Online Supplement

Long reads allow identification of longer, more complex variation at tandem repeat loci

Expansions of tandem repeats are known to cause multiple human diseases including myotonic dystrophy(Paulson, 2018). We tested the ability of long-read sequencing to resolve tandem repeat (TR) variation in the UDN cohort. To this end, we genotyped TR copy numbers with vamos (Ren et al., 2023) across 466,220 loci using LR-GS and ExpansionHunter across 174,260 loci using SR-GS (Dolzhenko et al., 2019) (**Figure 2b**). We used the recommended TR catalogs–a reference of where TRs exist in the genome and what the repeating motifs are–to run each tool and found they were widely characterizing different TRs; only 18% of loci were shared and, of those, 48% were annotated with different motifs (**Figure S2a**). Vamos defined their TR catalog from loci that varied in 32 haplotype-resolved PacBio HiFi genomes from the Human Genome Structural Variation Consortium and was designed to capture mostly complex variable number tandem repeats (VNTR) with motif lengths 7bp+, while the ExpansionHunter default catalog mainly captures short tandem repeats (STR) (2-6bp) that are small enough to be resolved with short reads. For the UDN cohort, we estimated the mean and variance of these TR copy numbers at each locus from both vamos (LR-GS) and ExpansionHunter (SR-GS) (**Table S2**). We discovered that vamos characterized loci that have higher mean TR copy numbers than ExpansionHunter, meaning LR-GS could detect longer repeats (**Figure 2b**). Further, the distribution of vamos TRs had a much longer tail for repeat motif length, mean repeat copy number, and repeat copy number variance than ExpansionHunter (**Figure S2c, S2d, S2e**). While almost all ExpansionHunter TRs had a mean repeat length of <100bp, vamos resolved TRs that had mean lengths well above 1kb (**Figure S2e**).

Direct comparison of LR-GS and SR-GS TRs can be impacted by their disparate TR catalog loci. To address this, we ran both tools using an identical TR catalog from STRchive (harrietdashnow.com/STRchive), a highly-curated set of 68 well-characterized and pathogenic TR loci. This allowed us to directly compare the LR-GS vamos and SR-GS ExpansionHunter on repeats with known disease-relevance. We calculated mean TR copy numbers across these 68 loci with both vamos and ExpansionHunter (**Table S3**). Vamos characterized all of the loci whereas ExpansionHunter only called 50 (73.5%) since some of these loci are longer than SR-GS fragment sizes. Of those called by short reads, we find a strong correlation across tools for the estimated mean TR copy number (Spearman correlation 0.833) (**Figure 2f**); however, there were a few repeats with larger discrepancies, most notably the CGC repeat in the gene XYLT1 that leads to the rare disease Desbuquois dysplasia-2 (ExpansionHunter mean=22, vamos mean=58), the AAGGG repeat in the *RFC1* gene that leads to Cerebellar ataxia with neuropathy and vestibular areflexia syndrome (ExpansionHunter mean=16, vamos mean=37), and the CCTG repeat in *CNBP* that leads to Myotonic dystrophy type 2 (ExpansionHunter mean=17, vamos mean=38) (**Table S3**). As vamos was developed for noisier long-read data, it is more permissive of assigning non-perfectly matching repeats to the input motif, particularly in nested repeats where the pathogenic motif is flanked by others at the same locus, which may explain why vamos calls higher repeat copy number means compared to ExpansionHunter. For each repeat locus, the correlation across individual haplotypes was more variable (median Spearman's correlation = 0.635, range= 0.163 – 0.850) (**Figure S2g**).

Identification of Rare Tandem Repeat Expansions

Similarly, we sought to assess TR distributions to identify rare tandem repeat expansions (TREs) in the UDN cohort, that is, extreme outliers of TR copy number. Using the ADRC reference for long-read sequencing (LR-GS), we developed an algorithm to call rare TREs by

thresholding on a standardized mean neighbor distance (MND) (see **Methods)**. This method evaluates how much longer on average an allele is from its K-nearest neighbors in the repeat copy number distribution. This statistic is designed to distinguish clearly separated and expanded alleles from those that are just at the tail of the distribution, enabling us to identify samples with rare repeat expansions (see schematic in **Figure 1b**). We apply this method to both long read and short read TR genotypes . Setting K = 0.5% of total allele number, we found a median of 14 TREs per individual, in comparison to median of 427 using Tukey's outlier approach based on empirical TR copy number distribution, greatly reducing the false positive TRE calls.

To compare the characteristics of LR-S and SR-GS TREs, we additionally assessed allele frequency and identified rare TREs from SR-GS with MND using genome-wide repeat catalog of 1000G as a reference. While ExpansionHunter detects more rare TREs from SR-GS that are STRs (2-5 bp) (median EH = 22, median vamos = 8), vamos detected more rare variable number tandem repeats (VNTR, 7+bp) TREs from LR-GS (median EH = 0, median vamos = 11). Further, vamos TREs were in loci with longer motif unit lengths (median = 22) compared to ExpansionHunter (median = 2) (**Figure S3c**). ExpansionHunter and vamos rare TREs had similar distributions of standardized MNDs, yet vamos was able to find outliers in repeats with longer mean lengths, and the expansion allele lengths were much larger in the LR-GS TREs compared to SR-GS data (**Figure S3d, S3e, S3f**).

Supplemental Methods

Allele Frequency Estimation

To assess SV allele frequency, we ran SVAFotate on the UDN SR Manta-Jasmine-merged callset, UDN LR-consensus callset, and IMD SR

SVTools-Parliament2-Jasmine-merged calls with 0.5 overlap fraction(Nicholas et al., 2022). For the SR callset, we saw high ascertainment of variants in the SVAFotate catalog, 88%, in comparison to only 8% for LR callset.

To further assess SVs not detected with SVAFotate, we genotyped all no-hit SVs with PARAGRAPH genotyper in 2,499 high-depth 1000 Genome unrelated individuals(Byrska-Bishop et al., 2022), allowing us to obtain allele frequency (AF) estimates for 97% of SR-GS variants and 92% of LR-GS variants. To get more robust AFs for ONT and address SVs outside of the SR-GS detection limit using PARAGRAPH, we used consensus SV calls from LR-GS of Stanford's Alzheimer's Disease Research Center (ADRC) and the Stanford aging and memory study (SAMS) (n=571) sequencing collection(Chemparathy et al., 2023). The ADRC contains 598 individuals sequenced on the same ONT PromethION platform to a median coverage of 60 Gb. An identical SV calling and merging workflow was run on ADRC. To get AFs, we ran a population Jasmine merge on all genomes from ADRC and UDN (n=667). From the population-merged VCF, AFs were estimated by dividing summed allele counts identified by total allele number in the merged VCF.

Tandem Repeat Copy Number Genotyping

For long-read sequencing data, the vamos efficient motif set was downloaded from (https://github.com/ChaissonLab/vamos/blob/master/snakefile/configs/vntr_region_motifs.e.bed.gz), and used as input into vamos (v1.3.6) to genotype tandem repeat (TR) copy numbers across across 467,000 TR loci (Ren et al., 2023). Vamos vcf files were converted to TSV files to report per sample and haplotype length of each TR (vcf INFO FIELD LEN_H1, or LEN_H2). We considered only repeat annotation lengths and not composition information.

For short-read sequencing data, TR copy numbers were estimated with ExpansionHunter (v5.0.0) (Dolzhenko et al., 2019). The recommended TR catalog for ExpansionHunter was downloaded from

(https://github.com/Illumina/RepeatCatalogs/blob/master/hg38/variant_catalog.json). The resulting output json was converted into a tsv file to report per sample per haplotype estimated repeat numbers and confidence intervals. TRs were again annotated based on their motif length. We calculated mean and variance per TR across both technologies. TRs were filtered to those that were detected in at least 50% of UDN individuals and detected in a minimum of 50 haplotypes. This resulted in 466,220 TR loci for vamos and 174,260 for ExpansionHunter. For comparing the ExpansionHunter variant catalog to the vamos TR catalog, we calculated genomic ranges based on coordinates found in their respective catalogs, and we flanked each region by 100bp on each side. The ExpansionHunter and vamos motif sets were then intersected with bedTools to determine if they were covering the same loci. For loci in both catalogs that did intersect, we further characterized if they contained the same motif. Vamos includes potentially multiple motifs at each TR in order to represent composition changes that occur within repeats, whereas ExpansionHunter only records a single motif. To compare if ExpansionHunter and vamos contain the same motif, we compared the ExpansionHunter motif with any motif in the vamos set, or any cyclic permutation of a motif in the vamos set (because, for instance, a CGC repeat and a CCG repeat would result in identical repeat sequence).

To characterize tandem repeat variation at known pathogenic loci, we downloaded the STRchive motif set from (github.com/hdashnow/STRchive/blob/main/data/hg38.STRchive-disease-loci.TRGT.bed). We analyzed the pathogenic motif set with ExpansionHunter and Vamos as described above. The resulting haplotype-level tsv files were concatenated across the cohort in each technology and then joined on TR_id. Since assignment of haplotypes in both tools is arbitrary, to compare haplotype-specific repeat copy numbers, we assign H1 to be the haplotype with the longer repeat count (max) and H2 to be the haplotype with the shortest repeat count (min). TRs were filtered if there was >50% missingness in individuals, removing 20 TRs from ExpansionHunter. For the remaining TRs, EH mean and vamos mean repeat copy numbers were calculated, and

per loci Spearman correlations between reassigned (max,min) haplotype repeat copy numbers were computed.

Mean Neighbor Distance Rare Tandem Repeat Expansion Calling

To detect rare and extreme tandem repeat expansions (TRE) at tandem repeat loci, we developed a method which involves calculating and thresholding on a mean neighbor distance (MND) statistic. This semiparametric approach uses a population reference (preferably of healthy controls) to jointly define tandem repeat copy number distributions together with any samples of interest. Then for each allele in the distribution, we calculate how far on average they are from their K-nearest neighbors (mean of the minimum K absolute differences). We then standardized this value dividing by the standard deviation of the original distribution. This statistic describes how many standard deviations away an allele is from their nearest K-neighbors. Here, if an allele was drawn from the same distribution as the rest of the alleles, then its nearest neighbors would be very close, and the standardized MND would be close to zero. Even if an individual was at the tail of the distribution, there would still be neighbors nearby and the standardized MND will still be close to zero. However, if an allele was extremely expanded, as is the case in repeat expansion disease, they will be sampled from a distinct distribution and should be clearly separated from the rest of the distribution. This means the k-nearest neighbors should be far away and the MND will be high. Thus, thresholding on a standardized MND can differentiate extreme expansions from just individuals who are near the tail of the distribution (see schematic in **Figure 1B**).

This MND method is analogous to a *Z*-score, but instead of calculating difference from an observation to the population mean, we replace the population mean with a local mean of the k-nearest neighbors of the observation. Thus, MND can be thought of as a topological or local *Z*-score. Additionally, to require the repeat expansion to be at the tail of the distribution and not

at an intermediate value between two mixtures, for instance, we also require the observation to be in the top K in order to be considered an expansion outlier (or bottom K in the case of rare extreme contractions). Finally, to avoid inflation of standardized MNDs from TRs with little variability, we forced a lower bound on the standard deviation of the original distribution by the expected standard deviation if the distribution was Poisson distributed, namely the standard deviation of the mean.

We observed that this method works well for detecting rare and extreme repeat expansions. Benefits of this model included its nonparametric estimation of the MND based only on nearest neighbors. This makes our method less susceptible to violations of underlying assumptions like the unimodality of data that is often assumed for standard *Z*-score scaling outlier methods, an assumption commonly broken by the multimodal distributions of many TRs. Also outlier calling methods like Tukey's assume a wide variance of the distribution in order to calculate interquartile ranges, however, many tandem repeat loci are pathological in this regard as the 25th and 75th percentile will have the same value, while still being variable in the remaining 25% of data points, thereby making any other value an outlier.

To call repeat expansions with this method there are two parameters to set. First, K to determine the number of nearest neighbors to consider when calculating the local *Z*-scores. As K increases, the local *Z*-score will approach the standard global *Z*-score and local information about the distribution will diminish. A default K parameter proportional to the total allele number in the reference population is recommended, for instance K = 0.5% of total alleles. This mirrors rare variants at a < 0.01 MAF, where if we see 1% or more of the alleles having the same or similar repeat number to a given allele, it will have a MND close to zero and will not be called a TRE. For determining ultra-rare events, a lower k threshold of 0.01% of total allele number can be used. Whereas more lenient tandem repeat expansion calling can be done at 2%. In the most extreme case, if K = 1, then an allele is only an outlier if no other allele is nearby, which should only be done in a rare disease setting if it is absolutely sure no other allele could have

the same pathogenic repeat. We show increasing K results in more TREs called on average per sample (**Supplemental Figure 3A**). The second parameter is the standardized MND threshold above which to call a sample an rare TRE. We used a default parameter of 2, but this can be increased in order to be more stringent and detect only the most extremely expanded of alleles. We also show that increasing this standardized MND threshold results in less and less TREs called per genome on average (**Supplemental Figure 3A**).

To benchmark our MND statistic to other outlier calling methods for tandem repeat expansion calling, we also called expansions from vamos output using two additional methods. We used the joint distribution of UDN combined with ADRC to perform all tandem repeat expansion calling. A Tukey's outlier method was applied by setting an outlier threshold per TR equal to the 75th percentile of the distribution + 1.5✕IQR of the distribution. Any allele that had a repeat count greater than this value was called an outlier. We also used a *Z*-score scaling method where repeat counts per tandem repeat were scaled to a mean of 0 and standard deviation of 1 by subtracting the population mean and dividing by the population standard deviation. A *Z*-score threshold of 3 was used to compare with Tukey's and MND threshold. We compared counts of tandem repeat expansions detected by each method from the UDN genomes and found MND drastically decreases the number of TREs compared to Tukey's method. Compared to the *Z*-score method, the MND method calls about half as many TREs as *Z*-scores (**Supplemental Figure 3C**).

SR-GS SV-calling in the CMG cohort

In the CMG cohort, we called rare SVs from the SR-GS data and We jointly called SVs using svtools (Larson et al., 2019) and Parliament2 (Zarate et al., 2020). After harmonizing variant calls with Jasmine, we called 11,360 deletions, 9,346 insertions, 3,820 inversions, and 592 duplications. Due to the challenges of calling rare SVs with a small cohort, we queried 1000

Genomes, CCDG, and gnomAD SV reference callsets with SVAFotate (Nicholas et al., 2022) and genotypes putative novel variants in 1000 Genomes SR-GS with PARAGRAPH (Chen et al., 2019) to annotate population SV allele frequencies (AF). Population database queries and PARAGRAPH enabled the annotation of AF in 97% of SVs.

Watershed-SV annotation features

SV-generated features

We one-hot encoded the type of rare SVs present nearby a gene like a presence-abscence vector. Allele frequency, length of the SV, copy numbers of variants could both be processed separately and uploaded as TSV or extracted from VCF file automatically. We log transformed length since it's a continuous positive feature in our feature encoding. AFractions of exons sequence affected by nearby SVs were calculated and aggregated using the maximum among SVs to depict the most deleterious impact. Similarly, binary features describing exon truncations by SV from either 5' or 3' of the genes were included. For noncoding variants, distances of variants to UFR and DFR were calculated, with the aggregated minimum among nearby SVs being used for gene-level feature curations; note, when variant is coding, these distances are 0.

VEP features

We one-hot encoded the Consequences from VEP into binary features. We replaced the exon_variant, intron_variant consequences from VEP with our own script because the VEP module can simultaneously call an SV being intronic and exonic, producing contradictory information for model training.

Regulatory element annotations

In order to capture the impact of rare SVs on various regulatory elements, we used annotation from multiple sources. Activity-by-contact(ABC) model predictions were extracted and aggregated from 78 cell lines and tissues, we considered any rare SVs nearby genes that overlap an ABC enhancer mapped to the given gene an ABC SV. We extracted cis-regulatory modules(CRMs) from REMAP2022 database (Hammal et al., 2022) and considered the maximum SV disrupted CRM score the remap score annotation for a gene. And we also collected information about tissue specificity of enhancers in the form of the number of primary tissues a given enhancer bed segment is detected in from enhancerAtlas 2.0 (Gao & Qian, 2020). Finally, the number of tissues a TAD boundary is detected in Wang et al (Wang et al., 2018) are collected to depict the impact of SV on regulations related to 3D genome organizations.

ChromHMM features

To collect chromHMM features, we selected 27 primary tissue/cell types from 127 epigenomes. We used the 25 state model generated by the Roadmap Epigenome consortium(Roadmap Epigenomics Consortium et al., 2015). We aggregated overlapping segments of the same state from 27 tissues into a single segment, with the count of tissues that the state is active in, forming modules of chromatin states. The higher the number of tissues, the more ubiquitously observed a state is.

Conservation Scores and other bigwig track features from UCSC Genome Browser

We used pybigwig(Ryan et al., 2016) to collect and summarize the score tracks of various conservation metrics, including LINSIGHT, CADD, PhastCON, GC-content, and percent

CpG in a disrupted CpG island from the UCSC Genome Browser(Karolchik et al., 2014). LINSIGHT, CADD, PhastCON have high sparsity in the scores, therefore, we selected the top 10 scores within the range of each SV, and took the average of the top 10 as the annotation for these score tracks. For GC-content, the mean of GC was taken. And for CpG, the max CpG percentage was considered.

Region-specific annotation aggregation

For region-specific annotation, conservation scores, regulatory annotations, chromHMM features are summarized and aggregated in the gene body, UFR, and DFR regions separately. Variant type information is also collected such that if an SV overlaps with the gene body and UFR, the SV segments that overlaps each regions will be considered separately, i.e. SVTYPE: DUP_UFR, DUP_gene_body. Other annotations, such as VEP, are aggregated as we did in non-region-specific aggregation.

References

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426–3440.e19.

Chemparathy, A., Guen, Y. L., Zeng, Y., Gorzynski, J., Jensen, T., Yang, C., Kasireddy, N., Talozzi, L., Belloy, M. E., Stewart, I., Gitler, A. D., Wagner, A. D., Mormino, E., Henderson, V. W., Wyss-Coray, T., Ashley, E., Cruchaga, C., & Greicius, M. D. (2023). A 3’UTR Insertion Is a Candidate Causal Variant at the TMEM106B Locus Associated with Increased Risk for FTLD-TDP. *medRxiv : The Preprint Server for Health Sciences*.

https://doi.org/10.1101/2023.07.06.23292312

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D. R., Schatz, M. C., Sedlazeck, F. J., & Eberle, M. A. (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, *20*(1), 291.

Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., … Eberle, M. A. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* , *35*(22), 4754–4756.

Gao, T., & Qian, J. (2020). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, *48*(D1), D58–D64.

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 Research*, *50*(D1), D316–D325.

Karolchik, D., Barber, G. P., Casper, J., Clawson, H., Cline, M. S., Diekhans, M., Dreszer, T. R., Fujita, P. A., Guruvadoo, L., Haeussler, M., Harte, R. A., Heitner, S., Hinrichs, A. S., Learned, K., Lee, B. T., Li, C. H., Raney, B. J., Rhead, B., Rosenbloom, K. R., … Kent, W. J. (2014). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, *42*(Database issue), D764–D770.

Larson, D. E., Abel, H. J., Chiang, C., Badve, A., Das, I., Eldred, J. M., Layer, R. M., & Hall, I. M. (2019). svtools: population-scale analysis of structural variation. *Bioinformatics* , *35*(22), 4782–4787.

Nicholas, T. J., Cormier, M. J., & Quinlan, A. R. (2022). Annotation of structural variants with reported allele frequencies and related metrics from multiple datasets using SVAFotate.

BMC Bioinformatics, *23*(1), 490.

Paulson, H. (2018). Repeat expansion diseases. *Handbook of Clinical Neurology*, *147*, 105–123.

Ren, J., Gu, B., & Chaisson, M. J. P. (2023). vamos: variable-number tandem repeats annotation using efficient motif sets. *Genome Biology*, *24*(1), 175.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330.

Ryan, D., Grüning, B., & Ramirez, F. (2016). pyBigWig 0.2. 4. *Cited on*, *3*.

Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M. N. K., Li, Y., Hu, M., Hardison, R., Wang, T., & Yue, F. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology*, *19*(1), 151.

Zarate, S., Carroll, A., Mahmoud, M., Krasheninina, O., Jun, G., Salerno, W. J., Schatz, M. C., Boerwinkle, E., Gibbs, R. A., & Sedlazeck, F. J. (2020). Parliament2: Accurate structural variant calling at scale. *GigaScience*, *9*(12). https://doi.org/10.1093/gigascience/giaa145