Supplementary information

Table of contents

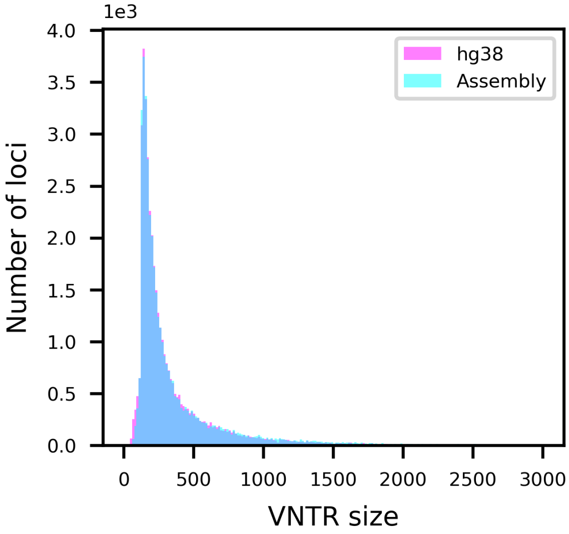
Supplemental Notes

Supplemental Note S1. File format of Supplementary Data S4.
The table of fine-mapped genes is saved in .tsv format, which is more convenient when used by command line utilities including awk and bioawk. Since each gene can have more than one likely-causal VNTR variant, some fields contain comma separated values and cannot be properly displayed in .csv format. The file can be opened in Google Sheets without error. To open in Excel, when importing the .tsv file, select "text" mode for the columns with VNTR coordinates (columns N, O, and P, and trailing columns if interested).

Supplemental Note S2. 1000 Genomes acknowledgement.
1000 Genomes acknowledgement for deep coverage of the extended 3202 genomes (or subset thereof):
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, NA06991, NA06993, NA06994, NA06995, NA06997, NA07000, NA07014, NA07019, NA07022, NA07029, NA07031, NA07034, NA07037, NA07045, NA07048, NA07051, NA07055, NA07056, NA07340, NA07345, NA07346, NA07347, NA07348, NA07349, NA07357, NA07435, NA10830, NA10831, NA10835, NA10836, NA10837, NA10838, NA10839, NA10840, NA10842, NA10843, NA10845, NA10846, NA10847, NA10850, NA10851, NA10852, NA10853, NA10854, NA10855, NA10856, NA10857, NA10859, NA10860, NA10861, NA10863, NA10864, NA10865, NA11829, NA11830, NA11831, NA11832, NA11839, NA11840, NA11843, NA11881, NA11882, NA11891, NA11892, NA11893, NA11894, NA11917, NA11918, NA11919, NA11920, NA11930, NA11931, NA11932, NA11933, NA11992, NA11993, NA11994, NA11995, NA12003, NA12004, NA12005, NA12006, NA12043, NA12044, NA12045, NA12046, NA12056, NA12057, NA12058, NA12144, NA12145, NA12146, NA12154, NA12155, NA12156, NA12234, NA12239, NA12248, NA12249, NA12264, NA12272, NA12273, NA12274, NA12275, NA12282, NA12283, NA12286, NA12287, NA12329, NA12335, NA12336, NA12340, NA12341, NA12342, NA12343, NA12344, NA12347, NA12348, NA12375, NA12376, NA12383, NA12386, NA12399, NA12400, NA12413, NA12414, NA12485, NA12489, NA12546, NA12707, NA12708, NA12716, NA12717, NA12718, NA12739, NA12740, NA12748, NA12749, NA12750, NA12751, NA12752, NA12753, NA12760, NA12761, NA12762, NA12763, NA12766, NA12767, NA12775, NA12776, NA12777, NA12778, NA12801, NA12802, NA12812, NA12813, NA12814, NA12815, NA12817, NA12818, NA12827, NA12828, NA12829, NA12830, NA12832, NA12842, NA12843, NA12864, NA12865, NA12872, NA12873, NA12874, NA12875, NA12877, NA12878, NA12889, NA12890, NA12891, NA12892]. These data were generated at the New York Genome Center with funds provided by NHGRI Grants 3UM1HG008901-03S1 and 3UM1HG008901-04S2.

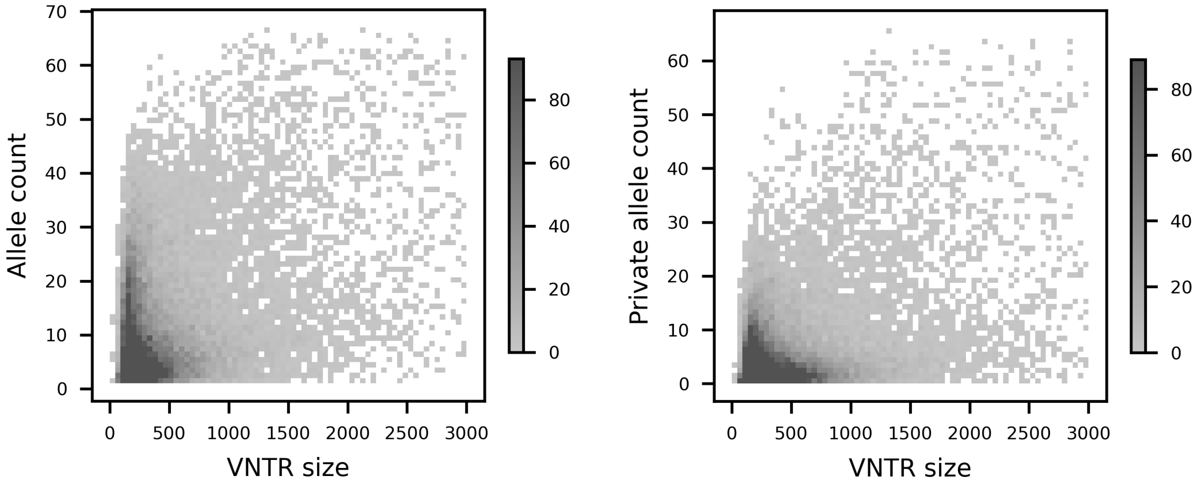
Supplemental Figures

Supplemental Fig. S1. Distribution of VNTR size in GRCh38 and in the 70 HGSVC assemblies.



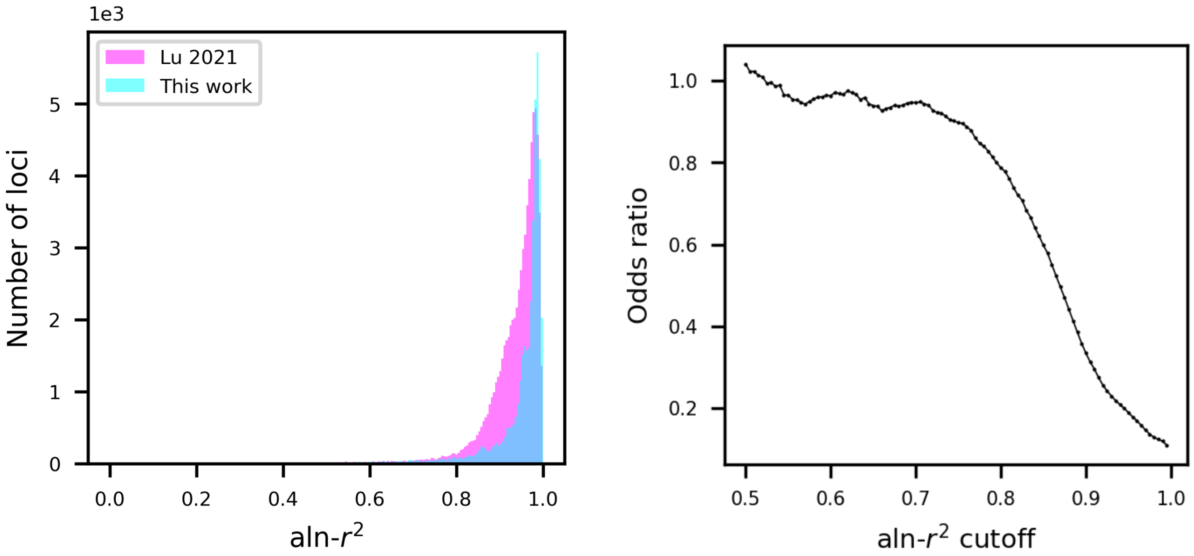
VNTR sizes greater than 3000 bp are not shown, contributing to 0.7% (281/40,314) of loci.

Supplemental Fig. S2. Distribution of VNTR size versus allele properties. Private alleles are those observed only once across the 70 haplotypes.



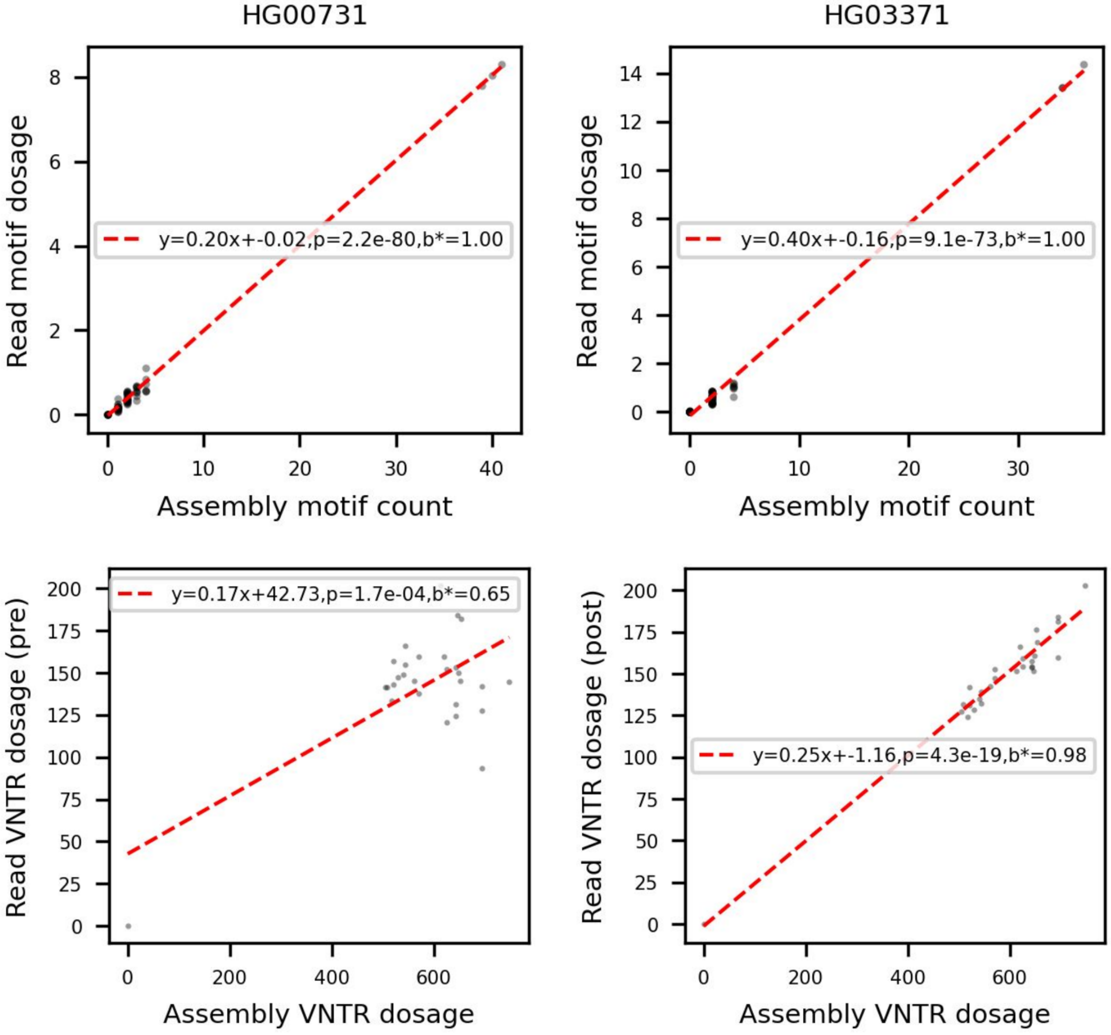
VNTR sizes greater than 3000 bp are not shown in both subpanels, contributing to 0.7% (281/40,314) of loci. Density greater than the limit shown in the color bar is clipped.

Supplemental Fig. S3. Distribution of aln-r^2 in this work and Lu 2021.



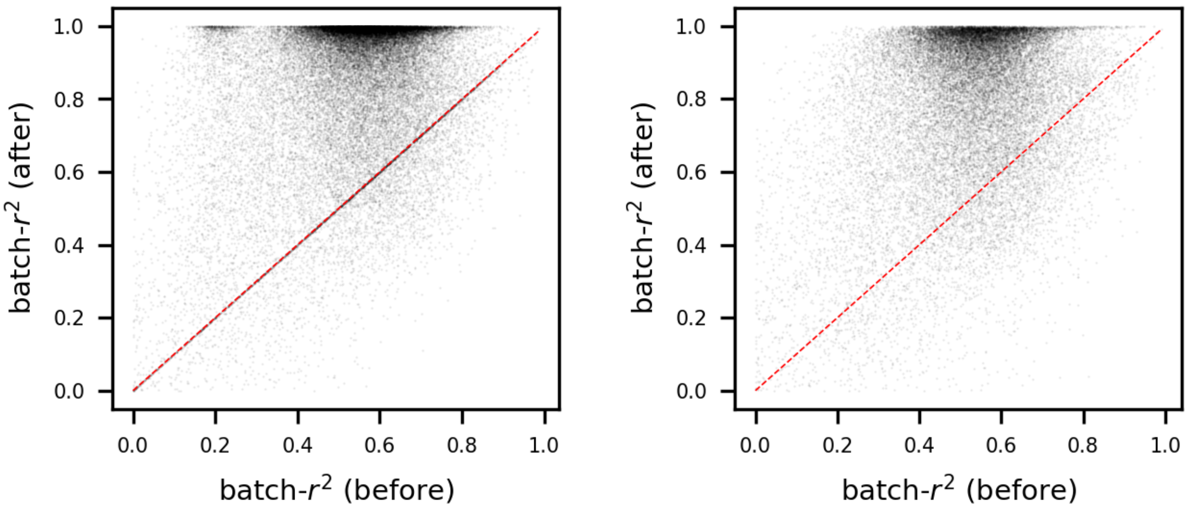
a, Comparing the distribution of aln-r^2 in both studies. **b,** Enrichment of high aln-r^2 loci in this work. Contingency tables with two variables, "≥ aln-r^2 cutoff" and "is from Lu 2021", were constructed to compute odds ratio.

Supplemental Fig. S4. Example of a locus with high mapping quality but low consistency in VNTR dosage across samples.



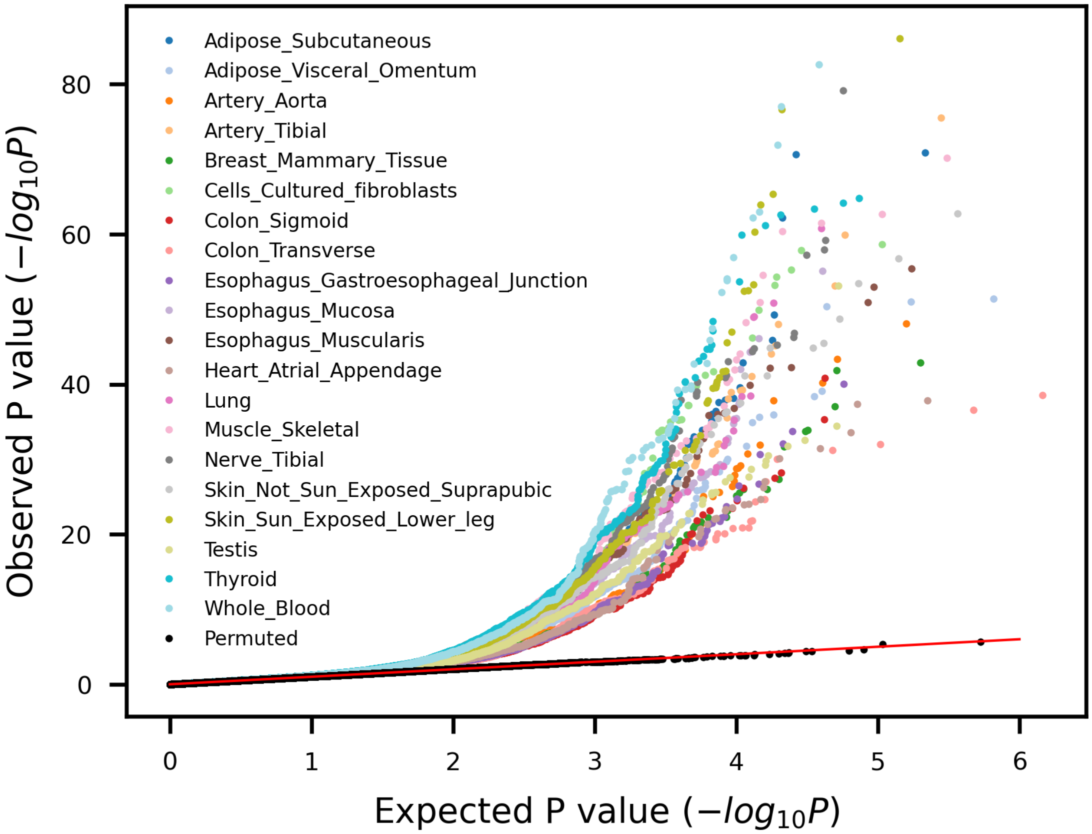
The VNTR at chr1:192,048,505-192,048,821 was selected to highlight the strong influence of bias, even when the average aln-r^2 was 0.99. The top two panels showed the two samples with two-fold difference in the slope of the regression lines. The batch-r^2 before correction (bottom left) is only 0.35 while the statistic after correction (bottom right) is 0.97.

Supplemental Fig. S5. Bias correction for VNTR dosage.



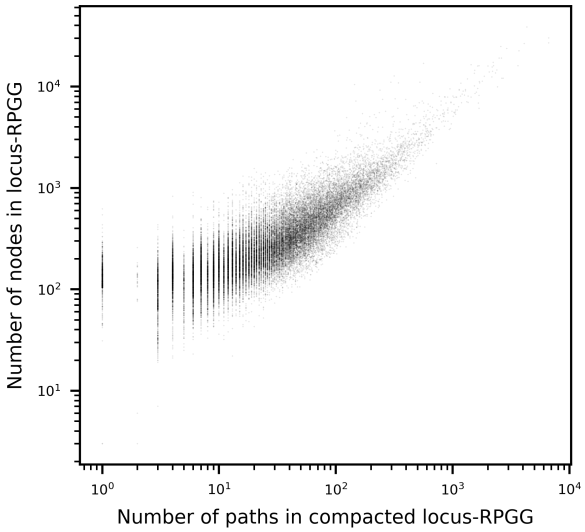
Bias correction on the HGSVC dataset (left panel). The batch-r^2 for 29,602 loci (75.7%) are shown. Bias correction on the HPRC dataset (right panel). The batch-r^2 for 27,499 loci (70.3%) are shown. The red dashed line indicates no improvement after bias correction.

Supplemental Fig. S6. eQTL mapping with VNTR length.



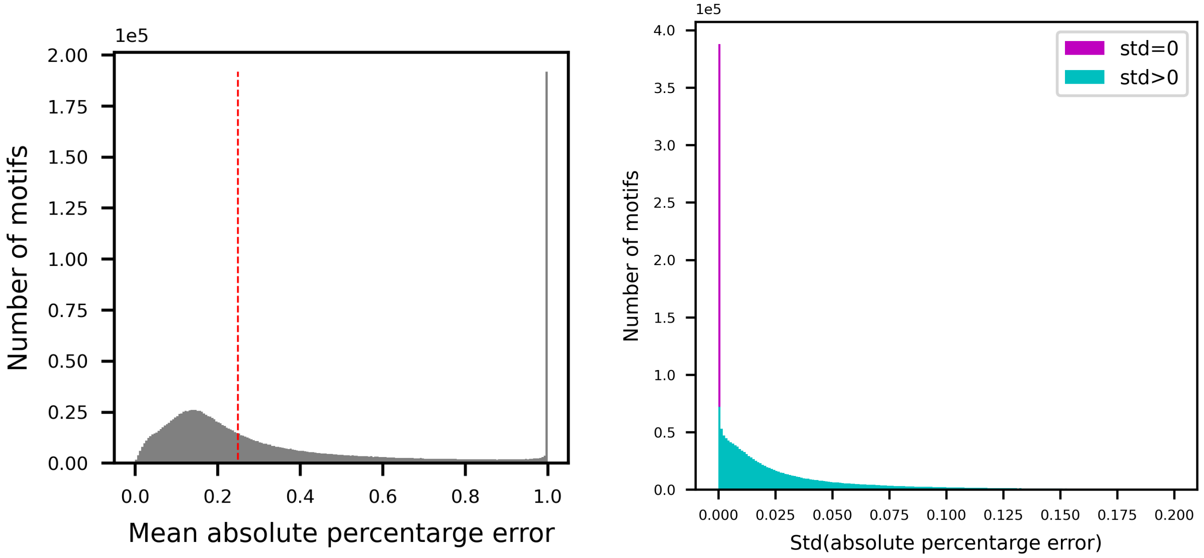
(A) Quantile-quantile plot of gene-level eVNTR discoveries using VNTR length across 20 human tissues from GTEx datasets. The expected P-values (x-axis) were drawn from Unif(0,1) and plotted against observed nominal association P-values.

Supplemental Fig. S7. Size of locus-RPGG before and after compaction.



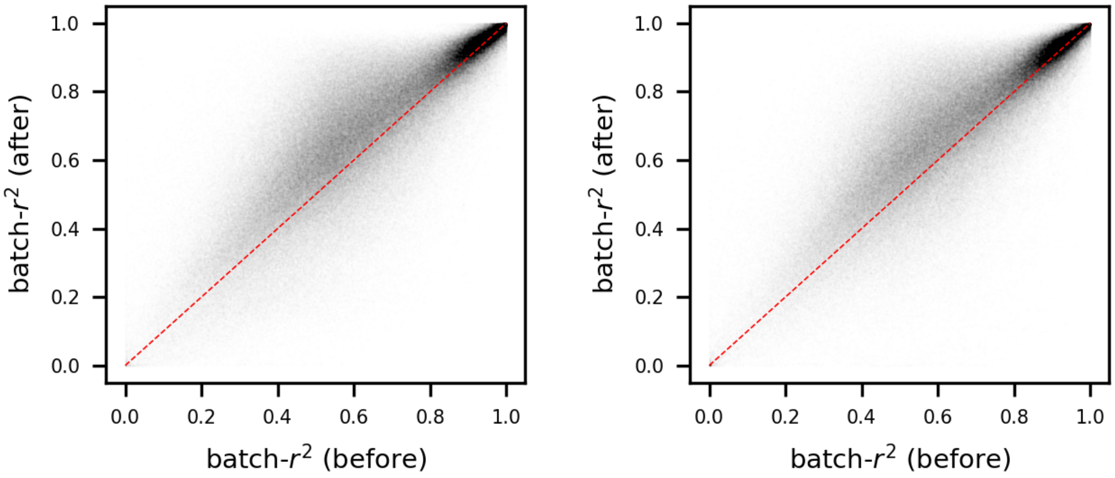
The number of nodes before compaction (y-axis) and the number of paths after compaction (x-axis) were compared for all 80,518 loci in the RPGG, both illustrated on a log scale.

Supplemental Fig. S8. Absolute percentage error of motifs.



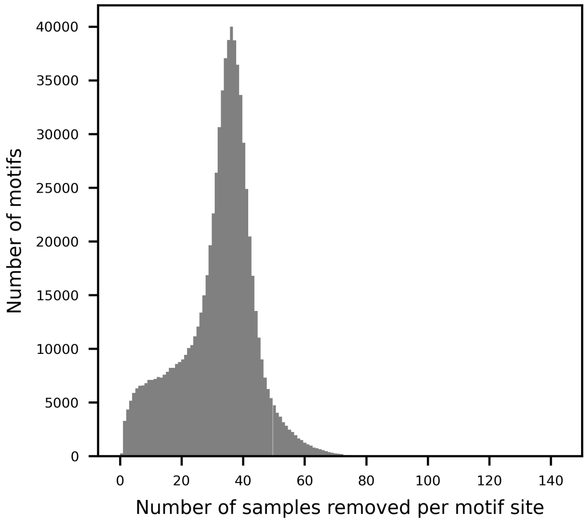
(Left) Distribution of motif-MAPE. For each motif, the mean absolute percentage error (MAPE) was computed by averaging the absolute percentage errors (APEs, Methods) across 35 genomes. A total of 1,724,959 motifs are shown in the plot, excluding the remaining 4.7% with MAPE greater than 1. (Right) Distribution of the standard deviation of motif-MAPE. Ground truth is measured from motif count in assemblies, and absolute percentage error is measured by regression from read depth on edges (motif) to assembly count. The spike at std=0 corresponds to motifs with only one observation across all haplotypes. Motifs with std > 0.2 are not shown, contributing to 7.8% (140,989/1,810,042) of motifs.

Supplemental Fig. S9. Bias correction on the HGSVC and HPRC dataset.



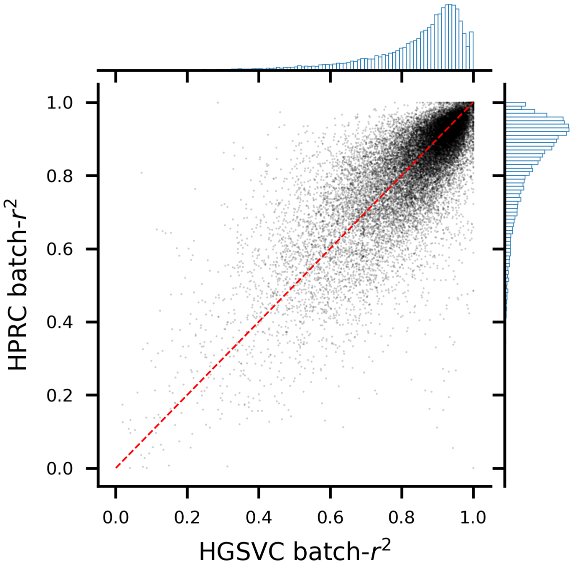
Motif batch-r^2 before and after bias correction were compared on the HGSVC (left panel) and the HPRC assemblies (right panel). VNTRs without invariant *k*-mers were removed from analysis; motifs in HPRC that do not overlap any *k*-mers in the same locus in HGSVC were removed as well, giving a 86.9% (n=668,961) and 76.8% (n=590,873) ascertainment rate in HGSVC and HPRC. The red dashed lines indicate x=y.

Supplemental Fig. S10. Distribution of the number of outlier samples across all motifs.



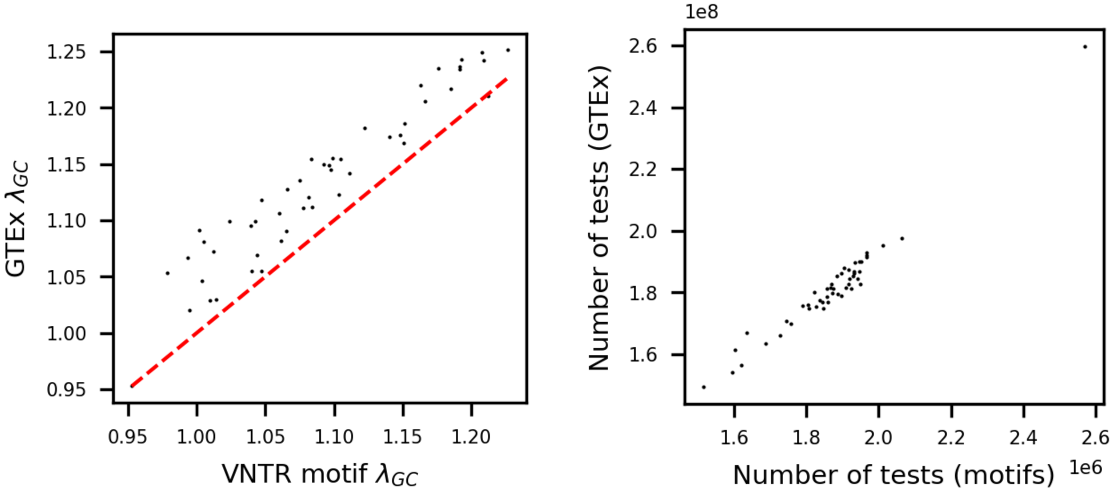
For each motif, samples with motif dosage greater than two standard deviations away from the mean were removed. Data points were collected from the final set of 769,616 motifs and 813 GTEx samples.

Supplemental Fig. S11. Distribution of eMotif batch-r^2.



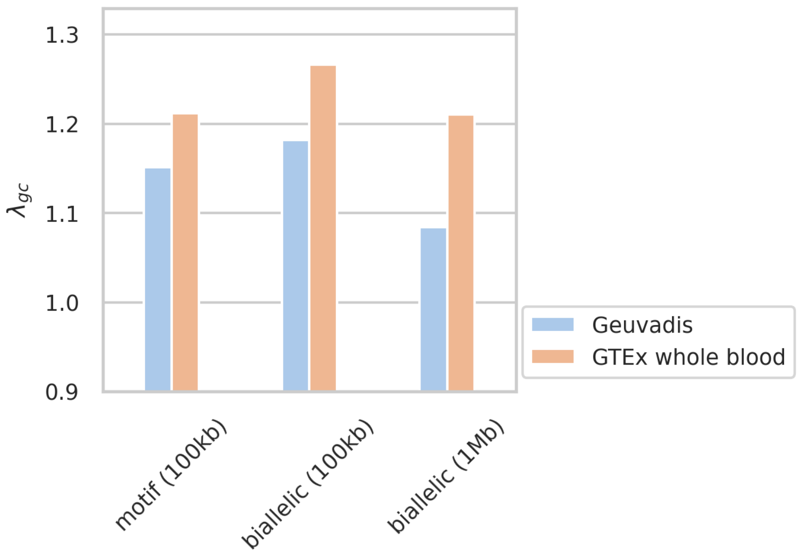
VNTRs without invariant *k*-mers were removed from analysis; motifs in HPRC that do not overlap any *k*-mers in the same locus in HGSVC were removed as well, giving a 95.5% ascertainment rate for 25,031 eMotif. The red dashed lines indicate x=y. Each bin in the marginal histogram spans 0.01.

Supplemental Fig. S12. Distribution of λ_{GC} in GTEx.



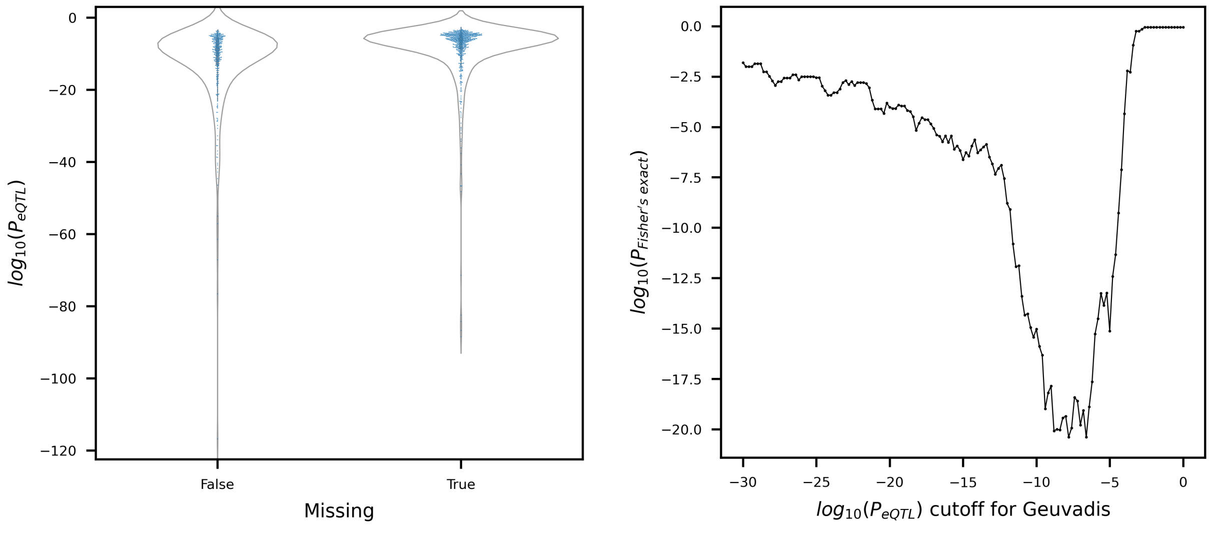
(Left panel) The λ_{GC} from GTEx variants in each tissue is shown on the y-axis, computed using the P-values from all tests originally released by the GTEx Consortium. The λ_{GC} from VNTR motifs in each GTEx tissue is shown on the x-axis. (Right panel) The number of tests in each tissue is compared.

Supplemental Fig. S13. λ_{GC} from biallelic variants versus VNTR motifs.



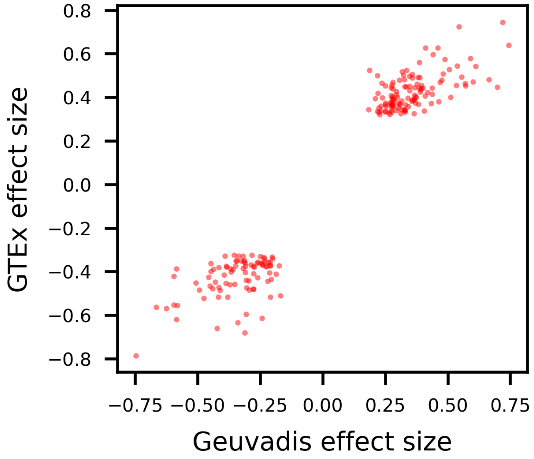
The λ_{GC} computed from different variants (biallelic versus VNTR motifs), different datasets (GTEx versus Geuvadis), and different cis-windows (100 kb versus 1Mb) are shown.

Supplemental Fig. S14. Distribution of eMotif P-values from Geuvadis missing in GTEx.



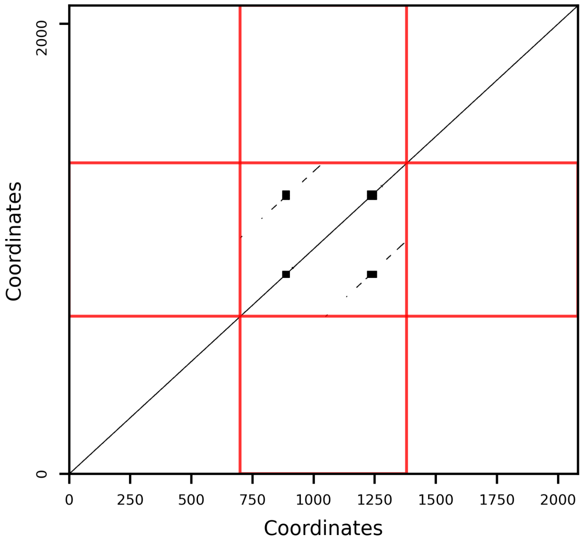
a, Distribution of the eMotif P-values from Geuvadis. The x-axis denotes whether the eMotifs are missing from the GTEx whole blood. **b,** Enrichment of missing eMotifs with lower significance. Contingency tables with two variables, "< P_{eQTL} cutoff" and "is missing", were constructed. Significance was computed with one-sided Fisher's exact test.

Supplemental Fig. S15. Correlation of eMotif effect sizes between Geuvadis and GTEx EBV-transformed lymphocytes.



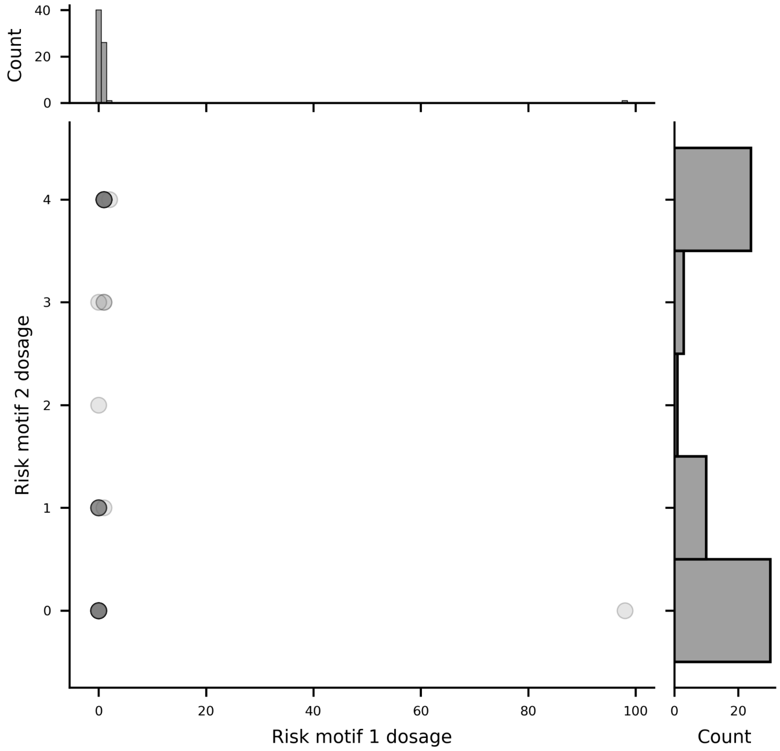
Only eGenes significant in both datasets were shown. Each pair of gene and motif that has the same/opposite sign across datasets were colored in red/black.

Supplemental Fig. S16. Dot plot analysis of AVPR1A RS3 VNTR.



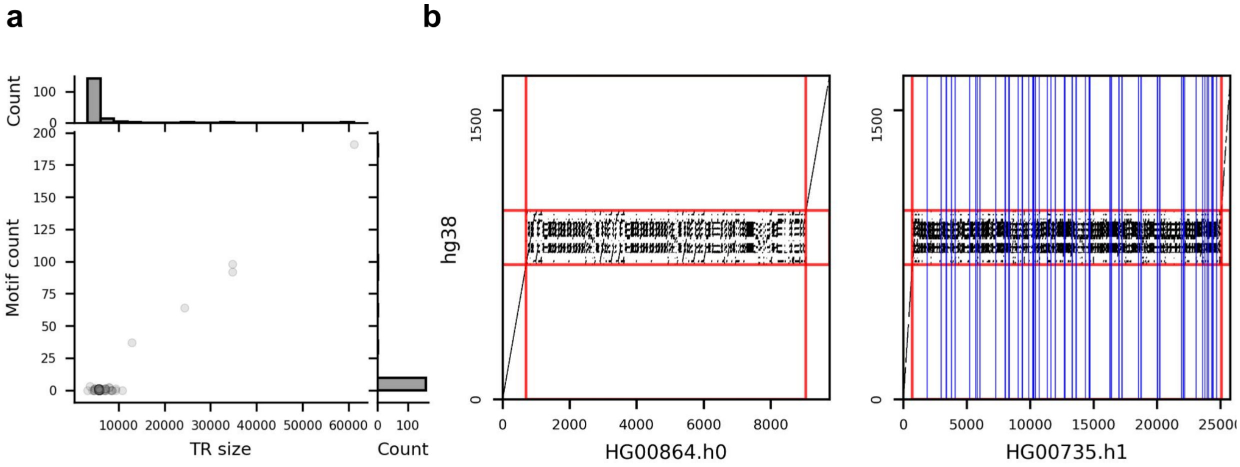
GRCh38 sequence at chr12:63,156,354-63,156,429 was plotted against itself. Each dot represents an exact matching of a 13-mer. The boundaries of 700 bp flanking sequences were indicated with red lines.

Supplemental Fig. S17. Frequency of CACNA1C risk motifs across the 35 HGSVC assemblies.



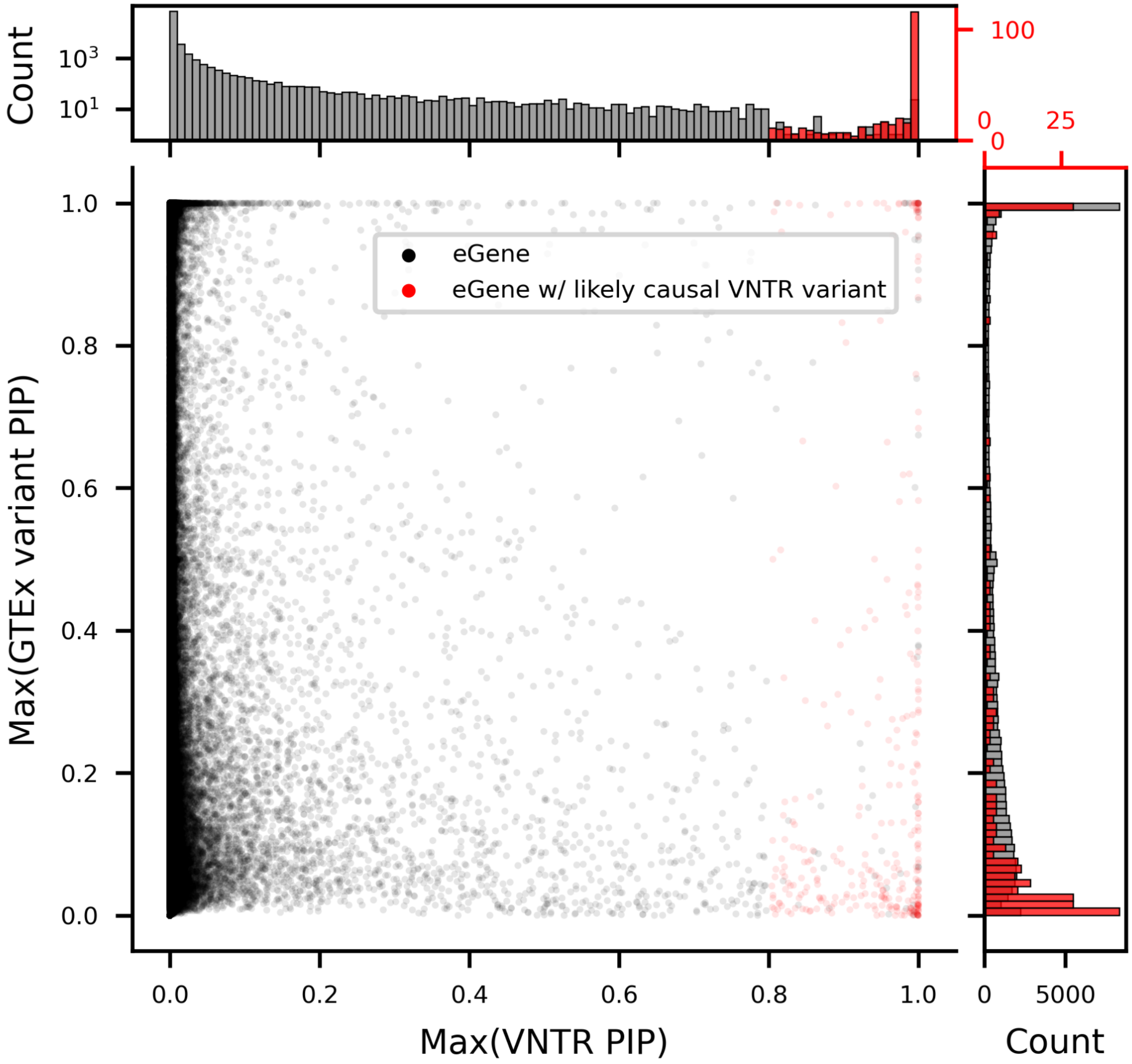
The occurrence of risk motif 1 CAACCACACGATCCTGACCTT and risk motif 2 CCCTGACCTTACTAGTTTACGA were counted considering both the motif and its reverse complement. Bin sizes for both marginal distributions are one.

Supplemental Fig. S18. CACNA1C VNTR allele diversity.



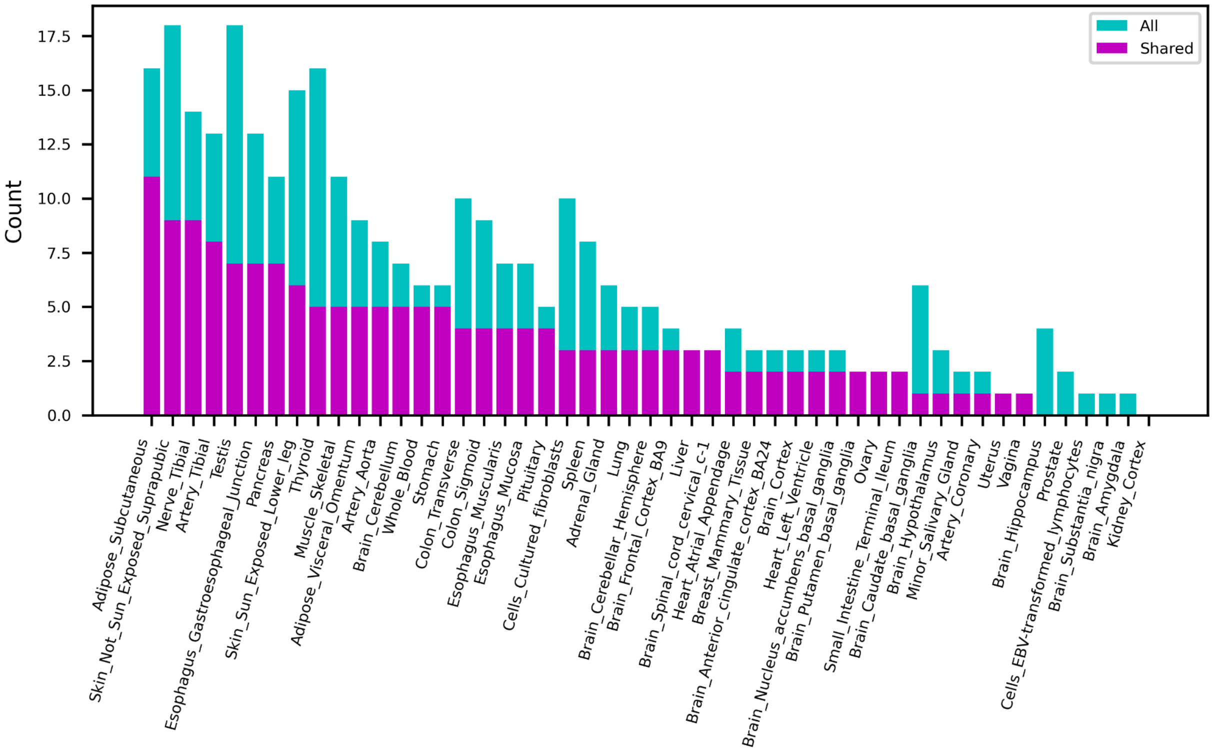
a, The count of risk motif 1 CAACCACACGATCCTGACCTT versus VNTR size. **b**, Two divergent haplotypes were compared with the location of risk motif 1 shown as blue lines. Red lines indicate the boundaries of the VNTR, with a 700 bp flank on each side.

Supplemental Fig. S19. Distribution of posterior inclusion probability (PIP).



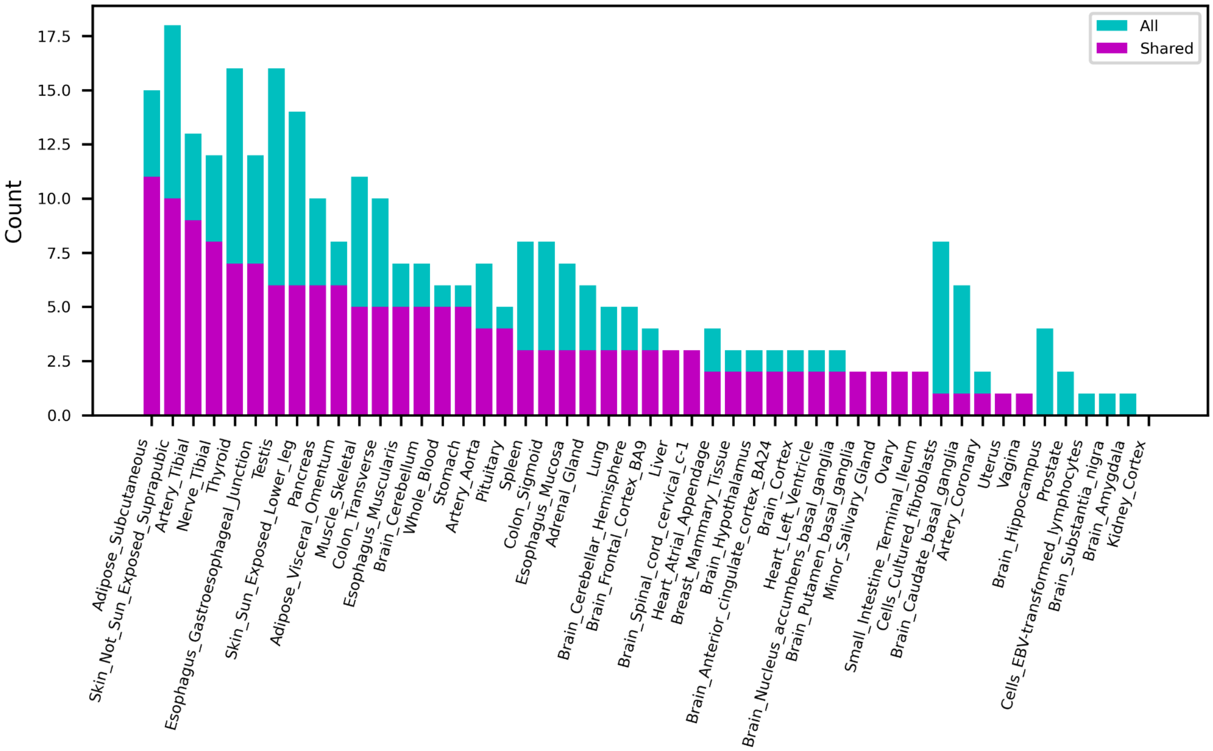
VNTR variants (motif or full VNTR length) with PIP greater than 0.8 were called likely causal. Each dot is from a fine-mapped gene in a tissue. The maximal PIPs of all VNTR/GTEx variants nearby a gene are shown on the x-axis/y-axis. Histograms show the marginal distribution of each variable; the counts of each subset of the data points are shown in histogram in different scales and colors. All counts are shown in linear scale except that the black histogram on top is shown in log scale.

Supplemental Fig. S20. Number of eGenes with likely causal eMotifs.



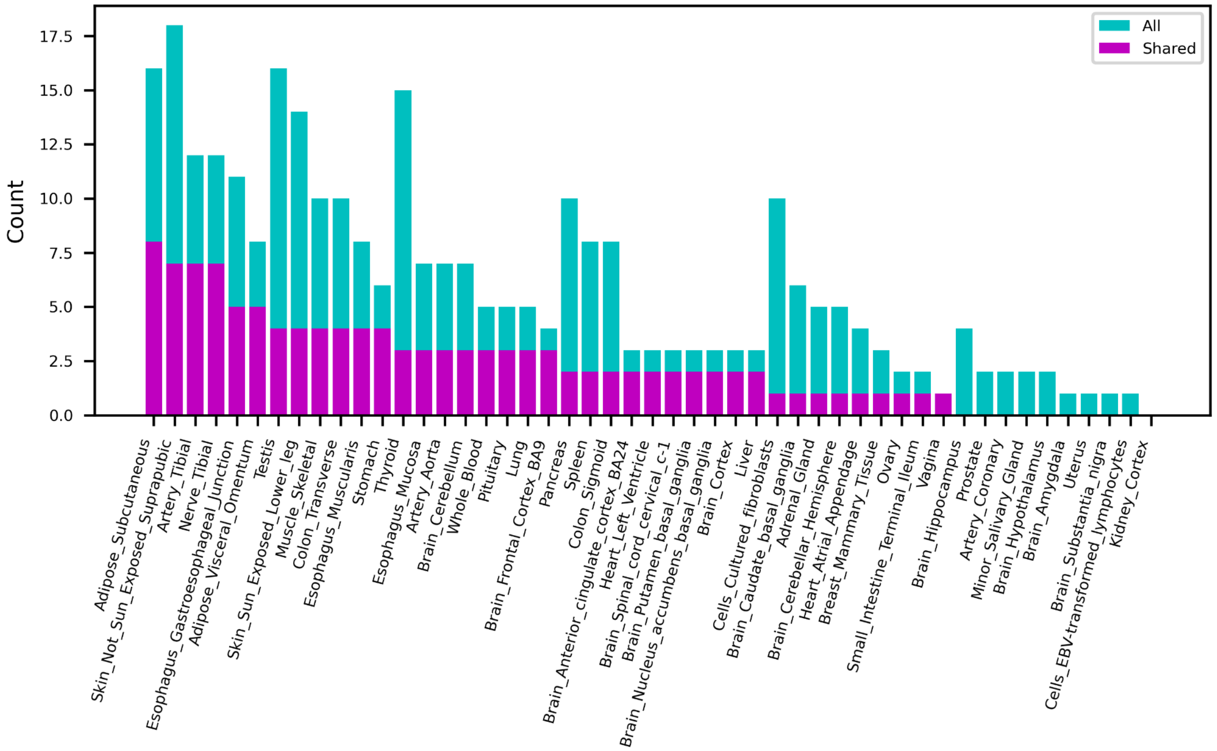
Each eGene reported by eQTL mapping was fine-mapped using all SNPs and eMotifs among the 100 kb window. eGenes of which the highest eMotif posterior inclusion probability (PIP) was greater than 0.8 were reported as containing likely causal eMotifs. An eGene was considered shared across tissues (magenta) if it was reported in at least one other tissue.

Supplemental Fig. S21. Number of eVNTRs with likely causal eMotifs.



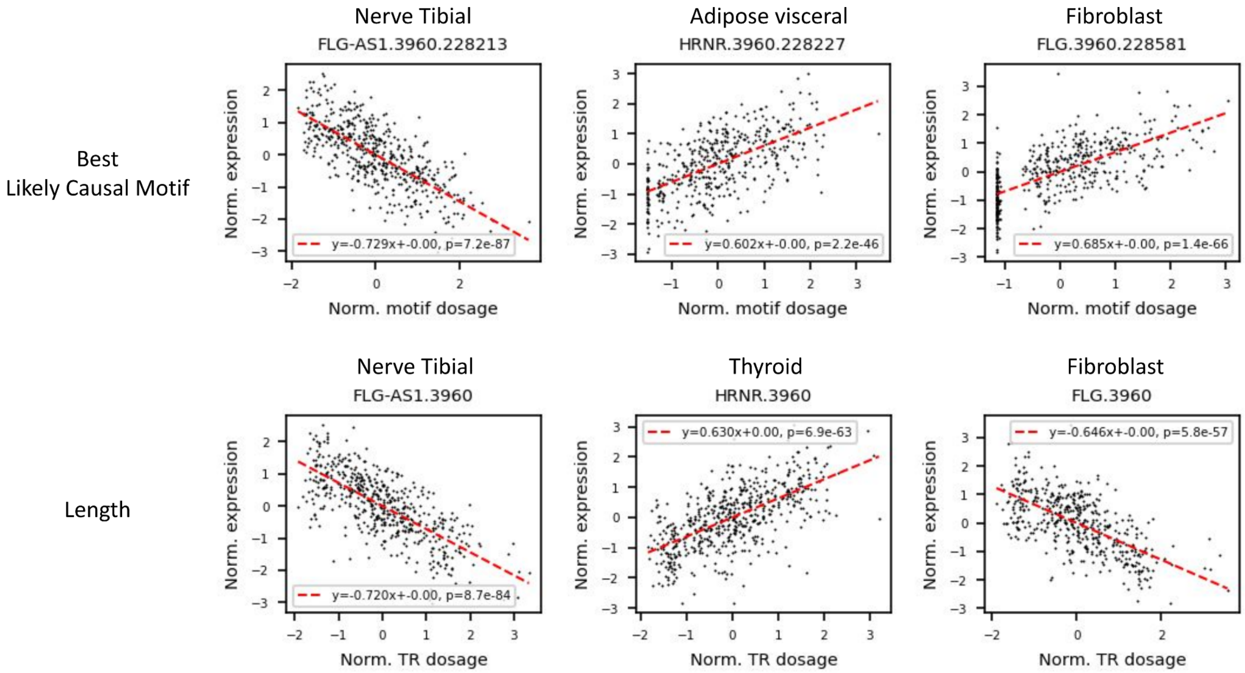
Each eGene reported by eQTL mapping was fine-mapped using all SNPs and eMotifs among the 100 kb window. eGenes of which the highest eMotif posterior inclusion probability (PIP) was greater than 0.8 were reported as containing likely causal eMotifs. An eVNTR was considered shared across tissues (magenta) if any eMotif in the VNTR was reported in at least one other tissue.

Supplemental Fig. S22. Number of likely causal eMotifs.



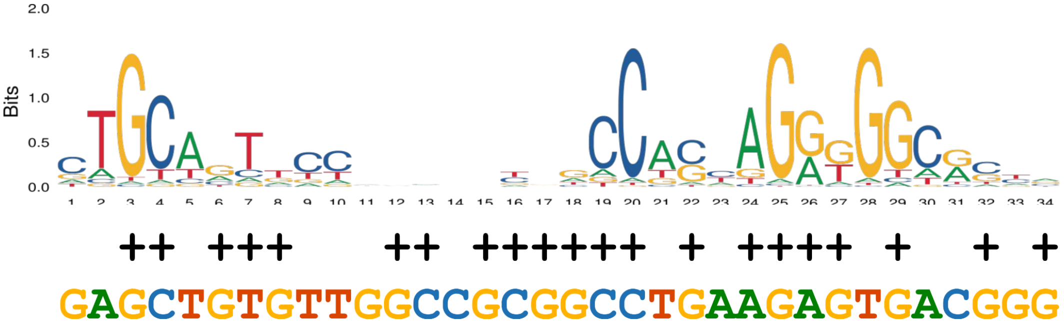
Each eGene reported by eQTL mapping was fine-mapped using all SNPs and eMotifs among the 100 kb window. eGenes of which the highest eMotif posterior inclusion probability (PIP) is greater than 0.8 were reported as containing likely causal eMotifs. An eMotif was considered shared across tissues (magenta) if it was reported in at least one other tissue.

Supplemental Fig. S23. Correlation between HRNR/FLG/FLG-AS1 expression and HRNR repeat.q



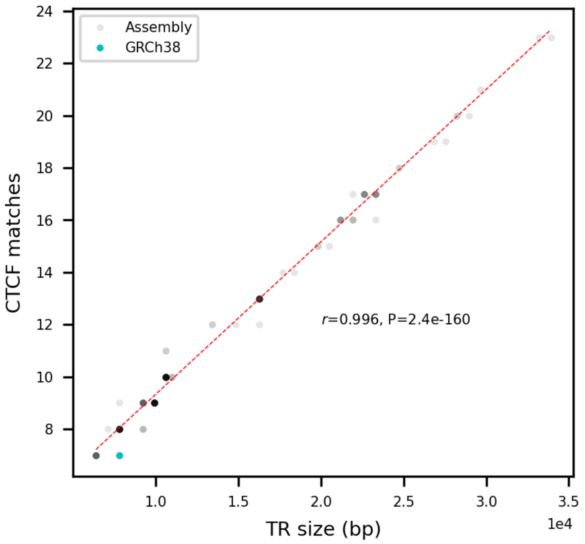
The best likely causal motif from HRNR repeat across all tissues is shown for each gene in the top three panels. The tissue with the most significant length association with each gene is shown on the bottom three panels.

Supplemental Fig. S24. Alignment of CTCF motif to the HRNR repeat.



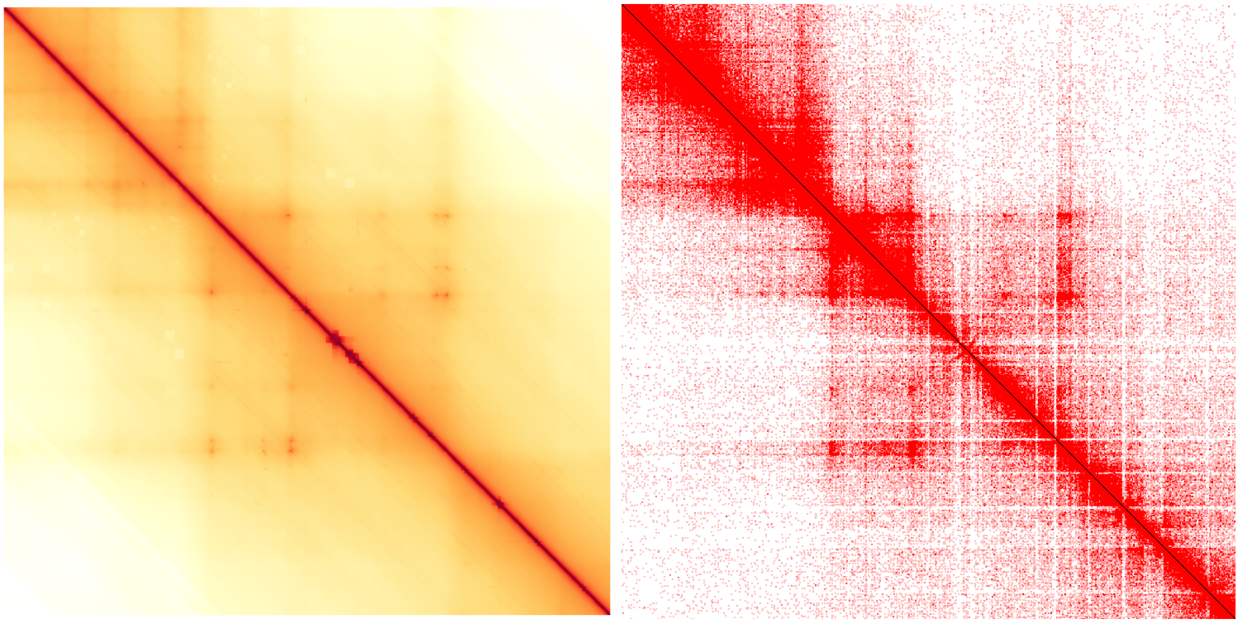
The HRNR repeat on GRCh38 (chr1:152,213,243-152,221,044) was aligned to the 34 bp, two-core CTCF motif (MA1929.1) using the MAST module (Bailey and Gribskov 1998) in the MEME suite. The best alignment at chr1:152,220,381 with a P-value of 4.6e-5 is shown.

Supplemental Fig. S25. Correlation between HRNR repeat size and Number of predicted CTCF sites.



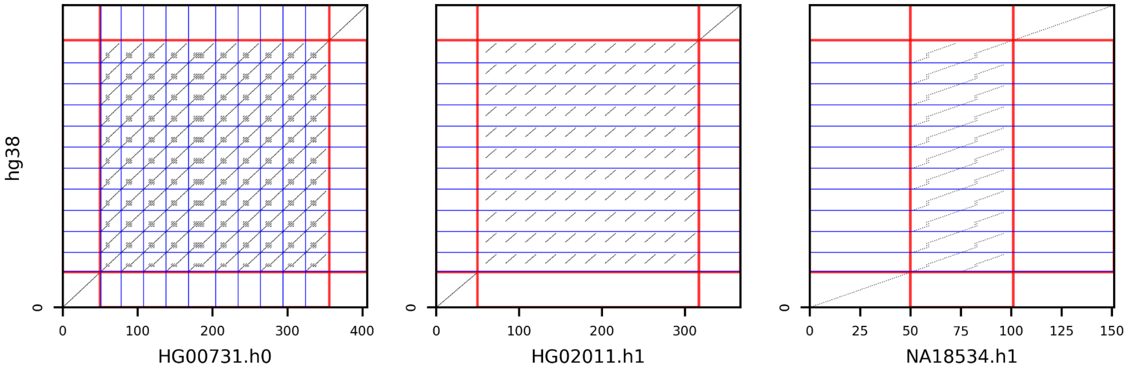
The HRNR repeat size for all HGSVC and HPRC haplotypes are shown on the x-axis. Number of CTCF sites in each haplotype was predicted using FIMO with a cut off of P < 1e-4. Red dashed line indicates the regression line, with the Pearson's *r* and the associated P-value shown.

Supplemental Fig. S26. Akita's prediction versus the micro-C map from the UCSC Genome Browser.



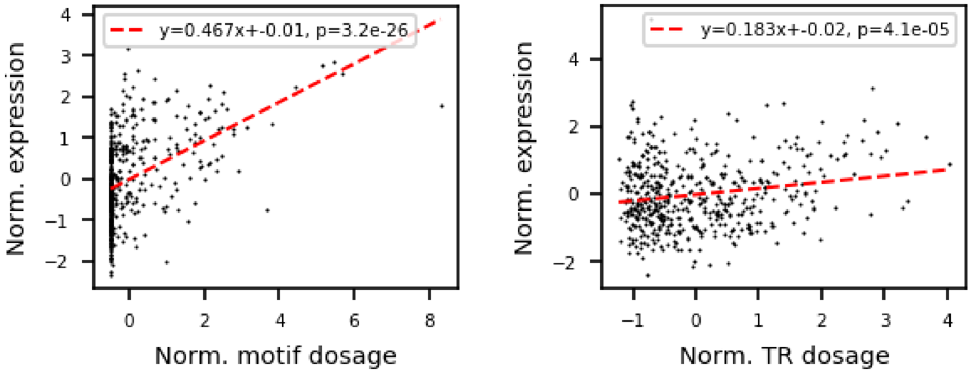
Akita's prediction for chr1:151,693,880-152,740,407 is compared with the HFFc6 micro-C map from UCSC Genome Browser. Akita's initial output "log(observed/expected)" was transformed to an "observed" map to be visually compared with the raw micro-C map. The "expected" map was generated by averaging corrected counts in the region over increasing genomic distances, using cooltools observed_over_expected (Open2C et al. 2022).

Supplemental Fig. S27. Allelic diversity of the *RNF213* promoter VNTR.



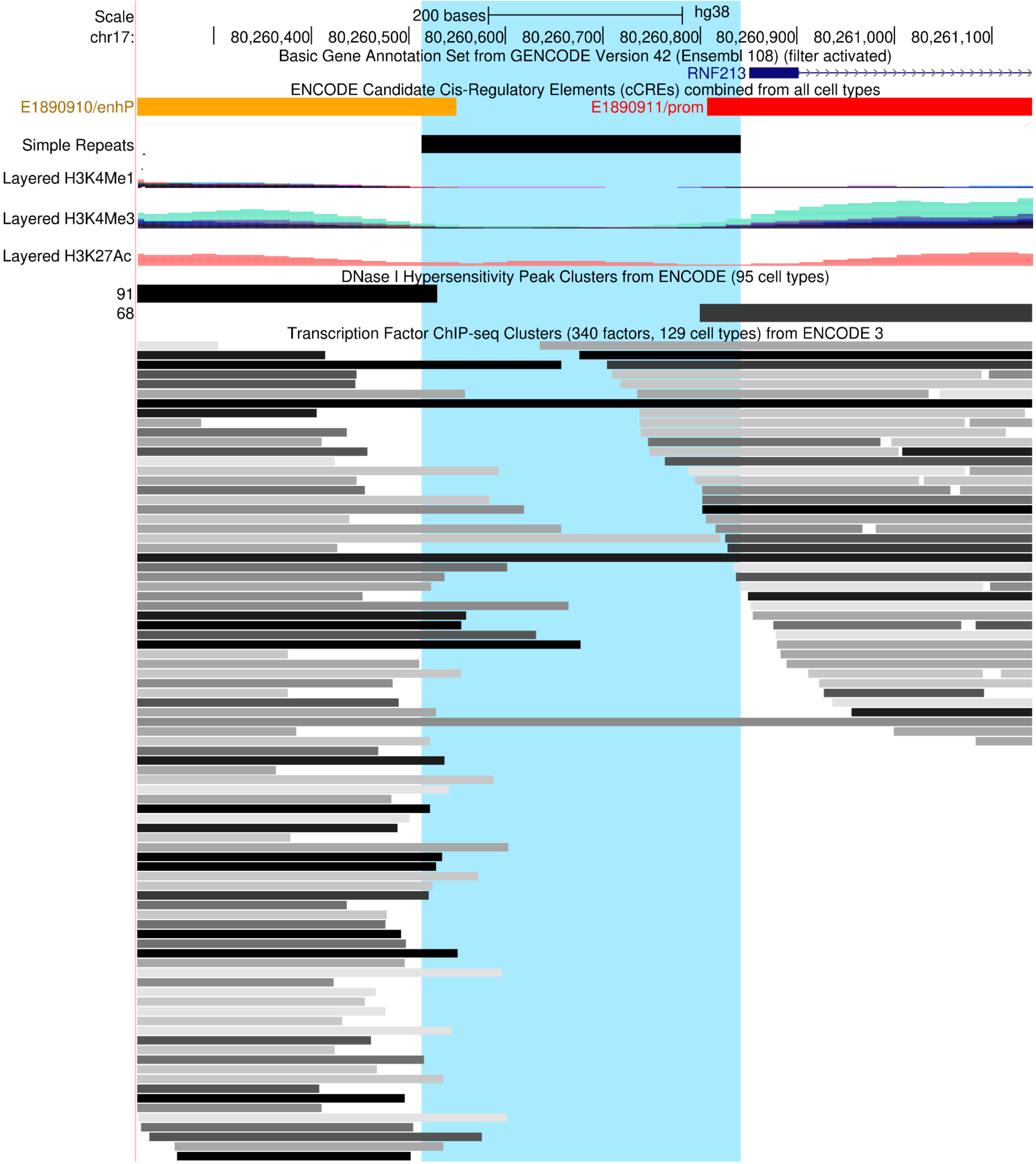
The dot plot of three divergent haplotypes are shown to demonstrate repeat expansion with different motifs. Blue lines represent the red repeat unit in Figure 5. Each dot represents a match of 21-mers. Red lines represent the boundaries of VNTR, with a flank of 50 bp on each side.

Supplemental Fig. S28. Effect of VNTR variants on *RNF213* expression.



(Left panel) Significant association of the likely causal motif GCGGGGCCGGCGGCGGCGGCGG in the promoter VNTR (chr17:80,260,506-80,260,846) with *RNF213* expression. Outliers are not removed to demonstrate the strong effect from extreme motif expansion. The P-value and effect size for the motif after outlier removal were 6.3×10^{-16} and 0.37, respectively. (Right panel) Mild effect from the repeat size of the *RNF213* VNTR. Expression in GTEx thyroid is shown for both panels.

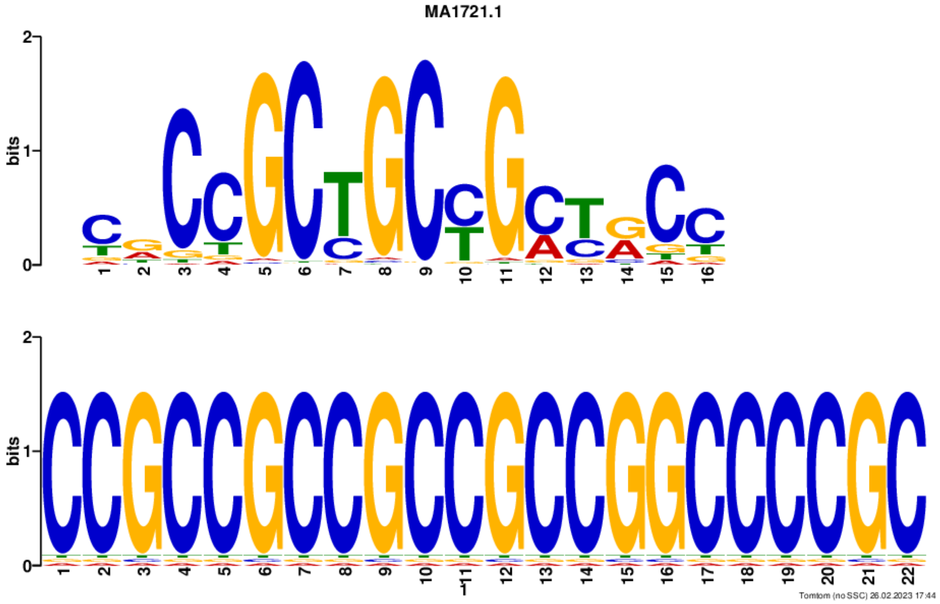
Supplemental Fig. S29. UCSC Genome Browser view of the *RNF213* VNTR.



Scale
200 bases
hg38
chr17:
80,260,400
80,260,500
80,260,600
80,260,700
80,260,800
80,260,900
80,261,000
80,261,100
Basic Gene Annotation Set from GENCODE Version 42 (Ensembl 108) (filter activated)
RNF213
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
E1890910/enhP
E1890911/prom
Simple Repeats
Layered H3K4Me1
Layered H3K4Me3
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
91
68
Transcription Factor ChIP-seq Clusters (340 factors, 129 cell types) from ENCODE 3

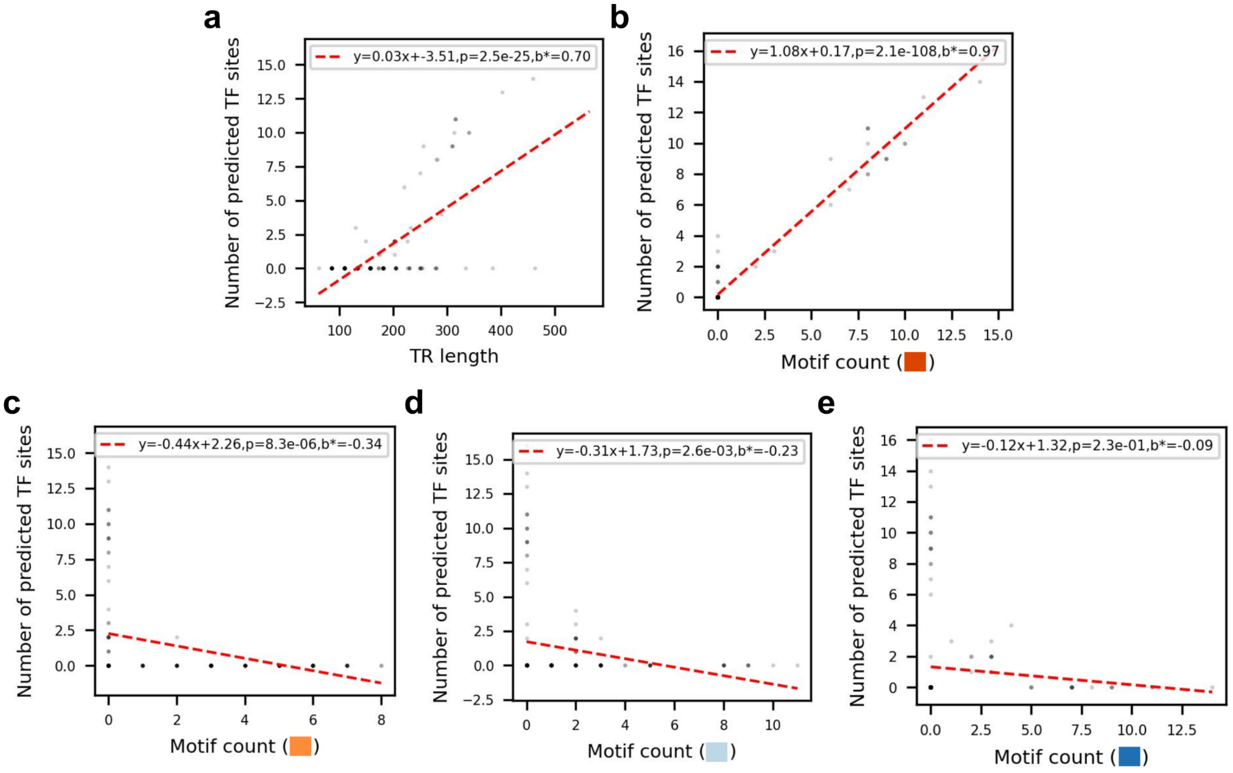
The *RNF* promoter repeat at chr17:80,260,506-80,260,846 was highlighted in light blue.

Supplemental Fig. S30. Best matching transcription factor motif for ZNF93 likely causal motif.



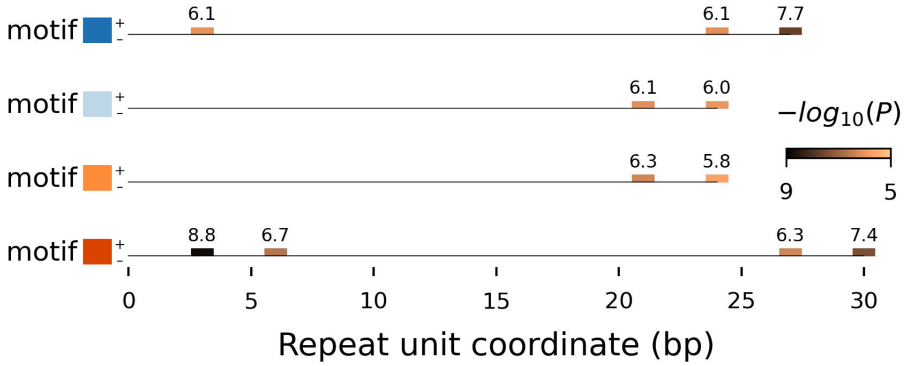
The likely causal motif CCGCCGCCGCCGCCGGCCCCGC (corresponding to the red repeat unit in Fig. 6) was aligned against the vertebrate (in *silico* and in *vivo*) database using TOMTOM with default parameters. The top hit is shown, a ZNF93 binding motif (JASPAR ID: MA1721.1) with p-value = 8.98e-07, E-value = 1.86e-03, and q-value = 3.68e-03.

Supplemental Fig. S31. Correlation between the number of predicted ZNF93 binding sites and *RNF213* VNTR variants.



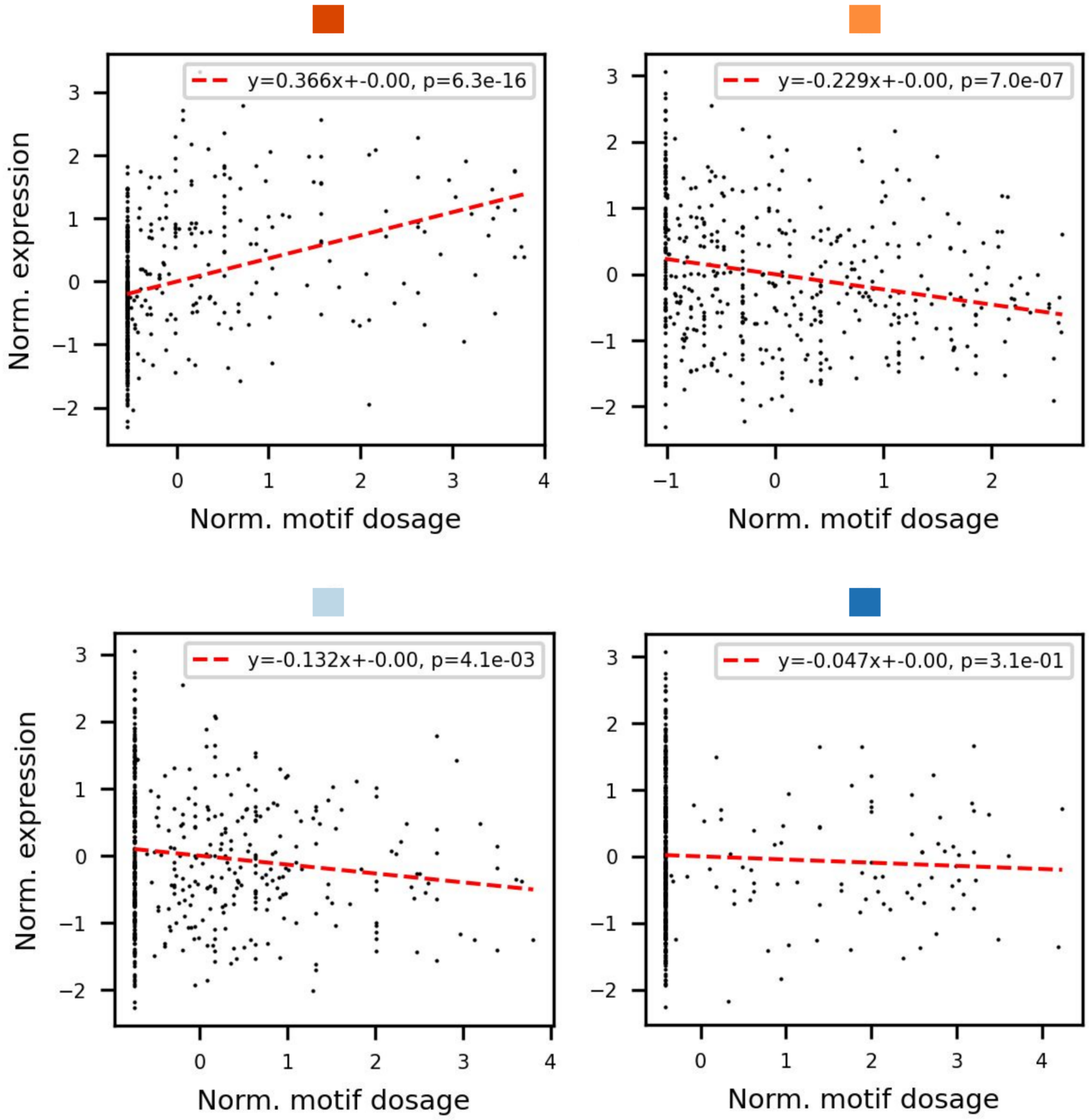
ZNF93 binding sites were predicted using FIMO with the MA1721.1 matrix and a cutoff of P<1e-8. Repeat length (a) and repeat unit counts for each of the efficient motifs annotated by vamos (b-e) are shown . Data points are collected from 35 HGSVC and 48 HPRC assemblies.

Supplemental Fig. S32. Predicted ZNF93 binding sites for each VNTR motif.



ZNF93 binding sites (MA1721.1) across the four efficient motifs annotated by vamos were predicted using FIMO with a cutoff of P < 10^{-5}. To detect binding sites that span the boundary of repeat units, two-copy units were taken as inputs. Only one repeat unit is shown such that periodic binding sites are removed. The significance ($–log_{10}P$) of the match is shown above each binding site.

Supplemental Fig. S33. Correlation between *RNF213* expression level and *RNF213* VNTR repeat units.



Each *RNF213* VNTR motif in the locus-RPGG that has a unique match with a two-copy unit (among the four repeat units annotated by vamos) was assigned as representative of the repeat unit. The VNTR motif with the lowest P-value in explaining gene expression in thyroid is shown for each repeat unit.

Supplemental Tables

Supplemental Table S1. Additional disease-relevant tandem repeats included in our reference set.

Gene	Associated/Causal TR Region	Disease	Note
ACAN	chr15:88855423-88857301	OSTD	coding
AR	chrX:67545317-67545419	SBMA	coding
ATN1	chr12:6936717-6936775	DRPLA	coding
ATXN1	chr6:16327634-16327724	SCA1	coding
ATXN2	chr12:111598950-111599019	SCA2	coding
ATXN3	chr14:92071011-92071052	SCA3	coding
ATXN7	chr3:63912685-63912716	SCA7	coding
ATXN8	chr13:70,139,352-70,139,429	SCA8	noncoding
AVPR1A	chr12:63149772-63149849	EXTB	intron_AVR
AVPR1A	chr12:63153304-63153366	EXTB	5UTR_RS1
AVPR1A	chr12:63156354-63156429	EXTB	5UTR_RS3
C9orf72	chr9:27573485-27573546	ALS	intron 1
CACNA1A	chr19:13207859-13207898	SCA6	coding
CACNA1C	chr12:2255791-2256090	bipolar, SCZ	intron 3
CEL	chr9:133071170-133071694	Monogenic diabetes	coding
CNBP	chr3:129172568-129172736	DM2	intron 1
CSTB	chr21:43776293-43776322	PMP1A	5' UTR
DMPK	chr19:45770205-45770266	DM1	3' UTR
DRD4	chr11:639989-640194	OCD, ADHD	coding
FMR1	chrX:147912037-147912111	FXS, FXTAS	5' UTR
FXN	chr9:69087917-69087957	FA	noncoding
GP1BA	chr17:4933823-4933963	ATF in stroke	coding
HIC1	chr17:2052378-2053079	MCC	promoter
HTT	chr4:3074877-3074940	HD	coding
IL1RN	chr2:113130529-113130872	stroke, CAD	intron

INS	chr11:2161570-2161976	T1D, T2D, Obesity	promoter
JPH3	chr16:87604283-87604329	HDL2	intron
MAOA	chrX:43655101-43655205	bipolar	promoter
MMP9	chr20:46013363-46013426	Kawasaki	coding
MUC1	chr1:155188676-155192051	MCKD1	coding
MUC21	chr6:30986289-30987497	DPB	coding
NACA	chr12:56717496-56718739	atrial fibrillation	coding
PER3	chr1:7829888-7830126	bipolar	coding
PPP2R2B	chr5:146878728-146878759	SCA12	noncoding
SLC6A3	chr5:1393582-1393985	ADHD, PK	3' UTR
SLC6A3	chr5:1414387-1414518	ADHD, PK	intron 8
SLC6A4	chr17:30221385-30221590	BPSD, AZ	intron
SLC6A4	chr17:30237128-30237450	OCD, anxiety, SCZ	promoter
TBP	chr6:170561907-170562017	SCA17	coding
TCHH	chr1:152111363-152112285	Male pattern baldness score	coding

ALS, amyotrophic lateral sclerosis. DM, myotonic dystrophy. DRPLA, dentatorubral-pallidoluysian atrophy. FTD, frontotemporal dementia. FXS, fragile X syndrome. FXTAS, fragile X tremor-ataxia syndrome. HD, Huntington disease. HDL2, Huntington disease-like 2. SBMA, spinobulbar muscular atrophy. SCA, spinocerebellar ataxia. SCZ, schizophrenia. FA, Friedreich ataxia. DPB, respiratory bronchioles, causing a progressive suppurative and severe obstructive respiratory disorder . AZ, Alzheimer’s disease. MCC, metastatic colorectal cancer. PMP1A, progressive myoclonic epilepsy 1A. PK, Parkinson’s disease. OSTD, osteochondritis dissecans. EXTB, externalizing behaviors. CAD, coronary artery disease.

Supplemental Table S2. eMotif discoveries for disease-relevant genes.

Gene	eGene			eGene-eTR pair			
	Rep.[1]	# of tissues	# of eTRs	Rep[2]	Tissue	Disease match[3]	eMotif (positive/negative effect)[4]
PER3	v	3	1	v	Cells_Cultured_fibroblasts Whole_Blood		
TCHH	v	11	1				
MUC1	v	6	1	v	Artery_Tibial Cells_Cultured_fibroblasts Esophagus_Mucosa Whole_Blood	?	-0.17:CCACGGTGTCACCTCGGCCCC -0.17:ACACCAGGCCGGCCCCGGGCT -0.17:GCCCACGGTGTCACCTCGGCC -0.16:CCCACGGTGTCACCTCGGCCC -0.16:CAGCCCACGGTGTCACCTCGG -0.16:CACGGTGTCACCTCGGCCCCG -0.16:CACCAGGCCGGCCCCGGGCTCCA -0.16:CCGGCCTGGTGTCCGGGGCCG -0.16:AGCCCACGGTGTCACCTCGGC -0.16:ACCTCGGCCCCGGACACCAGG -0.16:GACACCAGGCCGGCCCCGGGC -0.16:GTGTCCGGGGCCGAGGTGACA -0.16:CCGGGGCCGGCCTGGTGTCCG -0.16:CTGGTGTCCGGGGCCGAGGTGAC -0.16:CCCGGGGCCGGCCTGGTGTCC -0.16:GCCCCGGACACCAGGCCGGCC -0.16:ACGGTGTCACCTCGGCCCCGG -0.16:GTGTCACCTCGGCCCCGGACA -0.16:CGGTGTCACCTCGGCCCCGGA -0.15:CGGCCTGGTGTCCGGGGCCGA -0.15:GGTGTCACCTCGGCCCCGGAC -0.15:CCTCGGCCCCGGACACCAGGC -0.15:CCGGACACCAGGCCGGCCCCG -0.15:CCAGCCCACGGTGTCACCTCG -0.15:CGGCCCCGGGCTCCACCGCCC -0.15:CCCGGACACCAGGCCGGCCCC -0.15:CTCGGCCCCGGACACCAGGCC -0.15:AGGCCGGCCCCGGGCTCCACC -0.15:CCGGCCCCGGGCTCCACCGCC -0.15:GCCGGCCTGGTGTCCGGGGCC -0.15:GCCGGCCCCGGGCTCCACCGC -0.15:CAGGCCGGCCCCGGGCTCCAC -0.15:CCCAGCCCACGGTGTCACCTC -0.14:AGGTGACACCGTGGGCTGGGG -0.14:CCCCCAGCCCACGGTGTCACC -0.14:CCCCGGACACCAGGCCGGCCC -0.14:CGGTGGAGCCCGGGGCCGGCC -0.14:CTCCACCGCCCCCCCAGCCCA
DRD4	v	28	1	v	Colon_Transverse Esophagus_Mucosa Esophagus_Muscularis Nerve_Tibial Skin_Not_Sun_Exposed_Suprapubic Testis		
INS	v	7	1				
CACNA1C	v	16	1	v	Brain_Cerebellar_Hemisphere Brain_Cerebellum Cells_Cultured_fibroblasts Colon_Transverse Esophagus_Muscularis	v	-0.45:AAGGTCAGGATCGTGTGGTTG -0.41:CCCTGACCTTACTAGTTTACGA -0.18:ATCACACGATCCTGACCTGAC -0.18:AATCACACGATCCTGACCTGA -0.17:CACACGATCCTGACCTGACTA -0.17:ACACGATCCTGACCTGACTAG -0.17:AAACTAGTCAGGTCAGGATCGTG -0.17:TAGTCAGGTCAGGGTCGTGTAATTGTAAACTAGTCAGGTCA -0.17:ATCCTGACCTGACTAGTTTAC -0.17:CAACCACACGATCCTGACCTGA -0.17:CTGACCTTACTAGTTTACAACCACACGA -0.17:GATCCTGACCTGACTAGTTTA -0.17:CAATCACACGATCCTGACCTG -0.17:ACCACACGATCCTGACCTGAC -0.16:TCCTGACCTGACTAGTTTACA -0.16:CTGACTAGTTTACAATCACAC -0.16:ACCTGACTAGTTTACAATCAC -0.16:ATTGTAAACTAGTCAGGTCAG -0.16:GATTGTAAACTAGTCAGGTCA -0.16:CCTGACTAGTTTACAATCACA

							-0.16:GACCTGACTAGTTTACAATCA -0.16:AGTCAGGTCAGGATCGTGTGG -0.16:AGTCAGGTCAGGATCGTGTGA -0.16:CGTGTGGTTGTAAACTAGTCA -0.16:GACTAGTTTACAACCACACGA -0.16:CAGGTCAGGATCGTGTGATTCTAAAC -0.15:AAGGTCAGGATCGTGTGATTTTAAACTAGCAAGGTCAGGAT -0.15:GACCTGACTAGTTTACAACCA -0.15:GGTTGTAAACTAGTCAGGTCA -0.15:ACCTGACTAGTTTACAACCAC -0.15:CTGACCTGACTAGTTTACAAC -0.15:ACCACACGACCCTGACCTGACT -0.15:ACTAGTTTACAATCACAGGAT +0.16:AGGATCGTGTGGTTGTAAACCAG +0.18:AAACTAGCAAGGTCAGGGTCTTGTGATTG +0.19:CAAGGTCAGGATCGTGTGGTCG +0.21:CTTGCTAGTTTACAATCACAAGAC +0.24:ACCCTGACTTGACTAGTTTACAATCACAC
ATN1	v	33	1	v	Adipose_Subcutaneous Adipose_Visceral_Omentum Artery_Aorta Artery_Coronary Artery_Tibial Brain_Cerebellar_Hemisphere Brain_Cerebellum Brain_Cortex Breast_Mammary_Tissue Cells_Cultured_fibroblasts Colon_Sigmoid Colon_Transverse Esophagus_Gastroesophageal_Junction Esophagus_Mucosa Esophagus_Muscularis Heart_Atrial_Appendage Heart_Left_Ventricle Liver Lung Muscle_Skeletal Nerve_Tibial Ovary Pancreas Pituitary Prostate Skin_Not_Sun_Exposed_Suprapubic Skin_Sun_Exposed_Lower_leg Small_Intestine_Terminal_Ileum Spleen Stomach Testis Thyroid Vagina	v	+0.59:CAGCAGCAGCAGCAGCAGCAGCA
NACA	v	1	1				
AVPR1A	v	5	3	v	Adipose_Subcutaneous Cells_Cultured_fibroblasts		
				v	Artery_Tibial		
				v	Artery_Coronary Artery_Tibial		
ATXN2	v	1	1				
ATXN8		0	0				
ATXN3		0	0				
ACAN	v	3	1				
JPH3	v	9	1	v	Testis		
HIC1	v	2	1				
GP1BA	v	17	1				
SLC6A4	v	2	2	v	Esophagus_Muscularis Nerve_Tibial		
SLC6A4	v	2	2				

CACNA1A	v	11	1				
DMPK		0	0				
IL1RN	v	9	1	v	Breast_Mammary_Tissue Cells_Cultured_fibroblasts Skin_Not_Sun_Exposed_Suprapubic Skin_Sun_Exposed_Lower_leg		
MMP9	v	2	1				
CSTB	v	39	1				
ATXN7	v	9	1	v	Brain_Cerebellar_Hemisphere Esophagus_Muscularis Testis	v	-0.33:GCAGCAGCAGCAGCAGCAGCAGC
CNBP		0	0				
HTT	v	13	1	v	Muscle_Skeletal		
SLC6A3	v	17	2				
PPP2R2B	v	12	1				
ATXN1	v	2	1				
MUC21	v	5	1	v	Esophagus_Mucosa Lung Nerve_Tibial	v	-0.21:CACAGCCACCAACTCTGACTCCAGCACA -0.20:ACCTCCAGTGGGGCTAGCACAGCCACCAACTCTGA -0.20:CAGCACAACCTCCAGTGGGGTCAGCAC -0.20:ACTGCCACCAACTCTGAGTCTAG -0.20:CCAACTCTGAGTCTAGCACACTCTCCAGTGGGGCCAGCACA -0.19:GACTCAGAGTTGGTGGCAGTGCTGGC -0.19:GCCACCAACTCTGAGTCTAGCACA -0.19:AGCACAACCTCCAGTGGGGCTAGCAC -0.19:ACAGCCACCAACTCTGAGTCTAG -0.19:GGGGCCAGCACTGCCACCAAC -0.19:ATCCCACTGGACACTGTGCTAGACTCAGAGTTGG -0.18:ACTGCCACCAACTCTGAGTCCAG -0.18:AGCACAGCCACCAACTCTGACT -0.18:GAGTTGGTGGCAGTGCTGGCCC -0.17:CTGTGCTGGCCCCACTGGAGA -0.15:ACCTCCAGTGGGGCCGGCACA +0.15:ACCCAGAGTTGGTGGCTGTGCTG +0.16:AGGTCGTGCTGGACTCAGAGGTGGTGGCTGTGCTGGCCCCA +0.16:CAGAGTTGGTGACTGTGCTGGCCCCACTGGAG +0.17:CACAACCTCCAGTGGGGCCAGCAC
TBP	v	8	1				
C9orf72	v	35	1	v	Adipose_Subcutaneous Adipose_Visceral_Omentum Artery_Aorta Artery_Tibial Brain_Anterior_cingulate_cortex_BA24 Brain_Caudate_basal_ganglia Brain_Cerebellar_Hemisphere Brain_Cerebellum Brain_Cortex Brain_Frontal_Cortex_BA9 Brain_Nucleus_accumbens_basal_ganglia Brain_Putamen_basal_ganglia Breast_Mammary_Tissue Colon_Sigmoid Colon_Transverse Esophagus_Gastroesophageal_Junction Esophagus_Mucosa Esophagus_Muscularis Heart_Atrial_Appendage Heart_Left_Ventricle Lung Muscle_Skeletal Nerve_Tibial Pancreas Pituitary Prostate Skin_Not_Sun_Exposed_Suprapubic Skin_Sun_Exposed_Lower_leg Small_Intestine_Terminal_Ileum	v	-0.22:CCGCCCCGGGCCCGCCCCCGGGCCCGCCCCGACCACGCCCCGGCCCCGGCCCC +0.45:CCGGCCCCGGCCCCGGCCCCGGC +0.46:CCCCGGCCCCGGCCCCGGCCCC +0.46:CACGCCCCGGCCCCGGCCCCGGC +0.46:GCCCCGGCCCCGGCCCCGGCC

					Spleen Stomach Testis Thyroid Whole_Blood		
FXN	v	6	1				
CEL	v	5	1	v	Brain_Hypothalamus Pancreas	v	+0.20:CCCGGAGTCACCCGTGGGCGG +0.29:ACTCCGGGGCCCCCCCTGTGACCCCCACGGGTGACTC
MAOA	v	4	1	v	Muscle_Skeletal Thyroid Whole_Blood	?	-0.23:CCAGTACCGGCACCGGCACCAGTAC -0.18:CCGGCACCGGCACCAGTACCC -0.18:ACCGGCACCGGCACCAGTACC -0.17:ACTGGTGCGGGTACTGGTGCCGGTGCCG -0.17:CCCGCACCAGTACCGGCACCGGCACC -0.16:GCACCAGTACCCGCACCAGTACC -0.15:CCAGTACCCGCACCAGTACCGGCACC
AR	v	2	1				
FMR1	v	2	1				

[1] eGene replicated in eMotif discoveries using genome-wide P-value cutoffs.

[2] eGene-eTR pair replicated in eMotif discoveries using genome-wide P-value cutoffs. For eGene with multiple disease-relevant repeats, the order shown here is the same as the one listed in Table S1.

[3] The eGene-eTR tissue matches disease ontology. The ”?” denotes cases where the exact disease mechanism is unclear.

[4] Only eMotifs from eGene-eTR pairs matching disease ontology are shown. The largest effect size for each unique motif is shown.

Supplemental Table S3. eQTL mapping using a genome-wide P-value cutoff.

Tissue	P-value cutoff*
Kidney_Cortex	2.86E-05
Brain_Substantia_nigra	6.64E-05
Vagina	9.62E-05
Brain_Amygdala	9.64E-05
Uterus	1.06E-04
Cells_EBV-transformed_lymphocytes	1.25E-04
Brain_Spinal_cord_cervical_c-1	1.36E-04
Minor_Salivary_Gland	1.57E-04
Brain_Hippocampus	1.83E-04
Brain_Hypothalamus	1.99E-04
Brain_Anterior_cingulate_cortex_BA24	2.03E-04
Small_Intestine_Terminal_Ileum	2.05E-04
Ovary	2.19E-04

Brain_Putamen_basal_ganglia	2.60E-04
Artery_Coronary	2.79E-04
Liver	2.80E-04
Brain_Frontal_Cortex_BA9	2.84E-04
Prostate	3.37E-04
Brain_Nucleus_accumbens_basal_ganglia	3.44E-04
Brain_Caudate_basal_ganglia	3.46E-04
Adrenal_Gland	4.10E-04
Brain_Cerebellar_Hemisphere	4.46E-04
Brain_Cortex	4.51E-04
Pituitary	4.87E-04
Spleen	5.20E-04
Stomach	5.29E-04
Pancreas	6.43E-04
Brain_Cerebellum	6.49E-04
Colon_Transverse	6.78E-04
Colon_Sigmoid	6.79E-04
Breast_Mammary_Tissue	7.13E-04
Esophagus_Gastroesophageal_Junction	7.34E-04
Heart_Left_Ventricle	7.74E-04
Heart_Atrial_Appendage	8.11E-04
Artery_Aorta	9.59E-04
Testis	9.69E-04
Adipose_Visceral_Omentum	9.74E-04
Lung	1.05E-03
Skin_Not_Sun_Exposed_Suprapubic	1.23E-03
Esophagus_Mucosa	1.25E-03
Esophagus_Muscularis	1.26E-03
Adipose_Subcutaneous	1.34E-03
Artery_Tibial	1.44E-03
Cells_Cultured_fibroblasts	1.46E-03

Muscle_Skeletal	1.47E-03
Skin_Sun_Exposed_Lower_leg	1.48E-03
Nerve_Tibial	1.58E-03
Whole_Blood	1.59E-03
Thyroid	1.71E-03

* False discovery rate controlled using the Benjamini–Hochberg procedure taking all P-values from all tests as input. The genome-wide P-value cutoff was used to allow for more than one eMotif to be reported as significantly associated with a gene.

References

Bailey TL, Gribskov M. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**: 48–54.

Open2C, Abdennur N, Abraham S, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko A, Imakaev M, Oksuz BA, Venev SV. 2022. Cooltools: enabling high-resolution Hi-C analysis in Python. *bioRxiv* 2022.10.31.514564. https://www.biorxiv.org/content/10.1101/2022.10.31.514564v1.abstract (Accessed February 28, 2023).