The motif composition of variable-number tandem repeats impacts gene expression

Tsung-Yu Lu[1], Mark J.P. Chaisson[1,*]
[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, USA.
[*]corresponding author.

Abstract
Understanding the impact of DNA variation on human traits is a fundamental question in human genetics. Variable number tandem repeats (VNTRs) make up roughly 3% of the human genome but are often excluded from association analysis due to poor read mappability or divergent repeat content. While methods exist to estimate VNTR length from short-read data, it is known that VNTRs vary in both length and repeat (motif) composition. Here, we use a repeat-pangenome graph (RPGG) constructed on 35 haplotype-resolved assemblies to detect variation in both VNTR length and repeat composition. We align population scale data from the Genotype-Tissue Expression (GTEx) Consortium to examine how variations in sequence composition may be linked to expression, including cases independent of overall VNTR length. We find that 20,834 VNTRs are associated with nearby gene expression through motif variations, of which only 5.1% associations are accessible from length. Fine-mapping identifies 273 genes to be likely driven by variation in certain VNTR motifs and not overall length. To demonstrate the utility of association using VNTR motifs, we examine the intronic VNTR of *CACNA1C*, which has been reported to be associated with schizophrenia risk through motif variation. We show that in healthy populations a previously identified schizophrenia risk motif is associated with decreased expression of *CACNA1C*, and detect an additional novel motif with similar effect.

Introduction
Variable number tandem repeats (VNTRs) are repetitive DNA sequences with the size of a repeat unit greater than six nucleotides. The copy number of a repeat unit is hypervariable due to its susceptibility to replication slippage caused by strand mispairing between the same[1] or across haplotypes[2]. At the sequence level, single nucleotide variations or short indels are also prevalent along a repeat sequence and can greatly expand the number of identified alleles relative to classification by length[3]. Altogether, copy number variations, SNVs and short indels contribute to the full spectrum of VNTR polymorphism. Missing heritability[4] that cannot be explained by SNVs can be partially attributed to VNTR polymorphisms[5–7]. Accumulating evidence indicates that VNTRs are associated with a diverse array of human traits and are casual to several diseases at the copy number level or sequence level[8–10]. Furthermore, significant enrichment of VNTRs in subtelomeric genes that are mostly expressed in the brain suggests further exploration of their roles in shaping behavioral/cognitive polymorphisms and modulating neurological disease risks[11].

VNTR length polymorphism can modulate human traits through several mechanisms, including changing the number of protein domains[12], the distance between gene and gene regulators[13], the number of regulator binding sites[14], and the number of CpG sites[15,16]. Abundant associations between repeat copy number and human traits have been widely reported[10] and provide insights to functional annotations. However, it is impossible to fully understand the biological functions of VNTRs without examining variation at the sequence level. For example, a single cytosine insertion in *MUC1* VNTR was identified to be causal to medullary cystic kidney disease type 1 by adding a premature stop in translation[17]. In addition, certain repeat motifs in *CACNA1C* but not the total repeat copy number were reported to be associated with schizophrenia risk by tuning gene expression activity[18]. In both cases, long-read sequencing such as single-molecule real-time sequencing or capillary sequencing has been useful to resolve the full sequence of VNTRs and yield meaningful clinical interpretations.

Currently large-scale sequencing efforts use high-throughput short-read sequencing (SRS)[19], however, VNTR analysis with SRS suffers from ambiguity in read alignment, allelic bias of reference and the hypermutability of repeat sequences. Single-nucleotide and small indel variant calls from VNTR regions using short-read alignments are error-prone and blacklisted by ENCODE[20]. Recently, several methods have been developed specifically to estimate VNTR length from short-read data using Hidden Markov Models[9], read-depth[21], and repeat-pangenome graphs[7]. These approaches have found an association between estimated VNTR length and gene expression. In this study, we use a reference pangenome graph (PGG) to reduce allelic bias when mapping short reads to a reference, and to improve variant inference for motif composition. The PGG is a graph-based data structure that summarizes sequence variations from a collection of samples by representing variants as alternative paths or "bubbles" from the reference[22]. One of the most common implementations of PGG is to use a sequence graph. In this model, each node represents an allele; each edge points to a downstream allele; a traversal through the graph matches an observed haplotype. This allows sequencing reads to be placed more accurately across the genome and significantly improves variant calling accuracy in regions containing SVs[23–26], with the majority of which coming from indel events within VNTRs[25]. However, variant calling remains challenging for multiallelic VNTR regions as the position of calls varies[25]; an extra processing step is needed to reveal the multiallelic property of a locus.

Another commonly used graph model is the de Bruijn graph (dBG). The main distinction is that each node is a unique *k*-mer derived from one or more *k*-base substrings present in the input sequences. By augmenting with additional haplotype or distance information, dBG-based models have been useful in genome assembly[27,28] and variant calling[29,30]. Furthermore, this formulation groups all occurrences of repetitive k-mers across input sequences by construction, which can be a particularly desirable property when studying the biological implication of VNTR motifs.

By leveraging the advantages of PGG and dBG, genotyping VNTR from SRS samples at a population scale has been made possible with danbing-tk[7]. The method constructs a repeat-pangenome graph (RPGG) that consists of disjoint locus-RPGGs, each representing a

single VNTR locus and encoding observed VNTR alleles with a dBG. Read mapping to RPGG reveals the coverage of each *k*-mer and can be accumulated as a copy number estimate, allowing associations between repeat copy number and human trait to be identified.

In this work, we extend the application of the danbing-tk to examine the association between each path in the graph, or VNTR "motif", and gene expression using the complete read-mapping output i.e. the coverages of all *k*-mers. We show that motif usage/repeat count can be accurately estimated in a RPGG, and that this may be used to compare motif composition between individuals. We map genomes sequenced by GTEx to discover associations between motif composition and gene expression that are independent of VNTR length. Fine-mapping of these loci finds novel causal links where motif composition is likely to modulate changes in gene expression.

Results

Repeat pangenome graphs enable accurate profiling of motif composition

We developed an extended computational analysis pipeline based on the previously published danbing-tk method to map read depth and identify eQTLs from individual paths in an RPGG (Fig. 1). The RPGG is constructed using 35 haplotype-resolved assemblies including three trios released by the Human Genome Structural Variation Consortium (HGSVC)[31]. Orthologous boundaries of 80,478 VNTR loci were annotated using danbing-tk[7] from a set of 84,411 VNTRs (Methods). We further augment the VNTR annotations with additional 40 clinically relevant loci (Supplemental table S1), giving a total of 80,518 loci (Supplemental Data S1) for subsequent analyses. The VNTRs annotated have a mean length of 685 bp across assemblies versus 665 bp in GRCh38 (Fig. 2a, Supplemental Fig. S1). Among the 70 haplotypes, a VNTR has an average of 7.4 alleles per locus when defining an allele based on exact length (Supplemental Fig. 2). Each locus has an average of 3.0 alleles that are observed only once, denoted as private allele count, and 597 loci that have at least half the alleles ($N \geq 35$) as private (Supplemental Fig. S2). The number of alleles per locus is positively correlated with VNTR length ($\rho=0.36$). As VNTR length increases, the private allele count also increases (Supplemental Fig. S2), e.g. private allele count = 10.2 when VNTR length > 500 bp.

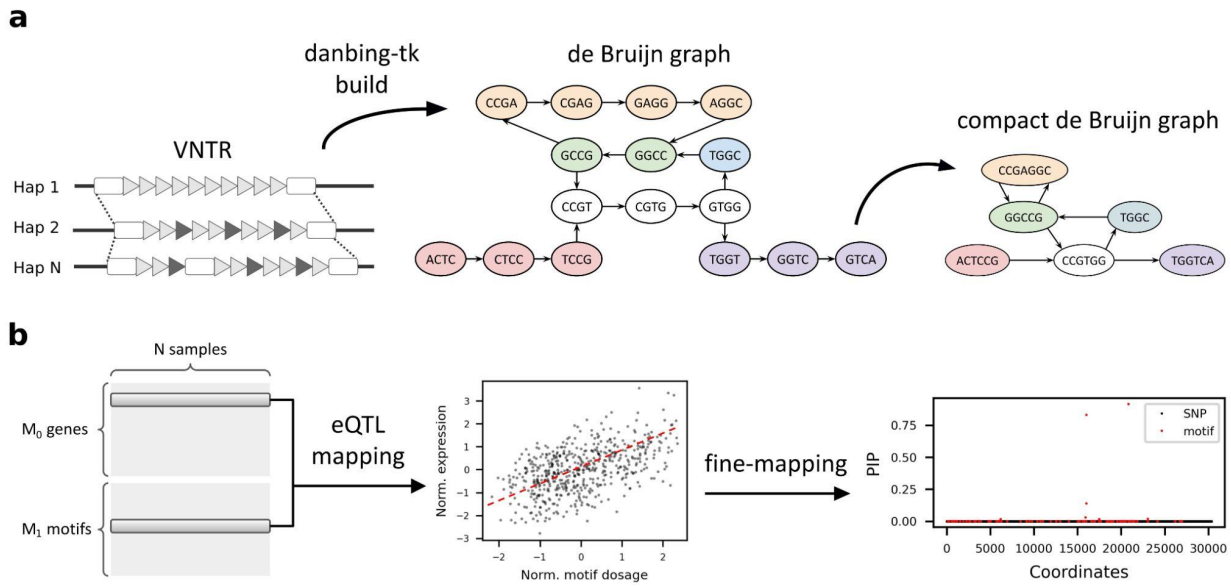

Figure 1. Methods overview. **a,** Estimating the dosages of VNTR motifs using a locus-RPGG. A locus-RPGG is built from haplotype-resolved assemblies by first annotating the orthology mapping of VNTR boundaries and then encoding the VNTR alleles with a de Bruijn graph (dBG), or locus-RPGG. A compact dBG is constructed by merging nodes on a non-branching path into a unitig, denoted as a motif in this context. Motif dosages of a VNTR can be computed by aligning short reads to an RPGG and averaging the coverage of nodes corresponding to the same motif. **b,** Identifying likely causal eMotifs. The dosage of each motif is tested against the expression level of a nearby gene. Genes in significant association with at least one motif (denoted as eMotif) are fine-mapped using susieR[32] in order to identify eMotifs that are likely causal to nearby gene expression. All GTEx variants (Methods) and lead motifs from each gene-VNTR pair are included in the fine-mapping model.

We used danbing-tk[7] to encode the allele information across haplotypes in an RPGG, consisting of 80,518 locus subgraphs, or locus-RPGGs. The graph has a total of 398,576,090 *k*-mers (*k*=21) or nodes, and 404,035,564 edges, with an average of 4,950 nodes in each locus-RPGG when including 700 bp flanking sequences on both sides. Each locus-RPGG has an average of 104 nodes or 10.6% nodes that are observed only in one assembly. The repeat region of RPGG (excluding flanking sequences) has 62,457,731 *k*-mers (Fig. 2b), which is 38% greater than the graph built from GRCh38 alone (n=45,148,309) and 38% greater than the smallest graph built from an assembly (HG00864, n=45,197,090). We evaluated the quality of the alignments to each locus-RPGG by measuring the consistency of *k*-mer counts from assemblies versus from short-read mapping, denoted as aln-r^2 (Methods). Overall, there is a slightly high aln-r^2 of 0.94±0.08 (Fig. 2c) compared to the aln-r^2 on the previously published 19 haplotype-resolved assemblies (0.93±0.08), with enrichment of loci with higher aln-r^2 (Supplemental Fig. S3). We had previously applied an aln-r^2 cutoff to limit computational requirements, however the

danbing-tk alignment method had a large, but constant factor improvement in runtime, and all 80,518 loci are genotyped in this study.

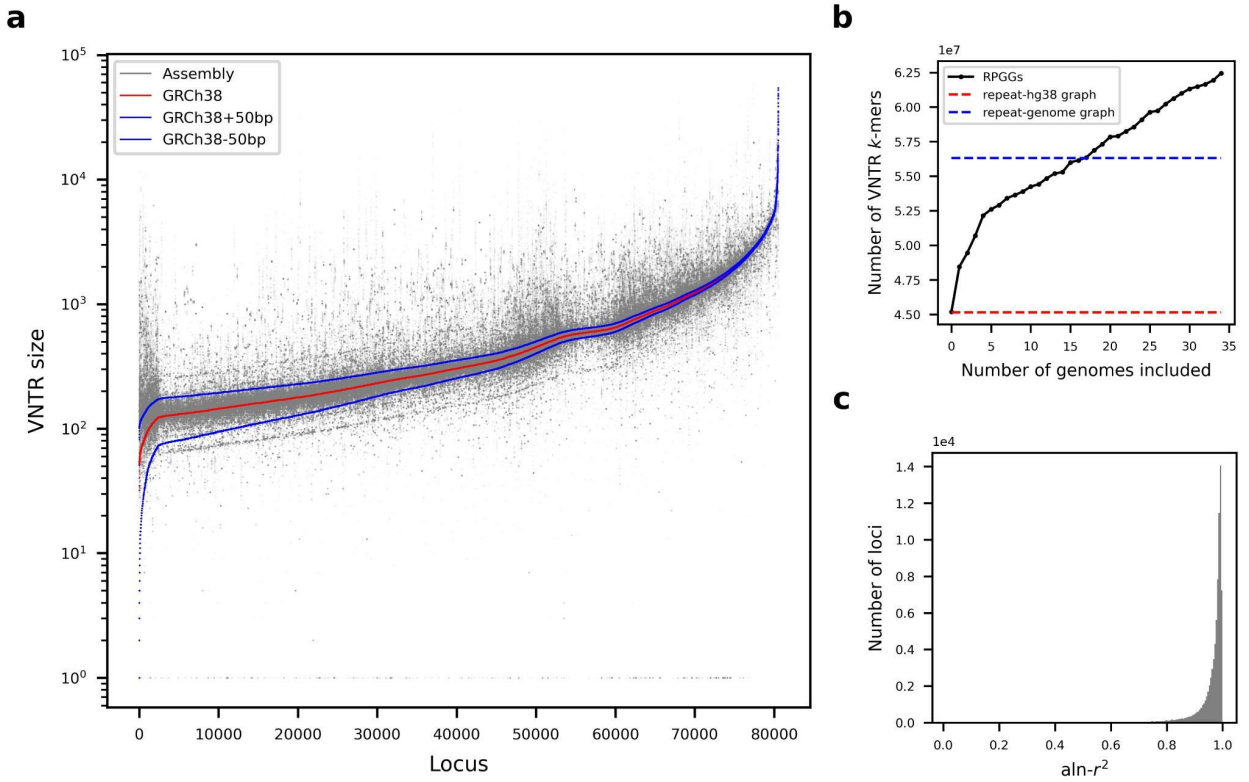

Figure 2. Characteristics of VNTRs and the RPGG. **a,** Size distribution of VNTR alleles across 35 HGSVC assemblies. Each dot represents the size of a VNTR locus in an assembly. The order of 80,518 VNTR loci were sorted according to size in GRCh38. **b,** Cumulative graph sizes. A total of 35 repeat-genome graphs were incrementally added to the RPGG in the order of their respective graph size. The red dash line denotes the size of the repeat graph built from GRCh38. The blue dash line denotes the average size of the graphs built from assemblies. **c,** Distribution of the aln-r^2 for all locus-RPGGs. The aln-r^2 of each locus was computed by regressing the assembly *k*-mer counts against the read *k*-mer counts from graph alignments.

VNTR motif composition has pervasive *cis*-effects on gene expression

Using the short-read alignment module in danbing-tk, we estimated the VNTR content of 80,518 loci as graph genotypes and processed 838 GTEx genomes[33] using ~12 cpu hours and ~29 Gb memory per sample. The read alignments to each subgraph are summarized as a vector of the number of reads mapped to each node/*k*-mer. When normalized by global read depth these represent mapping dosage used as input for *cis*-eQTL mapping.

In Lu 2021[7], *cis*-eQTL mapping using an RPGG was reported using the sum of the dosage vector for each locus-RPGG as an estimate of VNTR length. Replicating this approach on the 35-genome RPGG, we discovered 1,355 VNTRs in association with nearby gene

expression (denoted as eVNTR), which is 3.9-fold the number (N=346) reported from the previous RPGG built on 32,138 VNTRs[7]. Of the original eVNTRs, 84% (290/346) were reproduced in this version while the remaining 16% are enriched in discoveries with lower significance, in particular when nominal $P > 10^{-5}$ (odds ratio = 0.819, Fisher's exact $P = 1.4 \times 10^{-9}$, Supplemental Fig. S4).

In addition to VNTR length, our previous work[7] identified motifs enriched in certain populations sequenced by the 1000 Genomes Project Consortium[34]. We hypothesized that differential motif usage across individuals, possibly independent of overall VNTR length variation, can modulate nearby gene expression. To test this, we converted each locus-RPGG to a compact de Bruijn graph, and considered each path as a locus to test in eQTL mapping. The original RPGG contains on average 776 nodes in each locus-RPGG, which is reduced to 55 paths (referred to as motifs hereafter) after compaction (Supplemental Fig. S5), with a total of 4,456,881 motifs.

To ensure the quality of read-mapping to each motif in each locus-RPGG, we evaluate the "mappability" of each motif by measuring the consistency between the dosage from short reads and the dosage from the ground-truth assemblies using mean absolute percentage error (MAPE, Methods). We removed 49.8% (2,219,780/4,456,881) of the motifs with MAPE ≥ 0.25 (Fig. 3a). The number of motifs with zero variance in absolute percentage error (n=764,354) is equivalent to the number paths private to one genome among the 35 HGSVC assemblies (Supplemental Fig. S6) and were retained for subsequent analyses. Similar to eQTL mapping on SNVs where homozygous variants are removed, we avoid testing "invariant" motifs that appear the same number of times in a repeat across all assemblies. This further removes 25.5% (571,537/2,237,101) of motifs. By construction, our RPGG could contain loci with multiple VNTRs in close proximity but spaced apart by short flanking sequences. We removed additional 0.1% (2,005/1,665,564) motifs that are derived from those sequences and could be simply explained by SNVs, leaving a total of 1,663,559 motifs for eQTL mapping. For each motif, we consider only samples whose dosages are within two standard deviations from the mean, avoiding the discoveries of associations whose effects are mainly driven by outliers. This on average removes 34 from 813 of the samples per locus (Supplemental Fig. S7, Methods).

We ran *cis*-eQTL mapping for each VNTR motif and discovered 20,834 eVNTRs, including 53,209 motifs associated with nearby gene expression, denoted as eMotifs (Fig. 3b, Supplemental Data S2). While 78% (1,053/1,355) of the eVNTRs discovered using length estimates were also reported using motif dosage, 95% (19,781/20,834) of the eVNTRs discovered from motif were undetectable with length-based eQTL mapping. To assess the reproducibility of our methods, we performed eQTL mapping on the Geuvadis dataset[35] and compared the discoveries with the GTEx results. We found that 68.4% (1,923/2,811) of the eMotifs and 92.9% (2,397/2,579) of the eVNTRs from Geuvadis were also observed in at least one GTEx tissue. Unreplicated eGene-eVNTR pairs (ePairs) tend to have lower significance in Geuvadis, especially when $P > 2.5 \times 10^{-9}$ (odds ratio = 3.2, Fisher's exact $P = 4.2 \times 10^{-39}$, Supplemental Fig. S8). When comparing the whole blood tissue from GTEx and the

lymphoblastoid from Geuvadis, the effect sizes from the two datasets had a correlation coefficient of 0.80 (Fig. 3c). Among the replicated ePairs, 91% (800/878) had the same sign of effect (Fig. 3d), suggesting motif variations as a common explanatory variable in gene expression.

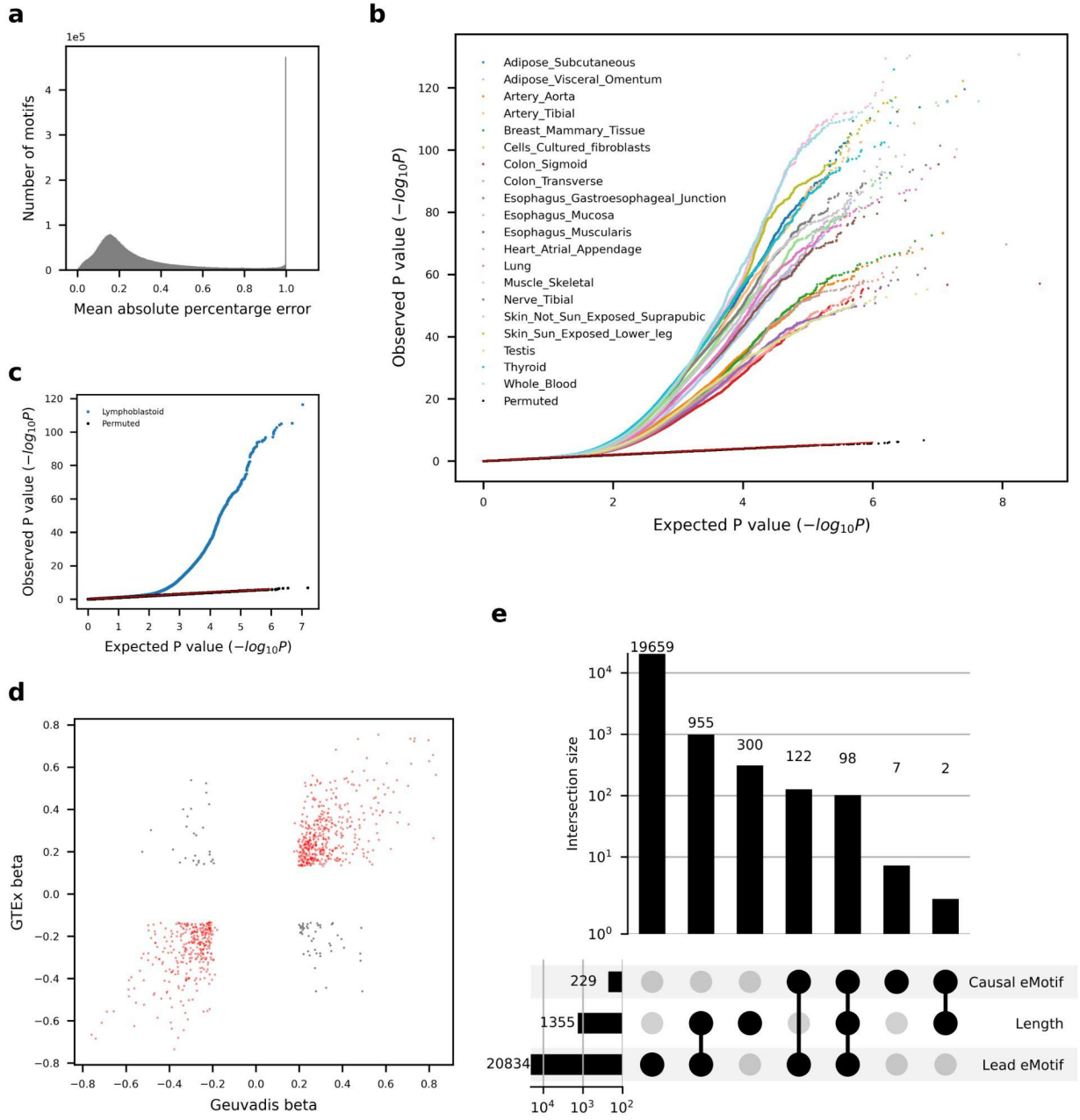

Figure 3. *cis*-eQTL mapping of VNTR motifs. **a**, Distribution of motif-MAPE. For each motif, the mean absolute percentage error (MAPE) was computed by averaging the absolute percentage errors (APEs, Methods) across 35 genomes. A total of 4,456,881 motifs were shown in the plot. **b,** Quantile-quantile plot of gene-level eMotif discoveries across 20 human tissues from GTEx

datasets. The expected P-values (x-axis) were drawn from Unif(0,1) and plotted against observed nominal P-values obtained from *t*-test on the slope of each linear model consisting of expression (response variable) versus motif dosage (explanatory variable). **c-d.** Replication on Geuvadis dataset. **c,** The quantile-quantile plot shows the observed P-value of each association test (two-sided t-test) versus the P-value drawn from the expected uniform distribution. Black dots indicate the permutation results from the top 5% associated (gene, motif) pairs. **d,** Correlation of eMotif effect sizes between Geuvadis and GTEx whole blood tissue. Only eGenes significant in both datasets were shown. Each pair of gene and motif that has the same/opposite sign across datasets were colored in red/black. **e.** VNTR-centric view of gene-level eQTL discoveries and fine-mapping. Lead eMotif denotes any VNTRs that are associated with gene expression through at least one eMotif under gene-level discoveries. Length denotes any VNTRs that are associated with gene expression through length under gene-level discoveries. Causal eMotif denotes any VNTRs of which a motif passes the fine-mapping procedure with a posterior inclusion probability ≥ 0.8 while being a significant eQTL under genome-wide P-value cutoff. Motifs with the lowest P-value for each VNTR-gene pair are included in the fine-mapping model.

Disease relevance of eMotifs

Among the 40 additional disease-relevant tandem repeats (matched with 36 genes) included in the RPGG, 18 of them including *C9orf72*, *CACNA1C*, *CSTB*, *DRD4* and *MUC21* (Supplemental Table S2) were identified as eQTLs and were associated with their original disease-linked genes. Additionally, at least one eVNTR was detected for 17 of the 18 remaining genes and was different from the originally annotated disease-relevant tandem repeat.

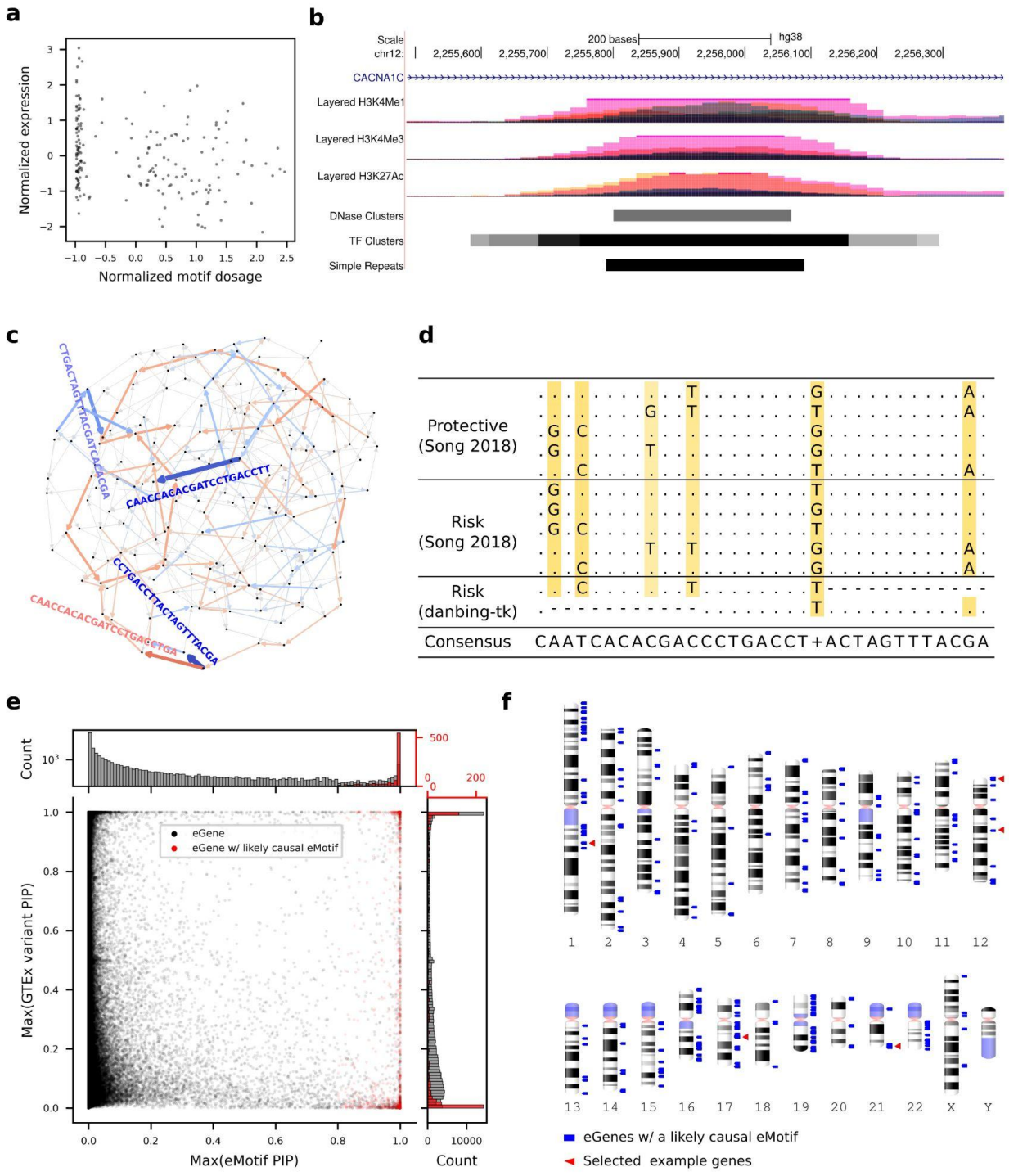

Figure 4. Disease-relevant genes and fine-mapping. **a-d,** Identification of risk motifs for *CACNA1C* expression. The motifs in *CACNA1C* VNTR at chr12:2,255,789-2,256,088 were analyzed. **a,** Association of *CACNA1C* VNTR motif CAACCACACGATCCTGACCTT with gene expression. **b,** Genome browser view of *CACNA1C* VNTR. **c,** Graph visualization of motif effect sizes from the *CACNA1C* VNTR. Each edge denotes a motif and is colored blue/red if

having a negative/positive effect on gene expression. Color saturation and edge width both scale with the absolute value of effect size. The sequence of a motif is shown parallel to the edge and colored in dark blue if having a significant effect or colored in light red/blue if borderline significant. **d,** Multiple sequence alignment of known *CACNA1C* risk motifs[18] and the likely causal eMotifs reported in this study. **e,** Distribution of posterior inclusion probability (PIP). eMotifs with PIP greater than 0.8 were called likely causal. Each dot is from a fine-mapped gene in a tissue. The maximal PIPs of all eMotifs/GTEx variants nearby a gene are shown on the x-axis/y-axis. **f,** Ideograms of selected eGenes and eGenes with a likely causal eMotif. The locations of all 510 eGenes with a likely causal eMotif are shown. The locations of *AVPR1A*, *CACNA1C*, *ITGB2*, *FCGR3B*, *FCGR3A* and *KANSL1* are shown in the selected example genes track.

We investigated two examples of motifs associated with disease to see if they had associations with expression in healthy individuals. Landefeld et al.[36] report that the "RS1" VNTR at the 5'UTR of *AVPR1A* is associated with externalizing behaviors while Vollebregt et al.[37] report that the "RS3" VNTR but not RS1 is associated with childhood onset aggression. No association between VNTR length of RS1 nor RS3 with *AVPR1A* expression was found, however we found eMotifs for *AVPR1A* in healthy individuals that correspond to the RS1 VNTR $(CTAT)_5TTAT(CTAT)_4$ ($b=-0.26$, $P=1.0\times10^{-4}$) and the RS3 VNTR $C(AG)_{10}$ ($b=0.21$, $P=2.1\times10^{-4}$). The RS3 VNTR (chr12:63,156,354-63,156,429) has a nested repeat structure with an average size of 701 bp in assemblies. It consists of two slightly divergent copies that each carries a highly repetitive (CT)nTT(CT)n(GT)n core motif at chr12:63,156,354-63,156,394 and chr12:63,156,701-63,156,751 (Supplemental Fig. S9). Other VNTR annotation approaches might consider this region as two separate VNTRs of which the length of the core motif is associated with *AVPR1A* expression.

The decreased expression of *CACNA1C* was known to be a risk factor for schizophrenia and has been reported to be associated with several 30 bp risk motifs at chr12:2,255,789-2,256,090 based on a case-control study[18]. Here, we found that the expression of *CACNA1C* in brain cerebellar hemisphere was associated with a risk eMotif CAACCACACGATCCTGACCTT (denoted as motif 1, $b=-0.44$, $P=1.2\times10^{-8}$, Fig. 4a). The eMotif covers five of the six mutation sites (Fig. 4b) and is novel to all of the risk motifs reported previously[18]. In addition, we were able to replicate findings from the case-control study[18] in healthy populations. The risk eMotif CCTGACCTTACTAGTTTACGA ($b=-0.40$, $P=1.5\times10^{-7}$, denoted as motif 2) that covers two of the mutation sites (Fig. 4b) matches two of the risk motifs and none of the protective motifs, indicating the prevalence of risk-modulating motifs even among healthy populations. When examining the frequency of the two risk eMotifs in the 35 HGSVC assemblies, 28 haplotypes carry motif 1 while 38 haplotypes carry motif 2 (Supplemental Fig. S10). Most of the individuals carry only few copies of the risk motifs except for the haplotype 2 of HG02818 which has 98 copies of motif 1. Annotating the locus-RPGG with the eQTL effect sizes also reveals that motifs with minor risk and protective effects are

pervasive within the locus, e.g. CTGACTAGTTTACGATCACACGA (b=−0.29, P=1.6×10^{-4}) and CAACCACACGATCCTGACCTGA (b=0.29, P=2.7×10^{-4}) (Fig. 4c). The size of this VNTR locus in GRCh38 is underrepresented with only 301 bp compared to an average size of 6,247 bp across assemblies. In addition, immense histone modification, DNase clusters and TF clusters signals can also be found in this locus (Fig. 4d), necessitating future investigations to fully understand the regulation mechanism of *CACNA1C*.

To further narrow down eMotifs that are likely causal to gene expression, we used susieR[32] to fine-map the 1Mb *cis*-window of each eGene. We discovered 273 out of 19,965 eGenes of which the highest eMotif posterior inclusion probability (PIP) is greater than 0.8, suggesting the likely causal roles of these eMotifs (Fig. 4e-f, Supplemental Data S3). The expression of these 273 eGenes are likely modulated by 560 eMotifs, or equivalently 229 eVNTRs (Fig. 3e). On average, 81.4% of the eGenes are shared across tissues while 82.4% of the eVNTRs and 53.1% of the eMotifs are observed across multiple tissues (Supplemental Figs. S11-S13).

Table 1. Selected examples of disease-related genes with likely causal eMotifs.

eGene	Disease	Likely causal eMotif [1]	eVNTR coordinate of the likely causal eMotif	Tissue(s) with a likely causal eMotif	Number of eMotifs	Beta [1]	PIP [1]	Position	Histone/ DNase/ TF [2]	Ref/asm TR size [3]
ITGB2	Leukocyte-adhesion deficiency syndrome, Immunologic deficiency syndromes, Atherosclerosis	ACCCTGGATGCCT GTGGGCTGCCTTC CTCACC	chr21:44,928,811-44,929,048	Exposed skin, Unexposed skin	2	0.27	0.83	exon 1	+/+/+	237/264
FCGR3B	Aggressive periodontitis, Lupus nephritis, Rheumatoid arthritis	TACTTGGTGACATG ATTGTGAGAATAAG CTCTGGCGA	chr1:161,545,578-161,545,729	Exposed skin, Whole blood, Lung, Atrial appendage, Left ventricle	2	0.44	1.00	intron of FCGR3A	+/+/+	151/126
FCGR3A	Immunodeficiency 20, Recurrent viral infections, Systemic lupus erythematosus, Alloimmune neonatal neutropenia	TACTTGGTGACATG ATTGTGAGAATAAG CTCTGGCGA	chr1:161,545,578-161,545,729	Whole blood, Unexposed skin	2	0.26	0.99	intron	+/+/+	151/126
KANSL1	Koolen-de Vries syndrome	AGCCCTGTCTCTAC AAAAAATACAAAAT TTAGGC	chr17:46,217,604-46,217,944	Nerve tibial, Thyroid, Exposed skin, Unexposed skin, Lung, Breast, Adipose visceral	24	0.46	0.90	5'UTR	-/+/+	340/5248

[1] The one with the lowest q-value in *cis*-eQTL mapping if there are likely causal eMotifs in multiple tissues.
[2] Denotes whether the VNTR coordinate on GRCh38 overlaps with the histone H3K4Me1 H3K4Me3, H3K27Ac mark, the DNase clusters, or the TF clusters track in Genome Browser.
[3] The VNTR size in GRCh38 (ref) versus the average size in assemblies (asm).

In search of eGene versus likely causal eMotif pairs that could have potential clinical implications, we first ranked eGenes by their fold change in disease enrichment using GS2D[38]. eGenes with associated disease matching the fine-mapped tissues and of which at least one eVNTR located within the gene body were retained. The top four example genes (Table 1) have important roles in immune response (*ITGB2*, *FCGR3B* and *FCGR3A*), cell-cell interactions (*ITGB2*), or histone modification (*KANSL1*). Most haplotypes in the HGSVC assemblies carry only up to two copies of the listed motifs (Supplemental Fig. S14), except for the motif of

ITGB2, which has 1-7 copies in observed haplotypes. Most eVNTRs in the HGSVC assemblies have similar size to GRCh38 except for the *KANSL1* VNTR at chr17:46,217,604-46,217,944, which is approximately 15 times of the sizes in GRCh38.

Discussion

Genomic variant discovery serves to link genetic and phenotypic variation. Using gene expression as a phenotypic measure, diverse classes of variation have been found to have an effect on gene expression including single-nucleotide variants [33], structural variation [31,39,40], STRs [41], and VNTRs [9]. Here we show that in addition to association of VNTR length with expression, a more nuanced measurement of VNTR variation that takes into account sequence composition reveals eMotifs that influence gene expression.

Overall, we find 20,834 VNTR loci containing at least one eMotif. In contrast, previous studies that used associations based on length estimate alone ranged between 163-2,980 eVNTRs [7,9,21], with the number roughly correlating with the number of loci each study profiled. Although more tests per VNTR locus are performed, the fine-mapping analysis finds that the majority of variants are linked with nearby eQTLs. After applying fine-mapping, 229 (1.1%) eVNTRs contain motifs determined as causal. In contrast, 0.18% of the 4.3M eQTL variants discovered in the GTEx (v8) are fine-mapped[33].

We observe that most eVNTRs have different motifs positively and negatively associated with expression of the same nearby gene. Most eQTL mapping pipelines are based on biallelic variants. When encoding a variant, the reference allele is usually treated as zero while the alternative allele is treated as one. Alternatively, this can be viewed as encoding the alternative allele as its copy number, which is simply one for a biallelic variant, while keeping the reference allele as zero. When the same encoding method is applied to a VNTR locus consisting of a reference motif and an alternative motif, the only difference is that the alternative allele becomes a continuous value representing the adjusted motif depth, and may take on a positive or negative association depending on the relation to the reference motif.

This study profiles 80,518 VNTR loci, a 2.5-fold increase over our previous analysis of VNTR variation using repeat-pangenome graphs[7] that is largely attributable to the high-quality haplotype-resolved assemblies used to construct the pangenome. The size of the graph increases sequentially with the number of assemblies included in the graph, and is consistent with the increasing number of structural variants discovered in VNTRs by whole-genome alignment[31]. The inclusion of additional genomes from large-scale sequencing projects such as the Human Pangenome Reference Consortium will yield an improved estimate of saturation of VNTR variation.

The use of repeat-pangenome graphs in this study differs from other implementations of pangenom-graphs including those constructed by progressive whole-genome alignment[25] and variant inclusion[42], both of which preserve haplotype information from the genomes or variants used to construct the pangenome graph. While systematic analysis of variation in VNTRs and association with expression has not yet been conducted using these approaches, we anticipate the

repeat-pangenome graph will provide complementary analysis. In particular, variant genotyping in graphs that preserve haplotype as implemented by Giraffe[42] and PanGenie[43] corresponds to associating read data with haplotypes (paths in a pangenome graph) covering variants. These approaches provide highly accurate genotyping of variants shared with the graph, however hypervariable VNTR sequences are more likely to have differences from genomes represented in the graph, and additional analysis is required to quantify motif usage in addition to genotype.

The implementation of the pangenome as a de Bruijn graph is an elegant approach to identifying the composition of identical motif repeats, however small differences in motif composition can make the graph complex, and additional development is necessary to identify graph topologies that naturally reflect VNTR repeat composition. One result of this complexity is our number of eMotifs that are deemed likely causal using fine mapping is possibly an underestimate. Many motifs have highly correlated read dosage, however we use a conservative approach of considering each motif as an independent variable for fine mapping. Future development that merges similar motifs to the same edge both aggregate depth otherwise split on several edges, and reduce the correlated motifs tested during fine mapping.

In summary, this study demonstrates how VNTR composition has a pervasive influence on gene expression, and highlights the need to profile variation in complex, repetitive regions of the genome. We anticipate this approach will be useful for future expression and association studies.

Methods

Repeat pangenome graph construction

A set of 88,441 VNTR coordinates were retrieved from danbing-tk v1.3[7]. The VNTR set was obtained by (1) detecting VNTRs over the five haplotype-resolved assemblies (AK1, HG00514, HG00733, NA19240, NA24385) released by Lu and Chaisson [7] using Tandem Repeat Finder, (2) selecting for VNTRs with size between 100 bp and 10 kbp and motif size > 6 bp, and (3) applying danbing-tk to the VNTRs in the five genomes to identify 88,441 loci with proper orthology mapping. Next, we downloaded the 35 haplotype-resolved assemblies released by HGSVC[31]. VNTR annotations on the 35 assemblies and the corresponding RPGG were generated using the build module of danbing-tk, giving a total of 80,518 loci.

VNTR genotyping

Whole-genome sequencing (WGS) samples for HGSVC assemblies were retrieved from 1000 Genomes Project phase 3[34]. WGS samples for eQTL mapping were retrieved from GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)[33] and Geuvadis[35] with a total of 879 and 445 samples respectively. VNTRs were genotyped using danbing-tk v1.3 with options "-ae -kf 4 1 -gc 85 -k 21 -cth 45". The output *k*-mer counts were adjusted by the coverage of each sample before subsequent analyses.

Alignment quality analysis

The aln-r^2 statistic was used to evaluate how well a VNTR can be genotyped. It is the r^2 computed by regressing the *k*-mer counts from assemblies against the counts from reads aligned to the locus-RPGG. Since VNTRs were genotyped using the RPGG, any read *k*-mers not present in the original assembly were ignored.

Graph compaction and motif dosage computation

Locus-RPGG built from k=21 contains abundant contiguous paths without branches. It is desirable to reduce the number of nodes to be tested in eQTL mapping by merging nodes on this type of path. This is essentially a problem of converting dBGs to compact dBGs where nodes on a non-branching path are merged into a unitig, or referred to as a motif in this context. For each motif, we recorded the mapping relation from its constituent nodes and computed the motif dosage by averaging the *k*-mer counts from constituent nodes. In practice, when given a matrix of VNTR genotype where each column represents a *k*-mer in a locus-RPGG and each row represents a sample, the matrix of motif dosages can be simply computed by column operations using the mapping relations.

Quality control of motifs

To ensure the quality of motifs tested in eQTL mapping, we applied three filters to remove motifs (1) with mean absolute percentage error (MAPE) > 0.25, (2) with dosage invariant across HGSVC haplotypes, or (3) that were derived from an inter-VNTR region (denoted as "flank-like") but were included in the repeat region of a locus-RPGG due to the distance between the upstream and downstream VNTRs being < 700 bp. For the first filter, we computed mean absolute percentage error (APE) for each motif by measuring the error size of each motif to the linear fit for aln-r^2. Formally, let $\mathbf{x} = (x_1, x_2, \ldots, x_P)$ be the motif dosages from assemblies and $\mathbf{y} = (y_1, y_2, \ldots, y_P)$ be the motif dosages from short reads, where P is the number of motifs in the locus-RPGG. For each genome g, $\hat{y}_g$ is the fitted value for the dosage of a motif from the linear fit between $\mathbf{x}$ and $\mathbf{y}$. The MAPE of a motif can be computed as follows:

$$MAPE = \frac{1}{N}\sum_{g=1}^{N}\frac{|\hat{y}_g - y_g|}{y_g}$$

, where N is the number of genomes with the motif. For the second filter, a motif was removed if the dosage of a motif was the same across all 70 haplotypes. The dosage was set to zero if the motif was not present in a haplotype. For the third filter, any *k*-mers derived from the inter-VNTR regions before the VNTR merging step were extracted. Any motifs overlapping with these *k*-mers were removed.

eQTL mapping

Gene expression data was processed as previously described[7] unless stated otherwise. Briefly, normalized gene expression matrices and covariates of all tissues were retrieved from GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)[33]. Confounding factors were removed

using covariates including sex, sequencing platform, amplification method, PEER factors, and top 10 principal components (PCs) from the joint SNP matrix with 1KGP samples. Residualization of the gene expression matrices was done with the following formula:

$$Y = (I - C(C^T C)^{-1} C^T)Y',$$

where Y is the residualized expression matrix; Y′ is the normalized expression matrix; I is the identity matrix; C is the covariate matrix where each column corresponds to a covariate mentioned above.

For eQTL mapping using motif dosage, samples with motif dosage being two standard deviations away from the mean were removed for each motif. For eQTL mapping using VNTR length, samples with VNTR length being three standard deviations away from the mean were removed for each VNTR. The motif dosages, VNTR lengths, and the residualized expression counts for the remaining samples were z-score normalized before testing for association.

For gene-level *cis*-eQTL discoveries, the P-value of each test was adjusted according to the number of variants tested for each gene using Bonferroni correction. The minimal P-values from all the tests against each gene were extracted and controlled at 5% false discovery rate using the Benjamini–Hochberg procedure. Only one eMotif (if using motif dosage) or one eVNTR (if using VNTR length) was reported for each eGene. For *cis*-eQTL discoveries using a genome-wide P-value cutoff, all P-values from all tests were recorded and controlled at 5% false discovery rate using the Benjamini–Hochberg procedure. The P-value cutoffs range from 1.2×10^{-5} (Kidney) to 1.4×10^{-3} (Thyroid), depending on the power in each tissue (Supplemental Table S3). A VNTR that contains an eMotif was also regarded as an eVNTR. Consequently, an eVNTR can be associated with gene expression through length or motif dosage depending on the type of tests performed.

Fine-mapping

To evaluate whether the eMotifs are causal to gene expression, we used susieR[32] to fine-map the *cis*-window around the transcription start site of each gene. All variants in GTEx's catalog (GTEx_Analysis_2017-06-05_v8_WholeGenomeSeq_838Indiv_Analysis_Freeze.vcf.gz), including SNVs, indels or structural variants, were extracted if within the 1 Mb *cis*-windows. For each tissue, all motifs that have the lowest P-value for each gene-VNTR pair were extracted if within the 100 kb *cis*-windows. The extracted GTEx variants and motifs were taken as input for fine-mapping. Susie was run using L=5 to allow up to five sets of causal variants within the whole region. Motifs with posterior inclusion probability (PIP) ≥ 0.8 while passing the genome-wide P-value cutoff as described in the previous section were reported as likely causal eMotifs.

Data availability

Assemblies and corresponding short-read data are available from https://www.internationalgenome.org/data-portal/data-collection/hgsvc2 .

The RPGG constructed from genome assemblies, source code, and eQTL tables are available at https://sandbox.zenodo.org/record/1036159#.YjNifH_MJH5, https://zenodo.org/record/5093660#.YjNivGRlAWM and https://github.com/ChaissonLab/danbing-tk/releases/tag/v1.3 .
GTEx samples (phs000424.v8.p2) can be accessed from https://www.gtexportal.org/home/datasets .

Acknowledgement
This work has been funded by NHGRI U24HG007497 and NHGRI R01HG011649. We would like to thank Dr. Nicholas Mancuso for helpful comments and suggestions.

References

1. Viguera, E., Canceill, D. & Ehrlich, S. D. Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.* **20**, 2587–2595 (2001).
2. Jeffreys, A. J. *et al.* Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* **6**, 136–145 (1994).
3. Novroski, N. M. M., King, J. L., Churchill, J. D., Seah, L. H. & Budowle, B. Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Sci. Int. Genet.* **25**, 214–226 (2016).
4. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
5. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
6. Mitra, I. *et al.* Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**, 246–250 (2021).
7. Lu, T.-Y. & Chaisson, M. J. P. Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. *Nat. Commun.* **12**, 1–12 (2021).
8. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of

structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).

9. Bakhtiari, M. *et al.* Variable number tandem repeats mediate the expression of proximal genes. *Nat. Commun.* **12**, 2075 (2021).
10. Mukamel, R. E. *et al.* Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).
11. Linthorst, J. *et al.* Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. *Transl. Psychiatry* **10**, 369 (2020).
12. Desseyn, J.-L., Aubert, J.-P., Porchet, N. & Laine, A. Evolution of the Large Secreted Gel-Forming Mucins. *Mol. Biol. Evol.* **17**, 1175–1184 (2000).
13. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224–238.e15 (2018).
14. Tsuge, M. *et al.* A variable number of tandem repeats polymorphism in an E2F-1 binding element in the 5′ flanking region of SMYD3 is a risk factor for human cancers. *Nat. Genet.* **37**, 1104–1107 (2005).
15. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
16. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
17. Kirby, A. *et al.* Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat. Genet.* **45**, 299–303 (2013).
18. Song, J. H. T., Lowe, C. B. & Kingsley, D. M. Characterization of a Human-Specific

Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. *Am. J. Hum. Genet.* **103**, 421–430 (2018).

19. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
20. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).
21. Garg, P. *et al.* Pervasive cis effects of variation in copy number of large tandem repeats on local DNA methylation and gene expression. *Am. J. Hum. Genet.* **108**, 809–824 (2021).
22. Eizenga, J. M. *et al.* Pangenome Graphs. *Annu. Rev. Genomics Hum. Genet.* **21**, 139–162 (2020).
23. Chen, S. *et al.* Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
24. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).
25. Li, H., Feng, X. & Chu, C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* **21**, 265 (2020).
26. Sirén, J. *et al.* Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *Cold Spring Harbor Laboratory* 2020.12.04.412486 (2020) doi:10.1101/2020.12.04.412486.
27. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
28. Muggli, M. D. *et al.* Succinct colored de Bruijn graphs. *Bioinformatics* **33**, 3181–3187 (2017).

29. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).

30. Narzisi, G. *et al.* Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun Biol* **1**, 20 (2018).

31. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).

32. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* vol. 82 1273–1300 (2020).

33. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

34. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

35. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

36. Landefeld, C. C. *et al.* Effects on gene expression and behavior of untagged short tandem repeats: the case of arginine vasopressin receptor 1a (AVPR1a) and externalizing behaviors. *Transl. Psychiatry* **8**, 1–10 (2018).

37. Evidence for association of vasopressin receptor 1A promoter region repeat with childhood onset aggression. *J. Psychiatr. Res.* **140**, 522–528 (2021).

38. Fontaine, J. F. & Andrade-Navarro, M. A. Gene Set to Diseases (GS2D): disease enrichment analysis on human gene sets with literature data. *Genomics and Computational Biology* vol. 2 33 (2016).

39. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.*

49, 692–699 (2017).

40. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
41. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
42. Sirén, J. *et al.* Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
43. Ebler, J. *et al.* Pangenome-based genome inference. doi:10.1101/2020.11.11.378133.