Method

Integration of transcriptomics and long-read genomics prioritizes structural variants in rare disease

Tanner D. Jensen,[1,14] Bohan Ni,[2,14] Chloe M. Reuter,[3,4] John E. Gorzynski,[1,3,4] Sarah Fazal,[5] Devon Bonner,[3,6] Rachel A. Ungar,[1] Pagé C. Goddard,[1] Archana Raja,[3,4] Euan A. Ashley,[3,4] Jonathan A. Bernstein,[3,7] Stephan Zuchner,[5] Undiagnosed Diseases Network, Michael D. Greicius,[8] Stephen B. Montgomery,[1,9,10] Michael C. Schatz,[2] Matthew T. Wheeler,[3,4,11] and Alexis Battle[2,12,13]

[1]*Department of Genetics, Stanford University, Stanford, California 94305, USA;* [2]*Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA;* [3]*Center for Undiagnosed Diseases, Stanford University, Stanford, California 94305, USA* [4]*Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, California 94305, USA;* [5]*Dr. John T. Macdonald Foundation Department of Human Genetics and John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, Florida 33136, USA;* [6]*Department of Pediatrics, Division of Medical Genetics, Stanford University School of Medicine, Stanford, California 94304, USA;* [7]*Department of Pediatrics, Stanford University School of Medicine, Stanford, California 94304, USA;* [8]*Department of Pediatrics, Stanford University School of Medicine, Stanford, California 94304, USA;* [9]*Department of Pathology;* [10]*Department of Biomedical Data Science, Stanford University, Stanford, California 94305, USA;* [11]*GREGoR Stanford Site, Stanford University, Stanford, California 94305, USA;* [12]*Department of Biomedical Engineering;* [13]*Department of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland 21218, USA*

Rare structural variants (SVs)—insertions, deletions, and complex rearrangements—can cause Mendelian disease, yet they remain difficult to accurately detect and interpret. We sequenced and analyzed Oxford Nanopore Technologies long-read genomes of 68 individuals from the undiagnosed disease network (UDN) with no previously identified diagnostic mutations from short-read sequencing. Using our optimized SV detection pipelines and 571 control long-read genomes, we detected 716 long-read rare (MAF < 0.01) SV alleles per genome on average, achieving a 2.4× increase from short reads. To characterize the functional effects of rare SVs, we assessed their relationship with gene expression from blood or fibroblasts from the same individuals and found that rare SVs overlapping enhancers were enriched (LOR = 0.46) near expression outliers. We also evaluated tandem repeat expansions (TREs) and found 14 rare TREs per genome; notably, these TREs were also enriched near overexpression outliers. To prioritize candidate functional SVs, we developed Watershed-SV, a probabilistic model that integrates expression data with SV-specific genomic annotations, which significantly outperforms baseline models that do not incorporate expression data. Watershed-SV identified a median of eight high-confidence functional SVs per UDN genome. Notably, this included compound heterozygous deletions in *FAM177A1* shared by two siblings, which were likely causal for a rare neurodevelopmental disorder. Our observations demonstrate the promise of integrating long-read sequencing with gene expression toward improving the prioritization of functional SVs and TREs in rare disease patients.

[Supplemental material is available for this article.]

Long-read sequencing technology has improved in recent years in terms of accuracy and throughput (Mahmoud et al. 2019; Kovaka et al. 2023). This has unlocked a large reservoir of previously inaccessible variation, especially structural variant (SV), repeat expansions, and other complex variants (Audano et al. 2019; Chaisson et al. 2019). In the clinical setting, however, exome and whole-genome short-read sequencing (SR-GS) remain the dominant approaches to facilitate diagnoses of rare diseases. While SR-GS increases the diagnostic yield of various rare diseases by 5%–20% over exome sequencing, the diagnostic rates for rare diseases remain below 50%. While some fraction of these undiagnosed cases are likely due to non-Mendelian and/or nongenetic causes, the relatively low diagnostic rate underscores the need to explore long-read genome sequencing (LR-GS) as a new tool to increase the detection of pathogenic variants (Hiatt et al. 2021; Sanford Kobayashi et al. 2022). The potential clinical utility of LR-GS is especially apparent for rare SVs, which are known to have a substantial impact on genome function (Wojcik et al. 2023), yet are systematically missed by SR-GS (Audano et al. 2019; Chaisson et al. 2019). Furthermore, among rare SVs discovered by SR-GS, many variants' impacts on genes are hard to interpret without

[14]**These authors contributed equally to this work.**
Corresponding authors: smontgom@stanford.edu, mschatz@cs.jhu.edu, wheelerm@stanford.edu, ajbattle@jhu.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279323.124. Freely available online through the *Genome Research* Open Access option.

additional functional validations, making rare disease diagnostic efforts difficult (Marwaha et al. 2022).

Gene expression data have been shown to help prioritize rare SNVs contributing to rare genetic diseases, particularly using approaches that identify individuals with extreme expression compared to the rest of the population (expression outliers) (Frésard et al. 2019; Ferraro et al. 2020; Montgomery et al. 2022; Li et al. 2023). Previous studies have shown a stronger enrichment of expression outliers having nearby rare SVs, suggesting a possible opportunity to prioritize rare functional SVs systematically (Chiang et al. 2017; Scott et al. 2021). Although numerous cases of rare SVs disrupting gene expression have been identified independently in the context of rare disease diagnostics, there is currently no systematic approach to effectively prioritize candidate disease SVs by integrating RNA-seq and rare SVs genomic annotations. Notably, while STRVCTVRE (Sharo et al. 2022), CADD-SV (Kleinert and Kircher 2022), and PhenoSV (Xu et al. 2023) can prioritize putative pathogenic SVs, STRVCTRVE only scores coding SVs, CADD-SV do not explicitly identify the affected gene and is trained to predict SVs under selection constraints but not regulatory SVs, and all tools have limited capability in finding functional SVs that modulate gene expression. In addition to SVs, variation at tandem repeat loci contributes to gene expression differences (Gymrek et al. 2016; Bakhtiari et al. 2021). Repeat expansions at these loci have been implicated in a variety of rare neurological and neuromuscular diseases (Depienne and Mandel 2021). Because these expansions can potentially span thousands of base pairs, they are difficult to detect with SR-GS, but LR-GS methods have shown promise in detecting them (Miyatake et al. 2022).

In light of the current limitations regarding SV detection and assessment of functional impact, we propose two modalities for improving SV prioritization in rare disease diagnostics. First, we hypothesize that augmenting patient data sets with long-read sequencing will improve the detection of clinically informative rare and private SVs, large indels (30 bp–50 bp) and complex variants where detection from SR-GS is more limited. Second, we recognize a major need to develop new methods that can prioritize rare SVs based on both genomic and functional -omics evidence, thereby providing better clues for clinicians to evaluate variants' pathogenicity to rare diseases. Long-read sequencing has previously been shown in diagnostic contexts to find missing disease-causing variants not revealed by short reads, increasing the number of rare coding SVs detecetd (Merker et al. 2018; Miller et al. 2021; Cohen et al. 2022).

Addressing these needs, we performed long-read sequencing with Oxford Nanopore Technologies (ONT) on 68 individuals from the undiagnosed disease network (UDN) with preexisting short-read genome sequence data to study the improvement in variant detection recall from SR-GS. Then, using blood or fibroblast RNA-seq data from the same patients, we characterize the impact of LR-GS SVs on gene expression and investigate the utility of integrating these data to predict variant function. To address this challenge, we developed the Watershed-SV model and pipeline. Watershed-SV extends Watershed (Ferraro et al. 2020), a probabilistic variant prioritization model integrating multiple -omic signals and SNV genomic annotations, with several new SV-related annotation features to prioritize functional rare SVs. We trained and evaluated Watershed-SV models on GTEx expression outliers and SR-GS SV callset and found Watershed-SV outperforms the WGS-only baseline models in prioritizing coding and noncoding variants. When applied to the UDN patient data set, Watershed-SV further provided additional disease-relevant SV candidates compared to CADD-SV and provided direct insights into the gene disruption mechanisms.

Results

Nanopore long-read sequencing of an exome-negative rare disease cohort

To assess the utility of LR-GS and SV prioritization for rare disease diagnosis, we performed ONT (R9.4.1) LR-GS on 68 individuals from the UDN with a spectrum of clinical features who had inconclusive genetic testing using short-read genome sequencing (SR-GS) exome or genome (Fig. 1A,B). Across the cohort, we achieved a median sequencing throughput of 61 Gb (range: 25 Gb–133 Gb), corresponding to an estimated aligned coverage of 18× (range: 7.1–33×), and a median read length N50 of 19.2 kb (range 5 kb–35 kb) (Supplemental Fig. S1A). After read quality control, most samples had a median read quality score above 12.5 (range 9.9–15.1), corresponding to a median 93.6% read identity aligning to GRCh38 (Supplemental Fig. S1B; Supplemental Table S1).

Long reads detect more insertions and deletions than called from short reads

To compare the LR-GS genomes to the previously generated SR-GS data, we called SVs (SVs) and large indels from both. Using a multi-algorithm consensus among Sniffles2 (Smolka et al. 2024), cuteSV (Jiang et al. 2020), SVIM (Heller and Vingron 2019), facilitated by variant merging with Jasmine-SV (Kirsche et al. 2023), we detected 77,596 large indels (30 bp to 50 bp) and 120,950 SVs (>50 bp) from LR-GS, 2.1 times and 3.2 times more, respectively, than were detected from standard clinical SR-GS SV calling with Illumina's Manta caller (Figs. 1B and 2A). Benchmarking our SV calling on a truth set from Genome in a Bottle sample HG002, our LR-GS consensus calling pipeline yielded an F1 score of 0.9, showing high accuracy of LR-GS-detected variants, compared to 0.45 from SR-GS, driven predominantly by a low recall (Supplemental Fig. S1A). From LR-GS, we observed a balanced median number of 18,375 deletions (8,788,30–50 bp, 9,807,50 bp+) and 18,906 insertions (6,064,30–50 bp, 12,815,50 bp+) per individual, which is 2.2 times and 3.6 times, respectively, the median number identified by SR-GS (Supplemental Fig. S1C). We further detected a smaller median number of 422 duplications and 35 inversions per individual from LR-GS, while SR-GS detected more at a median number of 524 and 233, respectively, potentially due to false positive calls from Manta-only SR-GS compared to confident consensus from LR-GS. In accordance with previous reports (Sudmant et al. 2015), the total number of SVs identified per genome differed by self-reported ancestry, consistent with previous reports of population genetic diversity (Supplemental Fig. S2B).

Stratifying by length, we see LR-GS identifies more variants in the 30 bp–50 bp window for both insertions and deletions. LR-GS also identifies many more insertions across the entire length spectrum, particularly above 1 kb, where SR-GS reads are too short to resolve any insertion sequence. For deletions, we also observe higher counts for LR-GS up to 10 kb, after which we start to observe similar numbers from short reads (Fig. 2A). This is because such large deletions are readily detected by a lack of mapping reads (short or long), whereas detecting large insertions benefits the most from long reads, as they are more likely to be partially mapped to the flanking sequences (Smolka et al. 2024). We find the length distribution of inversions and duplications to be

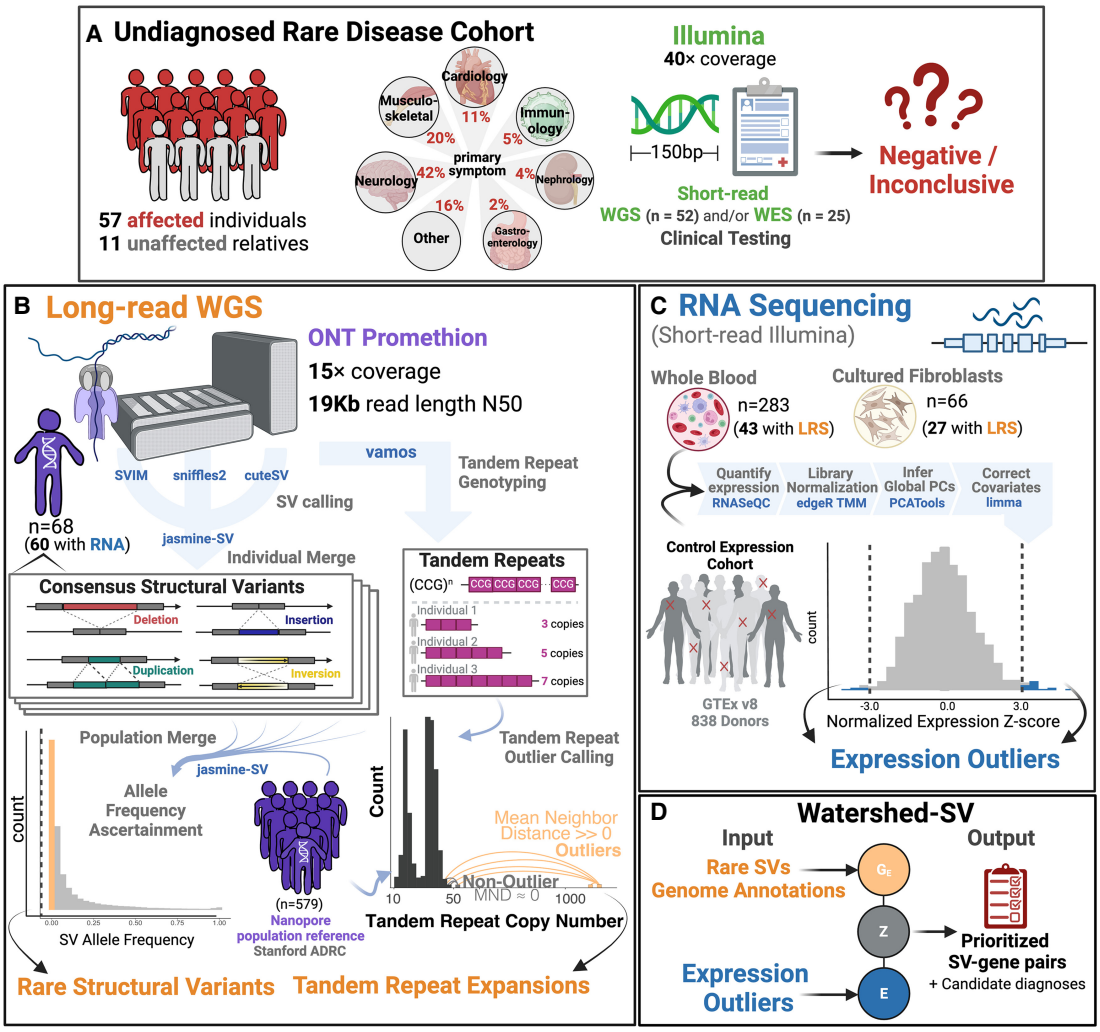

Figure 1. Undiagnosed patient cohort description and pipeline overview. Cohort description: (*A*) Patients were recruited from the UDN for a long-read sequencing (LR-GS) study. These included 57 affected individuals and 11 unaffected family members from a wide range of primary symptom categories, including Neurology, musculoskeletal, and cardiology. Patients had previous short-read genetic testing with Illumina that was negative or inconclusive. (*B*) Long-read Pipeline Overview: individuals were sequenced on R9.4 flowcells on the ONT PromethION. Consensus SVs were called by merging SVs across individual callers and keeping those that showed multialgorithm support. A population merge of the UDN genomes together with the Stanford ADRC population reference of 579 nanopore genomes, allowed ascertainment of robust allele frequencies for SVs. Rare SVs were filtered and intersected with overlapping genome annotations to input into Watershed. Vamos was used on a catalog of polymorphic tandem repeats to genotype tandem repeat copy numbers. A mean neighbor distance-based outlier calling method was used to define extreme repeat expansions. (*C*) RNA sequencing expression outlier pipeline: transcriptome data from the UDN was processed by quantifying expression, combining with tissue-matched controls from GTEx, normalizing for library size and composition bias, and correcting for batch effects and hidden factors. Expression outliers of the normalized data were input into Watershed. (*D*) Watershed-SV integrates signals from rare SVs and overlapping genome annotations to predict variants with large functional effects. High-scoring watershed variants are prioritized and curated per patient for disease relevance.

roughly similar between LR-GS and SR-GS, though we do detect slightly more from short reads (Supplemental Fig. S1D).

As expansions of tandem repeat loci are known to cause multiple Mendelian diseases, we also compared tandem repeat calling in both long and short reads. We found that long reads genotyped repeats with both larger motif lengths and a larger number of repeat copies (Fig. 2B; Supplemental Text).

Long-read population references enable frequency estimates for rare SVs and TREs

For rare disease diagnosis, it is critical to determine which variants are rare in the population. For LR-GS SV, this is a major challenge as existing short-read references, such as gnomAD, CCDG, and 1000 G, show low ascertainment of SVs discovered from LR-GS data (Supplemental Fig. S1F) (Sudmant et al. 2015; Abel et al. 2020; Collins et al. 2020). To address this issue, we used a technology-matched population reference of nanopore genomes generated from Stanford's Alzheimer's Disease Research Center (ADRC) and the Stanford Aging and Memory Study (SAMS) ($n = 571$) (Chemparathy et al. 2024). ADRC + SAMS nanopore genomes were processed with the same SV calling workflow and were merged using Jasmine-SV with the UDN callset; allele frequencies were estimated based on merged allele counts in the combined set. We observe the expected allele frequency distribution for SVs, finding 32.6% of variants are rare (minor allele frequency < 0.01)

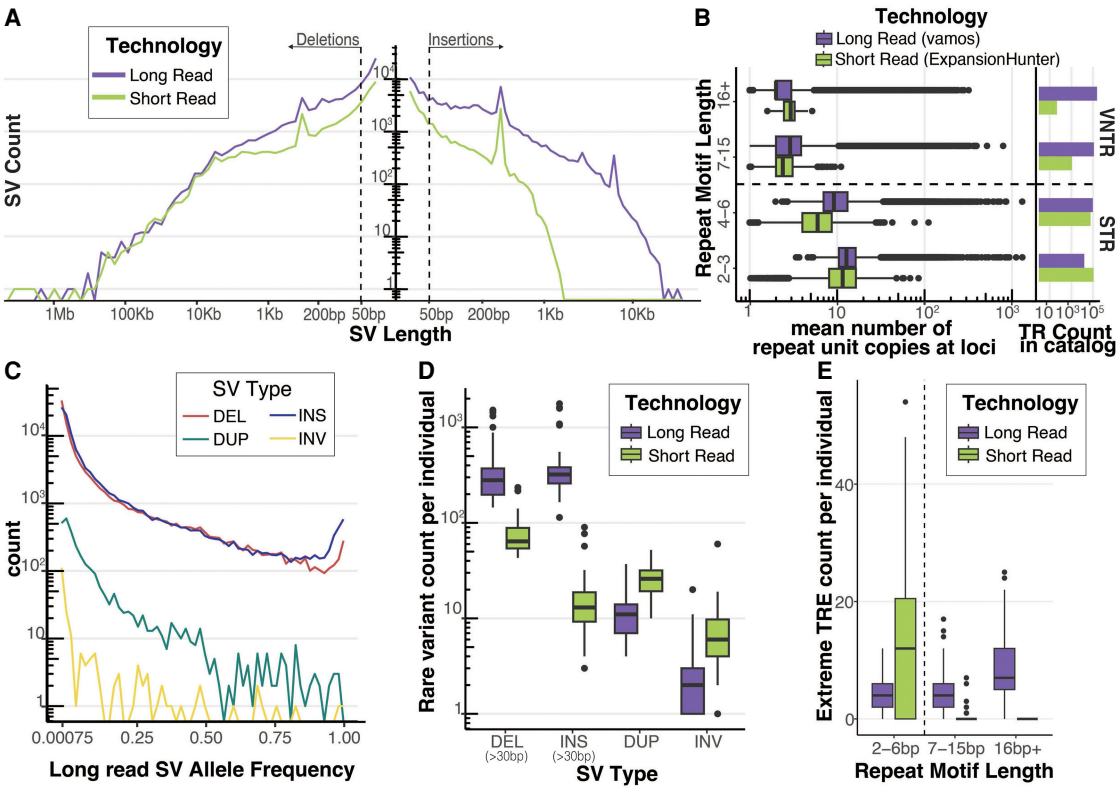

Figure 2. Long-read sequencing detects rare SVs and extreme tandem repeat expansions (TREs). (*A*) Length distribution of deletions and insertions detected by each technology on a log–log axis. SVs were called with a consensus SV calling pipeline including SVIM, cuteSV, and Sniffles2 for long reads and MantaSV calls were genotyped with paragraph for short reads. Dashed line represents 50 bp, the threshold for calling an indel an SV. (*B*) Mean tandem repeat copy numbers estimated from the UDN genomes stratified by repeat motif length. Short tandem repeats (STR) have repeat motifs between 2 bp and 6 bp. Variable number tandem repeats (VNTRs) have repeat motifs greater or equal to 7 bp. Vamos was used to genotype tandem repeat copy number in long reads and ExpansionHunter was used in short reads. Each tool used a different tandem repeat loci catalog to define TRs. Counts of TRs by repeat motif length bins present in the tools respective catalog is also plotted. (*C*) Allele frequency distribution of long-read discovered SVs from Jasmine-SV merge with ADRC genomes. ADRC provided a reference sample of 600 nanopore genomes to allow robust estimation of minor allele frequencies. (*D*) Count of rare SVs (MAF < 0.01), detected per individual stratified by SV Type and Technology. Short-read SVs were annotated with allele frequencies using SVAFotate and a lookup in gnomAD, CCDG, and 1000 G. (*E*) Count of extreme TRE detected per individual. Extreme TRE outliers in each technology were called by jointly estimating repeat copy number distribution of long-read vamos calls with the ADRC and of short-read ExpansionHunter calls with 1000 G, and then calculating for each allele its average distance from its *k*-nearest neighbors. Extreme TREs were defined as alleles with a standardized mean neighbor distance (MND) >2, with $k = 5$ for long reads and $k = 25$ for short reads.

(Fig. 2C; Supplemental Table S4). For SR-GS SVs, we observed a high genotyping rate of our Manta callsets with Paragraph and used them to determine the allele frequencies for SR-GS SV calls (Supplemental Fig. S1E), though variants in the 30 bp–50 bp window are not included in population references and therefore their frequency could not be ascertained from short reads. We filtered both LR-GS and SR-GS SVs down to rare variants with MAF < 0.01 and found each genome had a median burden of about 546 rare deletions and 490 rare insertions detected by LR-GS, whereas by SR-GS methods there was a median burden of 80.5 rare deletions and 11 rare insertions (Fig. 2D). We further observed a smaller burden of rare duplications and inversions at a median number of 11 and 2, respectively, from LR = GS, and 13.5 and 9 in SR-GS (Fig. 2D).

Using the combined population of over 600 genomes of our UDN and ADRC + SAMS cohort (Chemparathy et al. 2024), we improved our ability to filter LR-GS SVs for rare disease diagnosis. We detected a median of 716 rare SVs (>50 bp) per genome, whereas using short-read references (i.e., gnomAD) to ascertain rare variants, we would have detected over 17,000 per genome, making curation of these variants extremely difficult (Supplemental Fig. S1F). Genotyping LR-GS SVs in 1000 G short-read genomes with PARAGRAPH to ascertain allele frequencies yielded a median of 3,198 SVs per individual, still 4.4 times more than we discover from a technology-matched reference (Supplemental Fig. S1F).

Similarly, we sought to assess TR distributions to identify rare TREs in the UDN cohort, that is, extreme outliers of TR copy number. Using population references for LR-GS and SR-GS, we called TREs by thresholding on a standardized MND (see Methods; Supplemental Text). We identified a median number of 14 TREs per individual from LR-GS, and these long reads enabled the detection of more TREs in variable number tandem repeat (VNTR) loci (repeat motif length >7 bp) compared to short reads (Fig. 2E; Supplemental Fig. S3).

Transcriptome data identifies functional effects of rare SVs

To characterize the functional effects of rare SVs and extreme TREs detected by LR-GS, we used RNA-seq obtained from individuals in our UDN cohort. Most ($n = 60$) of the UDN LR-GS subset also had

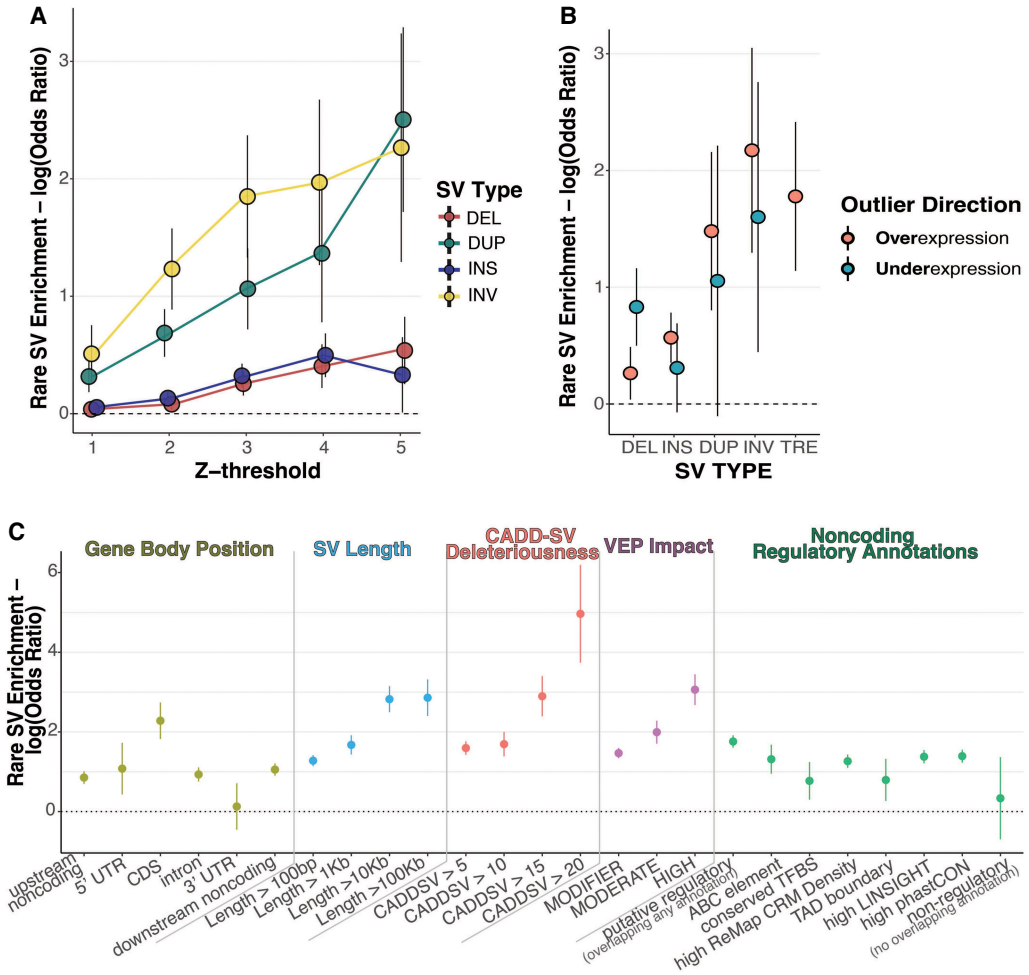

Figure 3. Rare long-read-discovered SVs are strongly enriched proximal to gene expression outliers (*A*) enrichment of rare SVs, stratified by type, within 10 kb of an expression outlier gene given the specified absolute *Z*-score threshold. Estimate of log odds ratio plotted with error bars representing standard errors of the estimate. (*B*) Directional enrichment for rare SVs within 10 kb of either over ($Z > 4$) or under ($Z < -4$) expression outliers. TRE-underexpression enrichment not plotted due to an insufficient number of rare TREs near underexpression outliers. (*C*) Enrichment of rare SVs, across all SV types, within 100 kb of expression outliers, stratified by genome and variant annotation categories. Gene body position displays enrichment of VEP annotated categories for SV location relative to the gene body of the expressed gene. If an SV overlaps multiple categories, it is assigned to the one with highest priority given the following ordering: CDS, 5′ UTR, 3′ UTR, intron, upstream noncoding, downstream noncoding. SV length and CADDSV deleteriousness display enrichment of rare SVs with length and CADDSV score respectively above the specified threshold. VEP impact displays enrichment of rare SVs with the given VEP impact category, where HIGH represents predicted loss-of-function variants. Finally, we display enrichment of SVs that overlap with noncoding regulatory annotations, including if it overlaps an ABC regulatory element linked to the expressed gene, a conserved transcription factor binding site (TFBS), a high density of ChIP-seq peaks defining conserved regulatory modules (CRM) from ReMap, a TAD boundary detected in multiple cell types, highly constrained LINSIGHT SNVs, or a highly conserved region by phastCons. We also display enrichments for SVs that overlapped any one of these annotations (putative regulatory SVs) and for SVs that do not overlap with any of these annotations (putative nonregulatory).

short-read RNA-seq data generated as part of their UDN enrollment, including ($n = 42$) samples from whole blood and ($n = 27$) from fibroblasts. We hypothesized that impactful SVs would drive outlier expression of nearby genes. We called gene expression outliers from blood and fibroblast tissues by jointly normalizing the UDN cohort with matched-tissue RNA-seq samples from GTEx, and adjusting for RIN, sex, batch, and global expression principal components as covariates, and observed a median of 139 outliers per individual in UDN blood samples (Fig. 1C; Supplemental Fig. S5A). We found strong enrichment at a log odds ratio (LOR) of 0.63 (adj. *P*-value = 4.8×10^{-6}) for rare SVs nearby (within 10 kb) of expression outlier genes. We observed enrichment across all types of SVs, with stronger enrichments for more stringent *Z*-score thresholds defining outliers (Fig. 3A). These observations are concordant with previous analysis in GTEx, where CNVs were enriched for proximity to genes with outlying expression (Chiang et al. 2017; Ferraro et al. 2020).

Using the improved variant discovery enabled by LR-GS, we expand upon previous results by analyzing insertions and TREs, which are difficult to call with SR-GS. We observed that genes from individuals with outlier expressions are significantly enriched for nearby rare insertions and TREs (*Z*-score threshold = 4, Insertion LOR (95% CI) = 0.496 (0.13–0.86), TRE 1.72 (0.47–3.0)) (Supplemental Fig. S5B). We observed the strongest enrichment for duplications and inversions, which may be due to the fact that these variants are on average much longer (INV mean = 103 kb, DUP mean = 18.4 kb) than deletions and insertions (DEL mean = 450, INS mean = 275). Stratifying enrichments by outlier

direction, we also observed that deletions were enriched for underexpression events (LOR (95% CI) = 0.83 (0.18–1.48)), and insertions, duplications, and TREs were enriched for overexpression events (insertion 0.568 (0.149–0.987), duplications 1.48 (0.153–2.81), TRE 1.78 (0.53–3.01)). Inversions were enriched for outliers in both directions, suggesting that large inversions may rearrange regulatory landscapes to increase or decrease the expression of nearby genes (Fig. 3B).

Previous work in GTEx based on SR-GS did not report significant expression outlier enrichments for insertion variants, which may be a reflection of the limited power of SR-GS to call insertions (Ferraro et al. 2020). To characterize differences in enrichments based on sequencing technology, we compared LR-GS enrichment to those calculated based on the UDN SR-GS SVs. We observed the LR-GS callsets had larger enrichments than SR-GS callsets, particularly for insertions where SR-GS insertions were not enriched like what was previously observed in GTEx, underscoring the utility of LR-GS for the discovery of novel functional insertions (Supplemental Fig. S5C). When considering other variant types, we observed that most rare SVs near outlier genes were detected uniquely by LR-GS (77.3% of deletions and 97.2% of insertions). Rare duplications found nearby outliers were largely detected by both technologies (57.1% shared). In contrast, rare inversions nearby outliers were frequently only detected by short reads (71.4%), but these variants had only modest enrichments, suggesting a higher false discovery rate of inversions from SR-GS (Supplemental Fig. S5D).

To characterize rare variant enrichments near expression outliers, we estimated SV enrichments stratified by several annotations, increasing our proximity-window size to 100 kb to capture more noncoding variants. Among gene body regions, we observed the strongest enrichment for SVs overlapping protein-coding sequence (LOR 2.28, adj. P-value = 1.5×10^{-05}). We also observed enrichments for noncoding SVs, including intronic (LOR 0.93, adj. P-value = 2.6×10^{-06}), upstream noncoding (LOR 0.85, adj. P-value = 1.6×10^{-06}), and downstream noncoding regions (LOR 1.05, adj. P-value = 1.5×10^{-10}). As the length of the SV increased, we observed enrichments up to log odds ratios of 2.86 (adj. P-value = 1.2×10^{-08}) for variants longer than 100 kb, indicating that variants that disrupt more sequence are more likely to drive outlier expression. Further, increasing thresholds for variant deleteriousness as predicted by CADD-SV (Kleinert and Kircher 2022) or VEP impact categories (McLaren et al. 2016) was also associated with increased enrichment. Variants with high CADD-SV score (>20) were enriched at a log odds ratio of 4.96 (adj. P-value = 1.3×10^{-03}) (Fig. 3C).

We then estimated enrichment for noncoding rare SVs using several regulatory annotations. We included the proximity to a coding SV as a covariate in all enrichment analyses to mitigate effects due to the tagging of coding SVs. We observed that SVs that intersect an activity-by-contact (ABC) mapped regulatory element (Nasser et al. 2021) were enriched for outlier expression of the ABC target gene (LOR 1.32, adj. P-value = 7.7×10^{-03}). Rare SVs were also enriched for outliers when overlapping a high density of ChIP-seq peaks from ReMap (Hammal et al. 2022) (LOR 1.27, adj. P-value = 1.1×10^{-12}), noncoding regions with evidence of negative selection from LINSIGHT variants (Huang et al. 2017) (LOR 1.38, adj. P-value = 1.6×10^{-15}), or highly conserved sequences by phastCons (Siepel et al. 2005) (LOR 1.39, adj. P-value = 2.2×10^{-16}). As a negative control, we curated a set of putatively nonregulatory SVs, noncoding SVs that did not overlap with any of the previous annotations or any conserved transcription factor binding motifs, TAD boundaries (Wang et al. 2018), ENCODE conserved *cis*-regulatory elements, or high-scoring CADD variants. These 519 nonregulatory rare SVs were not enriched proximal to expression outliers (LOR 0.33, adj. P-value = 1.00) (Fig. 3C). We then sought to validate these ABC-element enrichments in a larger sample size from GTEx. Notably, enhancer enrichments were observed for single-tissue outliers but not multitissue outliers (Supplemental Fig. S5E).

Watershed-SV integrates genomic annotations and transcriptomic signals to prioritize rare functional SVs

The co-occurrence of expression outliers with rare SVs provides the opportunity to jointly use genomic annotations and transcriptomic signals to prioritize rare functional SVs. To do so, we extended our method Watershed (Ferraro et al. 2020), originally created for SNVs, to develop Watershed-SV to assess SVs (Fig. 1D). Watershed-SV models individual gene expression jointly with genomic annotations of nearby SVs to infer the probability the individual harbors a high-impact regulatory SV (Watershed posterior probability). We included annotations characterizing the gene body and separate annotations for flanking sequence (up to either 10 kb or 100 kb from the gene body). We then trained the Watershed-SV models using individual gene pairs from the GTEx v8 SR-GS SV callset. We first developed three Watershed-SV models for gene expression in skeletal muscle, whole blood, and cell cultured fibroblasts tissues, which we refer to as tissue-specific Watershed-SV models. We then developed a multitissue Watershed-SV model using expression outliers defined from the median expression Z-score across 48 GTEx tissues. To assess the performance of the Watershed-SV models, we evaluated expression outlier predictions of the model using held-out pairs of individuals who share the same rare SV genotype near a particular gene (N2-pairs). We used the Watershed-SV posterior probability from one sample in the N2-pair as a prediction and the expression outlier status of the genotype-matched, held-out sample as the label. We compared the Watershed-SV models to a model that included genomic annotations but no expression information (WGS-only baseline) as well as CADD-SV. The 10 kb multitissue Watershed-SV model outperformed the WGS-only baseline (ΔAUPRC bound = (0.13, 0.27)) (Fig. 4A), demonstrating the utility of incorporating gene expression information in SV prioritization. Watershed-SV learned large feature importance scores for SV length, overlapping with exons, location in 5′ UTRs, CADD scores, and for disruption of *cis*-regulatory elements (Fig. 4B; Supplemental Fig. S8).

To capture longer range SV-gene interactions in the 100 kb Watershed-SV model, we included ABC (Nasser et al. 2021) enhancer-gene links among the annotations. We categorized annotations into three categories based on SV location: upstream flanking region (UFR) annotations, gene body annotations, and downstream flanking region (DFR) annotations. Watershed-SV learned especially large weights for UFR annotations, reflecting the known functional importance of regulatory elements near promoters. Among gene body annotations, we observed a high effect size CADD score and active transcription annotations (Fig. 4D; Supplemental Fig. S9). The 100 kb multitissue Watershed model also outperformed the baseline WGS-only model (ΔAUPRC bound = (0.12, 0.26)) and CADD-SV, although with a slightly lower performance than the 10 kb model, reflecting the challenges of annotating long-range regulatory elements (Fig. 4A). Further, when stratifying N2-pairs by minimum rare SV distance from genes, Watershed-SV still outperforms both WGS-only and CADD-SV and prioritizes more noncoding SVs near outliers (Fig. 4C; Supplemental Fig. S7). Taken together, we observed that Watershed-SV can

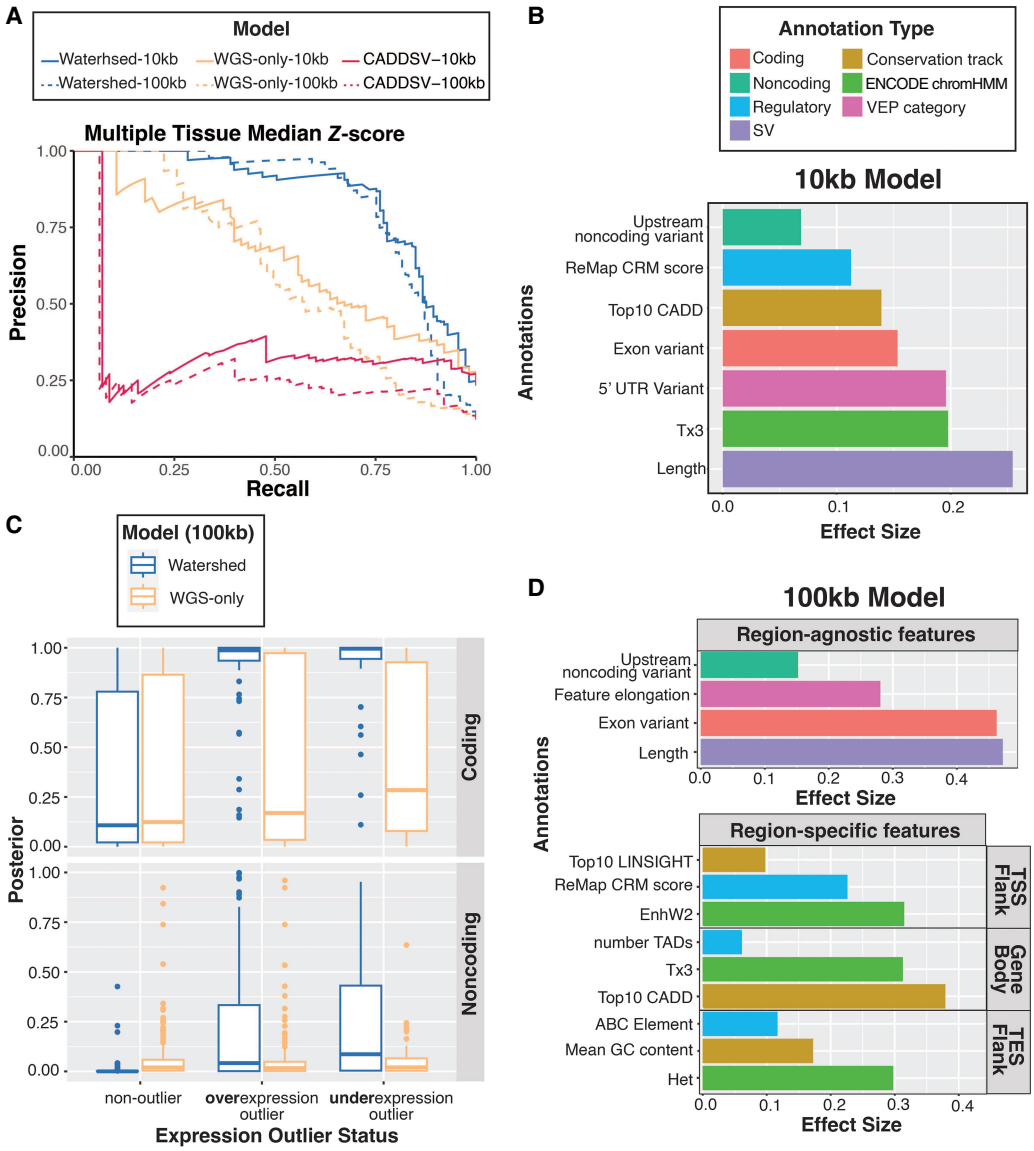

Figure 4. Watershed-SV improves the prioritization of rare SVs in healthy and muscular dystrophy cohort. (*A*) Precision-recall curves (PRC) of benchmark using held-out N2 pairs; We ran multitissue Watershed-SV using both 10 kb (solid) and 100 kb (dashed) distance limit as well as WGS-only model and CADD-SV with the same setup. (*B*) *Top* positive genomic annotation effect sizes (β) for seven major categories of the 10 kb multitissue Watershed-SV model. (*C*) Using a *Z*-score threshold of −3 and 3, we stratified 100 kb multitissue Watershed-SV model prediction on CMG muscular disorder data set posterior probabilities by under-, over-, and nonoutliers (column), and then by coding versus noncoding variants (row); each dot represent an gene-SV pair. (*D*) *Top* positive genomic annotation effect sizes for 100 kb multitissue Watershed-SV model. Seven annotation categories are grouped into region-specific (TSS/upstream Flank, Gene Body, TES/downstream Flank) and region-agnostic features. Region-specific features are separately aggregated for each SV, then collapsed to each gene by regions.

prioritize rare functional SVs affecting either coding or noncoding sequence.

Watershed-SV prioritizes SVs implicated in Mendelian muscular disorders

To demonstrate Watershed-SV's applicability to Mendelian disease diagnosis, we applied the method to a disease data set with previously reported functional SVs implicated in patient phenotypes. Specifically, we applied Watershed-SV, trained using GTEx SR-GS SV calls and multitissue transcriptomic outliers, to prioritize functional rare SVs from 26 patients with inherited muscular disease, among which two are diagnosed with SVs and six with SNVs from the Center for Mendelian Genomics (CMG) project (Cummings et al. 2017). In the CMG cohort, we called rare SVs from the SR-GS data (Supplemental Text) and called expression outliers using GTEx skeletal muscle as a reference expression baseline. We observed on average 318 outliers per individual in the CMG samples, in contrast to 94 outliers per GTEx control. We expect these differences are influenced by the relative sizes of the cohorts, differences in sample preparations, and other differences between studies (Supplemental Fig. S4A). Using the Watershed-SV pipeline, we identified both cases where previously described diagnostic variants are SVs (Cummings et al. 2017) and ranked

the two diagnostic SV-gene pairs as the most highly prioritized in respective patient data; both are cases of Duchenne muscular dystrophy (DMD) caused by inversions in the *DMD* gene (Supplemental Fig. S6E). Notably, the patient C4 inversion included the first 18 exons, resulting in the expression of a truncated transcript similar to other known cases of large inversion induced DMD (Geng et al. 2023); the C3 inversion inverted exon 51 and was associated with greatly reduced expression (Supplemental Fig. S6E).

Among 18 patients without prior genetic diagnosis, we observed four patients with Watershed-SV prioritized gene-SV pairs in muscular disorder-related genes, one gene-SV posterior above 0.5, and three above 0.9 (Supplemental Table S6). Among them, a female patient (N10) with undiagnosed Limb-Girdle Muscular Dystrophy carried a heterozygous rare deletion (gnomAD AF = 0.0003, Watershed-SV = 0.996, WGS-baseline = 0.78) disrupting the 3′ CDS-UTR region of *PIP5K1C*. Known disruptions of *PIP5K1C* can cause an autosomal recessive lethal contracture syndrome III (OMIM#611369) (Narkis et al. 2007), another type of severe muscular disorder. While there is not sufficient evidence to determine that this deletion alone underlies the patient's phenotype, it demonstrates Watershed-SV's ability to prioritize variants of interest for further genetic and functional investigation.

Watershed-SV prioritizes LR-GS SVs in rare disease

Knowing our method could prioritize diagnostic rare disease SVs from SR-GS, we investigate if we could prioritize likely functional LR-GS SVs that may underlie UDN patients' disease phenotypes. We trained a 100 kb Watershed-SV model that models per-tissue expression outliers from both blood and fibroblast using GTEx v8 data. We defined the Watershed-SV estimate as the maximum of the Watershed posterior probabilities for blood and fibroblast for each gene-SV pair. Out of 19,277 gene-SV pairs tested, 886 had a posterior of at least 60%, representing 4.6% of rare variants within 100 kb of the nearest gene.

To assess the variants that Watershed-SV could return for clinical review, we evaluated variants passing a Watershed-SV score threshold along with filters corresponding to basic approaches clinicians could use to prioritize potential disease-relevant variants. The filters encompass three general strategies: (1) filtering for variants near genes previously known to be related to the patients' symptoms, using HPO term matching (Zhao et al. 2020); (2) filtering for variants that have strong evidence of pathogenicity from WGS-based variant scoring, here using CADD-SV; and (3) prioritizing variants through RNA-seq outlier signals using expression outliers alone. Specifically, we compared Watershed-SV to CADD-SV (Kleinert and Kircher 2022), and evaluated whether Watershed-SV nominates different variants from CADD-SV, with these two scores evaluated alone and in combination with other filters. We compared a CADD-SV threshold of 10 to a Watershed posterior threshold of 0.6, corresponding to a roughly equivalent fraction (4.6%) of variants.

When comparing the most stringent set of filters, Watershed-SV combined with HPO term matching prioritized symptom-relevant functional gene-SVs for more patients ($n = 6$) than CADD-SV combined with expression outlier and HPO term matching ($n = 1$) (Fig. 5A)—that is, among genes with established relationships to the patient symptoms, more hits were identified using Watershed-SV than among a comparable size set prioritized by CADD-SV. In addition, the variant gene pairs prioritized by HPO term-matching along with either CADD-SV or Watershed-SV overlap minimally. When not constraining to the expression of outlier genes, filtering by HPO and Watershed-SV yields 18 variant-gene pairs, filtering by HPO and CADD-SV yields 12 variant-gene pairs, but the two sets only share a single variant-gene pair (Supplemental Fig. S10A). This suggests that Watershed-SV and CADD-SV contribute complementary information, with Watershed-SV optimized for prioritizing expression regulating variants and CADD-SV for constrained variants. Among all variants in UDN prioritized by either CADD-SV or Watershed-SV, we found an overwhelming majority coming from SVs and long indels (30 bp–50 bp) only identified in long read data (Fig. 5B).

Watershed-SV yielded a number of interesting variants for future analysis. In our patient set, 10 genes are designated as genes of interest, confident candidate genes, or diagnostic genes by the genetic counselors at Stanford Center for Undiagnosed Diseases. Among them, 8 have their candidate variants prioritization ranking within the top 5 of the given patients. Furthermore, in the CMG data set, diagnostic SVs for the two confirmed DMD positive examples are all ranked at the very top of the prioritization list. In one case, we identified a heterozygous CCG TRE in a pair of siblings with clinically diagnosed oculopharyngodistal myopathy (OPDM). The TRE is situated in the promoter and 5′ UTR of the *FAM193B* gene and was observed with overexpression outliers in both siblings' blood RNA samples (Karolchik et al. 2014). Previous studies have found CGG/CCG repeat expansions in the 5′ UTR of other genes to be causal for OPMD (Ishiura et al. 2019; Deng et al. 2020; Yu et al. 2021), and speculate repeat-associated non-ATG (RAN) translation of the CCG repeat and accumulation of toxic RAN proteins may be the mechanism. Fazal et al. (2024) predicted this repeat locus as pathogenic using RExPRT, a tool trained with tandem-repeat-specific annotations(), but was unable to resolve the full size of the repeat using SR-GS. Comparing LR-GS vamos (Ren et al. 2023) calls with ExpansionHunter (Dolzhenko et al. 2019) calls at the same loci, we saw that short reads underestimated the length of the expansion and cannot distinguish affected siblings from the unaffected mother (Supplemental Fig. S11A,B). Because *FAM193B* has not been annotated as a disease gene for OPDM, HPO-term-based filters failed to prioritize the variant. Both Watershed-SV (score = 0.63) and CADD-SV (score = 11) were able to prioritize the TREs (Fig. 5C). We observed similar tandem repeat numbers ($RN_A = 198$, $RN_B = 194$) for the two siblings and a smaller yet still expanded number in their unaffected parent (RN = 158), which could be consistent with a premutation allele that was expanded beyond a pathogenic threshold in both siblings (Fig. 5D). CGG repeat expansions have previously been associated with hypermethylation and gene silencing, yet with nanopore methylation calls in this case, we did not detect a change in methylation levels of the *FAM193B* promoter, consistent with the observed overexpression (Z-score = 10) (Supplemental Fig. S11D). Watershed-SV's integration of gene expression as evidence of regulatory disruption provided greater support for the prioritization and pathogenicity of this variant.

In a second case, we detected a set of compound heterozygous deletions separately affecting the 5′-UTR-CDS region and the 3′ CDS-UTR region of *FAM177A1* in another pair of siblings suffering from global developmental delay, macrocephaly, and seizures. CADD-SV computed a modest score of just 6 for one of the variants, while Watershed-SV prioritized both variants with strong posterior probabilities >0.9 (Fig. 5E). Notably, the siblings displayed strong underexpression in fibroblasts for *FAM177A1* in comparison to other UDN samples. While standard SR-GS also detected both variants, LR-GS enabled the phasing of these deletions

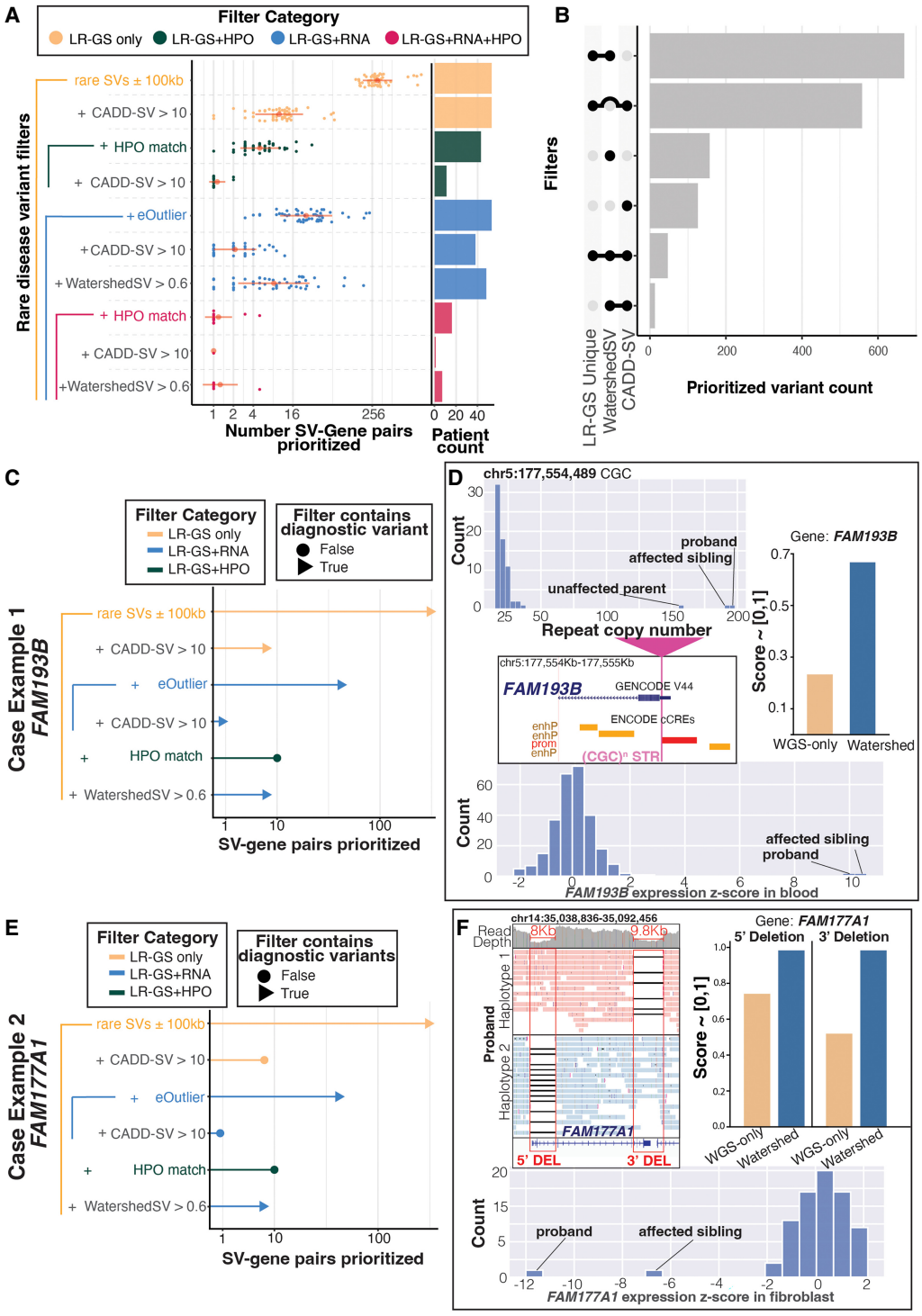

Figure 5. Watershed-SV prioritizes symptom-relevant functional rare SVs from UDN LR-GS data set. (*A*) Swarmplot for number of gene-SV pairs prioritized per individuals in the UDN LR-GS data set under different set of combined filters. There are four filter categories: LR-GS-only filters, LR-GS + HPO filters, LR-GS + RNA filters, and LR-GS + RNA + HPO filters, in increasing level of stringency due to increasing types of filters jointly applied; red dot represents the mean number of gene-SV pairs across individuals, red horizontal line represents standard deviation; *x*-axis is in $\log_2$ scale; the bar plot on the *right* shows number of samples with significant prioritizations. (*B*) UpSet plot depicting number of gene-SV pairs prioritized by Watershed-SV (posterior > 0.6), CADD-SV (score > 10), and whether the SV is uniquely identified using LR-GS. (*C*,*E*) Case example 1, rare TREs shared by both siblings, and case example 2, rare compound heterozygous deletions in siblings. Lollipop plot shows which set of filter includes the candidate diagnostic gene-SV pair (triangle) and which does not (circle), height of the lollipop represents number of gene-SV pairs prioritized in $\log_2$ scale. (*D*) Panels depict the TR copy numbers of the siblings and unaffected parent with less-expanded allele. The TRE loci is in 5′ UTR of *FAM193B*. Both Watershed-SV and CADD-SV can prioritize this but not WGS-only baseline model. Both siblings have extremely high overexpression *Z*-scores. (*F*) Panels depict the compound heterozygous deletions phased onto both alleles for *FAM177A1*, causing LOF of gene and thereby underexpression outliers. Only Watershed-SV succeeded at prioritizing both variants.

and showed they were *in trans*, disrupting both *FAM177A1* alleles (Fig. 5F). A study (Alazami et al. 2015) of disease variants in consanguineous families reported homozygous frameshift SNVs in *FAM177A1*(c.297_298insA) in individuals with similar neurologic symptoms to these siblings. Experimental validation study (Kohler et al. 2024) using a Zebrafish model further supported that biallelic LOF mutations on *FAM177A1* could potentially lead to neurodevelopmental problems via abnormal Golgi complex dynamics.

Discussion

Multitechnology, high-sequencing-depth references have shown that there are roughly 22,000–27,000 rare SVs per genome (Audano et al. 2019; Chaisson et al. 2019). We achieve similar numbers within our UDN samples using a multialgorithm consensus set from only low-depth (15× coverage) long-read nanopore data, yielding on average more than 2.4× of what was found with SR-GS callers, especially substantially more insertions in LR-GS across all size ranges. Meanwhile, we observed more inversions and duplications in SR-GS than in LR-GS, likely due both to higher false positives in the SR-GS callset and limitations in alignment-based SV calling from LR-GS. From our study, we also demonstrated that LR-GS-based reference panels (ADRC) significantly improved variant allele frequency estimates, stressing the need for high-quality, diverse ancestry population reference such as the ongoing LR-GS sequencing of the 1000 Genome Consortium or All of Us (Gustafson et al. 2024; Mahmoud et al. 2024). In addition, we also examined variation at TR loci, genotyping TR copy number from LR-GS at 2 times more loci than profiled routinely in SR-GS, and identifying longer TREs even up to 10 kb. However, our study did not investigate SNV and small indel calling with LR-GS due to limited read depth and the ONT error characteristics. We speculate that as error rate continues to drop, higher read-depth LR-GS could excel at calling both SNV, small variants, and SVs, enabling LR-only tests in the clinic.

Beyond variant discovery, our work investigates the functional impact of LR-GS rare SVs and TREs with expression outliers measured from RNA-seq. We reinforce previous findings of a strong enrichment of gene expression outliers nearby rare SVs (Chiang et al. 2017; Ferraro et al. 2020; Scottet al. 2021; Vanderstichele et al. 2024), and we also detected expression effects of additional variant types, including insertions and TREs, and enrichment for noncoding rare SVs that disrupt important enhancers predicted by the ABC enhancer model (Nasser et al. 2021). Future investigation in larger cohorts with disease-relevant transcriptomes and long-read genomes is expected to only further resolve the functional rare SV landscape. Relatedly, realignment of WGS and RNA data to a telomere-to-telomere (T2T) reference like CHM13 (Ungar et al. 2024), assembly-based SV callers like the Napu pipeline (Kolmogorov et al. 2023) and read-depth-based SV callers may enhance future studies with increased sensitivity and reduced false positive rate for detecting and interpreting other types of variation.

Finally, our work provides a framework, Watershed-SV (https://github.com/jasonbhn/Watershed-SV), to synthesize genomic annotations of SVs with transcriptomic signals to prioritize functional rare SVs, applicable to rare disease patients. We showed that Watershed-SV improves prioritization of rare SVs that affect gene expression over existing tools such as CADD-SV, and can identify candidate functional variants that would have been missed by such genome-only methods. Among the unique prioritizations from Watershed-SV in UDN LR-GS genomes, we successfully identified candidate diagnostic variants, suggesting the benefit of a dedicated model of rare SV's impact on gene expression. The Watershed-SV pipeline is applicable to any SR or LR-GS SV data sets with matched RNA-seq data. However, current Watershed-SV models were trained on SR-GS data from GTEx, likely not capturing the full functional SV landscape observed in LR-GS. Further, we have only investigated the impact of rare SVs on gene expression and our model currently does not jointly model rare SNV and SV impact. Future improvements may be possible from training on larger LR-GS data sets, and from incorporating additional disease-relevant tissue contexts, additional molecular outlier signals such as splicing, methylation available directly from nanopore sequencing, or proteomics (Li et al. 2023). Additionally, further improvements to the portability of Watershed-SV may be possible by developing outlier calling approach requiring only summary statistics instead of raw data from background data sets. Finally, future iterations of Watershed and Watershed-SV will also analyze transmission in trios or pedigrees of related individuals to further help variant prioritization.

In summary, we demonstrated the power of LR-GS in detecting functional and disease-relevant indels, TREs, and SVs using LR-GS genomes from UDN patients with negative exome or SR-GS. Altogether, our work is a demonstration of how LR-GS sequencing will broaden our understanding of the function of rare SVs and other rare variants in healthy and rare disease cohorts and provided a reusable framework for practically integrating LR-GS SV with various transcriptomic outlier signals to provide additional candidate diagnostic variants for undiagnosed rare diseases. With the increasingly context-specific RNA-seq and the rapidly expanding LR-GS resources, our analysis framework will be an even more powerful tool for rare SV prioritization including for disease.

Methods

Oxford nanopore sequencing

DNA was extracted from either whole blood ($n = 25$) or cultured fibroblast pellets ($n = 43$) for all UDN individuals using Qiagen's extraction protocol (Cat. ID:158023). Standard Nanopore ligation protocol was performed with LSK109 or LSK110 and loaded onto R9.4 flow cells. Data were base called with Guppy (v4.0.11), using the high-accuracy model (hac) with a Qscore filter of 7. Sequencing summary metrics were calculated and plotted with NanoPlot (De Coster and Rademakers 2023).

LR-GS SV calling

FASTQ files from base calling were aligned to the GRCh38 patch 13 reference (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.39) with minimap2 (2.26-r1175) (Li 2021, 2018): SVs were called with an ensemble algorithm combining three separate SV callers: Sniffles2 (v2.2) (Smolka et al. 2024), cuteSV (v2.1.0) (Jiang et al. 2020), and SVIM (v2.0.0) (Heller and Vingron 2019). We set minimum support for SVs to be at least three reads, min SV length to 30, and max SV size to 10 Mb.

After calling SVs with each tool, individual callers per sample were merged with Jasmine SV (v1.1.5) (Kirsche et al. 2023). Genotypes were decided for the post-merged consensus by taking a majority vote of the genotypes determined by each individual caller. We required that an SV be supported by at least two callers, creating the consensus set we used for downstream analysis. After merging with Jasmine, we also ran Iris module through Jasmine in order to refine insertion sequences. Allele frequencies were

annotated with population references and paragraph genotyping (Supplemental Text).

SR-GS SV calling

For SR-GS SV calling, we emulated the typical sequencing done in clinical WGS genetic testing. Current clinical standards for WGS clinical testing use the Illumina DRAGEN pipeline (Schobers et al. 2024) which calls SVs with Illumina's Manta software. We ran Manta (v1.6.0–2) with default parameters except setting a lower limit for SV scoring to 30 bp. Tough to limit false positive calls, we removed variants with QUAL < 125, annotated large (>10 kb) deletions and duplications with coverage fold change with Duphold. Variants that did not meet expected changes in coverage (<0.7× for deletions and >1.3× for duplications) were filtered. Finally, we annotated and filtered variants if they spanned centromere annotations for GRCh38, as these repetitive unassembled regions are common sources of false positive calls. Individual filtered Manta calls across the UDN cohort were merged with JasmineSV. From this merged set, we genotyped fully resolved SVs in 250,21,000 Genome SR-GS samples with PARAGRAPH (Chen et al. 2019) to estimate AFs and calculated variant lengths and counts per sample across SV types for 50 UDN individuals with both SR-GS and LR-GS data. Allele frequencies were annotated with SVAFotate.

Rare TRE calling in UDN

We applied our MND method (Supplemental Text) to vamos (from LR-GS genomes) and ExpansionHunter (from SR-GS genomes) tandem repeat genotypes from the UDN to call extreme TREs. In the first round of analysis, we calculated the MND on the default catalogs present for each tool. For the LR-GS vamos set, we jointly analyzed the data with ADRC vamos genotypes as a reference set, whereas for the SR-GS ExpansionHunter, we used the ExpansionHunter repeat catalog that was run on 1000 Genomes as a reference set, downloaded here (https://github.com/Illumina/RepeatCatalogs/blob/master/hg38/genotype/1000 genomes/1kg.gt .hist.tsv.gz). We set $K = 0.5\%$ of total allele number, corresponding to $K = 5$ in the vamos set, and $K = 25$ in the ExpansionHunter set. We set a standardized MDN threshold of 2 above which to call an allele as an extreme TRE. In a secondary round of analysis, we called TREs specifically from the STRchive catalog of pathogenic repeats (Supplemental Table S3). We used the same vamos and ExpansionHunter output as described above to input into the MND TRE calling script. For ExpansionHunter, because the 1000 Genomes reference repeat catalog did not contain all of these pathogenic repeats, we used a reference set of an additional 54 short-read UDN genomes that were also called with ExpansionHunter (bringing total reference size for SR-GS to 104). To account for the smaller reference size of ExpansionHunter, we decreased k for calling from the pathogenic repeats to $k = 5$ for the ExpansionHunter set (~2% of the total allele number).

Outlier calling against a healthy population background in matching tissues

We collected RNA-seq STAR-mapped BAM files from 24 skeletal muscle biopsy in the CMG Muscular Disease data set (Cummings et al. 2017) and 66 fibroblast and 283 blood samples in the UDN data set. To better identify aberrant expression from healthy samples, we selected GTEx RNA samples from skeletal muscle, cell cultured fibroblast, and whole blood as the best matching tissues controls. To reduce variation due to sample quality issues, we filtered all RNA-seq based on RNA Integrity Number (RIN) >5.5. To quantify gene-level expression, we used RNA-SeQC to generate read count and TPM (Graubert et al. 2021) with GENCODE (Frankish et al. 2019) models matching the version of the STAR alignment BAMs. We limited analysis to genes with at least six reads in at least 20% samples in a tissue. GLOBINclear was used on cDNA library to maximize the number of nonglobin and disease-relevant mRNA to be detected in the UDN blood sample. This created a library compositional shift between GTEx whole blood and UDN globin-depleted blood. We generated TMM adjusted TPM values with edgeR (Robinson et al. 2010), accounting for the shift in RNA composition that TPM normalization itself cannot account for (Robinson and Oshlack 2010). Subsequently, we estimated expression PCs, which was previously shown (Zhou et al. 2022) to well-represent technical hidden covariates. TMM-TPM are log transformed using $\log_2$ (TPM + 2), then scaled to mean of 0 and standard deviation of 1. In addition to 60 PCs (Ferraro et al. 2020), we also accounted for batch, RIN, and biological sex using linear model, and scaled the expression residual from the model to obtain expression Z-scores. We removed global outliers following previous protocols (Ferraro et al. 2020).

Expression outlier enrichment nearby rare SVs analysis

To test for enrichment of expression outliers (eOutliers) nearby rare SVs we used a multivariate logistic regression framework. From 33 UDN individuals with LR-GS and blood transcriptomics, we set the proper background by restricting to genes with at least one outlier among these 33 individuals. To test the enrichment of SV types, we predicted outlier status at various Z-score thresholds by indicator variables for whether the gene had a rare SV of a given type within 10 kb (Outlier ~ DEL + INS + INV + DUP). We also tested outlier status against genes having an extreme TRE and ran a direction stratified model to separately test underexpression and overexpression. For testing enrichments across rare variant annotations, we used an absolute Z-score threshold of $z = 4$ and increased window size to 100 kb so we could capture more noncoding variants. Annotations for enrichment analysis were consistent with those that were used as features for WatershedSV (details below). To create ABC annotations, we downloaded cell-type-specific ABC elements from (ftp.broadinstitute.org/outgoing/lincRNA/ABC/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz) and subset them to those from primary, unstimulated cell types from blood, fibroblast, or muscle cell types (Nasser et al. 2021). Coordinates for ABC elements across tissues were then merged with plyranges after grouping by ABC element class (promoter, intergenic, genic) and target gene (Lee et al. 2019). These coordinates were then lifted over to GRCh38 (hg38) using CrossMap. VEP gene body categories were made exclusive using the following prioritization schema [exon variant, 5′-UTR variant, 3′-UTR variant, intron variant, upstream noncoding variant, downstream noncoding variant]. Any SV overlapping multiple features was assigned to highest priority category, making gene body region categories orthogonal. Gene body region enrichments were then calculated (Outlier ~ upstream_noncoding_variant + five_prime_UTR_variant + exon_variant + intron_variant + three_prime_UTR_variant + downstream_noncoding_vari ant). For length and CADD-SV Kleinert and Kircher 2022) enrichments, we ran multiple models, setting an increasing threshold for determining which rare SVs to indicate (Outlier ~ has_rare_SV_above_threshold). CADD-SV scores were generated using their recommended Snakemake pipeline (Kleinert and Kircher 2022). VEP impact categories enrichments were calculated with the model Outlier ~ has_MODIFIER_SV + has_MODERATE_SV + has_HIGH_SV.

Finally, we compiled noncoding regulatory annotations. For continuous variables (LINSIGHT, phastCons, REMAP CRM density), binary high/low categories were assigned based on median value. We set regulatory annotation features to 0 if SV also overlaps an exon, so they won't spuriously capture coding variant effects. We defined putative-regulatory SVs as those that overlap *any* noncoding annotation. We also create putative-nonregulatory variants as a control, variants that do not overlap any of the regulatory annotations and additionally do not overlap any high CADD-scoring SNVs. For each regulatory annotation, we compute enrichments by the following model: Outlier ~ has_coding_SV + has_SV_with_regulatory_annotation, controlling for gene effect from coding variants. We adjusted *P*-values from all enrichment testing using Bonferroni method.

Watershed-SV annotation collection pipeline

Given the increased complexity of annotating SVs relative to SNVs, Watershed-SV includes a more comprehensive set of relevant annotations: VEP (McLaren et al. 2016) variant category and protein-coding sequence consequence, SV-specific annotations like variant types and length, ChromHMM (Ernst and Kellis 2012; Roadmap Epigenomics Consortium et al. 2015) states, and conservation scores, among others. Since many SVs affect a range within the genome, leading to overlaps with multiple annotations from the same source, we summarized the impact of these SVs on nearby genes that aligns with increased functional impact for a given annotation. Since there could be more than one rare SV nearby a given gene, we further aggregated annotations according to Supplemental Table S5, highlighting the strongest impact of variants nearby gene. Both gene level predictions and gene-variant level predictions can be computed for an individual from the trained model. We describe each individual annotation in detail in Supplemental Text.

Further, two annotation collection methods are provided. First, we have a mode that adds user specified flanking sequence to each end of a given gene, then considers the aggregated impact of all rare SVs overlapping these regions as the genomic annotation. In this version, variant in gene body, UFR, and DFR are aggregated together, regardless of the specific location of variant overlap. In another mode, we separately consider nearby the gene body, UFR, and DFR regions to better isolate the impact of rare SVs on different regions that might be contributing to outlier gene expression. In both modes, annotations are only collected from the overlapping region between an SV and a gene with its +/− flanking sequences to limit the impact to *cis*-regulatory.

Evaluating Watershed-SV using N2-pairs in GTEx samples

For evaluating the Watershed model, we largely followed the N2-pair evaluation framework developed previously (Ferraro et al. 2020). To further account for variability within a single locus, specifically CNVs with multiple possible rare copy numbers, we required that copy numbers must at least match in the direction of copy number change from the population mode copy number. To generate a gene-level score for CADD-SV, we took the maximum CADD-SV score of SVs associated with each N2-pair individual gene pair. To account for noise in the gene expression outlier *Z*-scores, we evaluated using stringent outlier signals threshold at $Z > 3$ but relaxed the outlier threshold to a *P*-value threshold of 0.05 in the second individual in N2 pairs. To evaluate the performances of models on noncoding variants, we performed the same analysis but limited to only N2-pairs with noncoding rare SVs nearby. We further relaxed the *P*-value threshold to 0.1 due to the weakened impact of noncoding variants in comparison to coding variants (Supplemental Fig. S5C). To study the impact of rare SV distance to Watershed-SV and CADD-SV performances, we grouped multitissue median *Z*-score N2-pairs into 3 bins by minimum SV distance (0 bp, 1 bp–50 kbp, 50 kb–100 kb) away from gene, and proceed to generate precision-recall curve for each bin. Bootstrapping of AUPRC is performed identically as mentioned in previous literature.

Applying Watershed-SV models to prioritize functional rare SVs in the inherited muscular disease data set

SVs were called using svtools (Larson et al. 2019) and Parliament2 (Zarate et al. 2020). svtools jointly infers the presence of SVs based on LUMPY (Layer et al. 2014) breakpoints and CNVnator (Abyzov et al. 2011) copy number changes. Due to the small cohort size, this approach is underpowered in calling rare duplications or deletions. We used Parliament2 on each sample, which merges variant calls from multiple callers for a single individual, to maximize recall of rare and private SVs that failed to be called from SVtools. Due to the high false positive rate of SV calling in short-read data, we required an SV to have the support from at least two different callers (Zarate et al. 2020). Subsequently, we merged individual sample VCFs using Jasmine. Finally, we harmonized the Parliament2 and SVtools callset using Jasmine. Variant allele frequencies were ascertained using SVAFotate and PARAGRAPH as described earlier. We collected annotations for rare SVs found in the data set, combined with skeletal muscle expression outliers to run Watershed-SV prediction mode based on the GTEx 100 kb region-specific median *Z*-score model.

Applying Watershed-SV models to prioritize rare SVs and extreme TR expansions in UDN data set

We applied both the median *Z*-score model and the tissue-Watershed model that was trained jointly with blood and fibroblast outlier signals to the patient data. For rare SVs, predictions were performed as described in the previous section. For extreme TR expansions, we cast them to the type of duplication and calculated the length as the total length of the expanded repeat. To evaluate whether Watershed-SV identified symptom-relevant rare disease variant gene pairs, as well as to evaluate performance against CADD-SV, we designed a set of filters to simulate a range of diagnostic filters that genetic counselors would use. The filters encompass three general approaches: matching of nearby gene with the patient's symptoms, variants that have strong evidence of pathogenicity from whole-genome sequencing-based variant prioritization methods, and variant prioritization leveraging RNA outlier signals. We compared CADD-SV to Watershed-SV to assess how CADD-SV scores for the same set of variants nearby 100 kb of genes with measured expression in blood or fibroblast compared to Watershed-SV scores. We note that CADD-SV is a variant centric prioritization approach and does not model the variant impact on a nearby gene. Furthermore, the CADD-SV score is on a Phred Scale, whereas Watershed-SV reports a probability. The same CADD-SV score is used for two gene-SV pairs that are the same SV but different gene in Watershed-SV. We used a Watershed threshold of 0.6, corresponding to ~10% of the total variants, and a roughly equivalent threshold of 10 in CADD-SV, corresponding to the top 10% pathogenicity. To compare whether variants prioritized are relevant to disease phenotypes, we created HPO term matching using Phen2Gene (Zhao et al. 2020), considering all genes for each patient with rank < 300 as recommended by authors.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI database of Genotypes and

Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) under accession numbers phs001232 and phs000424. Methods, scripts, and source annotations are available at GitHub (https://github.com/jasonbhn/Watershed-SV) and as Supplemental Code. We recommend execution of the pipeline through conda package watershed-sv (https://anaconda.org/dnachun/watershed-sv) or run using docker (ghcr.io/jasonbhn/watershed-sv: 0.1.9) following our instructions at GitHub.

Competing interest statement

S.B.M. is an advisor to BioMarin, MyOme, and Tenaya Therapeutics. A.B. is a co-founder of CellCipher, Inc, is a shareholder in Alphabet, Inc, and has consulted for Third Rock Ventures, LLC. E.A.A. is the founder of Personalis, Deepcell, Svexa, RCD Co, Parameter Health, an advisor for SequenceBio, Foresite Labs, PacBio, a nonexecutive director at AstraZeneca, hold stocks in Oxford Nanopore, Pacific Biosciences, AstraZeneca, and offers collaborative support in kind to Illumina, Pacific Biosciences, Oxford Nanopore.

Acknowledgments

We thank Benjamin Strober, Taibo Li for suggestions on Watershed-SV model, Joshua Weinstock and Rebecca Keener for editing of this manuscript and helpful conversations related to this work, and Daniel Nachun, Alex Miller, and Paul Petrowski for software engineering. Research reported in this manuscript was in part supported through the Undiagnosed Diseases Network (Award U01HG010218) and the GREGoR Consortium (Award U01HG011762). T.D.J. is supported by U01HG011762, T32HG000044. B.N. is supported by R35GM139580, U24HG010263, OT2OD034190, U01CA253481, R03CA272952, U01HG012069. J.E.G. is supported by U01HG010218, U01HG011762. S.F. is supported by 1R21HG013397, 5R01NS072248. C.M.R. is supported by U01HG010218, U01HG011762. D.E.B. is supported by U01HG011762, U01NS134358. R.A.U. is supported by U01HG011762. P.C.G. is supported by U01HG011762. A.N.R. is supported by U01HG010218. E.A.A. is supported by U01HG010218, U01HG011762 J.A.B. is supported by U01HG011762 and U01NS134358. S.Z. is supported by 1R21HG013397, 5R01NS072248. M.D.G. is supported by R35AG072290, P30AG066515, R01AG074339, R01AG048076. S.B.M. is supported by U01HG011762, U01AG072573, R01AG066490, R01MH125244, and U01HG012069. M.C.S. is supported by U24HG010263, OT2OD034190, U01CA253481, R03CA272952. M.W. is supported by U01HG010218, U01HG011762. A.B. is supported by R35GM139580, U01HG012069. This work utilized computing resources provided by the Stanford Genetics Bioinformatics Service Center, supported by National Institutes of Health Instrumentation Grant S10OD025082, and would not have been possible without the support of the Stanford SCG cluster system administrators.

Author contributions: T.D.J., B.N. did formal analysis, visualization, writing—original draft. C.M.R. and D.B. did data curation, investigation, and writing—review and editing. J.E.G., R.A.U., P.C.G. did investigation, methodology, and writing—review and editing. S.F., S.Z. did formal analysis and writing—review. A.R. did formal analysis and methodology. M.D.G. did resources, funding acquisition, and writing—review and editing. E.A.A., J.A.B., S.B.M., M.C.S., M.T.W., A.B. did conceptualization, funding acquisition, resources, supervision, and writing—review and editing.

References

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583:** 83–89. doi:10.1038/s41586-020-2371-0

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21:** 974–984. doi:10.1101/gr.114876.110

Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, Hijazi H, Alshammari M, Aldahmesh MA., Salih MA, et al. 2015. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep* **10:** 148–161. doi:10.1016/j.celrep.2014.12.015

Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al. 2019. Characterizing the major structural variant alleles of the human genome. *Cell* **176:** 663–675.e19. doi:10.1016/j.cell.2018.12.019

Bakhtiari M, Park J, Ding Y-C, Shleizer-Burko S, Neuhausen SL, Halldórsson BV, Stefánsson K, Gymrek M, Bafna V. 2021. Variable number tandem repeats mediate the expression of proximal genes. *Nat Commun* **12:** 2075. doi:10.1038/s41467-021-22206-z

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10:** 1784. doi:10.1038/s41467-018-08148-z

Chemparathy A, Le Guen Y, Zeng Y, Gorzynski J, Jensen TD, Yang C, Kasireddy N, Talozzi L, Belloy M, Stewart I, et al. 2024. A 3'UTR insertion is a candidate causal variant at the *TMEM106B* locus associated with increased risk for FTLD-TDP. *Neurol Genet* **10:** e200124. doi:10.1212/NXG.0000000000200124

Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, et al. 2019. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* **20:** 291. doi:10.1186/s13059-019-1909-7

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49:** 692–699. doi:10.1038/ng.3834

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. *Genet Med* **24:** 1336–1348. doi:10.1016/j.gim.2022.02.007

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581:** 444–451. doi:10.1038/s41586-020-2287-8

Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Reghan FA, Bolduc V, Waddell LB, Sandaradura SA, O'Grady GL, et al. 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* **9:** eaal5209. doi:10.1126/scitranslmed.aal5209

De Coster W, Rademakers R. 2023. Nanopack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39:** btad311. doi:10.1093/bioinformatics/btad311

Deng J, Yu J, Li P, Luan X, Cao L, Zhao J, Yu M, Zhang W, Lv H, Xie Z, et al. 2020. Expansion of GGC repeat in *GIPC1* is associated with oculopharyngodistal myopathy. *Am J Hum Genet* **106:** 793–804. doi:10.1016/j.ajhg.2020.04.011

Depienne C, Mandel J-L. 2021. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am J Hum Genet* **108:** 764–785. doi:10.1016/j.ajhg.2021.03.011

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. doi:10.1093/bioinformatics/btz431

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9:** 215–216. doi:10.1038/nmeth.1906

Fazal S, Danzi MC, Xu I, Kobren SN, Sunyaev S, Reuter C, Marwaha S, Wheeler M, Dolzhenko E, Lucas F, et al. 2024. RExPRT: a machine learning tool to predict pathogenicity of tandem repeat loci. *Genome Biol* **25:** 39. doi:10.1186/s13059-024-03171-4

Ferraro NM, Strober BJ, Einson J, Abell NS, Aguet F, Barbeira AN, Brandt M, Bucan M, Castel SE, Davis JR, et al. 2020. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369:** eaaz5900. doi:10.1126/science.aaz5900

Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47:** D766–D773. doi:10.1093/nar/gky955

Frésard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, Bonner D, Kernohan KD, Marwaha S, Zappala Z, et al. 2019. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med* **25:** 911–919. doi:10.1038/s41591-019-0457-8

Gao T, Qian J. 2020. Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res* **48:** D58–D64. doi:10.1093/nar/gkaa197

Geng C, Zhang C, Li P, Tong Y, Zhu B, He J, Zhao Y, Yao F, Cui LY, Liang F, et al. 2023. Identification and characterization of Two DMD pedigrees with large inversion mutations based on a long-read sequencing pipeline. *Eur J Hum Genet* **31:** 504–511. doi:10.1038/s41431-022-01190-y

Graubert A, Aguet F, Ravi A, Ardlie KG, Getz G. 2021. RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics* **37:** 3048–3050. doi:10.1093/bioinformatics/btab135

Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D, Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34:** 2061–2073. doi:10.1101/gr.279273.124

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, et al. 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* **48:** 22–29. doi:10.1038/ng.3461

Hammal F, de Langen P, Bergon A, Lopez F, Ballester B. 2022. ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res* **50:** D316–D325. doi:10.1093/nar/gkab996

Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35:** 2907–2915. doi:10.1093/bioinformatics/btz041

Hiatt SM, Lawlor JMJ, Handley LH, Ramaker RC, Rogers BB, Christopher PE, 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. doi:10.1016/j.xhgg.2021.100023

Huang Y-F, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* **49:** 618–624. doi:10.1038/ng.3810

Ishiura H, Shibata S, Yoshimura J, Suzuki Y, Qu W, Doi K, Almansour MA, Kikuchi JK, Taira M, Mitsui J, et al. 2019. Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nat Genet* **51:** 1222–1232. doi:10.1038/s41588-019-0458-z

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* **21:** 189. doi:10.1186/s13059-020-02107-y

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR., Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC genome browser database: 2014 update. *Nucleic Acids Res* **42:** D764–D770. doi:10.1093/nar/gkt1168

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. doi:10.1038/s41592-022-01753-3

Kleinert P, Kircher M. 2022. A framework to score the effects of structural variants in health and disease. *Genome Res* **32:** 766–777. doi:10.1101/gr.275995.121

Kohler JN, Legro NR, Baldridge D, Shin J, Bowman A, Ugur B, Jackstadt MM, Shriver LP, Patti GJ, Zhang B, et al. 2024. Loss of function of FAM177A1, a Golgi complex localized protein, causes a novel neurodevelopmental disorder. *Genet Med* **26:** 101166. doi:10.1016/j.gim.2024.101166

Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, Asri M, Alvarez Jerez P, Malik L, Dewan R, et al. 2023. Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat Methods* **20:** 1483–1492. doi:10.1038/s41592-023-01993-x

Kovaka S, Ou S, Jenike KM, Schatz MC. 2023. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nat Methods* **20:** 12–16. doi:10.1038/s41592-022-01716-8

Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, Layer RM, Hall IM. 2019. Svtools: population-scale analysis of structural variation. *Bioinformatics* **35:** 4782–4787. doi:10.1093/bioinformatics/btz492

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15:** R84. doi:10.1186/gb-2014-15-6-r84

Lee S, Cook D, Lawrence M. 2019. Plyranges: a grammar of genomic data transformation. *Genome Biol* **20:** 4. doi:10.1186/s13059-018-1597-8

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37:** 4572–4574. doi:10.1093/bioinformatics/btab705

Li T, Ferraro N, Strober BJ, Aguet F, Kasela S, Arvanitis M, Ni B, Wiel L, Hershberg E, Ardlie K, et al. 2023. The functional impact of rare variation across the regulatory cascade. *Cell Genomics* **3:** 100401. doi:10.1016/j.xgen.2023.100401

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019. Structural variant calling: the long and the short of it. *Genome Biol* **20:** 246. doi:10.1186/s13059-019-1828-7

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. doi:10.1038/s41467-024-44804-3

Marwaha S, Knowles JW, Ashley EA. 2022. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. *Genome Med* **14:** 23. doi:10.1186/s13073-022-01026-w

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The ensembl variant effect predictor. *Genome Biol* **17:** 122. doi:10.1186/s13059-016-0974-4

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, et al. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* **20:** 159–163. doi:10.1038/gim.2017.86

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. doi:10.1016/j.ajhg.2021.06.006

Miyatake S, Koshimizu E, Fujita A, Doi H, Okubo M, Wada T, Hamanaka K, Ueda N, Kishida H, Minase G, et al. 2022. Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *NPJ Genom Med* **7:** 62. doi:10.1038/s41525-022-00331-y

Montgomery SB, Bernstein JA, Wheeler MT. 2022. Toward transcriptomics as a primary tool for rare disease investigation. *Cold Spring Harb Mol Case Stud* **8:** 2. doi:10.1101/mcs.a006198

Narkis G, Ofir R, Landau D, Manor E, Volokita M, Hershkowitz R, Elbedour K, Birk OS. 2007. Lethal contractural syndrome type 3 (LCCS3) is caused by a mutation in *PIP5K1C*, which encodes PIPKI gamma of the phosphatidylinsitol pathway. *Am J Hum Genet* **81:** 530–539. doi:10.1086/520771

Nasser J, Bergman DT, Fulco CP, Guckelberger P, Doughty BR, Patwardhan TA, Jones TR, Nguyen TH, Ulirsch JC, Lekschas F, et al. 2021. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593:** 238–243. doi:10.1038/s41586-021-03446-x

Nicholas TJ, Cormier MJ, Quinlan AR. 2022. Annotation of structural variants with reported allele frequencies and related metrics from multiple datasets using SVAFotate. *BMC Bioinform* **23:** 490. doi:10.1186/s12859-022-05008-y

Quinlan AR. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinform* **47:** 11.12.1–34. doi:10.1002/0471250953.bi1112s47

Ren J, Gu B, Chaisson MJP. 2023. Vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol* **24:** 175. doi:10.1186/s13059-023-03010-y

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518:** 317–330. doi:10.1038/nature14248

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11:** R25. doi:10.1186/gb-2010-11-3-r25

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Sanford Kobayashi E, Batalov S, Wenger AM, Lambert C, Dhillon H, Hall RJ, Baybayan P, Ding Y, Rego S, Wigby K, et al. 2022. Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. *Sci Rep* **12:** 16945. doi:10.1038/s41598-022-20113-x

Schobers G, Derks R, Ouden Ad, Swinkels H, van Reeuwijk J, Bosgoed E, Lugtenberg D, Sun SM, Corominas Galbany J, Weiss M, et al. 2024. Genome sequencing as a generic diagnostic strategy for rare disease. *Genome Med* **16:** 32. doi:10.1186/s13073-024-01301-y

Scott AJ, Chiang C, Hall IM. 2021. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* **31:** 2249–2257. doi:10.1101/gr.275488.121

Sharo AG, Hu Z, Sunyaev SR, Brenner SE. 2022. StrVCTVRE: a supervised learning method to predict the pathogenicity of human genome structural variants. *Am J Hum Genet* **109:** 195–209. doi:10.1016/j.ajhg.2021.12.007

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily

conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050. doi:10.1101/gr.3715005

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* **42:**1571–1580. doi:10.1038/s41587-023-02024-y

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526:** 75–81. doi:10.1038/nature15394

Ungar RA, Goddard PC, Jensen TD, Degalez F, Smith KS, Jin CA, Undiagnosed Diseases Network, Bonner DE, Bernstein JA, Wheeler MT, et al. 2024. Impact of genome build on RNA-seq interpretation and diagnostics. *Am J Hum Genet* **111:** 1282–1300. doi:10.1016/j.ajhg.2024.05.005

Vanderstichele T, Burnham KL, de Klein N, Tardaguila M, Howell B, Walter K, Kundu K, Koeppel J, Lee W, Tokolyi A, et al. 2024. Misexpression of inactive genes in whole blood is associated with nearby rare structural variants. *Am J Hum Genet* **111:** 1524–1543. doi:10.1016/j.ajhg.2024.06.017

Wang Y, Song F, Zhang B, Zhang L, Xu L, Kuang D, Li D, Choudhary MNK, Li Y, Hu M, et al. 2018. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* **19:** 151. doi:10.1186/s13059-018-1519-9

Wojcik MH, Reuter CM, Marwaha S, Mahmoud M, Duyzend MH, Barseghyan H, Yuan B, Boone PM, Groopman EE, Délot EC, et al. 2023. Beyond the exome: what's next in diagnostic testing for Mendelian conditions. *Am J Hum Genet* **11:** 1229–1248. doi:10.1016/j.ajhg.2023.06.00

Xu Z, Li Q, Marchionni L, Wang K. 2023. PhenoSV: interpretable phenotype-aware model for the prioritization of genes affected by structural variants. *Nat Commun* **14:** 7805. doi:10.1038/s41467-023-43651-y

Yu J, Luan X-H, Yu M, Zhang W, Lv H, Cao L, Meng L, Zhu M, Zhou B, Wu X-R, et al. 2021. GGC repeat expansions in NOTCH2NLC causing a phenotype of distal motor neuropathy and myopathy. *Ann Clin Transl Neurol* **8:** 1330–1342. doi:10.1002/acn3.51371

Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, Schatz MC, Boerwinkle E, Gibbs RA, Sedlazeck FJ. 2020. Parliament2: accurate structural variant calling at scale. *GigaScience* **9:** giaa145. doi:10.1093/gigascience/giaa145

Zhao M, Havrilla JM, Fang L, Chen Y, Peng J, Liu C, Wu C, Sarmady M, Botas P, Isla J, et al. 2020. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform* **2:** lqaa032. doi:10.1093/nargab/lqaa032

Zhou HJ, Li L, Li Y, Li W, Li JJ. 2022. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* **23:** 210. doi:10.1186/s13059-022-02761-4

Received March 15, 2024; accepted in revised form January 6, 2025.

Integration of transcriptomics and long-read genomics prioritizes structural variants in rare disease

Tanner D. Jensen, Bohan Ni, Chloe M. Reuter, et al.

Genome Res. published online March 20, 2025
Access the most recent version at doi:10.1101/gr.279323.124