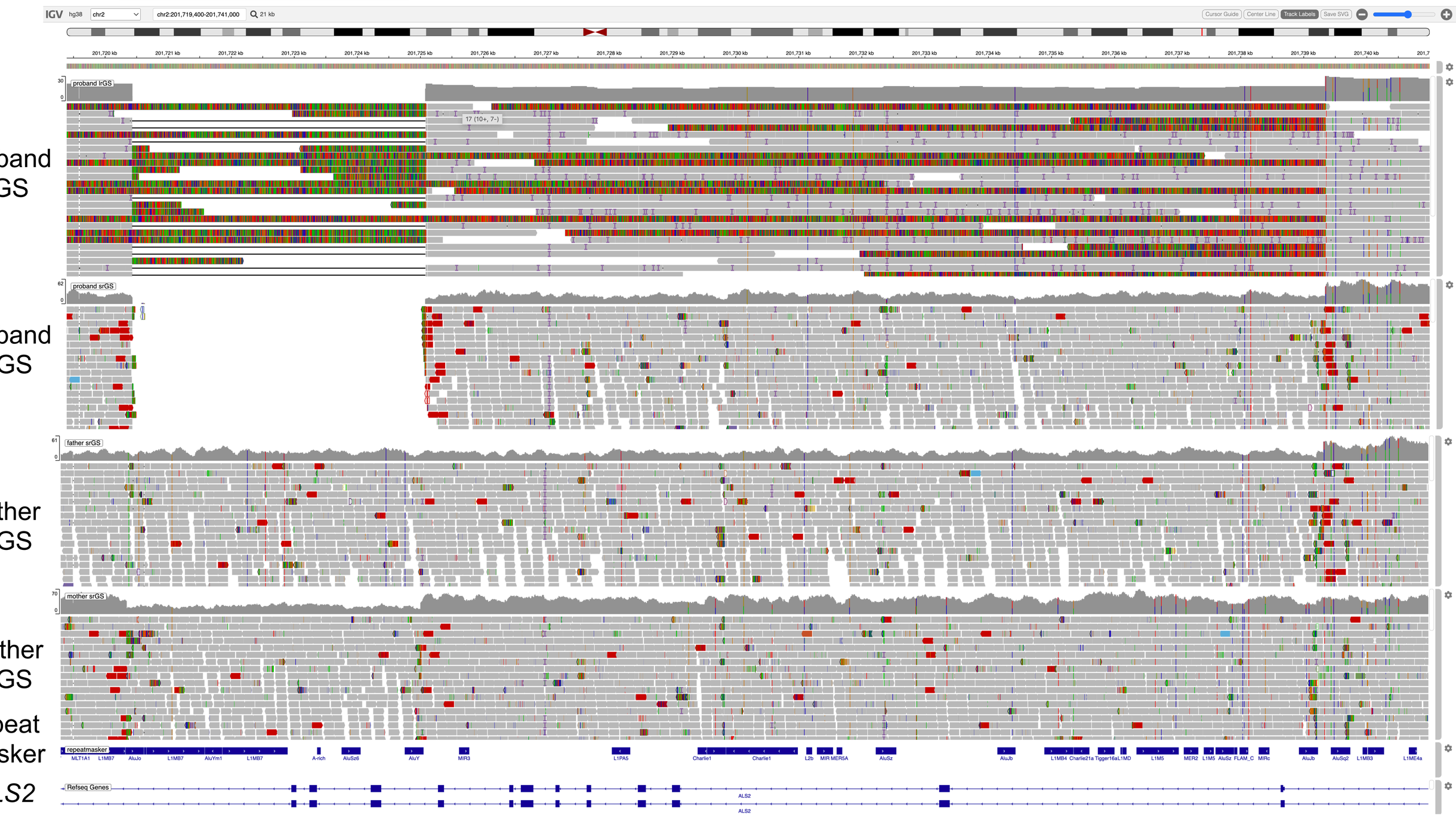
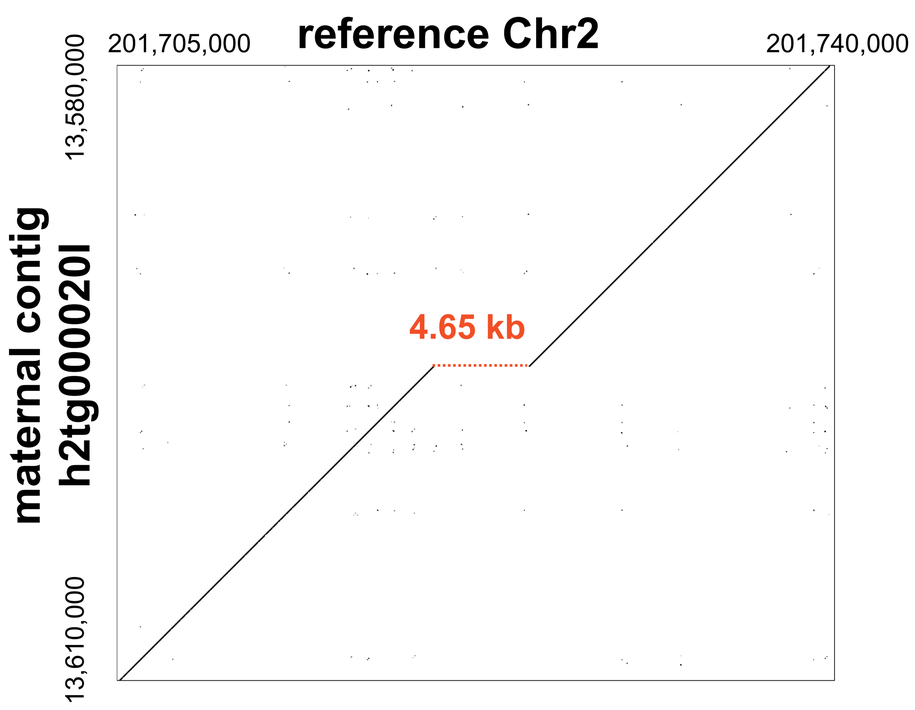


Figure 2





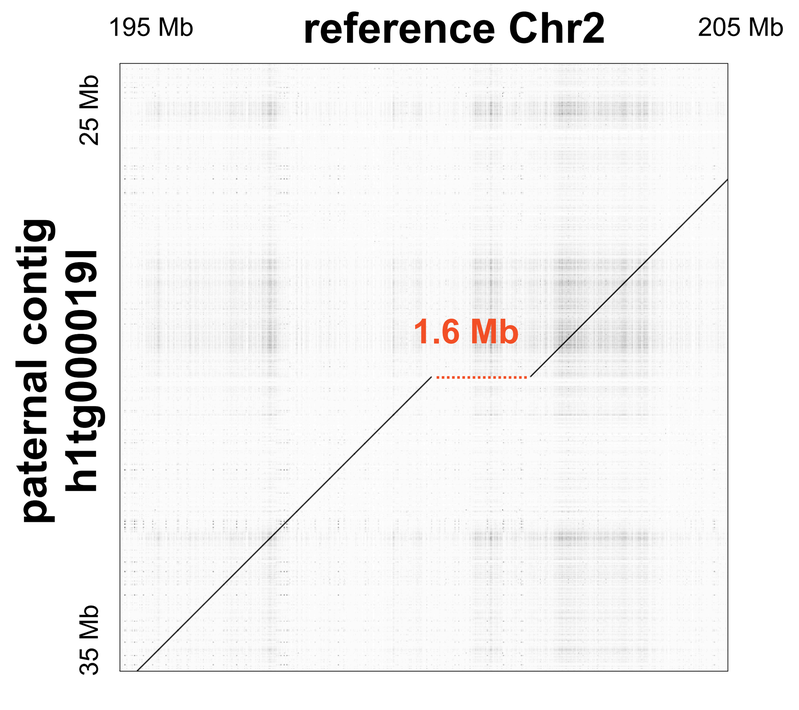


Figure 3