Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs

Tsung-Yu Lu, The Human Genome Structural Variation Consortium, Mark Chaisson

Abstract

Variable number tandem repeats (VNTR) are genetic loci composed of consecutive repeats of short segments of DNA, with hypervariable repeat count and composition across individuals. Many genes have coding VNTR sequences, and noncoding VNTR variants are associated with a wide spectrum of clinical disorders such as Alzheimer's disease, bipolar disorder and colorectal cancer. The identification of VNTR length and composition provides the basis for downstream analysis such as expression quantitative trait loci discovery and genome wide association studies. Disease studies that use high-throughput short read sequencing do not resolve the repeat structures of many VNTR loci because the VNTR sequence is missing from the reference or is too repetitive to map. We solve the VNTR mapping problem for short reads by representing a collection of genomes with a repeat-pangenome graph, a data structure that encodes both the population diversity and repeat structure of VNTR loci. We developed software to build a repeat-pangenome using haplotype-resolved single-molecule sequencing assemblies, and to estimate VNTR length and sequence composition based on the alignment of short read sequences to the graph. Using long-read assemblies as ground truth, we are able to determine which VNTR loci may be accurately profiled using repeat-pangenome graph analysis with short reads. This enabled measuring the global diversity of VNTR sequences in the 1000-Genomes Project, and the discovery of expression quantitative trait loci in the Genotype-Tissue Expression Project. This analysis reveals loci that have significant differences in length and repeat composition between continental populations. Furthermore, the repeat pangenome graph analysis establishes an association between previously inaccessible variation and gene expression. Taken together, these indicate that measuring VNTR sequence diversity with repeat-pangenome graphs will be a critical component of future studies on human diversity and disease.

Introduction

The human genome is composed of roughly 3% simple sequence repeats (SSRs) (I. H. G. S. Consortium and International Human Genome Sequencing Consortium 2001), loci composed of short, tandemly repeated motifs. These sequences are classified by motif length into short tandem repeats (STRs) with a motif length of six nucleotides or fewer, and variable-number tandem repeats (VNTRs) for repeats of longer motifs. SSRs are prone to hyper-mutability through motif copy number changes due to polymerase slippage during DNA replication (Viguera, Canceill, and Ehrlich 2001). Variation in SSRs are associated with tandem repeat disorders (TRDs) including amyotrophic lateral sclerosis and Huntington's disease (Gatchel and Zoghbi 2005), and VNTRs are associated with a wide spectrum of complex traits and diseases including attention-deficit disorder, Type 1 Diabetes and schizophrenia (Hannan 2018). While STR variation has been profiled in human populations (Mallick et al. 2016) and to find expression quantitative trait loci (eQTL) (Fotsing et al. 2019; Gymrek et al. 2016), and variation at VNTR sequences may be detected for targeted loci (Bakhtiari et al. 2018; Dolzhenko et al. 2019), the landscape of VNTR variation in populations and effects on human phenotypes are not yet examined genome-wide. Large scale sequencing studies including the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015), TOPMed (Taliun et al. 2019) and DNA sequencing by the Genotype-Tissue Expression (GTEx) project (G. Consortium and GTEx Consortium 2017) rely on high-throughput sequencing (SRS) characterized by SRS reads up to 150 bases. Alignment and standard approaches for detecting single-nucleotide variant (SNV) and indel variation (insertions and deletions less than 50 bases) using SRS are unreliable in SSR loci (Li et al., n.d.), and the majority of VNTR SV are missed using SV detection algorithms with SRS (Chaisson et al. 2019).

The full extent to which VNTR loci differ has been made more clear by single-molecule sequencing (LRS) and assembly. LRS assemblies have megabase scale contiguity and accurate consensus sequences (Koren et al. 2017; Chin et al. 2016) that may be used to detect VNTR variation. Nearly 70% of insertions and deletions

discovered by LRS assemblies greater than 50 bases are in STR and VNTR loci (Chaisson et al. 2019), accounting for up to 4 Mbp per genome. Furthermore, LRS assemblies reveal how VNTR sequences differ kilobases in length and by motif composition (Song, Lowe, and Kingsley 2018). Here we propose using a limited number of human LRS genomes sequenced for population references and diversity panels (Chaisson et al. 2019; Audano et al. 2019; Seo et al. 2016; Shi et al. 2016) to improve how VNTR variation is detected using SRS. It has been previously demonstrated that VNTR variation discovered by LRS assemblies may be genotyped using SRS (Hickey et al. 2020; Audano et al. 2019). However, the genotyping accuracy for VNTR SV is considerably lower than accuracy for genotyping other SV, owing to the complexity of representing VNTR variation and mapping reads to SV loci. Furthermore, genotyping tests for the presence of a known variant, and does not reveal the spectrum of copy number variation that exists in tandem repeat sequences. An alternative use of LRS assemblies is for population-specific references that improve SRS read mapping by adding sequences missing from the reference (Du et al. 2019; Shi et al. 2016). Because missing sequences are enriched for VNTRs (Audano et al. 2019), haplotype-resolved LRS genomes may help improve alignment to VNTR regions, as well as facilitate the development of a model to discover VNTR variation by serving as a ground truth.

The hypervariability of VNTR sequences prevents a single assembly from serving as an optimal reference. Instead, to improve both alignment and genotyping, multiple assemblies may be combined into a pangenome graph (PGG) (Hickey et al. 2020; Eggertsson et al. 2019; Garrison et al. 2018; Chen et al. 2019) composed of sequence-labeled vertices connected by edges such that haplotypes correspond to paths in the graph. Sequences shared by multiple haplotypes are stored in the same vertex, and genetic variation is represented by the branching structure of the graph. A conceptually similar construct is the repeat graph (Pavel A. Pevzner, Tang, and Tesler 2004), with similarly sequence-labeled vertices, except sequences repeated multiple times in a genome are represented by the same vertex. Graph analysis has been used to encode the elementary duplication structure of a genome (Jiang et al. 2007) and for multiple sequence alignment of

repetitive sequences with shuffled domains (Raphael et al. 2004), making them well-suited to represent VNTR sequences that differ in both motif copy number and composition. Here we propose the representation of human VNTR sequences as a repeat pangenome graph (RPGG), that encodes both the repeat structure and population diversity of VNTR loci by ensuring that the elementary repeat structure is represented by the graph topology, and with repeat haplotypes corresponding to paths with many cycles in the graph (Figure 1c).

The most straight-forward approach that combines a pangenome graph and a repeat graph is a de Bruijn graph on multiple sequences, and was the basis of one of the earliest representations of a pangenome by the Cortex method (Iqbal, Turner, and McVean 2013; Iqbal et al. 2012). The de Bruijn graph has a vertex for every distinct sequence of length *k* in a genome (*k*-mer), and an edge connecting every two consecutive *k*-mers, thus *k*-mers occurring in multiple genomes or in multiple times in the same genome are stored by the same vertex. While the Cortex method stores entire genomes in a de Bruijn graph, we construct a separate locus-RPGG for each VNTR and store a genome as the collection of locus-RPGGs, which deviates from the definition of a de Bruijn graph because the same *k*-mer may be stored in multiple vertices. We developed a tool suite, Tandem Repeat Genotyping based on Haplotype-derived Pangenome Graphs (danbing-tk) to identify VNTR boundaries in assemblies, construct RPGGs, align SRS reads to the RPGG, and infer VNTR motif composition and length in SRS samples. This enables the alignment of large SRS datasets into an RPGG to discover population genetics of VNTR loci, and to associate expression with VNTR variation.

Results.

Our approach to build RPGGs is to *de novo* assemble LRS genomes, and build de Bruijn graphs on the assembled sequences at VNTR loci, using SRS genomes to ensure graph quality. We used public LRS data for 19 individuals with diverse genetic backgrounds, including genomes from individual genome projects (Seo et al. 2016; Zook et al. 2020), structural variation studies (Chaisson et al. 2019), and diversity panel sequencing (Audano et al. 2019) (Figure 1a, Supplemental Material). Each genome was sequenced by either PacBio single

long read (SLR) between, or high-fidelity (HiFi) sequencing between 16 and 76-fold coverage along with matched 22-82-fold Illumina sequencing (Table 1). This data reflects a wide range of technology revisions, sequencing depth, and data type, however subsequent steps were taken to ensure accuracy of RPGG through locus redundancy and SRS alignments. We developed a pipeline that partitions LRS reads by haplotype based on phased heterozygous SNVs and assembles haplotypes separately by chromosome. When available, we used existing telomere-to-telomere SNV and phase data provided by Strand-Seq and/or 10X Genomics (Porubsky et al. 2017; Chaisson et al. 2019) with phase-block N50 size between 13.4-18.8 Mb. For other datasets, long-read data were used to phase SNVs. While this data has lower phase-block N50 (<0.5 - 6 Mb), the individual locus-RPGG do not use long-range haplotype information and are not affected by phasing switch error. Reads from each chromosome and haplotype were independently assembled using the flye assembler (Kolmogorov et al. 2019) for a diploid of 0.88-14.5Mb N50, with the range of assembly contiguity reflected by the diversity of input data.

In this study, the number of resolved VNTR loci is a more accurate measurement of useful assembly contiguity than N50 because a disjoint RPGG is generated for each VNTR locus. An initial set of 84,411 VNTR intervals (mean length = 420 bp in GRCh38) were annotated by Tandem Repeats Finder (TRF) (Benson 1999), and then mapped onto contig coordinates using pairwise contig alignments. Long VNTR loci tended to have fragmented TRF annotation, which can cause erroneous length estimates in downstream analysis. During locus assignment, danbing-tk expands boundaries and merges loci to ensure boundaries of all VNTRs are well-defined and harmonized across genomes (Methods) (Figure 1b). In practice, we found 13,497-54,491 (16-65%) of the VNTR loci are subject to boundary expansion depending on the quality of assembly and alignment, with a total of 83,021 regions (98%) expanded in at least one genome. After removing regions expanded by more than 5 kb, 20,126-81,908 VNTR intervals remained for each genome (mean length = 667 bp across 19 genomes plus GRCh38, Supplementary Material).

The RPGGs are constructed as the bi-directional de Bruijn graph of GRCh38 and the assembled haplotypes separately for each locus, along with 700 bases of flanking sequence. In a bi-directional de Bruijn graph, each distinct sequence of length k (k-mer) and its reverse complement map to a vertex, and each sequence of length k+1 connects the vertices to which the two composite k-mers map. The correspondence between k-mer and source genome are retained allowing annotation of reference and non-reference sequences in the RPGG. There was little effect on downstream analysis for values of k between 17 and 25, and so k=21 was used for all applications. To remove spurious vertices and edges from assembly consensus errors, SRS from genomes matching the LRS samples were mapped to the RPGG, and k-mers not mapped by SRS were removed from the graph (average of 11.1 per locus, Table 1). Using the number of vertices as a proxy for sampled genetic diversity, we find that 45% of the sequences novel with respect to GRCh38 are discovered after the inclusion of 19 genomes, with diversity linearly increasing per genome after the first 8 genomes are added to the RPGG (Figure 1c).

The alignment of a read to an RPGG may be defined by the path in the RPGG with a sequence label that has the minimum edit distance to the read among all possible paths. While several methods exist to find alignments that do not reuse cycles (Garrison et al. 2018; Rakocevic et al. 2019), alignment with cycles is a more challenging problem recently solved by the GraphAligner method to map long reads to pangenome graphs (Rautiainen, Mäkinen, and Marschall 2019). We evaluated this method with simulated error-free 150 base paired-end reads. Although >95% of reads were aligned, many had low quality; 62.8% of reads matched with at least 90% identity. We developed an alternative approach tuned for RPGG alignments in danbing-tk (Figure 1d) to map SRS reads to the RPGG in two passes, first by finding locus-RPGGs with a high number (>45 in each end) shared k-mers with reads, and next by threading the paired-end reads through the locus-RPGG, allowing for up to 2 differences and at least 50 bases aligned per read against the threaded path (Methods). A mapping quality value is calculated for each paired-end read that approximates the probability of misalignment to the incorrect locus. The danbing-tk alignments of simulated error-free 150 base paired-end reads were used to

measure an upper-bound for how well loci could be mapped, demonstrating that 96.8% (81,291/84,008) loci are mapped with >90% identity alignments. On real data, danbing-tk has modest computational requirements, using 14.5 GB of memory to map 150 base paired-end reads at 9.05 Mb/sec on 12 cores.

Alignment of SRS reads to the RPGG enables estimation of VNTR length and motif composition. The count of *k*-mers in SRS reads mapped to the RPGG are reported by danbing-tk for each locus. For M samples and L VNTR loci, the result of an alignment is L count matrices of dimension $M \times N_i$, where N_i is the number of vertices in the de Bruijn graph on the locus i, excluding flanking sequences. If SRS reads from a genome were sequenced without bias, sampled uniformly, and mapped without error to the RPGG, the count of a *k*-mer in a locus mapped by an SRS sample should scale by a factor of read depth with the sum of the count of the *k*-mer from the locus of both assembled haplotypes for the same genome. The accuracy of alignment and sequencing bias were measured by comparing the *k*-mer counts from five matched Illumina and LRS genomes (Figure 2a). In total, 34% (29,052/84,008) loci had a median or mean $r^2 \geq 0.8$ between SRS and assembly *k*-mer counts, and were marked as "valid" loci to carry forward for downstream diversity and expression analysis (Figure 2b). Valid had an average length of 499 bp, compared to 667 bp in the entire database (Figure 2c). VNTR loci that did not genotype well (invalid) were enriched for sequences that map within Alu (27,173), SVA (1,829), and other 36,011 mobile elements (Supplementary Material). Valid VNTRs did not show a strong GC bias (GC%=46.4%). The complete RPGG on valid loci contains 8,940,667 vertices, in contrast to the corresponding RPGG only on GRCh38, which has 6,158,697 vertices. We validate that the additional vertices in the RPGG are indeed important for accurately recruiting reads pertinent to a VNTR locus, using the *CACNA1C* VNTR as an example (Figure 2d). It is known that the reference sequence at this locus is truncated compared to the majority of the populations (319 bp in GRCh38 versus 5,687 bp averaged across 19 genomes). The limited sequence diversity provided by GRCh38 at this locus failed to recruit reads that map to paths existing in the RPGG but missing or only partially represented in the repeat-GRCh38 graph. A linear fit between the *k*-mers from mapped reads and the ground truth assemblies shows that there is a 2.3-fold gain in slope, or measured

read depth, when using RPGG compared to repeat-GRCh38 graph (Figure 2e). The *k*-mer counts in RPGG also correlate better with the assembly *k*-mer counts compared to repeat-GRCh38 graph (r^2 = 0.992 versus 0.916).

New genomes with arbitrarily complicated combinations of motifs and copy numbers at VNTR loci should still align to an RPGG as long as the motifs are represented in the graph. To measure this, leave-one-out analysis was used to evaluate how well novel genomes align to the RPGG and may have VNTR length accurately predicted from alignments. An RPGG was constructed with one LRS genome missing, SRS reads from the missing genome were mapped into the RPGG, and the predicted locus lengths were compared to the average length of the corresponding loci in the missing LRS assembly. The locus length is predicted as the adjusted sum of *k*-mer counts kms mapped from SRS sample s: $kms/(cov_s \times \hat{b})$, where cov_s is sequencing depth of s, $\hat{b}$ is a correction for locus-specific sampling bias (LSB). Because the SRS datasets used in this study during pangenome construction were collected from a wide variety of studies with different sample preparation and technology versions, there was no consistent LSB in either repetitive or nonrepetitive regions for samples from different sequencing runs (Supplementary Material). However, principal component analysis (PCA) of repetitive and nonrepetitive regions shows highly similar projection patterns, which enables using LSB in nonrepetitive regions as a proxy for finding the nearest neighbor of LSB in VNTR regions. Leveraging this finding, a set of 397 nonrepetitive control regions were used to estimate the LSB of an unseen SRS sample (Methods), giving a median length-prediction accuracy of 0.77 for 16 unrelated genomes (Figure 3a). Read depth of a repetitive region reflects variation in length when aligning short reads to a linear reference genome. However, direct inference of VNTR lengths from read depth has an accuracy of only 0.71. We also compared the performance of RPGG and repeat-GRCh38 graph in length prediction (Figure 3a,b). The overall error rate, measured with mean absolute percentage error (MAPE), of all loci are similar using either RPGG or repeat-GRCh38 graph (0.23 versus 0.22) and better than the read depth approach (0.27); however a 36% improvement is observed for the 4,596 loci poorly genotyped using repeat-GRCh38 graph (MAPE > 0.4).

Furthermore, genotyping using the RPGG shows a 24% decrease in MAPE for the 16 unrelated genomes, demonstrating the utility of the RPGG.

To use danbing-tk to explore global diversity of VNTR sequences and potential functional impact on variation, we aligned reads from 2,492 individuals from diverse populations sequenced at 30-fold coverage sequenced by the 1000-Genomes project (1KGP) (Fairley et al. 2020; 1000 Genomes Project Consortium et al. 2015), and 857 GTEx genomes (G. Consortium and GTEx Consortium 2017) to the RPGG. The fractions of reads from these studies that aligned to the RPGG were similar to the fraction of reads from SRS data from genomes in the RPGG; 1.32% of reads per genome from the 1KGP, and 1.02% reads per GTEx genome aligned to the RPGG, similar to 1.12% of matched SRS data. PCA on the LSB of both datasets showed the 1KGP and GTEx genomes as separate clusters in both repetitive and nonrepetitive regions (Supplementary Material), indicating experiment-specific bias that prevents cross data set comparisons. Consistent with the finding in previous leave-one-out analysis, genomes from the same study cluster together in the PCA plot of LSB, and so within each dataset and locus, *k*-mer counts from SRS reads normalized by sequencing depth were used to compare VNTR content across genomes.

The *k*-mer dosage: kms/cov, was used as a proxy for locus length to compare tandem repeat variation across populations in the 1KGP genomes. The 1KGP samples contain individuals from African (26.5%), East Asian (20.1%), European (19.9%), Admixed American (13.9%), and South Asian (19.5%) populations. When comparing the average population length to the global average length, the majority of loci, 52.3% (15,193/29,052) have differential length between populations ($P < 0.01$, one-way ANOVA), with similar distributions of differential length when loci are stratified by the average accuracy of length prediction as low (<80%), medium (80-90%), and high (>90%) (Figure 4a). Population differentiation was calculated using the V_{ST} statistic (Redon et al. 2006) on VNTR length (Figure 4b, Supplementary Material). The top two loci ranked

by V_{ST} are both intronic: an 864 base VNTR in *FREM1* (V_{ST}=0.47), and an 80 base locus in *EML1* (V_{ST}=0.42) (Figure 4c,d).

VNTR loci that are "fragile" may undergo hyper-expansion and are implicated as a mechanism of multiple diseases (Hannan 2018). To discover new potentially fragile loci, we searched the 1kg genomes for evidence of rare VNTR hyper-expansion. Loci were screened for individuals with extreme (>6 standard deviations) variation, and then filtered for deletions or unreliable samples (Methods) to characterize 103 loci as potentially fragile. These loci include 31 genes, that while not significantly reduced from the number expected by chance (p=0.26, permutation test), only one fragile locus overlaps a UTR (*ITGB1BP1*, p=0.033, permutation test). Furthermore, noting that tandem repeat expansion disorders are enriched for neurological disorders (Hannan 2018), we observed fragile loci in several genes that are expressed in the brain *CA10*, *DPP6*, *RIMBP2*, and *NEGR1,* and represent candidates for further study.

Alignment to an RPGG provides information about motif usage in addition to estimates of VNTR length because genomes with different motif composition will align to different vertices in the graph. To detect differential motif usage, we searched for loci with a *k*-mer that was more frequent in one population than another and not simply explained by a difference in locus length, comparing African (AFR) and East Asian populations for maximal genetic diversity. Lasso regression against locus length was used to find the *k*-mer with the most variance explained (VEX) in EAS genomes, denoted as the most informative *k*-mer (mi-kmer). Two statistics are of interest when comparing the two populations: the difference in the count of mi-kmers (kmc_d) and the difference between proportion of VEX (r_d^2) by mi-kmers. kmc_d describes the usage of an mi-kmer in one population relative to another, while r_d^2 indicates the degree that the mi-kmer is involved in repeat shrinkage or expansion in one population relative to another. We observe that 4,231 loci have significant difference in the usage of mi-kmers between the two populations (two-sided P < 0.01, bootstrap). Among these, the mi-kmers of 1,059 loci are crucial to length variation in the EAS but not in the AFR population (two-sided P

< 0.01, bootstrap) (Figure 4e, Supplementary Material). A top example of these loci with r^2 at least 0.8 in the EAS population was visualized with a heatmap of relative *k*-mer count from both populations, and clearly showed differential usage of cycles in the RPGG (Figure 4f).

Because the danbing-tk length estimates showed population genetic patterns expected for human diversity, we assessed whether danbing-tk alignments could detect VNTR variation with functional impact. Genomes from the GTEx project were mapped into the RPGG to discover loci that have an effect on nearby gene expression in a length-dependent manner. A total of 857/879 genomes passed quality filtering. Similar to the population analysis, the *k*-mer dosage was used as a proxy for locus length. Methods previously used to discover eQTL using STR genotyping (Fotsing et al. 2019) were applied to the danbing-tk alignments. In sum, 27,374 VNTRs within 100 kb to 45,217 genes were tested for association, with a total of 136,950 tests and approximately 3 VNTRs tested per gene. Using a gene-level FDR cutoff of 1%, we find 3,631 (2,504 aftering excluding fibroblast) eQTL (eVNTRs) (Figure 5a), among which 3,610 (99.4%) discoveries are novel (Supplemental data), indicating that the spectrum of association between tandem repeat variation and expression extends beyond the lengths and the types of SSR considered in previous STR (Mousavi et al. 2019) and VNTR (Bakhtiari et al. 2020) studies. Both positive and negative effect were observed among eVNTRs (Figure 5b). More eVNTRs with positive effect size were found than with a negative effect size (1,905 versus 1,726, binomial test, P<0.01), with an average effect of +0.250 (from +0.141 to +0.711) versus −0.225 (from −0.526 to −0.141), respectively. Additionally, the reference length of eVNTRs (mean=447 bp) tended to be longer than an average VNTR (mean=401 bp) (Wilcoxon–Mann–Whitney test, P<1E-25). eVNTRs are slightly more AT-rich than the whole VNTR set (43.2% versus 46.4%), and the distribution of the distances of an eVNTR to its nearest telomere is not distinct from other VNTRs (Supplementary Materials). Because many exons contain VNTR sequences, expression measured by read depth should increase with length of the VNTR, and there is an 1.9-fold enrichment of eVNTRs in coding regions as expected.

Among the top eVNTRs, an intergenic eVNTR at chr7:142799001-142799221, ranked 5th among all eVNTRs, has an overlap with both the signals of H3K27Ac modification and DNase digestion, and shows a strong negative correlation with *PRSS2* gene expression (effect size = −0.41, P = 5.4E-18, Figures 5b,c). *PRSS2* was shown to be a causal gene of hereditary pancreatitis through duplication of a ~605 kb region (Masson et al. 2008), and was suggested to facilitate tumor invasion through activating metalloproteinase and disruption of tight junctions (Vilen et al. 2012), which could be further exacerbated in both diseases by abnormal contraction of the repeat. We also find that the eVNTR at chr4:3305624-3305756 is associated with the Huntington disease gene, *HTT*, at 1% FDR (Figure 5d). The eVNTR overlaps with strong signal of H3K27Ac histone marker and DNase digestion and shows a positive effect size of 0.29, suggesting the risk of Huntington disease could potentially be modulated by another layer of molecular interaction.

In addition to the 29,052 VNTRs, we augmented our RPGG with 6 additional disease-related tandem repeats (Supplementary Material), with the *CNBP* STR identified as an eQTL across multiple tissues at 5% FDR (Figure 5e). The intronic eSTR at *CNBP* was found to be associated with several nearby genes (effect size between −0.20 and +0.25) beyond *CNBP*, with the largest effect size at *CNBP* gene, indicating a potential role of *CNBP* eSTR as a common regulator for local gene expression. While previous studies have shown an association between *CNBP* disease status and effect on nearby expression (Alwazzan et al. 1999; Thornton et al. 1997), this demonstrates an extension of effect to states of healthy individuals.

Discussion.

Previous commentaries have proposed that variation in VNTR loci may represent a component of undiagnosed disease and missing heritability (Hannan 2010), which has remained difficult to profile even with whole genome sequencing (Mousavi et al. 2019). To address this, we have proposed an approach that combines multiple genomes into a pangenome graph that represents the repeat structure of a population. This is supported by the software, danbing-tk and associated RPGG. We used danbing-tk to generate a pangenome from 19

haplotype-resolved assemblies, and applied it to detect VNTR variation across populations and to discover eQTL. This discovered 15,193 loci that have divergent lengths across populations, and 54 loci with $V_{ST} > 0.3$, indicating potential population structure of hypervariable loci, and 3,631 eVNTRs.

The structure of the RPGG can help to organize the diversity of assembled VNTR sequences, and to gauge the utility of RPGG representation over a single reference. In particular, 45 % of the graph structure is novel after the addition of 19 genomes relative to the RPGG on GRCh38. Combined with the observation that using the 19-genome RPGG gives an 24% decrease in accuracy prediction error, this indicates that much of the population can be mapped into the graphical representation of the GRCh38 reference, and that the pan-genomes add detail for more rare variation. Furthermore, when considering length prediction per-locus, accuracy may be improved using the repeat graph representation of a single genome (e.g. GRCh38), by improving how reads map to highly repetitive regions of the genome. With the availability of additional genomes sequenced through the Pangenome Reference Consortium (https://humanpangenome.org/) and the HGSVC (https://www.internationalgenome.org), combined with advanced haplotype-resolve assembly methods (Porubsky et al., n.d.), the spectrum of this variation will be revealed in the near future. While we anticipate that eventually the full spectrum of VNTR diversity will be revealed through LRS of the entire 1kg, the RPGG analysis will help organize analysis by characterizing repeat domains. For example, with our approach, we are able to detect 1,062 loci with differential motif usage between populations, which could be difficult to characterize using an approach such as multiple-sequence alignment of VNTR sequences from assembled genomes.

In contrast to other pangenome approaches (Garrison et al. 2018; Rakocevic et al. 2019), danbing-tk does not keep track of a reference (e.g. GRCh38) coordinate system. Furthermore, because it is often not possible to reconstruct a unique path in an RPGG, only counts of mapped reads are reported rather than the order of traversal of the RPGG. An additional caveat of our approach is that genotype is calculated as a

continuum of *k*-mer dosage rather than discrete units, prohibiting direct calculation of linkage-disequilibrium for fine-scale mapping (LaPierre et al., n.d.).

The rich data provided by danbing-tk and pangenome analysis provide the basis for additional association studies. While most analysis in this study focused on the diversity of VNTR length or association of length and expression, it is possible to query differential motif usage using the RPGG. The ability to detect motifs that have differential usage between populations brings the possibility of detecting differential motif usage between cases and controls in association studies. This can help distinguish stabilizing versus fragile motifs (Braida et al. 2010), or resolve some of the problem of missing heritability by discovering new associations between motif and disease (Song, Lowe, and Kingsley 2018).

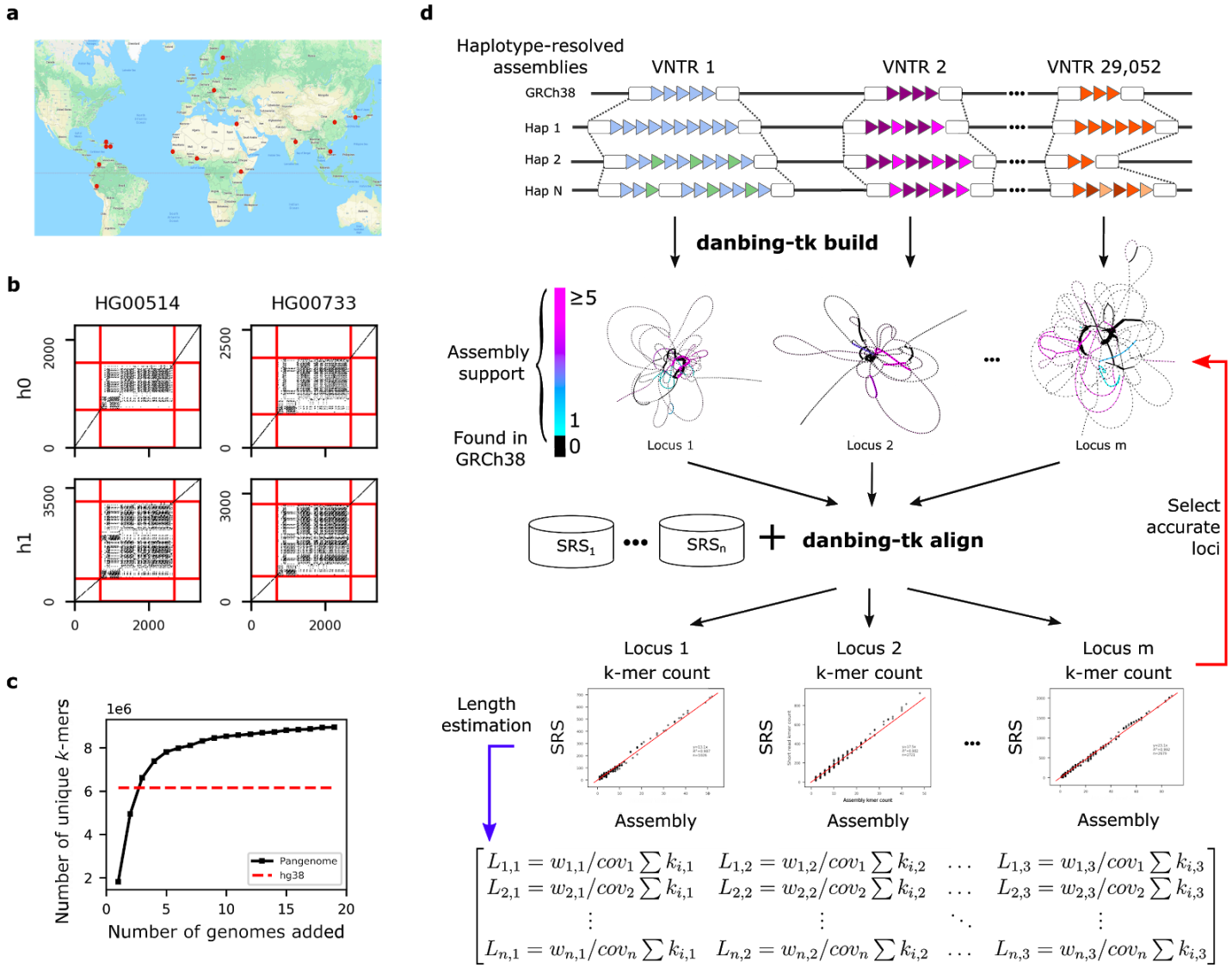

Figure 1. Sequence diversity of VNTRs in human populations.
a, Global diversity of SMS assemblies. **b,** Dot-plot analysis of the VNTR locus chr1:2280569-2282538 (SKI intron 1 VNTR) in genomes that demonstrate varying motif usage and length **c**, Diversity of RPGG as genomes are incorporated, measured by the number of *k*-mers in the graph. **d,** danbing-tk workflow analysis. (top) VNTR loci defined from the reference are used to map haplotype loci. Each locus is converted to a de Bruijn graph, from which the collection of graphs is the RPGG. The de Bruijn graphs shown illustrate sequences missing from the RPGG built only on GRCh38. The alignments may be either used to select which loci may be accurately mapped in the RPGG using SRS that match the assemblies (red), or may be used to estimate lengths on sample datasets (blue).

Genome	Continental population	Study	Cov	Assembly N50 (Mb)	Fraction of VNTR annotated	Ancestry
AK1	EAS	KG	54	0.88	0.783	Korean
HG00268	EUR	DP	67	3.51	0.935	Finnish
HG00512	**EAS**	**HGSVG**	**28**	**8.83**	**0.980**	**Han Chinese**
HG00513	**EAS**	**HGSVG**	**30**	**1.57**	**0.977**	**Han Chinese**

HG00514	EAS	HGSVG	31	1.32	0.914	Han Chinese
HG00731	AMR	HGSVG	31	2.18	0.978	Puerto Rican
HG00732	AMR	HGSVG	16	1.3	0.967	Puerto Rican
HG00733	AMR	HGSVG	46	6.88	0.965	Puerto Rican
HG01352	AMR	DP	68	5.97	0.953	Colombian
HG02059	EAS	DP	76	19.5	0.972	Vietnamese
HG02106	AMR	DP	57	0.88	0.583	Peruvian
HG02818	AFR	DP	56	0.66	0.732	Gambian
HG04217	SAS	DP	60	0.86	0.244	Telugu
NA12878	EUR	DP	54	4.67	0.943	Central European
NA19238	**AFR**	**HGSVG**	**23**	**2.64**	**0.971**	**Yoruban**
NA19239	**AFR**	**HGSVG**	**35**	**4.87**	**0.973**	**Yoruban**
NA19240	AFR	HGSVG	49	3.4	0.953	Yoruban
NA19434	AFR	DP	62	11	0.932	Luhya
NA24385	EUR	GIAB	54	1.32	0.949	Ashkenazim

Table 1. Source genomes for RPGG. Continental populations represented are East Asian (EAS), European (EUR), Admixed Amerindian (AMR), South Asian (SAS), and African (AFR). Coverage is estimated diploid coverage based on alignment to GRCh38. Assembly N50 is of haplotype-resolved assemblies. The fraction of VNTR annotated are all VNTR with at least 700 flanking bases assembled.

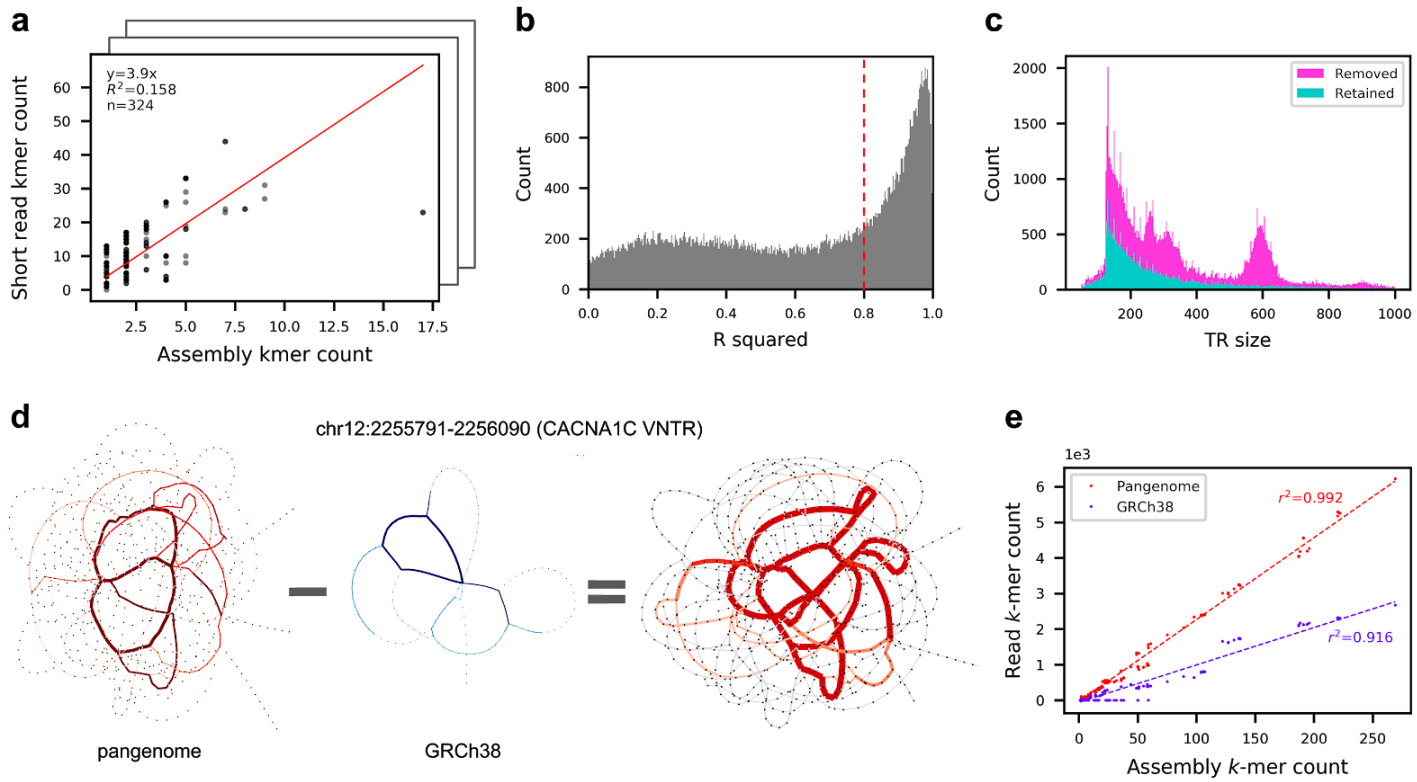

Figure 2. Mapping short reads to repeat-pangenome graphs. a, An example of evaluating the genotyping quality of a locus mapped by SRS reads. The genotyping quality is measured by the r^2 of a linear fit between the *k*-mer counts from the ground truth assemblies and from the mapped reads (Methods). **b,** Distribution of the genotyping quality scores of 84,411 loci. Loci with genotyping quality less than 0.8 when averaged across

samples are removed from downstream analysis (Methods). **c,** Distribution of VNTR lengths in GRCh38 removed or retained for downstream analysis. **d-e**, Comparing the read mapping results of the *CACNA1C* VNTR using RPGG or repeat-GRCh38 graph. The *k*-mer counts in each graph and the differences are visualized with edge width and color saturation (**d**). The *k*-mer counts from the ground truth assemblies are regressed against the counts from reads mapped to the RPGG (red) and repeat-GRCh38 graph (blue), respectively (**e**).

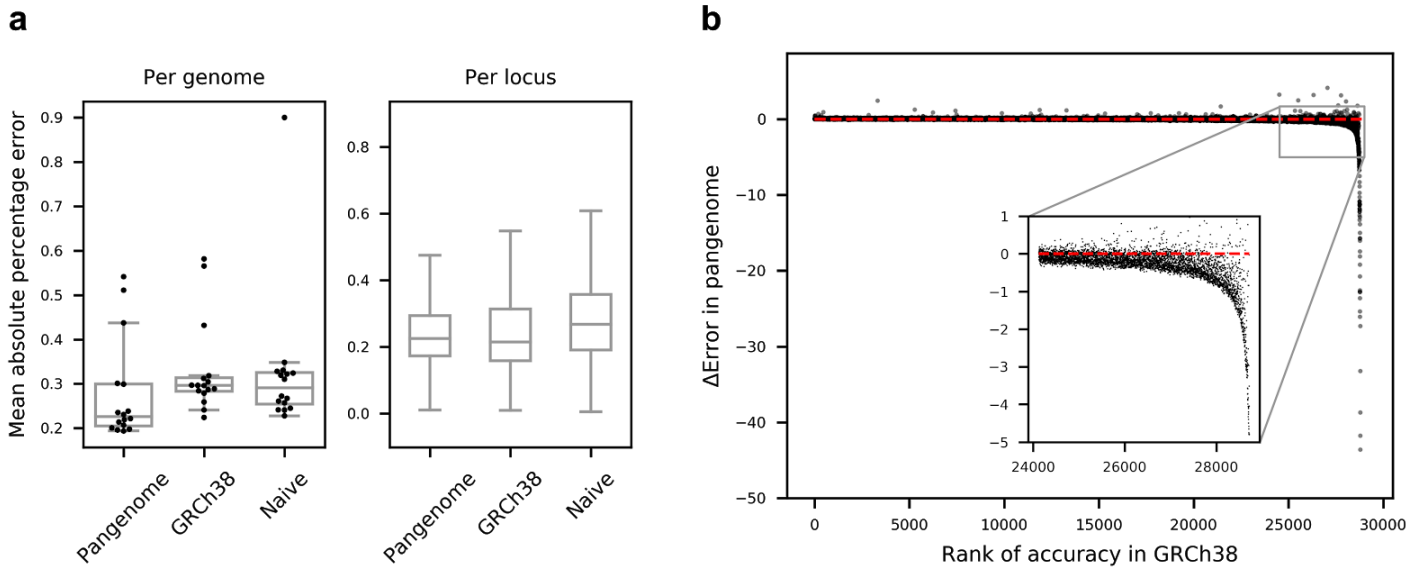

Figure 3. Read sampling bias and VNTR length prediction. a, Accuracies of VNTR length prediction measured for each genome (left) and each locus (right). Mean absolute percentage error (MAPE) in VNTR length is averaged across loci and genomes, respectively. Lengths were predicted based on RPGG, repeat-GRCh38 graph or naive read depth method, respectively. Boxes span from the lower quartile to the upper quartile, with horizontal lines indicating the median. Whiskers extend to points that are within 1.5 interquartile range (IQR) from the upper or the lower quartiles. **b,** Relative performance of RPGG versus repeat-GRCh38 graph. Loci are ordered along the x-axis by genotyping accuracy in repeat-GRCh38 graph. The y-axis shows the decrease in MAPE using RPGG versus repeat-GRCh38 graph. The subplot shows loci poorly genotyped (MAPE>0.4) in repeat-GRCh38 graph. The red dotted line indicates the baseline without any improvement.

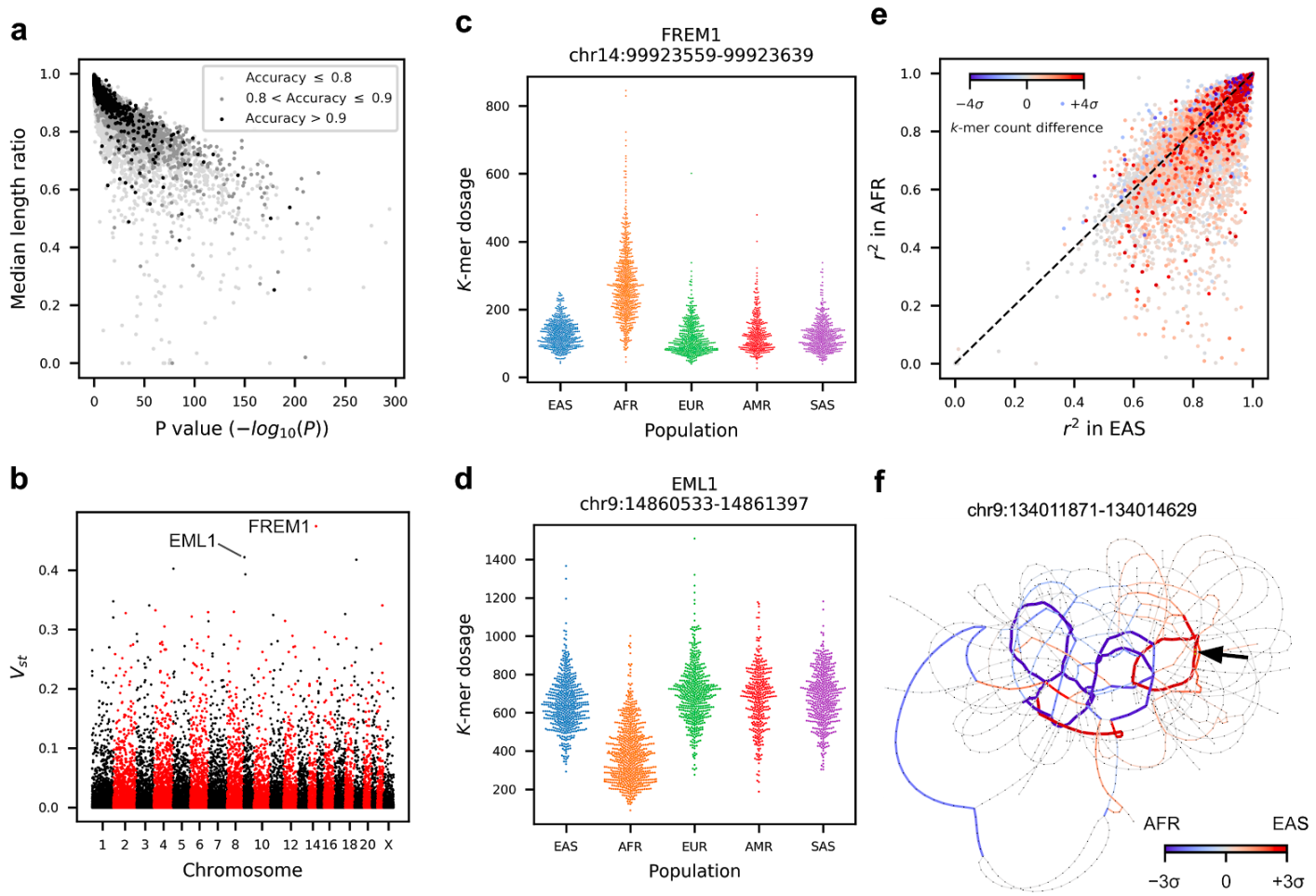

Figure 4. Population properties of VNTR loci. a, Ratios of median length between populations for loci with significant differences in average length. Loci are stratified by accuracy prediction (<0.8), medium (0.8-0.9), and high (0.9+). **b**, Manhattan plot of V_{ST} values. The two top loci are intronic *FREM1* and *EML1*. **c-d**, The distribution of estimated length via *k*-mer dosage in continental populations for *FREM1* and *EML1* VNTR loci. Each point is an individual. **e,** Differential usage and expansion of motifs between the EAS and AFR populations. For each locus, the proportion of variance explained by the most informative *k*-mer in the EAS is shown for the EAS and AFR populations on the x and y axes, respectively (left), colored by the difference of normalized counts of the *k*-mer. Dark red, blue appear more frequently in EAS, AFR populations, respectively. **f,** An example VNTR with differential motif usage. Edges are colored if the *k*-mer count is biased toward a certain population. The black arrow indicates the location of the *k*-mer that explains the most variance of VNTR length in the EAS population.

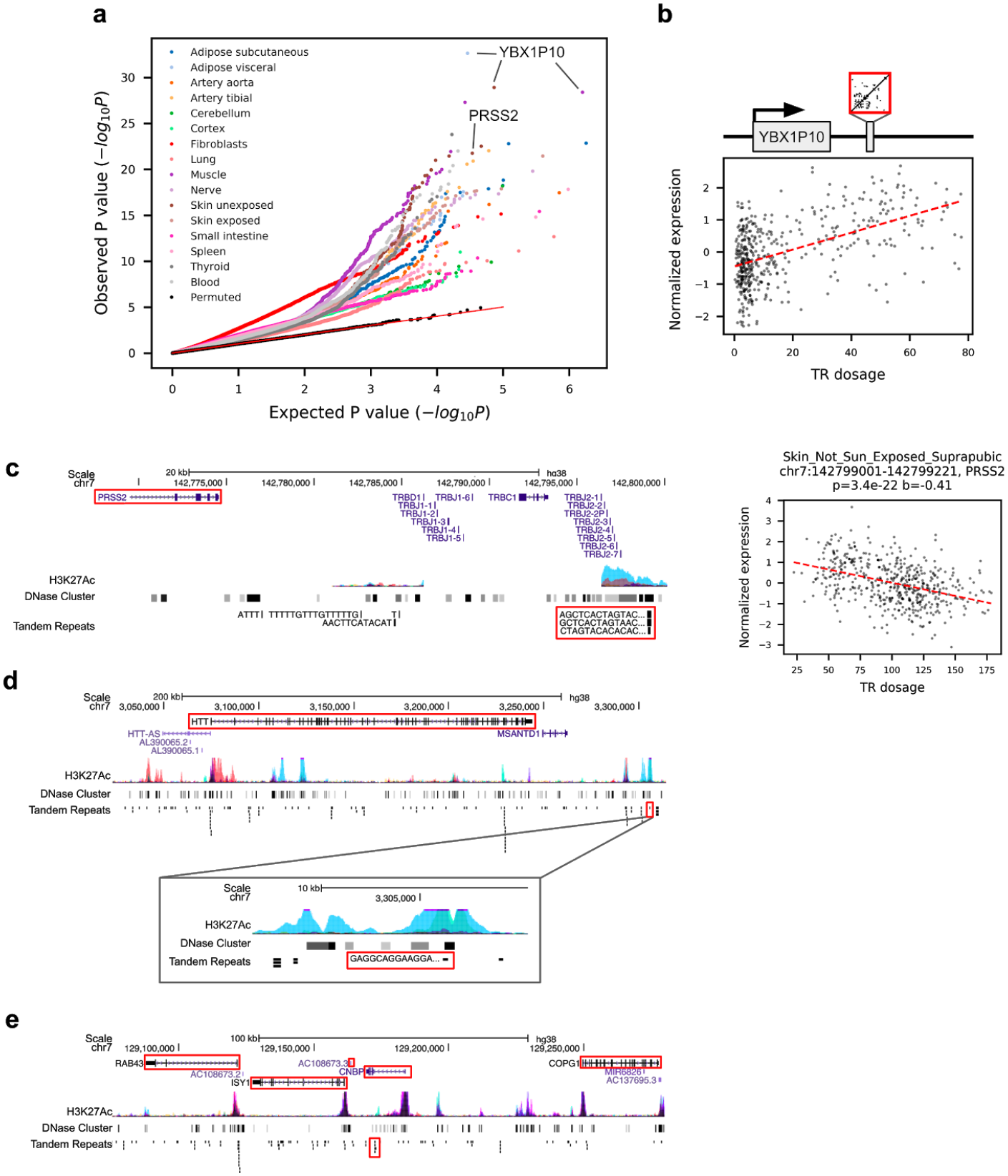

Figure 5. *cis*-eQTL mapping of VNTRs. a, eVNTR discoveries in 16 human tissues. The quantile-quantile plot shows the observed P value of each association test versus the P value drawn from the expected uniform distribution. Black dots indicate the permutation results from the top 5% associated (gene, VNTR) pairs in each tissue. The regression plots for *YBX1P10* and *PRSS2* are shown in b and c. **b,** An example of association

between VNTR lengths and gene expression values for (chr9:35,999,208-35,999,382, *YBX1P10*). The dot plot of the VNTR is shown in the red box. The x-axis shows VNTR lengths scaled by a locus-specific sampling bias. The y-axis shows z-score normalized expression counts. **c-e**, Genomic view of eVNTRs. The associations between (chr7:142799001-142799221, *PRSS2*) (c), (chr4:3305624-3305756, *HTT*) (d) and (chr3:129172568-129172736, *CNBP*) (e) are shown. Red boxes indicate the eGenes and eVNTRs in association with each other.

Materials and methods

Pangenome construction.

Initial discovery of tandem repeats: TRF v4.09 (option: 2 7 7 80 10 50 500 -f -d -h) (Benson 1999) was used to roughly annotate the SSR regions of five PacBio assemblies (AK1, HG00514, HG00733, NA19240, NA24385). The scope of this work focuses on VNTRs that cannot be resolved by typical short read sequencing methods. We selected the set of SSR loci with a motif size greater than 6 bp and a total length greater than 50 bp and less than 20 kbp. For each haplotype, the selected VNTR loci were mapped to GRCh38 reference genome to identify homologous VNTR loci. To maintain data quality, VNTR loci that could not be assigned homology were removed from datasets.

Boundary expansion of VNTRs: The biological boundaries of a VNTR are ill-defined; VNTRs with sparse recurring motifs or transition between different motifs or a nested motif structure often fail to be fully annotated by TRF. A misannotation of VNTR boundaries can cause erroneous length estimates. To avoid the propagation of this error to downstream analysis, we developed a boundary expansion algorithm to recover the proper boundary for each VNTR that is consistent between homologous loci across all haplotypes. The algorithm performs in three steps: (1) individual boundary expansion in each haplotype, (2) joint boundary expansion of the two haplotypes, and (3) quality check. In the first step, the algorithm expands the boundary of a VNTR region until there are no shared *k*-mers between a VNTR region and its flanking regions. Regions expanded by more than 5 kbp are removed. The size of the flanking regions is chosen to be 700 bp, which is approximately

the upper bound of the insert size of typical SRS reads. In the second step, a similar expansion is performed taking into account homology between haplotypes, similarly discarding loci if the expansion is more than 5 kbp. The genomic regions of all VNTR loci of all haplotypes were mapped back to GRCh38 and merged to represent the unified locations of all VNTRs in human populations.

Read-to-graph alignment: For the two haplotypes of an individual, three data structures are used to encode the information of all VNTR loci, including VNTRs and their 700 bp flanking sequences. The first data structure allows fast locus lookup for each *k*-mer (*k*=21) by hashing each canonical *k*-mer in the VNTRs and the flanking sequences to the index of the original locus. The second data structure enables graph threading by storing a bi-directional de Bruijn graph for each locus. The third data structure is used for counting *k*-mers originating from VNTRs. The read mapping algorithm maps each pair of Illumina paired-end reads to a unique VNTR locus in three phases: (1) In the *k*-mer set mapping phase, the read pair is converted to a pair of canonical *k*-mer multisets. The VNTR locus with the highest count of intersected *k*-mers is detected with the first data structure. (2) In the threading phase, the algorithm tries to map the *k*-mers in the read pair to the bi-directional de Bruijn graph such that the mapping forms a continuous path/cycle. To account for sequencing and assembly errors, the algorithm is allowed to substitute a limited number of nucleotides in a read if no matching *k*-mer is found in the graph. The read pair is determined feasible to map to a VNTR locus if the number of mapped *k*-mers is above an empirical threshold. (3) In the *k*-mer counting phase, canonical *k*-mers of the feasible read pair are counted if they existed in the VNTR locus. Finally, the read mapping algorithm returns the *k*-mer counts for all loci as mapped by SRS reads. Alignment timing was conducted on an Intel Xeon E5-2650v2 2.60GHz node.

Graph pruning and merging: Pan-genome representation provides a more thorough description of VNTR diversity and reduces reference allele bias, which effectively improves the quality of read mapping and downstream analysis. Considering the fact that haplotypes assembled from long read datasets are error prone in

VNTR regions, it is necessary to prune the graphs/*k*-mers before merging them as a pan-genome. We ran the read mapping algorithm with error correction disabled so as to detect *k*-mers unsupported by SRS reads. The three data structures were updated by deleting all unsupported *k*-mers for each locus. By pooling and merging the reference regions corresponding to the VNTR regions in all individuals, we obtained a set of "pan-reference" regions, each indicating a location in GRCh38 that is likely to map to a VNTR region in any other unseen haplotype. By referencing the mapping relation of VNTR loci across individuals, we encoded the variability of each VNTR locus by merging the three data structures across individuals.

Genotyping quality analysis: To evaluate the quality of the haplotype assemblies and the performance of the read mapping algorithm, VNTR *k*-mer counts in the original assemblies were regressed against those mapped from SRS reads. The r^2 of the linear fit was used as the genotyping quality score. To measure genotyping accuracy in the pan-genome setting, only the *k*-mer set derived from the genotyped individual was retained as the input for regression.

Data filtering: A final set of 29,052 VNTR regions was called by filtering based on genotyping quality. The quality of a locus was measured by the mean and median r^2 across individuals. Loci with both mean and median r^2 across individuals below 0.8 were removed from the final call set.

Expanding pan-genomes graphs: For the remaining 14 genomes (HG00268, HG00512, HG00513, HG00731, HG00732, HG01352, HG02059, HG02106, HG02818, HG04217, NA12878, NA19238, NA19239 and NA19434), VNTR loci are identified by mapping from the genomic regions of the good set, processed by boundary expansion, converted to graphs, pruned by short reads and merged with the first-generation

pan-genome graphs. The final pan-genome graphs were used to genotype large Illumina datasets, measure length prediction accuracy, analyze population structures and map eQTL.

Predicting VNTR lengths: Read depths at VNTR regions usually vary considerably from locus to locus. Furthermore, the sampling bias of different sequencing runs are also different, which limits our ability to genotype the accurate length of VNTRs. To account for this, we compute locus-specific biases (LSBs) b_s for each sample s, a tuple of (genome g, sequencing run) as follows:

$$b_s = \frac{kms_s}{cov_s \times L_g}$$

,where L_g is the ground truth VNTR lengths of 29,052 loci in genome g; kms_s is the sum of *k*-mer counts in each locus mapped by samples s; cov_s is the global read depth of sample s estimated by averaging the read depths of 397 unique regions without any types of repeats or duplications.

The ground truth VNTR length of a locus l in genome g is averaged across haplotypes:

$$L_{g,l} = \frac{1}{H}\sum_{h=1}^{H} L_{g,h,l}$$

,where H is the number of haplotype(s) in genome g, i.e. 2 for normal individuals and 1 for complete hydatidiform mole (CHM) samples.

With the above bias terms, the VNTR length of locus l in sample s can be computed by:

$$L_{s,l} = \frac{kms_{s,l}}{cov_s \times b_{\hat{s}}}$$

,where cov_s is same as described above; $b_{\hat{s}}$ is the estimated LSBs computed from sample $\hat{s}$ with ground truth VNTR lengths; $kms_{s,l}$ is the sum of *k*-mer counts of locus l mapped by sample s. We assume the LSBs that best approximates b_s come from samples within the same sequencing run. Without prior knowledge on the

ground truth VNTR lengths of s and therefore b_s, we determine the "closest" sample $\hat{s}$ w.r.t. s based on r^2 between the read depths, RD, of the 397 unique regions as follows:

$$\hat{s} = \underset{s', s' \in GT, s' \neq s}{\operatorname{argmax}} \; r^2(RD_{s'}, RD_s)$$

, where GT is the set of samples with ground truths and within the same sequencing run as s. We cross-validate our approach by leaving one sample out of the pan-genome database and evaluating the prediction accuracy on the excluded sample.

For comparison, VNTR lengths were also estimated by a read depth method. For each VNTR region, the read depth, computed with samtools bedcov -j, was divided by the global read depth, computed from the 397 nonrepetitive regions, to give the length estimate.

<u>Comparing with GraphAligner:</u> The GFA file for each locus-RPGG was generated with a pipeline of bcalm v2.2.3 (option: -kmer-size 21 -abundance-min 1) and stark v1.0 using the VNTR sequences from all assemblies as input. GFA files were then reindexed and concatenated with vg v1.24.0 to generate the RPGG for 29,052 loci. Error-free paired-end reads were simulated from all VNTR regions at 2x coverage with 150 bp read length and 600 bp insert size (300 bp gap between each end). Reads were aligned to the RPGG using GraphAligner v1.0.11 with option -x dbg --seeds-minimizer-length 13 --try-all-seeds. Reads with alignment identity > 90% were counted from the output gam file. To compare in a similar setting, danbing-tk was run with option -gc -thcth 117 -k 21 -cth 45 -rth 0.5 to assert >90% identity for all reads aligned, given that $(read_length - kmer_size + 1) \times 0.9 = 117$.

<u>V_{ST} calculation:</u> V_{ST} was calculated according to (Redon et al. 2006):

$$V_{ST}[i] = max(0, \frac{var_{all} - \frac{1}{n}\sum_{p \in P} var_p \times n_p}{var_{all}})$$

Top V_{ST} loci were considered as the top 0.1%.

Identifying fragile loci: A locus was annotated as a candidate for being fragile if at least one individual had outlying *k*-mer dosage ≥ six standard deviations above the mean, using population and locus specific summary statistics, no individuals had dosage less than 10 or a bimodal distribution was not detected (diptest v0.75-7, p > 0.9). Among this set, the number of times each genome appeared as an outlier was used to select a set of genomes with an over abundant contribution to fragile loci. Any candidate locus with an individual that was an outlier in at least four other loci was removed from the candidate list. The loci were compared to gencode v34, excluding readthrough, pseudogenes, noncoding RNA, and nonsense transcripts.

Identifying differential motif usage and expansion: We use the EAS population as the reference for measuring differential motif usage and expansion. Initially, a lasso fit using the statsmodel.api.OLS function in python statsmodel v0.10.1 (Seabold and Perktold 2010) was performed for each locus to identify the *k*-mer with the most variance explained (VEX) in VNTR lengths using the following formula: $y = Xb + \epsilon$, where $y \in \mathbb{R}^N$ is the VNTR length of N individuals in the EAS population; $X \in \mathbb{R}^{N \times M}$ is the *k*-mer dosage matrix for N individuals at M loci; $b \in \mathbb{R}^M$ is the model coefficient, and $\epsilon \sim N(0, \sigma^2)$ is the error term. The lasso penalty weight α was scanned starting at 0.9 with at a step size of −0.1 until at least one covariate has a positive weight or α is below 0.1. Loci with no covariate having positive weight are removed, leaving 29,051 loci for downstream analysis. The *k*-mer with the highest weight is denoted as the most informative *k*-mer (mi-kmer) for the locus.

To identify loci with differential motif usage between populations, we subtracted the median count of the mi-kmer of the AFR from the EAS population for each locus, denoted as kmc_d. The null distribution of kmc_d was estimated by bootstrap. Specifically, EAS individuals were sampled with replacement

$N_{EAS} + N_{AFR}$ times, matching the sample sizes of the EAS and AFR populations, respectively. The bootstrap statistics, kmc_d^*, were computed by subtracting the median count of the mi-kmer of the last N_{AFR} from the first N_{EAS} bootstrap samples for each locus. The estimated null distribution is then used to determine the threshold for calling a locus having significant differential motif usage between populations (two-sided P < 0.01).

To identify loci with differential motif expansion between populations, we subtracted the proportion of VEX by mi-kmer in the AFR from the EAS population, denoted as r_d^2. The null distribution of r_d^2 was estimated by bootstrap in a similar sampling procedure as kmc_d, except for subtracting the proportion of VEX by the mi-kmer in the last N_{AFR} from the first N_{EAS} bootstrap samples for each locus. The estimated null distribution is used to determine the threshold for calling a locus having significant differential motif expansion between populations (two-sided P < 0.01).

eQTL mapping

Retrieving datasets: WGS datasets of 879 individuals, normalized gene expression matrices and covariates of all tissues are accessed from the GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2).

Genotype data preprocessing: VNTR lengths are genotyped using daunting-tk with options: -gc -thcth 50 -cth 45 -rth 0.5. All the *k*-mer counts of a locus are summed and adjusted by global read depth and ploidy to represent the approximate length of a locus. Adjusted values are then z-score normalized as input for eQTL mapping.

Expression data preprocessing: The downloaded expression matrices are already preprocessed such that outliers are rejected and expression counts are quantile normalized as standard normal distribution. Confounding factors

such as sex, sequencing platform, amplification method, technical variations and population structure are removed prior to eQTL mapping to avoid spurious associations. Technical variations are corrected with the covariates, including PEER factors, provided by the GTEx Consortium. Population structures are corrected with the top 10 principal components (PCs) from the SNP matrix of all samples. Particularly, principal component analysis (PCA) was performed jointly on the intersection of the SNP sets from GTEx samples and 1KGP Omni 2.5 SNP genotyping arrays (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz). This is done by first using CrossMap v0.4.0 to liftover the SNP sites from Omni 2.5 arrays to GRCh38, followed by extracting the intersection of the two SNP sets using vcftools isec. The SNP set is further reduced by LD-pruning with plink v1.90b6.12 using the options: --indep 50 5 2, leaving a total of 757,000 sites. Finally, PCA on the joint SNP matrix was done by smartpca v13050. The normalized expression matrix are residualized with the above covariates using the following formula:

$$Y = (1 - H)Y'$$

$$H = C(C^T C)^{-1} C^T$$

, where Y is the residualized expression matrix; Y' is the normalized expression matrix; H is the projection matrix; C is the covariate matrix where each column corresponds to a covariate mentioned above. The residualized expression values are z-score normalized as the input of eQTL mapping.

<u>Association test</u>: VNTRs within 100 kb to a gene are included for eQTL mapping. Linear regression was done using the statsmodel.api.OLS function in python statsmodel v0.10.1 (Seabold and Perktold 2010) with expression values as the dependent variable and genotype values as the independent variable. Nominal P values are computed by performing *t* tests on slope. Adjusted P values are computed by Bonferroni correction on

nominal P values. Under the assumption of at most one causal VNTR per gene, we control gene-level false discovery rate at 1%. Specifically, the adjusted P values of the lead VNTR for each gene are taken as input for Benjamini-Hochberg procedure using statsmodels.stats.multitest.fdrcorrection v0.10.1. Lead VNTRs that passed the procedure are identified as eVNTRs.

Data availability

The overall analysis pipeline is delivered in a software package at https://github.com/ChaissonLab/danbing-tk .

1000 Genomes Acknowledgement:

The following cell lines/DNA samples were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: [NA06984, NA06985, NA06986, NA06989, NA06994, NA07000, NA07037, NA07048, NA07051, NA07056, NA07347, NA07357, NA10847, NA10851, NA11829, NA11830, NA11831, NA11832, NA11840, NA11843, NA11881, NA11892, NA11893, NA11894, NA11918, NA11919, NA11920, NA11930. NA11931, NA11932, NA11933, NA11992, NA11994, NA11995, NA12003, NA12004, NA12005, NA12006, NA12043, NA12044, NA12045, NA12046, NA12058, NA12144, NA12154, NA12155, NA12156, NA12234, NA12249, NA12272, NA12273, NA12275, NA12282, NA12283, NA12286, NA12287, NA12340, NA12341, NA12342, NA12347, NA12348, NA12383, NA12399, NA12400, NA12413,, NA12414, NA12489, NA12546, NA12716, NA12717, NA12718, NA12748, NA12749, NA12750, NA12751, NA12760, NA12761, NA12762, NA12763, NA12775, NA12776, NA12777, NA12778, NA12812, NA12813, NA12814, NA12815, NA12827, NA12828, NA12829, NA12830, NA12842, NA12843, NA12872, NA12873, NA12874, NA12878, NA12889, NA12890]. These data were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1. Data accession IDs are given in supplementary table S4.

(7571): 68–74.
Alwazzan, M., E. Newman, M. G. Hamshere, and J. D. Brook. 1999. "Myotonic Dystrophy Is Associated with a Reduced Level of RNA from the DMWD Allele Adjacent to the Expanded Repeat." *Human Molecular Genetics* 8 (8): 1491–97.
Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, Annemarie E. Welch, Max L. Dougherty, et al. 2019. "Characterizing the Major Structural Variant Alleles of the Human Genome." *Cell* 176 (3): 663–75.e19.
Bakhtiari, Mehrdad, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L. Neuhausen, Bjarni V. Halldórsson, Kári Stefánsson, Melissa Gymrek, and Vineet Bafna. 2020. "Variable Number Tandem Repeats Mediate the Expression of Proximal Genes." *bioRxiv*. https://doi.org/10.1101/2020.05.25.114082.
Bakhtiari, Mehrdad, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. 2018. "Targeted Genotyping of Variable Number Tandem Repeats with adVNTR." *Genome Research* 28 (11): 1709–19.
Benson, G. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research*. https://doi.org/10.1093/nar/27.2.573.
Braida, Claudia, Rhoda K. A. Stefanatos, Berit Adam, Navdeep Mahajan, Hubert J. M. Smeets, Florence Niel, Cyril Goizet, et al. 2010. "Variant CCG and GGC Repeats within the CTG Expansion Dramatically Modify Mutational Dynamics and Likely Contribute toward Unusual Symptoms in Some Myotonic Dystrophy Type 1 Patients." *Human Molecular Genetics* 19 (8): 1399–1412.
Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. "Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes." *Nature Communications* 10 (1): 1784.
Chen, Sai, Peter Krusche, Egor Dolzhenko, Rachel M. Sherman, Roman Petrovski, Felix Schlesinger, Michael Kirsche, et al. 2019. "Paragraph: A Graph-Based Structural Variant Genotyper for Short-Read Sequence Data." *bioRxiv*. https://doi.org/10.1101/635011.
Chin, Chen-Shan, Paul Peluso, Fritz J. Sedlazeck, Maria Nattestad, Gregory T. Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods* 13 (12): 1050–54.
Consortium, Gtex, and GTEx Consortium. 2017. "Genetic Effects on Gene Expression across Human Tissues." *Nature*. https://doi.org/10.1038/nature24277.
Consortium, International Human Genome Sequencing, and International Human Genome Sequencing Consortium. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature*. https://doi.org/10.1038/35057062.
Dolzhenko, Egor, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, et al. 2019. "ExpansionHunter: A Sequence-Graph-Based Tool to Analyze Variation in Short Tandem Repeat Regions." *Bioinformatics* 35 (22): 4754–56.
Du, Zhenglin, Liang Ma, Hongzhu Qu, Wei Chen, Bing Zhang, Xi Lu, Weibo Zhai, et al. 2019. "Whole Genome Analyses of Chinese Population and De Novo Assembly of A Northern Han Genome." *Genomics, Proteomics & Bioinformatics* 17 (3): 229–47.
Eggertsson, Hannes P., Snaedis Kristmundsdottir, Doruk Beyter, Hakon Jonsson, Astros Skuladottir, Marteinn T. Hardarson, Daniel F. Gudbjartsson, Kari Stefansson, Bjarni V. Halldorsson, and Pall Melsted. 2019. "GraphTyper2 Enables Population-Scale Genotyping of Structural Variation Using Pangenome Graphs." *Nature Communications*. https://doi.org/10.1038/s41467-019-13341-9.
Fairley, Susan, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. 2020. "The International Genome Sample Resource (IGSR) Collection of Open Human Genomic Variation Resources." *Nucleic Acids Research* 48 (D1): D941–47.
Fotsing, Stephanie Feupe, Jonathan Margoliash, Catherine Wang, Shubham Saini, Richard Yanicky, Sharona Shleizer-Burko, Alon Goren, and Melissa Gymrek. 2019. "The Impact of Short Tandem Repeat Variation

on Gene Expression." *Nature Genetics* 51 (11): 1652–59.
Garrison, Erik, Jouni Sirén, Adam M. Novak, Glenn Hickey, Jordan M. Eizenga, Eric T. Dawson, William Jones, et al. 2018. "Variation Graph Toolkit Improves Read Mapping by Representing Genetic Variation in the Reference." *Nature Biotechnology* 36 (9): 875–79.
Gatchel, Jennifer R., and Huda Y. Zoghbi. 2005. "Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles." *Nature Reviews. Genetics* 6 (10): 743–55.
Gymrek, Melissa, Thomas Willems, Audrey Guilmatre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J. Daly, et al. 2016. "Abundant Contribution of Short Tandem Repeats to Gene Expression Variation in Humans." *Nature Genetics* 48 (1): 22–29.
Hannan, Anthony J. 2010. "Tandem Repeat Polymorphisms: Modulators of Disease Susceptibility and Candidates for 'missing Heritability.'" *Trends in Genetics*. https://doi.org/10.1016/j.tig.2009.11.008.
———. 2018. "Tandem Repeats Mediating Genetic Plasticity in Health and Disease." *Nature Reviews. Genetics* 19 (5): 286–98.
Hickey, Glenn, David Heller, Jean Monlong, Jonas A. Sibbesen, Jouni Sirén, Jordan Eizenga, Eric T. Dawson, Erik Garrison, Adam M. Novak, and Benedict Paten. 2020. "Genotyping Structural Variants in Pangenome Graphs Using the vg Toolkit." *Genome Biology* 21 (1): 35.
Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. 2012. "De Novo Assembly and Genotyping of Variants Using Colored de Bruijn Graphs." *Nature Genetics* 44 (2): 226–32.
Iqbal, Zamin, Isaac Turner, and Gil McVean. 2013. "High-Throughput Microbial Population Genomics Using the Cortex Variation Assembler." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts673.
Jiang, Zhaoshi, Haixu Tang, Mario Ventura, Maria Francesca Cardone, Tomas Marques-Bonet, Xinwei She, Pavel A. Pevzner, and Evan E. Eichler. 2007. "Ancestral Reconstruction of Segmental Duplications Reveals Punctuated Cores of Human Genome Evolution." *Nature Genetics* 39 (11): 1361–68.
Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37 (5): 540–46.
Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive K-Mer Weighting and Repeat Separation." *Genome Research* 27 (5): 722–36.
LaPierre, Nathan, Kodi Taraszka, Helen Huang, Rosemary He, Farhad Hormozdiari, and Eleazar Eskin. n.d. "Identifying Causal Variants by Fine Mapping Across Multiple Studies." https://doi.org/10.1101/2020.01.15.908517.
Li, Heng, Jonathan M. Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, and Daniel MacArthur. n.d. "New Synthetic-Diploid Benchmark for Accurate Variant Calling Evaluation." https://doi.org/10.1101/223297.
Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations." *Nature* 538 (7624): 201–6.
Masson, Emmanuelle, Cédric Le Maréchal, Giriraj R. Chandak, Jérôme Lamoril, Stephane Bezieau, Swapna Mahurkar, Seema Bhaskar, D. Nageshwar Reddy, Jian-Min Chen, and Claude Férec. 2008. "Trypsinogen Copy Number Mutations in Patients with Idiopathic Chronic Pancreatitis." *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association* 6 (1): 82–88.
Mousavi, Nima, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. 2019. "Profiling the Genome-Wide Landscape of Tandem Repeat Expansions." *Nucleic Acids Research* 47 (15): e90.
Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. "An Eulerian Path Approach to DNA Fragment Assembly." *Proceedings of the National Academy of Sciences of the United States of America* 98 (17): 9748–53.
Pevzner, Pavel A., Haixu Tang, and Glenn Tesler. 2004. "De Novo Repeat Classification and Fragment Assembly." *Genome Research* 14 (9): 1786–96.

Porubsky, David, Peter Ebert, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Katherine M. Munson, Melanie Sorensen, et al. n.d. “A Fully Phased Accurate Assembly of an Individual Human Genome.” https://doi.org/10.1101/855049.

Porubsky, David, Shilpa Garg, Ashley D. Sanders, Jan O. Korbel, Victor Guryev, Peter M. Lansdorp, and Tobias Marschall. 2017. “Dense and Accurate Whole-Chromosome Haplotyping of Individual Genomes.” *Nature Communications* 8 (1): 1293.

Rakocevic, Goran, Vladimir Semenyuk, Wan-Ping Lee, James Spencer, John Browning, Ivan J. Johnson, Vladan Arsenijevic, et al. 2019. “Fast and Accurate Genomic Analyses Using Genome Graphs.” *Nature Genetics*. https://doi.org/10.1038/s41588-018-0316-4.

Raphael, Benjamin, Degui Zhi, Haixu Tang, and Pavel Pevzner. 2004. “A Novel Method for Multiple Alignment of Sequences with Repeated and Shuffled Elements.” *Genome Research* 14 (11): 2336–46.

Rautiainen, Mikko, Veli Mäkinen, and Tobias Marschall. 2019. “Bit-Parallel Sequence-to-Graph Alignment.” *Bioinformatics* 35 (19): 3599–3607.

Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. “Global Variation in Copy Number in the Human Genome.” *Nature* 444 (7118): 444–54.

Seo, Jeong-Sun, Arang Rhie, Junsoo Kim, Sangjin Lee, Min-Hwan Sohn, Chang-Uk Kim, Alex Hastie, et al. 2016. “De Novo Assembly and Phasing of a Korean Human Genome.” *Nature* 538 (7624): 243–47.

Shi, Lingling, Yunfei Guo, Chengliang Dong, John Huddleston, Hui Yang, Xiaolu Han, Aisi Fu, et al. 2016. “Long-Read Sequencing and de Novo Assembly of a Chinese Genome.” *Nature Communications* 7 (June): 12065.

Song, Janet H. T., Craig B. Lowe, and David M. Kingsley. 2018. “Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia.” *American Journal of Human Genetics* 103 (3): 421–30.

Taliun, Daniel, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, et al. 2019. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *bioRxiv*. https://doi.org/10.1101/563866.

Thornton, C. A., J. P. Wymer, Z. Simmons, C. McClain, and R. T. Moxley 3rd. 1997. “Expansion of the Myotonic Dystrophy CTG Repeat Reduces Expression of the Flanking DMAHP Gene.” *Nature Genetics* 16 (4): 407–9.

Viguera, E., D. Canceill, and S. D. Ehrlich. 2001. “Replication Slippage Involves DNA Polymerase Pausing and Dissociation.” *The EMBO Journal* 20 (10): 2587–95.

Vilen, Suvi-Tuuli, Juho Suojanen, Fidel Salas, Juha Risteli, Merja Ylipalosaari, Outi Itkonen, Hannu Koistinen, et al. 2012. “Trypsin-2 Enhances Carcinoma Invasion by Processing Tight Junctions and Activating ProMT1-MMP.” *Cancer Investigation* 30 (8): 583–92.

Zook, Justin M., Nancy F. Hansen, Nathan D. Olson, Lesley Chapman, James C. Mullikin, Chunlin Xiao, Stephen Sherry, et al. 2020. “A Robust Benchmark for Detection of Germline Large Deletions and Insertions.” *Nature Biotechnology*, June. https://doi.org/10.1038/s41587-020-0538-8.