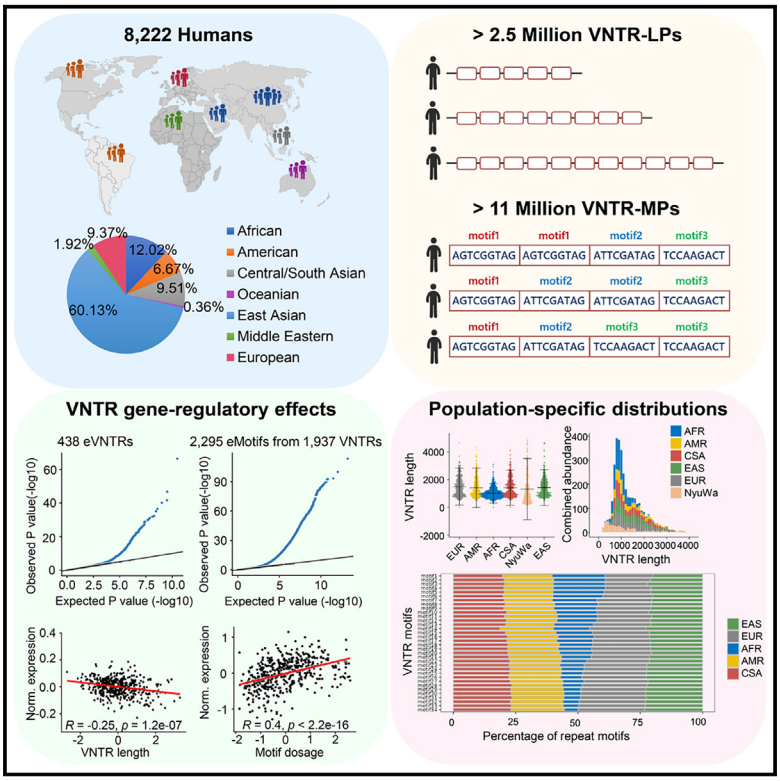
Cell Genomics

Article

Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits

Graphical abstract



Highlights

- Systematic study of VNTR polymorphisms in 8,222 high-coverage WGS genomes
- Identification of 2.5 M VNTR length polymorphisms and 11 M VNTR motif polymorphisms
- Identification of 438 eVNTRs and 2,295 eMotifs associated with gene expression
- Impact of VNTR polymorphisms on phenotypic traits and disease susceptibility

Authors

Sijia Zhang, Qiao Song, Peng Zhang, ..., Tingrui Song, Tao Xu, Shunmin He

Correspondence

xutao@ibp.ac.cn (T.X.),
heshunmin@ibp.ac.cn (S.H.)

In brief

Zhang et al. constructed a comprehensive genome-wide map of VNTR polymorphisms both in length and repeat composition across 8,222 high-coverage genomes, with over 2.5 M VNTR length and 11 M VNTR motif polymorphisms identified. This study will expand our knowledge of VNTR polymorphisms and their functional implications in human genetics.

Zhang et al., 2024, Cell Genomics *4*, 100699
December 11, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2024.100699

CellPress

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

Cell Genomics

CellPress
OPEN ACCESS

Article

Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits

Sijia Zhang,[1,2,5,6] Qiao Song,[1,6] Peng Zhang,[1] Xiaona Wang,[1] Rong Guo,[1] Yanyan Li,[1] Shuai Liu,[1] Xiaoyu Yan,[1] Jingjing Zhang,[1] Yiwei Niu,[1] Yirong Shi,[1] Tingrui Song,[1] Tao Xu,[2,3,4,*] and Shunmin He[1,2,7,*]

[1]Key Laboratory of Epigenetic Regulation and Intervention, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[3]National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
[4]Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan, Shandong 250117, China
[5]Department of Scientific Research, Jiangsu Cancer Hospital & Jiangsu Institute of Cancer Research & Affiliated Cancer Hospital of Nanjing Medical University, Nanjing, China
[6]These authors contributed equally
[7]Lead contact
*Correspondence: xutao@ibp.ac.cn (T.X.), heshunmin@ibp.ac.cn (S.H.)
https://doi.org/10.1016/j.xgen.2024.100699

SUMMARY

Variable number tandem repeat (VNTR) is a pervasive and highly mutable genetic feature that varies in both length and repeat sequence. Despite the well-studied copy-number variants, the functional impacts of repeat motif polymorphisms remain unknown. Here, we present the largest genome-wide VNTR polymorphism map to date, with over 2.5 million VNTR length polymorphisms (VNTR-LPs) and over 11 million VNTR motif polymorphisms (VNTR-MPs) detected in 8,222 high-coverage genomes. Leveraging the large-scale NyuWa cohort, we identified 2,982,456 (31.8%) NyuWa-specific VNTR-MPs, of which 95.3% were rare. Moreover, we found 1,937 out of 38,685 VNTRs that were associated with gene expression through VNTR-MPs in lymphoblastoid cell lines. Specifically, we clarified that the expansion of a likely causal motif could upregulate gene expression by improving the binding concentration of PU.1. We also explored the potential impacts of VNTR polymorphisms on phenotypic differentiation and disease susceptibility. This study expands our knowledge of VNTR-MPs and their functional implications.

INTRODUCTION

Variable number tandem repeats (VNTRs) are consecutive DNA characterized by repeating patterns, typically spanning from 7 to 100 base pairs.[1–3] In humans, VNTRs have been detected to comprise more than 5 million sequences in the telomere-to-telomere genome.[3–5] In the last decade, great progress has been made in genome-wide association studies (GWASs) in human genetics, with nearly 400,000 SNP-trait associations having been curated in the NHGRI-EBI GWAS Catalog.[6] However, the issue of "missing heritability" remains a significant concern that cannot be fully elucidated by SNPs but can be partially attributed to VNTR polymorphisms.[7,8] Recently, more and more attention has been paid to repeat polymorphism studies, and numerous striking studies have been carried out to investigate VNTR-related human traits and diseases. For example, Mukamel et al. analyzed 118 protein-altering VNTRs in 415,280 UK Biobank individuals and found significant associations of common VNTR variants with human height, hair morphology, and biomarkers of health.[9] They also examined 9,561 autosomal VNTR loci in 418,136 UK Biobank participants, revealing two non-coding VNTRs located at *TMCO1* and *EIF3H*, which significantly contributed to the risk of glaucoma and colorectal cancer.[10] Moreover, Cui et al. constructed a biobank-scale database of 0.86 million tandem repeats derived from 338,963 whole-genome sequencing (WGS) samples in diverse ancestries, termed TR-gnomAD, which provided an invaluable resource for tandem repeat interpretation.[11]

Currently, specific repeat motifs have been reported to regulate gene expression and be associated with disorder risks. Song et al.[12] demonstrated that a specific VNTR motif on the *CACNA1C* gene is associated with the risk of schizophrenia by modulating gene expression activity. Additionally, Kirby et al. revealed that the insertion of a single cytosine in the motif sequence of a long-coding VNTR in *MUC1* gene can cause medullary cystic kidney disease type 1.[13] However, since existing tools for short-read genotyping focus on a single motif and provide only a limited description of the complexity of tandem repeats, the functional implications of repeat motif polymorphisms have yet to be comprehensively examined. Recently, a newly

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

developed toolkit, danbing-tk,[14] has utilized the repeat-pangenome graph to estimate VNTR composition with short reads. This approach has been proven to be powerful for population studies and trait association in highly variable loci.[15] Given its capabilities, we chose this tool as the most suitable approach for investigating genome-wide VNTR and motif characteristics in our work.

Population-specific genomics are fundamental for genetic disease research and precision medicine. To date, several state-of-art population-scaled VNTR studies have been mainly based on the 1,000 Genome Project (1KGP)[14,16,17] or European ancestry,[15] where the proportion of Han Chinese people only accounts for 8% in the 1KGP. It is worth noting that Han Chinese people, as the largest ethnic group in East Asia, and even around the world, is far from underrepresented in human genetic studies.[18,19] Our previous works[20–24] have demonstrated the superiority of the Chinese-specific cohort NyuWa in discovering novel functional rare variants and selection signals. Considering the striking polymorphism and mutability of VNTRs between East Asians and Europeans, the construction of VNTR polymorphism patterns in Han Chinese people is needed to fill the missing diversity worldwide.

Here, we constructed a variant map of 38,685 genome-wide VNTRs and 916,938 VNTR motifs across 8,222 genomes, including 4,126 Chinese individuals (~30.5×, NyuWa) and 4,096 multiracial individuals (~33×, the 1KGP and Human Genome Diversity Project [HGDP]). In total, over 2.5 million VNTR length polymorphisms (VNTR-LPs) and 11 million VNTR motif polymorphisms (VNTR-MPs) were detected in our dataset. We identified that 839,211 (34.8%) VNTR-LPs and 2,982,456 (31.8%) VNTR-MPs were specific to the NyuWa population. Among these, 74.4% of VNTR-LPs and 95.3% of VNTR-MPs were classified as rare (frequency < 0.001). We subsequently conducted a systematic functional investigation of VNTRs and elucidated the regulatory role of a specific VNTR motif in gene expression. Additionally, we discovered VNTRs with significant differences in length and motif composition among continent populations, evaluating their impact on phenotypic differentiation and disease occurrence in different populations. This work extends the scale of VNTR variations in existing genetic research, provides new insights into the role of VNTRs in gene regulation, and serves as an important reference for future clinical research and genotype-phenotype association studies.

RESULTS

Large Chinese population cohort facilitated rare VNTR detection

In this work, we performed a genome-wide identification of 38,685 VNTRs and 916,938 VNTR motifs across 8,222 genomes, including 4,126 individuals from the NyuWa genome resource[22] and 4,096 individuals[25] from the high-coverage 1KGP and HGDP.[26] In total, over 2.5 million VNTR-LPs and 11 million VNTR-MPs were identified, with 2,411,625 VNTR-LPs and 9,387,220 VNTR-MPs detected in the NyuWa dataset as well as 1,753,496 VNTR-LPs and 8,131,582 VNTR-MPs detected in the 1KGP and HGDP datasets. Compared to the 1KGP and HGDP, we observed that 839,211 (34.8%) VNTR-LPs and 2,982,456 (31.8%) VNTR-MPs were unique to the NyuWa cohort (Figure 1A), with the majority of VNTR-LPs (74.4%) and VNTR-MPs (95.3%) being low frequency (Figures S1A, S1C, and S1D). Furthermore, we observed a high concordance rate of nearly 97.7% for VNTR-MPs between East Asian (EAS, subset of the 1KGP) and NyuWa, with an extra 4,310,484 (47.3%) exclusively detected in the NyuWa dataset (Figures 1A and S1B). For VNTR-LPs, we also observed a high concordance rate of nearly 98.3% between EAS (subset of the 1KGP) and NyuWa, with an extra 1,666,887 (69.1%) exclusively detected in the NyuWa dataset (Figures 1A and S1D). To evaluate the impact of sample size and batch effects on these observed discrepancies, we randomly sampled the NyuWa dataset five times, each time selecting a sample size matching that of the 1KGP.EAS population for motif polymorphism comparisons (Figure S2A). We also conducted pairwise comparisons between the NyuWa dataset and each subpopulation within 1KGP.EAS (Figure S2B), ensuring consistent sample sizes across all comparisons to control for subpopulation variations, to evaluate the similarities and discrepancies while controlling for the factor of subpopulation variations. The results indicated that, after randomly selecting a sample size equivalent to the EAS population from the NyuWa population, the overall consistency between the two datasets remained relatively high, and the number of loci specific to the NyuWa population decreased. However, substantial differences in consistency were observed when comparing the randomly sampled NyuWa population to the individual EAS subpopulations. These findings suggest that the distinctiveness of the NyuWa population is primarily attributable to population differences and sample size, with minimal influence from batch effects. At the k-mer level, which represents the basic composition of motifs, a substantial amount of k-mers were found in NyuWa dataset, with a frequency $\leq$0.01 (Figure 1B), highlighting the significant value of the deeply sequenced NyuWa genome resource for unveiling rare VNTR polymorphisms in human genomes.

Pervasive gene-regulatory effects of VNTRs and motif compositions

Tandem repeat polymorphisms can influence diverse phenotypes and susceptibility to diseases via their gene-regulatory effects.[27] Here, we applied 441 RNA sequencing (RNA-seq) data of lymphoblastoid cell lines (LCLs) from the Geuvadis project[28] to investigate associations between VNTR-MPs and gene expression levels. As distinct motif compositions have been proved to impact gene expression independently of VNTR length,[15] we utilized VNTR lengths and motif dosages, respectively, to identify expression quantitative trait loci (eQTLs). A total of 438 VNTRs (denoted as eVNTRs) (Figure 2A) and 2,295 motifs (denoted as eMotifs; corresponding to 1,937 VNTRs) (Figure 2B) exhibited significant associations at a gene-level false discovery rate (FDR) of 5%. To corroborate our discoveries, the available eVNTRs (Figure S3) and eMotifs (Figure S4) detected from 879 Genotype-Tissue Expression (GTEx) genomes[15] were used as a comparison, with the strongest correlation of effect size found between LCLs and Epstein-Barr virus

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

CellPress
OPEN ACCESS

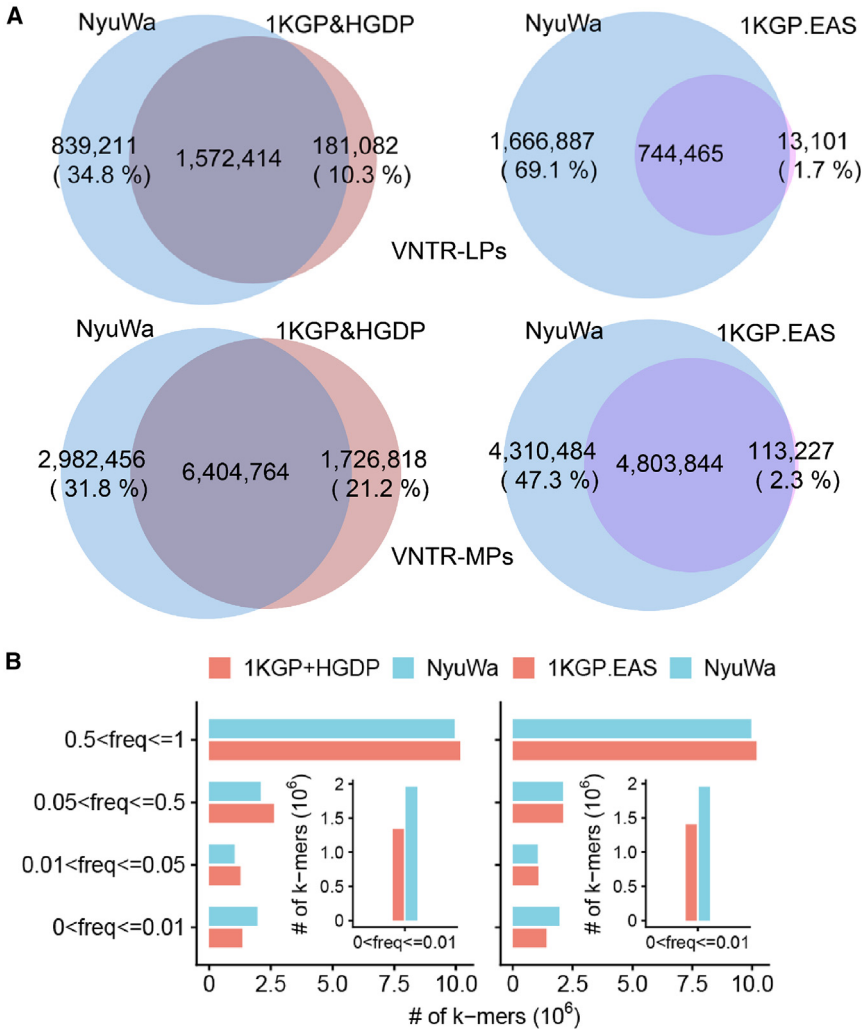


Figure 1. VNTR-LPs and VNTR-MPs identified in this study

(A) Comparison of VNTR length polymorphisms (VNTR-LPs; top) and VNTR motif polymorphisms (VNTR-MPs; bottom) identified in the NyuWa dataset with those identified in the 1KGP and HGDP datasets (left) and the East Asian samples from the 1KGP dataset (right).

(B) Bar plot of frequency distributions for k-mers identified in the NyuWa dataset with k-mers identified in the 1KGP and HGDP datasets (left) and the East Asian samples from the 1KGP dataset (right). Inset, the distribution of rare k-mers (0 < frequency [freq] $\leq$ 0.01) is shown at a finer scale.

(EBV)-transformed lymphocytes from GTEx. The Pearson correlation coefficient (R) of eVNTRs is 0.74 (Figure 2C) and that of eMotif is 0.97 (Figure 2D).

To further substantiate the regulatory influence of eVNTRs and eMotifs, we annotated them in the context of functional genomics data. We observed that eVNTRs were highly enriched in regions associated with active histone marks like H3K9ac, H3K4me3, and H3K27ac (Figure 2E) while eMotifs were more concentrated in peaks for H3K9me3 (Figure 2F). These findings were largely consistent with earlier investigations of expression short tandem repeats (eSTRs).[21,29] In addition, these eVNTRs and eMotifs exhibited enrichments in coding sequences (CDSs), promoters, and assay for transposase-accessible chromatin (ATAC) regions (Figures 2E and 2F), indicating their direct or intermediated regulatory impacts on gene expression. By performing analysis utilizing ChromHMM annotations, we found significant enrichments of eVNTRs in the flanking regions of transcription start sites (TSSs) and active TSSs (Figure 2G), while eMotifs not only enriched near TSSs but also showed a marked abundance in active enhancers (Figure 2H). These observations suggested that eMotifs shared similar genomic characteristics with eVNTRs, while the sequence specificity of eMotifs enables them to function as active enhancers. Overall, the preferential enrichments of eVNTRs and eMotifs in regulatory elements and accessible chromatin regions highlighted their vital roles in gene regulation.

A previous study reported that eSTRs associated with gene expression show a certain enrichment in evolutionarily conserved regions of the genome.[29] We here applied an evolutionary model, the LINSIGHT score,[30] to compare the conservation between eVNTRs and VNTRs unrelated to gene expression (considered control VNTRs). As expected, conservation scores within the ±2 kb window around each eVNTR were higher when compared to adjacent regions of control VNTRs (Figure 2I). Additionally, there were noteworthy peaks within the 1 kb region surrounding eVNTRs, indicating a higher degree of purifying selection in the vicinity of eVNTRs. In addition, genes corresponding to eMotifs were mainly involved in viral infection such as herpes simplex virus 1 infection and salmonella infection (Figure 2J), which can be explained by the functions of the LCLs we used. A recent study revealed that short tandem repeat (STR) expansion can increase the number of weak binding sites for transcription factors (TFs) to exert regulatory effects.[31] We therefore investigated the potential TFs that bind to 2,295 eMotif sequences. As a result, the most common TFs were from the zinc-finger family, followed by the E-twenty six (ETS)-domain, forkhead, interferon regulatory factor (IRF), and E2 promoter binding factor (E2F) families (Figure 2K). Focused on the eMotifs that were located in promoter regions, we found that the EWSR1-FLI1 fusion protein of the ETS family was the most effective TF bound to eMotifs (Figure S5). This protein can promote the expression of certain tumor-related genes, inhibit the expression of tumor-suppressor genes, and influence the growth, proliferation, invasion, and metastasis of tumor cells.[32] These results provide insight into the relationship between genotypes and molecular phenotypes and further our understanding of the impacts of VNTRs on gene expression and biological characteristics.

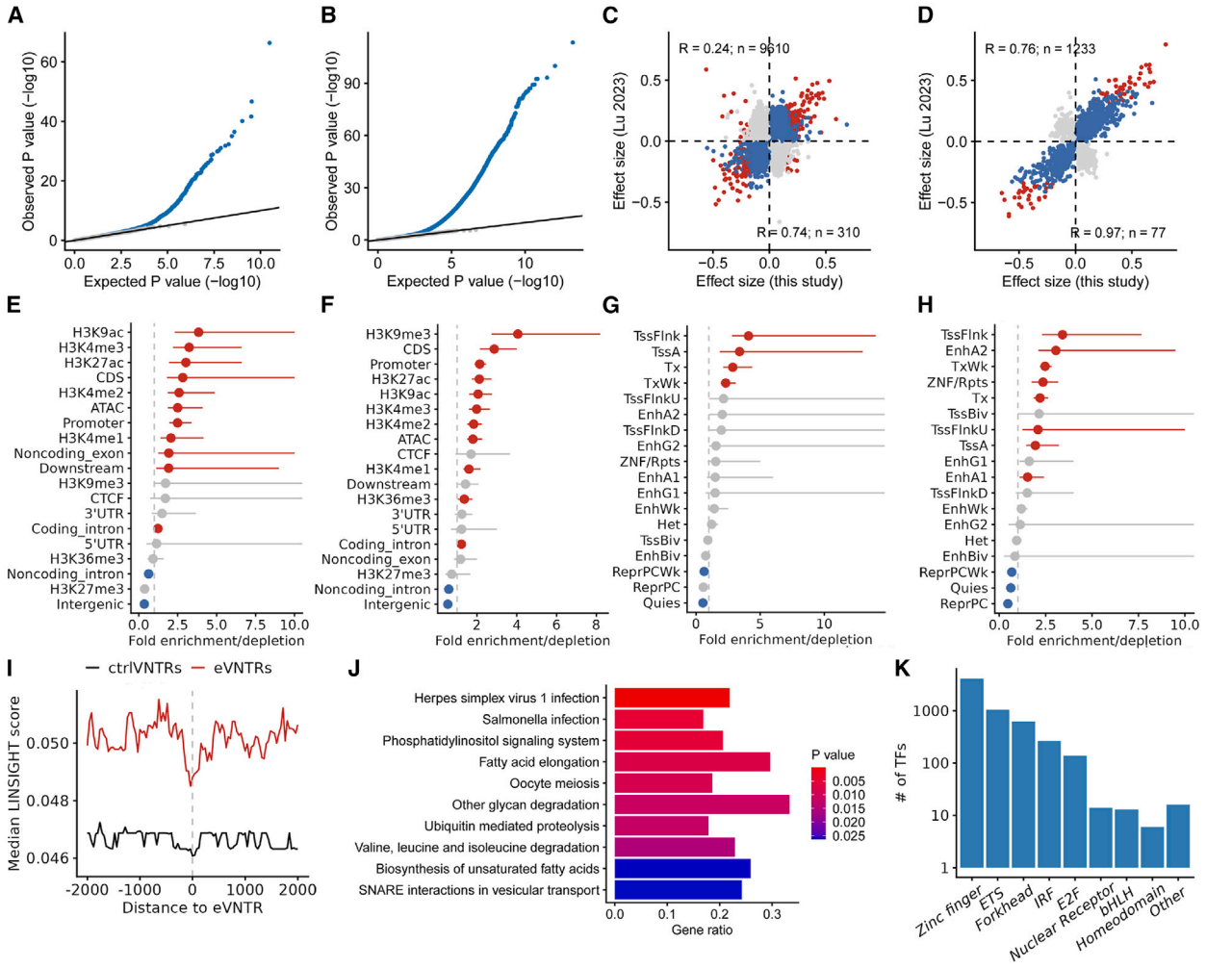


Figure 2. *Cis*-regulatory effects of eVNTRs and eMotifs in gene expression

(A and B) Quantile-quantile plot comparing observed *p* values for VNTR gene association tests (two-sided t test in linear model) versus the expected uniform distribution in eVNTR (A) and eMotif (B) analysis. The blue dots represent the observed association tests, and gray dots represent *p* values for permutation control. The black line gives the expected *p* value distribution under the null hypothesis of no association.

(C and D) Correlations of eVNTR (C) and eMotif (D) effect sizes identified in this study and a previous study by Lu et al.[15] The blue points indicate eVNTRs and eMotifs whose directions of effect were concordant in two studies, and gray points denote eVNTRs and eMotifs with discordant directions of effect in two studies. The eVNTRs and eMotifs detected in both studies are colored red regardless of the concordance of effect.

(E and F) Fold enrichment of eVNTRs (E) and eMotifs (F) in genome and epigenetic regions in the GM12878 cell line. A permutation test was repeated 1,000 times, and empirical *p* values were computed together with the enrichment values by GAT v.1.3.4. Points denote the enrichment values. Red and blue points denote significant enrichments or depletions ($p < 0.05$ after Benjamini and Hochberg correction), and bars show 95% confidence intervals.

(G and H) Fold enrichment of eVNTRs (G) and eMotifs (H) in chromatin states defined by ChromHMM in the GM12878 cell line. A permutation test was repeated 1,000 times, and empirical *p* values were computed together with the enrichment values by GAT v.1.3.4. Points denote the enrichment values. Red and blue points denote significant enrichments or depletions ($p < 0.05$ after Benjamini and Hochberg correction), and bars show 95% confidence intervals.

(I) Conservative evaluation of eVNTRs. We randomly selected 500 VNTRs unrelated to gene expression as controls. The median LINSIGHT scores in 50 bp windows were measured for visualization.

(J) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment for annotated genes of eMotifs.

(K) Classification of predicted binding transcription factors for eMotifs.

The regulatory functions of the causal eMotif *MAD1L1* in gene expression

To determine whether these eMotifs are likely causal to gene expression, we further performed fine-mapping analysis on eVNTRs and eMotifs separately. In total, three eVNTRs (Figure S6) and 14 eMotifs (Table S1) were identified, with the highest posterior inclusion probability (PIP) exceeding 0.8, and were thus considered as likely causal variants. Among these three eVNTRs, two (chr13:113,231,793 and chr7:19,40,924) were situated in the introns of *CUL4A* and *MAD1L1*, respectively, and the other one (chr22: 22,292,464) was located in the intergenic region. It is worth noting that the intronic expansion of VNTR *CUL4A* had

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

CellPress
OPEN ACCESS

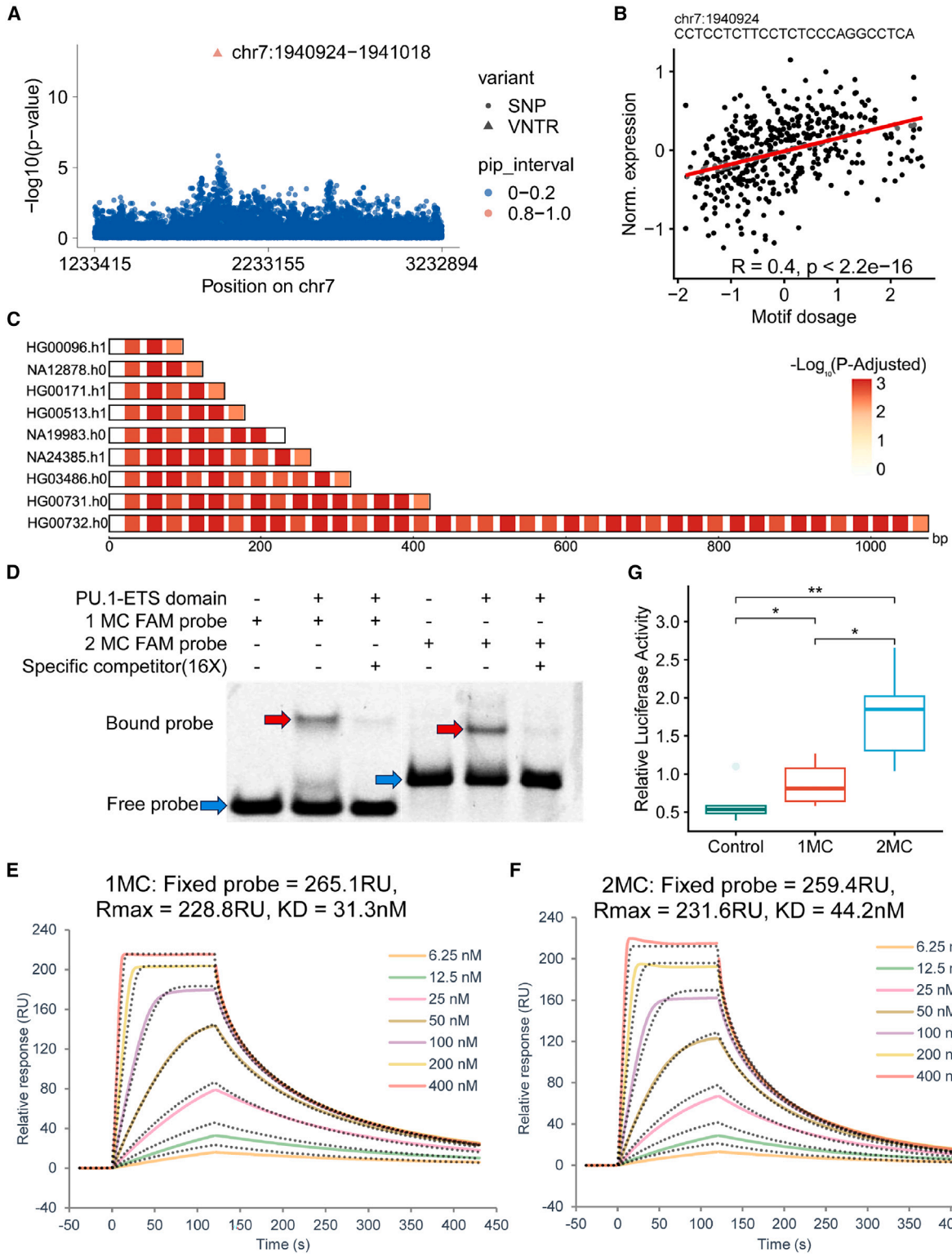


Figure 3. Gene regulation of specific eMotif *MAD1L1* polymorphism

(A) Fine-mapped eMotif *MAD1L1* (chr7:1,940,924). *p* value was obtained from eQTL association tests. Dots with diverse colors and shapes represent different PIP intervals.

(legend continued on next page)

been previously reported to alter mRNA splicing and be associated with erythrocyte traits.[9] As expected, alternative splicing events were also observed in our dataset (Figure S7) and led to a decrease in gene expression levels (Figure S8). Notably, the eVNTR loci (chr7:1,940,924) was found to be a significant eQTL for the first time, exerting likely causal effects on gene expression both by VNTR length ($p = 3.3 \times 10^{-11}$, PIP > 0.99) and motif composition ($p = 8.9 \times 10^{-14}$, PIP > 0.99) (Figure 3A). We found that the dosage of the eMotif "CCTCCTCTTCCT CTCCCAGGCCTCA" in this locus exhibited a strong association with the expression of *MAD1L1* ($p < 0.001$) (Figure 3B). In addition, histone modification, DNase I hypersensitivity clusters, TF clusters, and a distal enhancer (ENCODE ID: EH38E2529017) were also found in this locus (Figure S9). Prediction of TF binding sites revealed that PU.1, which is an activating TF during myeloid and B lymphoid cell development, was identified as the most significant TF binding to the eMotif CCTCCTCTTCCTCTCCC AGGCCTCA, with a *p* value of 0.0006, and this DNA-protein interaction was supported by the chromatin immunoprecipitation sequencing (ChIP-seq) data (GEO: GSE128834 and GSE128835) from ReMap.[33] The polymorphic validation of this locus in 35 haplotype-resolved assemblies from the Human Genome Structural Variation Consortium (HGSVC)[34] revealed that the longer VNTR *MAD1L1* was associated with an increased number of predicted PU.1 binding sites (Figure 3C). Given that enhancers can harbor clusters of TF motifs,[35] we hypothesized that the expansion of eMotif CCTCCTCTTCCTCTCCCAGGC CTCA might function as an active enhancer to improve the PU.1 binding concentration and thereby regulate nearby gene expression.

To validate the above hypothesis, we conducted electrophoretic mobility shift assay (EMSA), surface plasmon resonance (SPR), and dual-luciferase reporter (DLR) assays separately. As expected, specific DNA-protein complexes (bound probe) were clearly observed in binding reactions involving a one-copy motif (1CM) probe and a 2CM probe (Figure 3D). We further performed SPR experiments to quantify the binding affinities of PU.1 to these DNA probes. Biotin-modified probes 1CM and 2CM were immobilized on the streptavidin sensor chip with the same mass, leading to 1CM having twice the number of molecules as 2CM. From the experimentally determined maximum response level (Rmax), it can be observed that the fixed 1CM (Figure 3E) and 2CM (Figure 3F) interacted with equivalent quantities of PU.1, suggesting that 2CM bound twice the amount of PU.1 compared to 1CM on a per-molecule level. In addition, the reporter plasmid containing 2CM had significantly higher luciferase activities than that with 1CM in HEK293T cells in DLR assays (Figure 3G). Collectively, these results substantiated our hypothesis that the expansion of VNTR motif CCTCCTCTTCCTCTCCCAGGCCTCA in the locus chr7:1940,924 could upregulate the expression of *MAD1L1* by increasing interactions with PU.1. As a component of the mitotic spindle assembly checkpoint, *MAD1L1* plays a vital role in cell cycle control and tumor suppression. Due to the overexpression of *MAD1L1* having been widely investigated in numerous diseases and cancers,[36–39] this novel eMotif may exert *cis*-regulatory effects, contributing to disease and cancer susceptibility.

The relevance of hypervariable VNTR motifs to human traits

In modern humans, the emerging roles of hypervariable VNTRs have been found in human-specific traits, particularly in neural-related functions.[2,40] We firstly utilized the coefficient of variation (CV) to identify hypervariable motifs for each superpopulation (STAR Methods) and identified 180,177 motifs with highly mutable dosages shared in all superpopulations. Gene Ontology (GO) enrichment analysis suggested that genes enclosing these motifs play critical roles in neuron development and synapse transmission (Figure 4A). Consistently, these genes showed significant enrichment for those primarily expressed in the cerebral cortex ($p < 1 \times 10^{-48}$) (Figure 4B). These results suggested the involvement of highly variable VNTRs in various neurological phenotypes in humans. As it has been reported that VNTRs are powerful markers for studying population structure,[17] we performed a principal-component analysis (PCA) on motifs within high V_{ST} (a statistic that estimates population differentiation and varies from 0 to 1) VNTR motifs (STAR Methods). As expected, the results (Figure 4C) generally exhibited similar trends to those observed with SNPs,[22] structural variants (SVs),[41] and STRs.[21] PC1 effectively distinguished Africans, PC2 distinguished East Asians (including NyuWa), and PC4 and PC5 distinguished South Asians and Americans from the rest, respectively (Figure 4C). VNTRs have also participated in the formation of a wide variety of population-specific traits.[42] Focused on the VNTRs in protein-coding regions that may strongly shape human phenotypes, the genes enclosing these VNTRs were highly enriched in immunoglobulin (Ig)A levels and hair colors in GWAS Catalog traits (Figure S10). Specifically, two VNTRs within exons of *FADS2* and *TCF25* were detected to be significantly diverse across populations. The overall lengths of VNTR *FADS2* in Africans were significantly longer than those in other populations

(B) Correlations between the dosage of the eMotif *MAD1L1* and normalized expression of *MAD1L1*. The red line indicates the best fit under simple linear regression.

(C) Predicted number of PU.1 binding sites across 9 length-divergent haplotypes. Each haplotype was scanned for matches with a 20 bp PU.1 binding motif (MA0080.5) using FIMO with a cutoff of *p*-adjusted < 10^{-2}.

(D) EMSA analysis of the 1CM and 2CM double-stranded DNA (dsDNA; 5 μM, 1 μL) binding to PU.1 ETS-domain (0.17 μg/μL, 1 μL) with (+) or without (−) specific competitor (20 μM, 4 μL).

(E and F) Kinetic analysis of recombinant PU.1 ETS-domain binding to 1CM and 2CM dsDNA by surface plasmon resonance (SPR). The concentrations of the PU.1 ETS-domain protein are 6.25, 12.5, 25, 50, 100, 200, and 400 nM in each graph, respectively. Equilibrium and kinetic constants were calculated by a global fit to 1:1 Langmuir binding model. The gray dashed lines are curves fit to the data using the 1:1 binding model. RU, resonance units.

(G) Dual-luciferase assays of enhancer activity for the empty vector (control) and the plasmids with the insertion of 1CM and 2CM in the forward position. The luciferase activity of each construct was normalized against the activity of Renilla luciferase. Data are shown as the median (minimum to maximum), from six independent experiments for each construction. *p* values were calculated by a two-sided Wilcoxon rank-sum test. *$p < 0.05$ and **$p < 0.01$.

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

CellPress
OPEN ACCESS

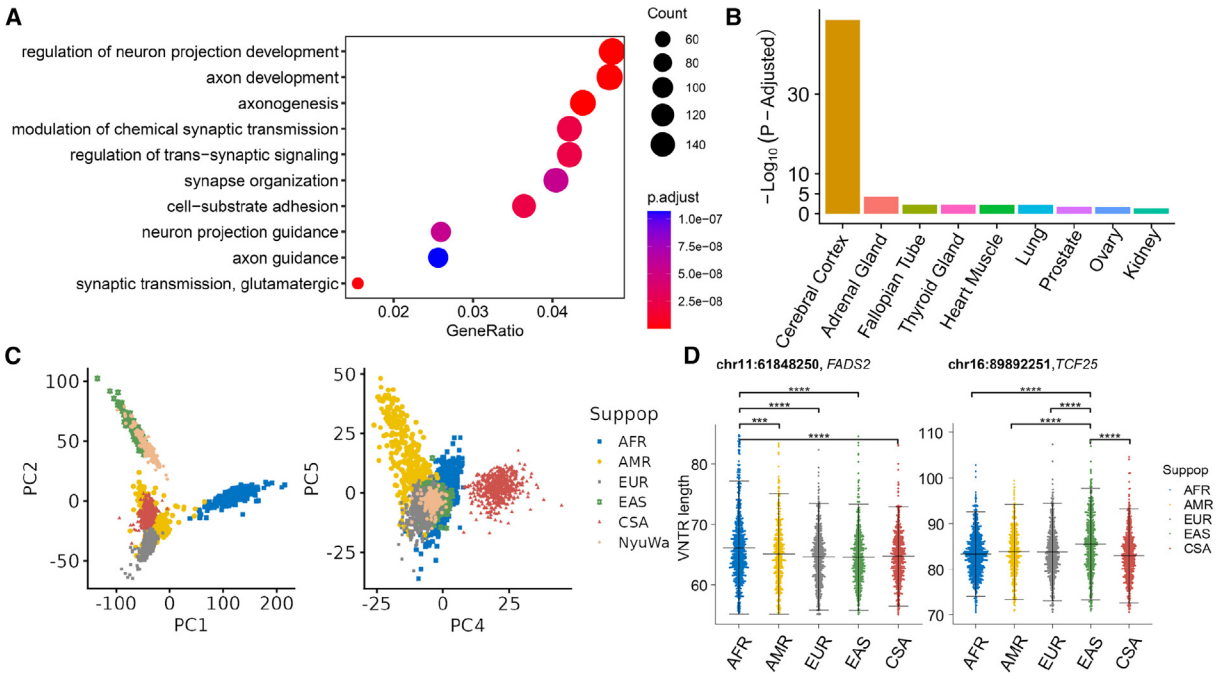


Figure 4. Hypervariable VNTR motifs within superpopulations associated with human phenotypes

(A) Gene Ontology (GO) enrichment analysis for annotated genes of hypervariable VNTR motifs in all superpopulations. The top ten items with significant *p* values are shown.

(B) Tissue-specific gene enrichment analysis of hypervariable VNTR motifs is shown in bar plot. The y axis represents the adjusted *p* value derived from hypergeometric test and corrected using the Benjamini and Hochberg correction by TissueEnrich v.1.16.0.

(C) Principal-component analysis of hypervariable VNTR motifs in all superpopulation samples and 200 randomly selected NyuWa samples. Shapes represent the superpopulation of each sample. Colors represent the population of each sample.

(D) Distributions of lengths for two phenotype-related VNTR significant differences across superpopulations. *p* values were computed using the two-sided Wilcoxon rank-sum test. ***$p < 0.001$ and ****$p < 0.0001$.

In (C) and (D), AFR denotes African superpopulation, AMR denotes American superpopulation, EAS denotes East Asian superpopulation, EUR denotes European superpopulation, and CSA denotes Central/South Asian superpopulation.

(Figures 4D and S11A). Given that African ancestry was consistently associated with higher serum IgA levels compared to other ancestries,[43] this suggests the potential involvement of the expansion of VNTR *FADS2* in regulating IgA levels in Africans. Another VNTR in the exon of *TCF25* was tagged by GWAS Catalog SNPs associated with brown/black hair color. This VNTR showed significantly longer lengths in East Asians than in any other population (Figures 4D and S11B), implying its potential impact on the formation of hair colors in East Asians.

The relevance of HSE VNTRs to diseases

To our knowledge, dozens of Mendelian disorders have been reported to implicate tandem repeat expansions.[44–46] Out of 38,685 VNTR sites, 360 overlapped with human-specific expansion (HSE) VNTRs identified by Course et al.[47] Considering that VNTRs can exhibit preferential expansions in specific populations,[48,49] we conducted pairwise length comparisons among five continental populations for 360 HSE VNTRs (Figure 5). In addition to the VNTR *DYNC2I1* reported by Course et al. being identified in our analysis, we also discovered two novel outliers linked to the genes *RPH3AL* and *VIPR2*. Over half of African Americans have exhibited loss of heterozygosity at *RPH3AL*, which is associated with the poor survival of African Americans with colorectal adenocarcinomas.[50] We observed shorter lengths of VNTR *RPH3AL* in Africans than in all other populations (Figure S12A). Consistently, the frequency of the repeat motifs at this locus varied among superpopulations, with a significant deficiency observed in Africans compared to other superpopulations (Figure S12B). This observation may suggest a potential functional relevance of the VNTR in *RPH3AL*, possibly involved in tumor progression. We identified another VNTR near the gene *VIPR2* that exhibited a noticeably shorter length (Figure S13A) and less frequency of motifs (Figure S13B) in East Asians compared to other ancestries. Loss of function in *VIPR2* has been reported to potentially compromise bipolar cell function and has been associated with high myopia in a Chinese Han cohort.[51] Collectively, these observations underscore the importance of further investigating the functional implications of these two VNTRs in disease risk, particularly in the context of genetic differences across populations.

DISCUSSION

In summary, we applied the state-of-art genotyper danbing-tk to identify genome-wide variants both in VNTR lengths and motif

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

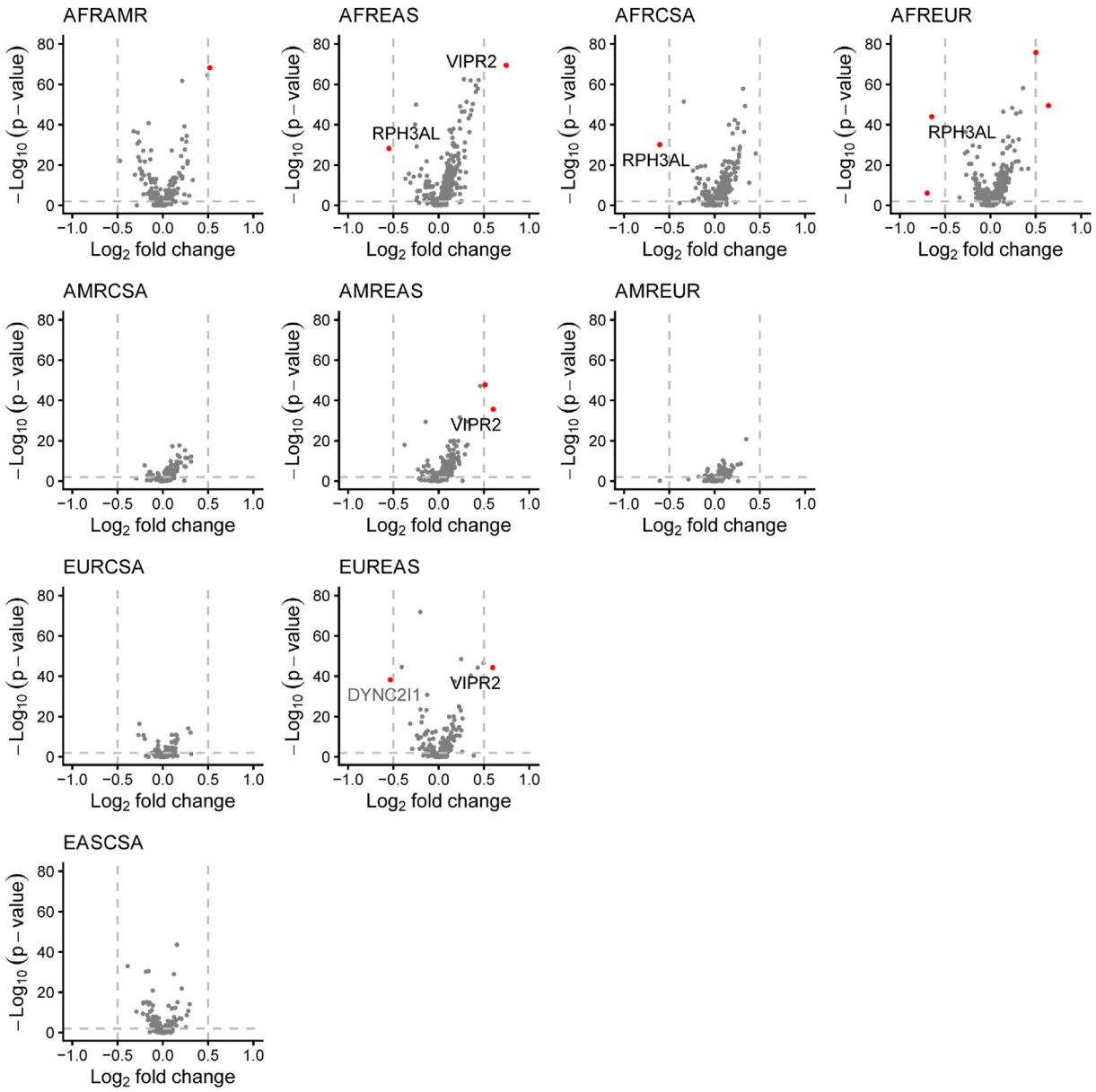


Figure 5. Length comparisons of 360 human-specific expansion VNTRs across five superpopulations

Volcano plots show pairwise comparisons of average lengths for 360 human-specific expansion VNTRs between superpopulations from the 1KGP and HGDP. VNTRs with significant length differences are shown in red, while two novel VNTRs (black) and one previously reported VNTR (gray) are labeled by the nearest gene or the gene in which they reside. AFR, African superpopulation; AMR, American superpopulation; EAS, East Asian superpopulation; EUR, European superpopulation; SAS, South Asian superpopulation. *p* values are reported for the two-sided Wilcoxon rank-sum test and adjusted using the Benjamini and Hochberg correction.

compositions for 8,222 worldwide genomes. We constructed the most recent and the largest genome-wide VNTR-MP map, identifying more than 11 million VNTR-MPs in these samples. We systematically investigated gene-regulatory effects of VNTR length and motif composition in LCLs and conducted in-depth interpretations of the regulation mechanism for a fine-mapped VNTR motif. Additionally, we explored the roles of hypervariable VNTRs in the divergence of human phenotypes as well as disease risks of VNTR expansions. These findings expanded our understanding of the functional consequences of VNTR polymorphisms and provided a subset of VNTRs with potential functional implications.

We integrated genomic data of 4,126 Han Chinese individuals from the NyuWa genomic resource and 4,096 individuals from the 1KGP and HGDP to construct a comprehensive VNTR polymorphism map, analyzing the variation characteristics of VNTRs in length and motif composition. To improve clarity of presentation, we measured the discrete copy numbers of each motif

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

using the motif dosage. The motifs with different copies across samples were deemed VNTR-MPs. By comparing VNTR-MPs between the NyuWa population and global populations, we found that 31.8% of the variations were unique to NyuWa, with 95.3% of them being rare variants. These results clearly demonstrate the significant value of the NyuWa genome resource in uncovering rare variations in humans.

By combining 441 RNA-seq data from the Geuvadis project, we identified a total of 438 VNTRs that affect gene expression through length variation, as well as 1,937 VNTRs that influence gene expression through motif dosage. We assessed the functional impact of these expression-related VNTRs based on genomic features, epigenetic modifications, evolutionary conservation, and TF binding preferences and found that they were enriched in regulatory elements, accessible chromatin areas, and conserved regions. Moreover, we performed fine-mapping analysis and identified a novel VNTR motif in the gene *MAD1L1* to be likely causal to gene expression. Further biological experiments we performed validated that the expansion of this motif can upregulate the expression of adjacent genes by increasing the binding concentrations of the TF PU.1. These results provide new insights into the role of VNTR variants in gene regulation.

By leveraging multiethnic data, we achieved effective population stratification and discovered VNTRs with significant differences in length and motif composition among continent populations. We generated two sets including exonic VNTRs and HSE VNTRs to investigate the potential impacts of VNTR polymorphisms on phenotypic differentiation and disease susceptibility across various populations. Our comprehensive analysis revealed four significant risk VNTRs. Specifically, we elucidated the associations of two VNTRs with IgA levels and hair color and two VNTRs with colorectal adenocarcinomas and high myopia, respectively. We believe that these findings could serve as an important reference for future clinical research and genotype-phenotype association studies.

Limitations of the study

Although this study has yielded valuable findings, it is crucial to take into account the following constraints. Firstly, as danbing-tk can only genotype diploid VNTR contents from sequencing read depth, the precise haploid variants in each allele remained unknown. This limited our understanding of specific allele differences between individuals. Secondly, the estimation based on sequencing depth can be influenced by batch effects across different cohorts, requiring more rigorous quality control to eliminate their impact. Thirdly, the VNTRs in non-reference (not included in the danbing-tk repeat-pangenome graph) loci have not been accounted for. In the future, we plan to upgrade more VNTR loci from the telomere-to-telomere (T2T) reference genome and explore the possibility of integrating various VNTR identification and genotyping tools in order to provide a more comprehensive understanding of VNTR variations in humans.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Shunmin He (heshunmin@ibp.ac.cn).

Materials availability

Plasmids pcDNA3.1-PU.1, pGL3-promoter-1CM, and pGL3-promoter-2CM are available from the lead contact upon request.

Data and code availability

- The DNA sequencing data of NyuWa samples used in this study have been deposited in the Genome Sequence Archive (GSA) in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number HRA004185 (https://ngdc.cncb.ac.cn/gsa-human/). These data are available under restricted access for privacy protection and can be obtained by application on the GSA database website (https://ngdc.cncb.ac.cn/gsahuman/) following the guidance of the "request data" section on this website. These data have also been deposited in the National Omics Data Encyclopedia (NODE) of the Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, under accession number OEP002803 (http://www.biosino.org/node). The user can register and login to this website and follow the guidance of the "request for restricted data" section to request the data. The user can contact the corresponding author to apply for permission to access the full list of VNTR genotyped data. The alignment files for the 1KGP dataset are available at https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data/and https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/additional_698-_related/. Genotype data for SNPs and insertions or deletions (indels) for the 1KGP dataset are available at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/. Aligned files for the HGDP are available at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGDP/. Haplotype-resolved assemblies from HGSVC2 are available at https://www.internationalgenome.org/data-portal/data-collection/hgsvc2. RNA-seq data of the human LCL from the Geuvadis consortium are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/geu-vadis/working/geuvadis_topmed/.
- All software used in this research are listed in the key resources table. We used Qualimap v.2.2.1 (http://qualimap.conesalab.org/), BCFtools v.1.5 (http://samtools.github.io/bcftools/), GATK (https://gatk.broadinstitute.org/hc/en-us), and SAMtools v.1.9 (http://www.htslib.org/) to process genome datasets. For VNTR genotype analysis and quality control, we used danbing-tk v.1.3 (https://github.com/ChaissonLab/danbing-tk). For PCA, we used the scikit-learn v.1.2.1 package in Python 3.10.9. For functional analysis of these VNTRs, we used featureCounts v.2.0.3 (https://subread.sourceforge.net/), the edgeR v.3.32.1 package in R v.4.2.3, and Enrichr v.3.2 (https://maayanlab.cloud/Enrichr/). For eVNTR and eMotif analysis, we used susieR v.0.12.35 (https://github.com/cran/susieR) and the glm function in R v.4.2.3. For TF binding site predictions, we used FIMO v.4.12.0 (https://meme-suite.org/meme/tools/fimo). For data visualization, we used ggplot2 v.3.4.4 in R v.4.2.3. For assembly of third-generation sequencing data, we used Canu v.2.2 (https://github.com/marbl/canu). Statistical information about VNTR-LPs and VNTR-MPs can be viewed on the website http://bigdata.ibp.ac.cn/NyuWaVNTR/. Main analysis codes and used files have been made publicly accessible on GitHub at https://github.com/Skye722/NyuWaVNTR and have been archived via Zenodo at the DOI: https://doi.org/10.5281/zenodo.14003180.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We thank the people who generously contributed samples to the NyuWa dataset. We thank Y.Y.C., Z.W.Y., and B.X.Z. (Institute of Biophysics, Chinese Academy of Sciences) for technical support with Biacore experiments. We thank R.Z.M. for technical assistance with EMSA experiments. Data analysis and computing resources were supported by the Center for Big Data Research in Health (http://bigdata.ibp.ac.cn), Institute of Biophysics, Chinese Academy of Sciences. This work was supported by the Strategic Priority Research

Program of the Chinese Academy of Sciences (XDB38040300), the National Key R&D Program of China (2021YFF0703701, 2021YFF0704500, and 2022YFC3400405), the 14th Five-year Informatization Plan of the Chinese Academy of Sciences (CAS-WX2021SF-0203), the National Natural Science Foundation of China (32200478 and 32470660), the China Postdoctoral Science Foundation (2022M713311 and GZC20232899), and the National Genomics Data Center, China.

AUTHOR CONTRIBUTIONS

T.X. and S.H. conceptualized and supervised the project. S.Z., Q.S., X.W., R.G., X.Y., J.Z., Y.L., S.L., Y.N., Y.S., and T.S. conducted analyses. T.X. contributed to sample collection and data generation. Q.S. and X.W. performed the biological experiments. S.Z., Q.S., P.Z., T.X., and S.H. drafted the manuscript, and all the primary authors reviewed, edited, and approved manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Subjects and specimens
  - Cell lines
- METHOD DETAILS
  - Electrophoretic mobility shift assay (EMSA)
  - Surface plasmon resonance (SPR) measurements
  - Cell culture, reporter gene constructs, and dual-luciferase reporter assays
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Datasets processing
  - Genome-wide VNTR genotyping
  - Quality control of VNTR length and motif dosage sets
  - Comparisons of VNTR length and motif polymorphism
  - Identification of eVNTR and eMotif
  - Enrichment analysis
  - Fine-mapping
  - Prediction of TF binding sites with eMotifs
  - Hypervariable VNTR motifs calculation
  - Principal component analysis
  - Comparison of lengths for human-specific expansion VNTRs across superpopulations
  - Validation using nanopore targeted sequencing

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100699.

Received: April 8, 2024
Revised: August 31, 2024
Accepted: November 1, 2024
Published: November 27, 2024

REFERENCES

1. Vergnaud, G., and Denoeud, F. (2000). Minisatellites: mutability and genome architecture. Genome Res. *10*, 899–907. https://doi.org/10.1101/gr.10.7.899.
2. Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Human Genome Structural Variation Consortium; Warren, W.C., Pollen, A.A., Chaisson, M.J.P., and Eichler, E.E. (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. USA *116*, 23243–23253. https://doi.org/10.1073/pnas.1912175116.
3. Chaisson, M.J.P., Sulovari, A., Valdmanis, P.N., Miller, D.E., and Eichler, E.E. (2023). Advances in the discovery and analyses of human tandem repeats. Emerg. Top. Life Sci. *7*, 361–381.
4. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663–675.e19. https://doi.org/10.1016/j.cell.2018.12.019.
5. Linthorst, J., Meert, W., Hestand, M.S., Korlach, J., Vermeesch, J.R., Reinders, M.J.T., and Holstege, H. (2020). Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain. Transl Psychiat *10*, 369. https://doi.org/10.1038/s41398-020-01060-5.
6. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. *51*, D977–D985. https://doi.org/10.1093/nar/gkac1010.
7. Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet. *26*, 59–65.
8. Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K.E., and Gymrek, M. (2021). Patterns of de novo tandem repeat mutations and their role in autism. Nature *589*, 246–250. https://doi.org/10.1038/s41586-020-03078-7.
9. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Zheng, Y., McCarroll, S.A., and Loh, P.R. (2021). Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science *373*, 1499–1505. https://doi.org/10.1126/science.abg8289.
10. Mukamel, R.E., Handsaker, R.E., Sherman, M.A., Barton, A.R., Hujoel, M.L.A., Mccarroll, S.A., and Loh, P.R. (2023). Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. Cell *186*, 3659–3673.e23. https://doi.org/10.1016/j.cell.2023.07.002.
11. Cui, Y., Ye, W., Li, J.S., Li, J.J., Vilain, E., Sallam, T., and Li, W. (2024). A genome-wide spectrum of tandem repeat expansions in 338,963 humans. Cell *187*, 1–6.
12. Song, J.H.T., Lowe, C.B., and Kingsley, D.M. (2018). Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. Am. J. Hum. Genet. *103*, 421–430. https://doi.org/10.1016/j.ajhg.2018.07.011.
13. Kirby, A., Gnirke, A., Jaffe, D.B., Barešová, V., Pochet, N., Blumenstiel, B., Ye, C., Aird, D., Stevens, C., Robinson, J.T., et al. (2013). 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.
14. Lu, T.Y., and Human Genome Structural Variation Consortium; and Chaisson, M.J.P. (2021). Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs. Nat. Commun. *12*, 4250. https://doi.org/10.1038/s41467-021-24378-0.
15. Lu, T.Y., Smaruj, P.N., Fudenberg, G., Mancuso, N., and Chaisson, M.J.P. (2023). The motif composition of variable number tandem repeats impacts gene expression. Genome Res. *33*, 511–524. https://doi.org/10.1101/gr.276768.122.
16. Bakhtiari, M., Park, J., Ding, Y.C., Shleizer-Burko, S., Neuhausen, S.L., Halldórsson, B.V., Stefánsson, K., Gymrek, M., and Bafna, V. (2021). Variable number tandem repeats mediate the expression of proximal genes. Nat. Commun. *12*, 2075. https://doi.org/10.1038/s41467-021-22206-z.
17. Eslami Rasekh, M., Hernández, Y., Drinan, S.D., Fuxman Bass, J.I., and Benson, G. (2021). Genome-wide characterization of human minisatellite

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

CellPress
OPEN ACCESS

VNTRs: population-specific alleles and gene expression differences. Nucleic Acids Res. *49*, 4308–4324. https://doi.org/10.1093/nar/gkab224.

18. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. *85*, 762–774.
19. Gao, Y., Zhang, C., Yuan, L., Ling, Y., Wang, X., Liu, C., Pan, Y., Zhang, X., Ma, X., Wang, Y., et al. (2020). PGG. Han: the Han Chinese genome database and analysis platform. Nucleic Acids Res. *48*, D971–D976.
20. Niu, Y., Teng, X., Zhou, H., Shi, Y., Li, Y., Tang, Y., Zhang, P., Luo, H., Kang, Q., Xu, T., and He, S. (2022). Characterizing mobile element insertions in 5675 genomes. Nucleic Acids Res. *50*, 2493–2508. https://doi.org/10.1093/nar/gkac128.
21. Shi, Y., Niu, Y., Zhang, P., Luo, H., Liu, S., Zhang, S., Wang, J., Li, Y., Liu, X., Song, T., et al. (2023). Characterization of genome-wide STR variation in 6487 human genomes. Nat. Commun. *14*, 2092. https://doi.org/10.1038/s41467-023-37690-8.
22. Zhang, P., Luo, H., Li, Y., Wang, Y., Wang, J., Zheng, Y., Niu, Y., Shi, Y., Zhou, H., Song, T., et al. (2021). NyuWa Genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. Cell Rep. *37*, 110017. https://doi.org/10.1016/j.celrep.2021.110017.
23. Luo, H., Zhang, P., Zhang, W., Zheng, Y., Hao, D., Shi, Y., Niu, Y., Song, T., Li, Y., Zhao, S., et al. (2023). Recent positive selection signatures reveal phenotypic evolution in the Han Chinese population. Sci. Bull. *68*, 2391–2404.
24. Liu, S., Luo, H., Zhang, P., Li, Y., Hao, D., Zhang, S., Song, T., Xu, T., and He, S. (2024). Adaptive Selection of Cis-regulatory Elements in the Han Chinese. Mol. Biol. Evol. *41*, msae034.
25. Koenig, Z., Yohannes, M.T., Nkambule, L.L., Zhao, X., Goodrich, J.K., Kim, H.A., Wilson, M.W., Tiao, G., Hao, S.P., Sahakian, N., et al. (2024). A harmonized public resource of deeply sequenced diverse human genomes. bioRxiv, 2023.01.23.525248. https://doi.org/10.1101/2023.01.23.525248.
26. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.
27. Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298. https://doi.org/10.1038/nrg.2017.115.
28. Lappalainen, T., Sammeth, M., Friedländer, M.R., t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511. https://doi.org/10.1038/nature12531.
29. Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29. https://doi.org/10.1038/ng.3461.
30. Huang, Y.F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624. https://doi.org/10.1038/ng.3810.
31. Horton, C.A., Alexandari, A.M., Hayes, M.G.B., Marklund, E., Schaepe, J.M., Aditham, A.K., Shah, N., Suzuki, P.H., Shrikumar, A., Afek, A., et al. (2023). Short tandem repeats bind transcription factors to tune eukaryotic gene expression. Science *381*, eadd1250. https://doi.org/10.1126/science.add1250.
32. Yu, L., Davis, I.J., and Liu, P. (2023). Regulation of EWSR1-FLI1 function by post-transcriptional and post-translational modifications. Cancers *15*, 382.
33. Hammal, F., de Langen, P., Bergon, A., Lopez, F., and Ballester, B. (2022). ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Res. *50*, D316–D325. https://doi.org/10.1093/nar/gkab996.
34. Ebert, P., Audano, P.A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M.J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science *372*, eabf7117. https://doi.org/10.1126/science.abf7117.
35. Isbel, L., Grand, R.S., and Schübeler, D. (2022). Generating specificity in genome regulation through transcription factor sensitivity to chromatin. Nat. Rev. Genet. *23*, 728–740. https://doi.org/10.1038/s41576-022-00512-6.
36. Sun, Q., Zhang, X., Liu, T., Liu, X., Geng, J., He, X., Liu, Y., and Pang, D. (2013). Increased expression of Mitotic Arrest Deficient-Like 1 (MAD1L1) is associated with poor prognosis and insensitive to Taxol treatment in breast cancer. Breast Cancer Res Tr *140*, 323–330. https://doi.org/10.1007/s10549-013-2633-8.
37. Ryan, S.D., Britigan, E.M.C., Zasadil, L.M., Witte, K., Audhya, A., Roopra, A., and Weaver, B.A. (2012). Up-regulation of the mitotic checkpoint component Mad1 causes chromosomal instability and resistance to microtubule poisons. Proc. Natl. Acad. Sci. USA *109*, E2205–E2214. https://doi.org/10.1073/pnas.1201911109.
38. Lima, K., and Machado-Neto, J.A.J.D. (2018). MAD1L1 (Mitotic Arrest Deficient 1 like 1), *7*, p. 3.
39. Avram, S., Mernea, M., Mihailescu, D.F., Seiman, C.D., Seiman, D.D., and Putz, M.V. (2014). Mitotic Checkpoint Proteins Mad1 and Mad2-Structural and Functional Relationship with Implication in Genetic Diseases. Curr Comput-Aid Drug *10*, 168–181. https://doi.org/10.2174/1573409910666140410124315.
40. Kim, K., Bang, S., Yoo, D., Kim, H., and Suzuki, S. (2019). De novo emergence and potential function of human-specific tandem repeats in brain-related loci. Hum. Genet. *138*, 661–672. https://doi.org/10.1007/s00439-019-02017-5.
41. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81. https://doi.org/10.1038/nature15394.
42. Garg, P., Jadhav, B., Lee, W., Rodriguez, O.L., Martin-Trujillo, A., and Sharp, A.J. (2022). A phenome-wide association study identifies effects of copy-number variation of VNTRs and multicopy genes on multiple human traits. Am. J. Hum. Genet. *109*, 1065–1076. https://doi.org/10.1016/j.ajhg.2022.04.016.
43. Liu, L., Khan, A., Sanchez-Rodriguez, E., Zanoni, F., Li, Y., Steers, N., Balderes, O., Zhang, J., Krithivasan, P., LeDesma, R.A., et al. (2022). Genetic regulation of serum IgA levels and susceptibility to common immune, infectious, kidney, and cardio-metabolic traits. Nat. Commun. *13*, 6859. https://doi.org/10.1038/s41467-022-34456-6.
44. Macdonald, M., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., and Groot, N. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell *72*, 971–983.
45. Crawford, H., Scerif, G., Wilde, L., Beggs, A., Stockton, J., Sandhu, P., Shelley, L., Oliver, C., and McCleery, J. (2021). Genetic modifiers in rare disorders: The case of fragile X syndrome. Eur. J. Hum. Genet. *29*, 173–183.
46. DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J., Nicholson, A.M., Finch, N.A., Flynn, H., Adamson, J., et al. (2011). Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of Causes Chromosome 9p-Linked FTD and ALS. Neuron *72*, 245–256. https://doi.org/10.1016/j.neuron.2011.09.011.
47. Course, M.M., Sulovari, A., Gudsnuk, K., Eichler, E.E., and Valdmanis, P.N. (2021). Characterizing nucleotide variation and expansion dynamics in human-specific variable number tandem repeats. Genome Res. *31*, 1313–1324. https://doi.org/10.1101/gr.275560.121.

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

48. Durinovic-Belló, I., Wu, R.P., Gersuk, V.H., Sanda, S., Shilling, H.G., and Nepom, G.T. (2010). Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin. Genes Immun. *11*, 188–193.

49. Motzo, C., Contu, D., Cordell, H.J., Lampis, R., Congia, M., Marrosu, M.G., Todd, J.A., Devoto, M., and Cucca, F. (2004). Heterogeneity in the magnitude of the insulin gene effect on HLA risk in type 1 diabetes. Diabetes *53*, 3286–3291. https://doi.org/10.2337/diabetes.53.12.3286.

50. Jia, X., Shanmugam, C., Katkoori, V., Jhala, N., Callens, T., Messiaen, L., Grizzle, W., and Manne, U. (2007). Loss of heterozygosity at 17p13. 3 and 17p13. 1 loci is associated with poor survival of African Americans with colorectal adenocarcinomas. Cancer Epidemiol. Biomark. Prev. *16*, B101.

51. Zhao, F., Li, Q., Chen, W., Zhu, H., Zhou, D., Reinach, P.S., Yang, Z., He, M., Xue, A., Wu, D., et al. (2022). Dysfunction of VIPR2 leads to myopia in humans and mice. J. Med. Genet. *59*, 88–100. https://doi.org/10.1136/jmedgenet-2020-107220.

52. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.

53. Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics *32*, 292–294.

54. Danecek, P., and McCarthy, S.A. (2017). BCFtools/csq: haplotype-aware variant consequences. Bioinformatics *33*, 2037–2039.

55. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

56. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. bioinformatics *25*, 2078–2079.

57. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/gr.229102.

58. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

59. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformatics *26*, 139–140.

60. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/nar/gkw377.

61. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J R Stat Soc B *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.

62. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.

63. Smit, A.F. (2004). Repeat-masker open-3.0. http://www.repeatmasker.org.

64. Zhou, H.J., Li, L., Li, Y., Li, W., and Li, J.J. (2022). PCA outperforms popular hidden variable inference methods for molecular QTL mapping. Genome Biol. *23*, 210–217. https://doi.org/10.1186/s13059-022-02761-4.

65. Heger, A., Webber, C., Goodson, M., Ponting, C.P., and Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. Bioinformatics *29*, 2046–2048. https://doi.org/10.1093/bioinformatics/btt343.

66. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

67. Zhang, J., Zhang, Y., You, Q., Huang, C., Zhang, T., Wang, M., Zhang, T., Yang, X., Xiong, J., Li, Y., et al. (2022). Highly enriched BEND3 prevents the premature activation of bivalent genes during differentiation. Science *375*, 1053–1058. https://doi.org/10.1126/science.abm0730.

68. Munde, M., Poon, G.M.K., and Wilson, W.D. (2013). Probing the Electrostatics and Pharmacological Modulation of Sequence-Specific Binding by the DNA-Binding Domain of the ETS Family Transcription Factor PU.1: A Binding Affinity and Kinetics Investigation. J. Mol. Biol. *425*, 1655–1669. https://doi.org/10.1016/j.jmb.2013.02.010.

69. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). GENCODE 2021. Nucleic Acids Res. *49*, D916–D923.

70. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. *83*, 132–139. https://doi.org/10.1016/j.ajhg.2008.06.005.

71. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492. https://doi.org/10.1038/nprot.2017.124.

72. Boix, C.A., James, B.T., Park, Y.P., Meuleman, W., and Kellis, M. (2021). Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature *590*, 300–307. https://doi.org/10.1038/s41586-020-03145-z.

73. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173.

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Phosphate buffer saline (PBS)	Roche	Cat#11666789001
Dulbecco's Modified Eagle Medium (DMEM)	Gibco	Cat#C11995500BT
Fetal Bovine Serum	Invitrogen	Cat#26140095
Trypsin-EDTA	Gibco	Cat#25300054
Penicillin/streptomycin	Corning	Cat#30-002-CI
Lipofectamine 3000	Invitrogen	Cat#L3000015
PAGE Gel Fast Preparation Kit	EpiZyme	Cat#PG110
EMSA/Gel-Shift Binding Buffer,5X	Beyotime	Cat#GS005
EMSA/GEL-SHIFT Loading Buffer (Colorless, 10X)	Beyotime	Cat#GS006
P20	Cytiva	Cat#BR100054
EDTA	Invitrogen	Cat#AM9912
NaCl (5 M)	Invitrogen	Cat#AM9760G
di-Sodium hydrogen phosphate dodecahydrate	Aladdin	Cat#D431179
Sodium phosphate monobasic monohydrate	Aladdin	Cat#S431207
His_6- PU.1 ETS protein lyophilized powder	Zoonbio Biotechnology	N/A
Critical commercial assays		
Dual-Luciferase Reporter Assay System	Promega	Cat#E1910
Deposited data		
NyuWa WGS data	Zhang et al.[20–24]	https://ngdc.cncb.ac.cn/gsa-human/browse/HRA004185
1000 Genomes Project WGS data	Byrska-Bishop et al.[52]	https://www.internationalgenome.org/data-portal/data-collection/30x-grch38
VNTR-LP & VNTR-MP Statistics	This manuscript	http://bigdata.ibp.ac.cn/NyuWaVNTR/
Human Genome Diversity Project WGS data	Bergström et al.[26]	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGDP/
HGSVC2 long-read assemblies	Ebert et al.[34]	https://www.internationalgenome.org/data-portal/data-collection/hgsvc2
Geuvadis	Lappalainen et al.[28]	https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/geuvadis/working/geuvadis_topmed/
danbing-tk reference VNTR set	Lu et al.[15]	https://github.com/ChaissonLab/danbing-tk
Experimental models: Cell lines		
HEK293T cell	ATCC	Cat#CRL-11268
Oligonucleotides		
5′FAM 1CM dsDNA	Tsingke Biotech	N/A
5′Biotin 1CM dsDNA	Tsingke Biotech	N/A
1CM dsDNA	Tsingke Biotech	N/A
5′FAM 2CM dsDNA	Tsingke Biotech	N/A
5′Biotin 2CM dsDNA	Tsingke Biotech	N/A
2CM dsDNA	Tsingke Biotech	N/A
Recombinant DNA		
pcDNA3.1-PU.1	Tsingke Biotech	N/A
pGL3-promoter-1CM	Tsingke Biotech	N/A
pGL3-promoter-2CM	Tsingke Biotech	N/A
pGL3-promoter	Tsingke Biotech	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pRL-TK	BNCC	Cat#BNCC354583
Software and algorithms		
Qualimap	Okonechnikov et al.[53]	http://qualimap.conesalab.org/
Bcftools	Danecek et al.[54]	http://samtools.github.io/bcftools/
GATK	DePristo et al.[55]	https://gatk.broadinstitute.org/hc/en-us
SAMtools	Li et al.[56]	http://www.htslib.org/
danbing-tk	Lu et al.[15]	https://github.com/ChaissonLab/danbing-tk
The Human Genome Browser at UCSC	Kent et al.[57]	https://genome.ucsc.edu/index.html
Scikit-learn	Scikit-learn	https://www.scikitlearn.com.cn/
FeatureCounts	Liao et al.[58]	https://subread.sourceforge.net/
EdgeR	Robinson et al.[59]	https://bioconductor.org/packages/release/bioc/html/edgeR.html
Enrichr	Kuleshov et al.[60]	https://maayanlab.cloud/Enrichr/
SusieR	Wang et al.[61]	https://github.com/cran/susieR
FIMO	Grant et al.[62]	https://meme-suite.org/meme/tools/fimo
ggplot2	ggplot2	https://cran.r-project.org/web/packages/ggplot2/index.html
RepeatMasker	Smit et al.[63]	http://repeatmasker.org
PCAForQTL	Zhou et al.[64]	https://github.com/heatherjzhou/PCAForQTL
GAT	Heger et al.[65]	http://code.google.com/p/genomic-association-tester.
Canu	Koren et al.[66]	https://github.com/marbl/canu
R	R Core	https://www.r-project.org/
Python	Python Software Foundation	https://www.python.org/
Main analysis scripts and used files in this paper	This manuscript	https://doi.org/10.5281/zenodo.14003180
Other		
Nanopore sequencing service	Beijing Polyseq Biotech Co. Ltd.	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Subjects and specimens

Whole blood DNA samples from 4,129 Chinese individuals, including those with diabetes and those without, were collected for this study. The research was approved by the Medical Research Ethics Committee at the Institute of Biophysics, Chinese Academy of Sciences. All participants provided written informed consent. They were informed that their samples would be used for genomic research conducted by the Chinese Academy of Sciences. The consent form was designed for individuals aged 30 to 70, encompassing both patients and healthy individuals with full legal capacity. Participants willingly provided their blood samples, medical history, and signed the consent form. Their personal information was strictly confidential. The participants had the right to refuse to donate samples or to withdraw from the study at any point.

Cell lines

HEK293T cells were cultured in Dulbecco's Modified Eagle's Medium with 100 μg/ml of penicillin, 100 μg/mL of streptomycin, and 10% fetal bovine serum (FBS). The cells were grown at a temperature of 37°C with 5% CO_2 in a humidified incubator.

METHOD DETAILS

Electrophoretic mobility shift assay (EMSA)

Our EMSA experiments were modified based upon Zhang J. et al.[67] EMSAs were performed using 5′-FAM-labeled double-stranded DNA (dsDNA) probes and purified PU.1 ETS domain protein (amino acid residue 165–270). Briefly, the indicated recombinant His_6-PU.1 ETS proteins were expressed in *Escherichia coli* strain BL21 (DE3) and purified by His-tag affinity purification. For each DNA-protein binding reaction, the 5′-FAM-labeled dsDNA probes (5 μM, 1 μL) were incubated with 0.17 μg/μL His_6-PU.1 ETS protein (1 μL) in EMSA/Gel-Shift Binding Buffer (Beyotime, GS005) at a temperature of 25°C for 1 h. Additionally, for competitive binding assays, unlabeled probes were added to the reaction mixtures at an indicated molar 16-fold excess compare to the FAM labeled probes. The

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

CellPress
OPEN ACCESS

6% native polyacrylamide gel was pre-electrophoresed run at 160 V in 0.5× TBE buffer at 4°C for 2 h, then 10 μL DNA-protein mixture with 1.1 μL of gel shift loading buffer (Beyotime, GS006) were subjected to gel and run at 100 V in 0.5× TBE buffer at 4°C for 1.5 h. The gels were scanned by Typhoon 7000 (GE Healthcare). Complementary DNA oligonucleotides of 1CM and 2CM were synthesized by Beijing Tsingke Biotech Company and labeled with FAM at the 5′ end. The nucleotide sequences of double-stranded oligonucleotides were as follows: 5′-CCTCCTCTTCCTCTCCCAGGCCTCA-3′, and 5′- CCTCCTCTTCCTCTCCCAGGCCTCACCTCCTCTTCC TCTCCCAGGCCTCA-3′ for the 1CM and 2CM probe sequence, respectively.

Surface plasmon resonance (SPR) measurements

The kinetics of PU.1 ETS binding to DNA probes and purified PU.1 proteins were evaluated by SPR technology on a Biacore 8000 instrument (GE Healthcare). Briefly, biotin-labeled 1CM and 2CM DNA probes were captured on an SA sensor chip (Cytiva). To correct for instrumental and concentration effect, a blank flow cell was used as blank control. Solutions for SPR experiments were prepared with 25 mM sodium phosphate buffer (Na_2HPO_4/NaH_2PO_4) of pH 7.4, 1 mM EDTA, 0.05% P20, and 300 mM NaCl at 25°C.[68] PU.1 ETS-domain protein with various concentrations were injected over the 1CM and 2CM probes surface and blank flow cell at a flow rate of 100 μL/min for 120s, and dissociated in running buffer for 120 s. Equilibrium and kinetic constants were calculated using a global fit to 1:1 Langmuir binding model with the Biacore 8000 evaluation software (GE Healthcare).

Cell culture, reporter gene constructs, and dual-luciferase reporter assays

We used transient transfection to investigate the effects of the VNTR motif sequence as a predictive enhancer on the transcription activity of reporter genes. HEK293T cells were cultured in Dulbecco's Modified Eagle's Medium with 100 μg/ml of penicillin, 100 μg/mL of streptomycin, and 10% fetal bovine serum (FBS). The cells were grown at a temperature of 37°C with 5% CO_2 in a humidified incubator. The Firefly luciferase reporter constructs were 37°C with 5% CO_2 in a humidified incubator. The Firefly luciferase reporter constructs were generated by inserting 1CM and 2CM sequence upstream of the promoter of pGL3-promoter (Promega). For transient transfection experiments, cells were seeded onto 24-well plates (20,000 cells per well), and simultaneously transfected with constructed luciferase vector containing the vector DNA (250 ng) with different alleles and PU.1 expression plasmids (250 ng). As an internal standard, all plasmids were co-transfected with 50 ng of the pRL-TK Renilla luciferase plasmid (Promega) using Lipofectamine 3000 (Invitrogen, USA) according to manufacturer's protocol. The pGL3-promoter vector without an insert was used as a negative control. 24-h post-transfection, Firefly and Renilla luminescence was measured by Dual-Luciferase Reporter Assay System (Promega) according to manufacturer's protocol. The relative luciferase activity was calculated by normalized against empty pGL3-promoter. Independent triplicate experiments were conducted for each plasmid.

QUANTIFICATION AND STATISTICAL ANALYSIS

Datasets processing

The NyuWa Genome resource contained whole-genome sequencing data for 4,129 participants that were recruited from 25 administrative divisions across the China. Among them, 4,013 unrelated samples were reported in our previous study[20–22] and 116 samples were newly sequenced. All participants provided written informed consent. Genomic DNA was extracted and sequenced according to the standard protocols of Illumina on HiSeq X10 platform or NovaSeq 6000, and the sequencing reads were paired-end 150 nt. The mean depth of NyuWa cohort evaluated by Qualimap v2.2.1[53] was about 30.5×. The raw sequencing reads were processed and mapped to GRCh38 followed GATK[55] Best Practices Workflows Germline short variant discovery pipeline. The sexes of all samples were inferred by BCFtools v1.5[54] guess-ploidy module. We also employed 3,166 high-coverage samples from 1KGP and 930 sample from HGDP were employed. The CRAM files of 1KGP samples were downloaded from https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/data/and https://ftp-trace.ncbi.nlm.nih.gov/1000genomes/ftp/1000G_2504_high_coverage/additional_698_related/with reads aligned to GRCh38 and each CRAM file was then converted to BAM file using SAMtools v1.9.[56] The HGDP CRAM files on GRCh38 were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGDP/and converted to BAM files using SAMtools v1.9. More detailed sample information can be found in Table S2.

Genome-wide VNTR genotyping

We adapted a repeat-pangenome graph-based method danbing-tk v1.3[14] to genotype VNTRs. For VNTR length calculation of short-read sequencing (SRS) data, we downloaded reference files "pan.tr.mbe.v2.bed.gz" and "RPGG.tar.gz" from https://zenodo.org/records/5093660. The file "pan.tr.mbe.v2.bed.gz" contained VNTR coordinates of the 35 HGSVC genomes over 80,518 loci, and "RPGG.tar.gz" was the repeat-pangenome graph built from the VNTR annotations. Using "danbing-tk align" module with the options "-gc 80 -ae -kf 4 1 -cth 45 -k 21 -qs pan -fai/dev/stdin -p 4", we obtained the k-mer counts of each VNTR locus for each sample. To detect the deviation of an observed read depth from the expected value within VNTR regions, we computed locus-specific bias (LSBs) (see "Methods" in Lu et al.[14]) of 80,518 VNTR loci and 397 non-VNTR loci for each sample as the condition of VNTR length

estimation. To facilitate comprehension, we re-explained the definition of LSBs here. LSBs is a tuple of (genome g, sequencing run), as follows:

$$b_s = \frac{kms_s}{cov_s \times L_g} \quad \text{Equation (1)}$$

where L_g is the ground truth VNTR lengths of 80,518 loci in genome g; kms_s is the sum of k-mer counts in each locus mapped by sample s; cov_s is the global read depth of sample s estimated by averaging the read depths of 397 non-VNTR regions without any types of repeats or duplications. The ground truth VNTR length of a locus l in genome g is averaged across haplotypes:

$$L_{g,l} = \frac{1}{H}\sum_{h=1}^{H} L_{g,h,l} \quad \text{Equation (2)}$$

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.

We then used 'bedcov' module of SAMtools to get the coverage of 80,518 VNTR regions and 397 non-VNTR regions. Taken LSBs, coverage files and bed files of these VNTR and non-VNTR regions as input, we ultimately utilized the script 'kmc2length.py' to estimate the VNTR lengths for all 8,225 samples.

For the motif dosage calculation, we followed the method described by Lu et al.[15] The corresponding toolkit and files are available from both the https://github.com/ChaissonLab/danbing-tk and our website http://bigdata.ibp.ac.cn/NyuWaVNTR. The original reference file contains a total of 4,456,881 motifs. This file listed all VNTR loci and their corresponding motifs, where the first column was the serial number, the second column was the index value of each VNTR locus and each number represents a VNTR locus, and the third column listed all motifs observed at each locus across 35 HGSVC2 assemblies. For each sample, we extracted k-mers by sliding the 21-bp window on each motif and averaged their k-mer counts to obtain the dosage of each motif. For the 35 HGSVC2 assemblies, we initially utilized the 'danbing-tk build' module to calculate k-mer counts. Following this, we applied the identical procedure to calculate motif dosages for the assemblies. To facilitate the comparability of polymorphic motif dosage between various datasets, we normalized the motif dosage in each individual by dividing it by the reference length of the motif, and rounded it to the nearest integer, referred to as VNTR motif polymorphisms (VNTR-MPs).

Quality control of VNTR length and motif dosage sets

To ensure the reference VNTR set mainly consisted of simply tandem repeats instead of VNTRs in mobile elements, we removed 40,199 loci which overlapped with mobile elements detected by RepeatMasker v4.1.2-p1 (http://repeatmasker.org). We mainly focused on 22 autosomes and kept 38,866 loci for subsequent analysis. As danbing-tk suggested that bias correction was needed for the stochastic nature of read coverage or unknown technical biases, we performed the bias correction on VNTR lengths of 8,225 samples and calculated batch-r^2 to measure the performance of the correction. Batch-r^2 was defined as the correlation coefficient r^2 between VNTR size in 35 HGSVC assemblies and estimated VNTR length in corresponding 35 short-read sequencing genomes. After correction, an evident improvement of VNTR length correlation was observed (Figure S14A). The number of VNTR loci with correlation coefficient exceeding 0.8 was increased from 9,430 to 14,517. Furthermore, we examined the estimate rate of each VNTR locus and each sample. The 'locus estimation rate' indicates the proportion of genotypes that have been successfully called among all samples for each locus. The 'Sample estimation rate' indicates the proportion of genotypes that have been successfully called among all loci for each sample. Total 38,685 loci with estimate rate above 80% (Figure S14B) and 8,222 samples with estimate rate above 50% (Figure S14C) were retained.

We further applied three steps to control the quality of motif dosage set. We firstly used mean absolute percentage error (MAPE) to evaluate the relative error between the dosage from short reads and the dosage from the long reads assemblies for each motif, where lower MAPE indicated more accurate prediction of motif dosage for short reads. 1,161,890 motifs exhibited MAPE values less than 0.5 (Figure S15A), which were retained for subsequent analyses. Then, 148,245 motifs with dosage invariant across HGSVC2 haplotypes was removed. We further measured the estimate rate of each motif across 8,222 genomes and total 96,707 with rate below 20% were removed (Figure S15B). After filtering, we also observed a large consistency in motif dosage between the SRS data and assemblies (Figure S15C), with over 74% of motifs (680,216/916,938) exhibiting the batch-r^2 above 0.8. As a result, the final motif dosage set consisting of 916,938 motifs was generated. The set of 916,938 high-quality VNTR motifs can be obtained from our website (http://bigdata.ibp.ac.cn/NyuWaVNTR).

Comparisons of VNTR length and motif polymorphism

As danbing-tk was sensible for modeling read-depth-based (continuous) estimates of diploid VNTR length and VNTR motif content, the genotypes of VNTR length and VNTR motifs were real-valued quantities rather than integer-valued numbers. To improve clarity of presentation, we rounded VNTR length to the nearest integer and VNTR with different discrete lengths across samples were deemed as VNTR length polymorphisms (VNTR-LPs). As for VNTR motif polymorphisms (VNTR-MPs), we transformed each motif dosage into the copy number by dividing the motif dosage by the length of the motif sequence and rounding to the nearest integer, and VNTR motifs with diverse copy numbers across samples were deemed as VNTR-MPs. We have constructed a website to allow users to

Please cite this article in press as: Zhang et al., Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits, Cell Genomics (2024), https://doi.org/10.1016/j.xgen.2024.100699

query and browse polymorphism information across 8,222 samples via http://bigdata.ibp.ac.cn/NyuWaVNTR. To evaluate the VNTR-MPs consistency between NyuWa and 1KGP.EAS, we performed a random sampling from the NyuWa dataset for ten times, each time selecting a number of samples equivalent to that of the 1KGP.EAS population. We also we conducted a pairwise comparison between the NyuWa dataset and each subpopulation within 1KGP.EAS, ensuring that the sample sizes remain consistent across all comparisons. In addition, we excluded the EAS samples from 1KGP and HGDP (termed as non EAS-1KGP & HGDP) to examine the divergence of VNTR-LPs and VNTR-MPs between NyuWa and non-Asian population.

Identification of eVNTR and eMotif

Public RNA-seq data of 462 unrelated human lymphoblastoid cell line (LCL) samples from Geuvadis consortium were downloaded from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/geuvadis/working/geuvadis_topmed/, where 441 samples from African and European populations were included in 8,222 genomes. We processed 441 RNA-seq data as previously described.[21] In brief, the featureCounts v2.0.3[58] was applied to count reads of genes annotated in GENCODE v34.[69] Then, we applied edgeR v3.32.1[59] to filter lowly expressed genes by "filterByExpr" function and generated log2 normalized FPKM values. To eliminate spurious associations caused by confounding factors, we used known covariates including gender and top 10 principal components (PCs) from the population structure analysis to detect hidden factors for eQTL mapping. The population structure was analyzed by performing principal component analysis (PCA) of SNP genotypes with 441 samples. The SNPs were selected with minor allele frequency $\geq$0.05 and 27 known long-range LD regions[70] were pruned for PCA. After performing PCA by adding known covariates using PCAForQTL v0.1.0,[64] we chose 25 hidden covariates based on the proportion of variance explained of PCs (Figure S16A) and correlations between covariates (Figure S16B). Together with gender, 10 population structure factors and 25 hidden factors, the gene expression matrix was adjusted by a linear model. As a result, fully processed, filtered, and adjusted expression matrix in 441 samples was obtained for subsequent eQTL identification.

For association test between gene expression and dosages, only VNTRs and corresponding motifs within 100 kb of expressed genes were selected. Then, we masked outliers if the VNTR or motif dosage being three or two standard deviations away from the mean for each sample, separately. We further removed each VNTR or motif if there were more than 50% outlier dosages across 441 samples. The final VNTR length matrix contained 23,283 VNTRs and motif dosage matrix contained 556,402 motifs and these two matrixes were normalized by *Z* score. For each gene-dosage pair, the linear regression was fit by applying "glm()" function in R v4.2.3 between gene expression and the VNTR length or motif dosage. *p*-values were obtained from *t*-test on the slope and were adjusted using a Bonferroni correction. We picked the minimal *p*-value of multiple associations at 5% false discovery rate (FDR) for each gene and used "p.adjust" function with Benjamini–Hochberg method. Only one VNTR or one motif was related to each expressed gene. We then controlled the adjusted *p*-values with gene-level FDR at 5% to filter out all significant gene-dosage pairs. As a result, we obtained 543 VNTR-gene pairs corresponding to 439 VNTRs and 2,508 motif-gene pairs corresponding to 1,937 VNTRs. All genes in pairs were denoted as eGenes, and all VNTRs or motifs found in pairs were regarded as eVNTRs or eMotifs separately.

Enrichment analysis

The gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and GWAS Catalog traits enrichment analysis were performed with Enrichr v3.2.[60] The gene set enclosing the VNTRs was taken as input and significant level was set with "p.adjust" value over 5%.

The genomic context enrichment analysis of epigenetic features for eVNTRs and eMotifs were conducted by using available data of GM12878 cell line. We downloaded the CTCF peaks and histone mark peaks of ChIP-seq and ATAC-seq from ENCODE portal with the following accession code: ENCFF796WRU (CTCF), ENCFF998CEU (H3K4me3), ENCFF291DHI (H3K27me3), ENCFF981JOU (H3K9ac), ENCFF023LTU (H3K27ac), ENCFF321BVG (H3K4me1), ENCFF432EMI (H3K36me3), ENCFF283LNH (H3K4me2), ENCFF725UFY (H3K9me3), ENCFF748UZH (ATAC-seq). We also used GM06990 and GM12865 cell lines to provide comparable validations. All data were obtained with the following accession code: ENCFF239QAE (GM06990 DNase-seq), ENCFF031WEA (GM06990 CTCF), ENCFF307CEJ (GM06990 H3K27me3), ENCFF253XQQ (GM06990 H3K4me3), ENCFF884ZOW (GM06990 H3K36me3), ENCFF754VPH (GM12865 DNase-seq), ENCFF541DDH (GM12865 CTCF) and ENCFF438JMP (GM12865 H3K4me3). Additionally, 18 chromatin states data of these three cell lines defined by ChromHMM[71] were downloaded from EpiMap.[72]

We further adopted GAT v1.3.4[65] to quantify the enrichment of eVNTRs and eMotifs in these epigenetic and genomic features. After getting the fold enrichment value and empirical *p*-values in each analysis, *p*-values were adjusted by "p.adjust" function with "BH" method in R and a significant level was set to 5%.

Fine-mapping

To determine whether eVNTRs and eMotifs are likely to represent casual effects on gene expression, we applied the R package susieR v0.12.35[61] to perform fine-mapping analyses. We downloaded 1KGP high-coverage phased genetic data from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/. The genotypes of SNP and indel variants within 1 Mbp window size near the transcription start site of each eGene were extracted. We generated the genotype matrix and was taken together with eVNTRs or eMotifs as input for fine-mapping. We also downloaded the SNP eQTLs

in 445 Geuvadis samples from http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/QTS000013/QTD000110/QTD000110.all.tsv.gz and compared the *p*-values between SNPs and VNTRs. Totally, three eVNTRs and fourteen eMotifs which PIP >0.8 were assigned a high posterior probability of causality. To characterize these likely causal eVNTRs and eMotifs, we utilized UCSC Genome Browser track on GRCh38 for functional annotations.

Prediction of TF binding sites with eMotifs

TF binding sites of eMotifs were predicted with FIMO v4.12.0.[62] We downloaded distribution matrix of binding sites for all transcription factors of homo sapiens from CORE set in JASPAR[73] database.

Hypervariable VNTR motifs calculation

To identify the hypervariable VNTR motifs within each superpopulation, we computed the coefficient of variation (CV) value for every motif within each superpopulation, and motif with top 30% CV were considered as hypervariable VNTR motifs. CV was calculated as $CV = \frac{S_p}{M_p} \times 100\%$, where S_p represents the standard deviation of motif dosage among all individuals in the superpopulation, and M_p represents the average of motif dosage. Before calculation, motif with callrate lower than 50% in any superpopulation were eliminated.

Principal component analysis

We firstly calculated V_{ST} to represent VNTR motif dosages with significant divergence among superpopulations.

$$V_{ST} = 1 - \frac{\sum_P N_P V_P}{N_T V_T}$$

The result of this formula represents the degree of difference in motif dosage at a locus between superpopulations, where V_T represents the variance of motif dosage among all unrelated individuals, V_P represents the variance of motif dosage among all unrelated individuals in each superpopulation, N_T represents the number of all unrelated individuals, N_P represents the number of all unrelated individuals in each superpopulation. According to danbing-tk, values below 1 are excluded during the variance calculation, any computed V_{ST} below 0 are treated as 0. For a population, the VNTR motifs with V_{ST} more than three standard deviations above the mean were considered as high V_{ST} motifs.

We then used high V_{ST} motifs to perform PCA. PCA model was trained on samples from the 1KGP through the PCA function from the sklearn.decomposition module in Python. To match the sample number of NyuWa populations to the sample number of 1KGP populations, 200 randomly selected individuals from NyuWa dataset and all samples from 1KGP datasets were projected onto the PCA models. Sampling procedure was repeated three times to ensure the stability of the PCA results.

Comparison of lengths for human-specific expansion VNTRs across superpopulations

To identify human-specific expansion (HSE) VNTRs with significant different in mean length across superpopulations, we compared 467 HSE VNTRs from Course et al.[47] with 38,685 loci in our dataset and found 360 overlapped ones. For each locus, we used Wilcoxon rank-sum test to compare the VNTR length distributions between each two superpopulations. We adjusted the *p*-values using the "p.adjust" function in R with "BH" method. Comparisons with adjusted $p < 0.01$ were considered to be significant. The results were shown using volcano plots.

Validation using nanopore targeted sequencing

To validate our findings using nanopore targeted sequencing, we followed the approach outlined in the danbing-tk tool. Five samples from the NyuWa dataset were selected and sent to Beijing Polyseq Biotech Co. Ltd. for sequencing, targeting five specific loci. The sequencing was conducted using the PolyseqOne nanopore sequencer. The target loci were first amplified using PCR, with an average amplification region size of 2,000 base pairs (bp). The sequencing results yielded an average read length of 1,000 bp and an average depth ranging from 1,000× to 2,000×.

Post-sequencing, all reads with a quality score below Q10 were filtered out. The remaining sequencing data were assembled using Canu[66] with the parameters: genomeSize = 100,000, readSamplingCoverage = 100, contigFilter = "2 0 1.0 0.5 0", and -nanopore-raw. The assembled sequences were subsequently processed using the GoodPanGenomeGraph.snakefile workflow from the danbing-tk tool to extract k-mer counts at each locus. Loci with sequencing depths below 100× or above 10,000× were excluded from further analysis. For the remaining loci, k-mer counts were normalized by dividing by the sequencing depth to estimate VNTR lengths from the third-generation sequencing data.

The Pearson correlation coefficient was calculated between these third-generation sequencing results and the corresponding second-generation sequencing genotype data, demonstrating that the danbing-tk method is relatively well-suited for analyzing the NyuWa reference genome. Detailed statistical results are presented in Table S3.

Cell Genomics, Volume *4*

Supplemental information

Genome-wide investigation of VNTR motif polymorphisms in 8,222 genomes: Implications for biological regulation and human traits

Sijia Zhang, Qiao Song, Peng Zhang, Xiaona Wang, Rong Guo, Yanyan Li, Shuai Liu, Xiaoyu Yan, Jingjing Zhang, Yiwei Niu, Yirong Shi, Tingrui Song, Tao Xu, and Shunmin He

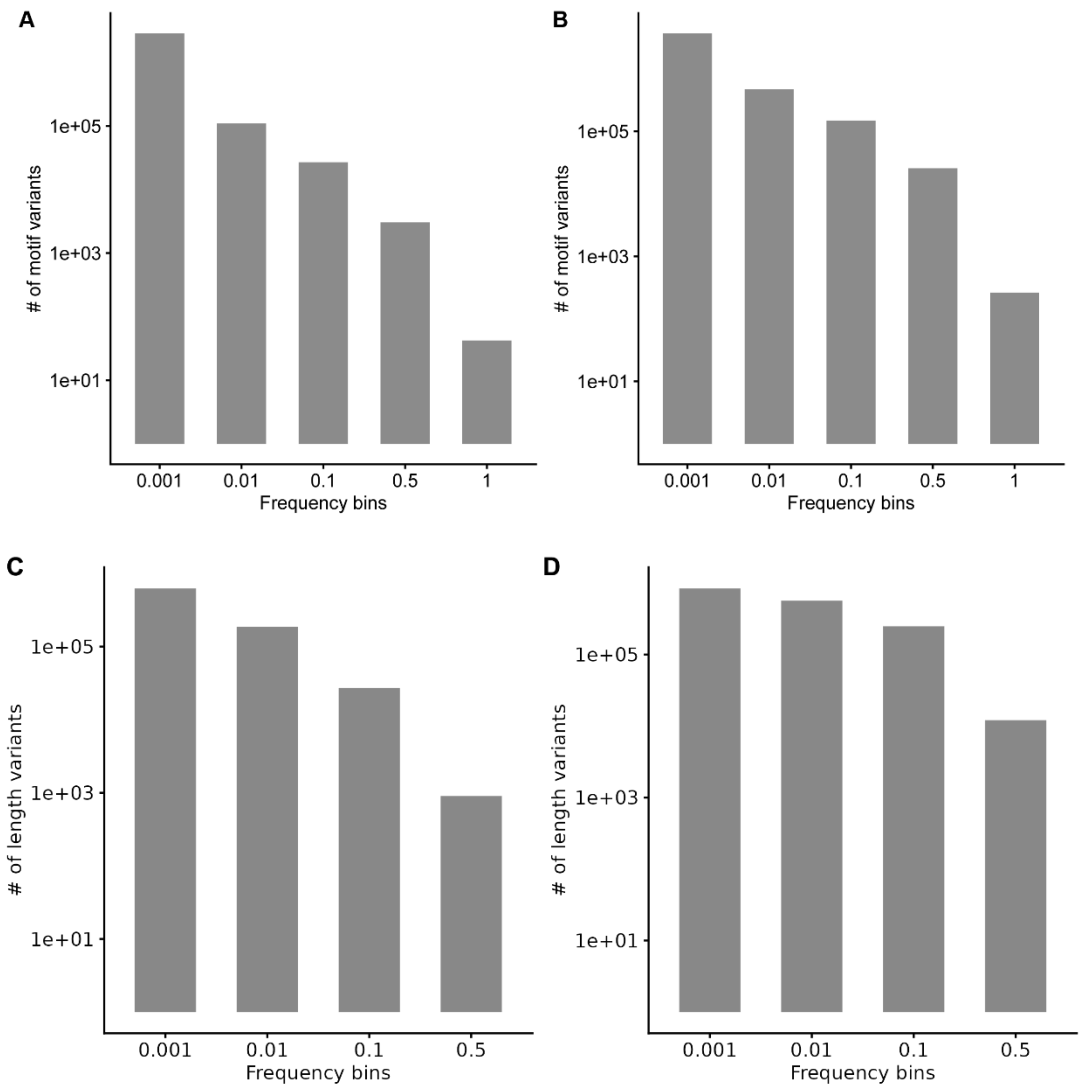


Figure S1. Frequency distribution of NyuWa-specific VNTR-MPs and VNTR-LPs, Related to Figure 1.

(A) Frequency distribution of NyuWa-specific VNTR-MPs when compared to the 1KGP and HGDP datasets. **(B)** Frequency distribution of NyuWa-specific VNTR-MPs when compared to the EAS population from the 1KGP. **(C)** Frequency distribution of NyuWa-specific VNTR-LPs when compared to the 1KGP and HGDP datasets. **(D)** Frequency distribution of NyuWa-specific VNTR-LPs when compared to the EAS population from the 1KGP.The allele frequencies were cut into five bins: [0, 0.001), [0.001, 0.01), [0.01, 0.1), [0.1, 0.5), and [0.5, 1).

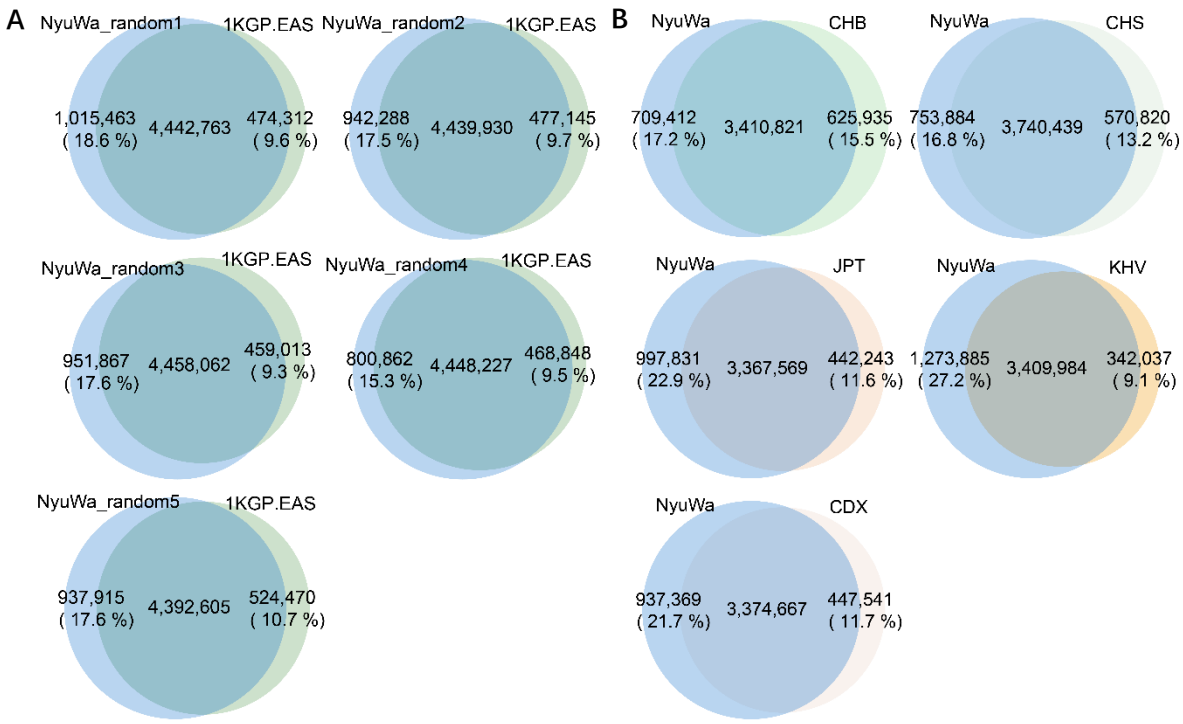


Figure S2. Comparison of VNTR motif polymorphisms (VNTR-MPs) between random-selected NyuWa subsets and other datasets, Related to Figure 1.

(A) Comparison between random selected NyuWa subsets and EAS population from 1KGP. **(B)** Comparison between random selected NyuWa subsets and five subpopulations of EAS population from 1KGP.

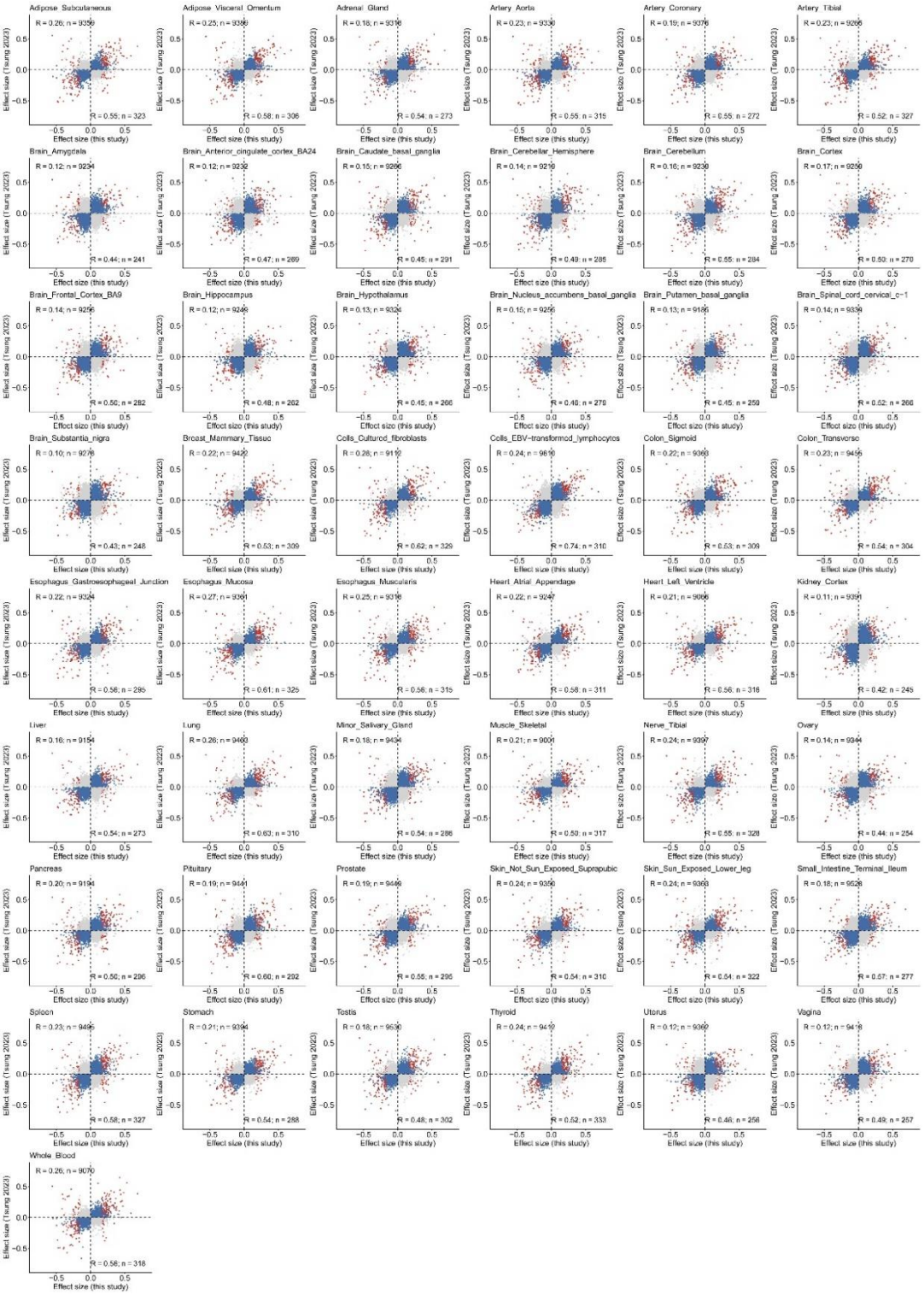


Figure S3. Effect size of eVNTRs identified in this study and a previous study by Lu et al [1], Related to Figure 2.

Correlations of the effect size of eVNTRs identified in this study and a previous study by Lu et al. in GTEx dataset. Each subplot represents a tissue in GTEx. The blue points indicate eVNTRs whose directions of effect were concordant in two studies, and the gray points denote eVNTRs with discordant directions of effect for that eVNTR. The eVNTRs detected in both studies are colored red, regardless of the concordance of effect.

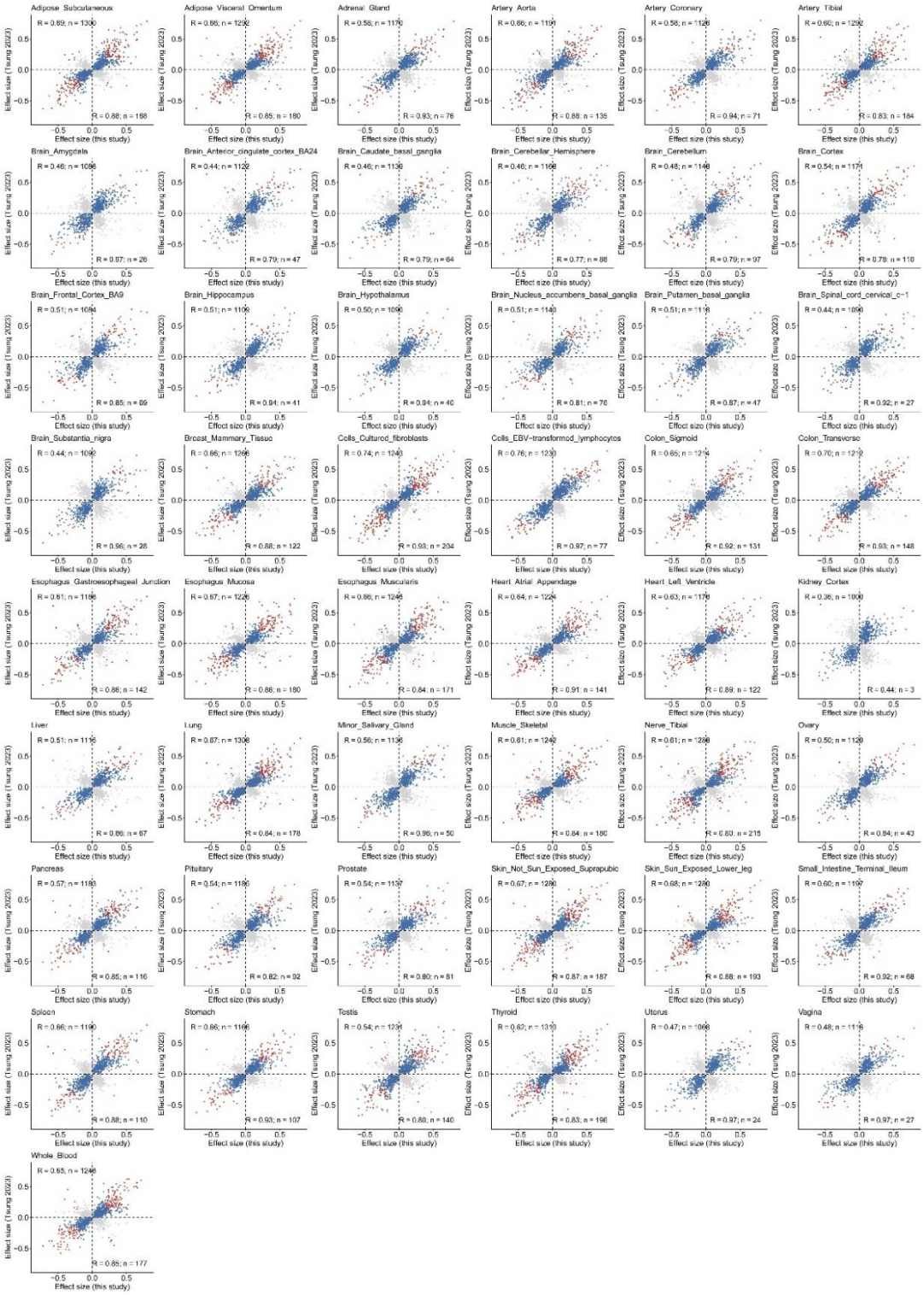


Figure S4. Effect size of eMotifs identified in this study and a previous study by Lu et al[1], Related to Figure 1.

Correlations of the effect size of eMotifs identified in this study and a previous study by Lu et al. in GTEx dataset. Each subplot represents a tissue in GTEx. The blue points indicate eMotifs whose directions of effect were concordant in two studies, and the gray points denote eMotifs with discordant directions of effect for that eMotifs. The eMotifs detected in both studies are colored red, regardless of the concordance of effect.

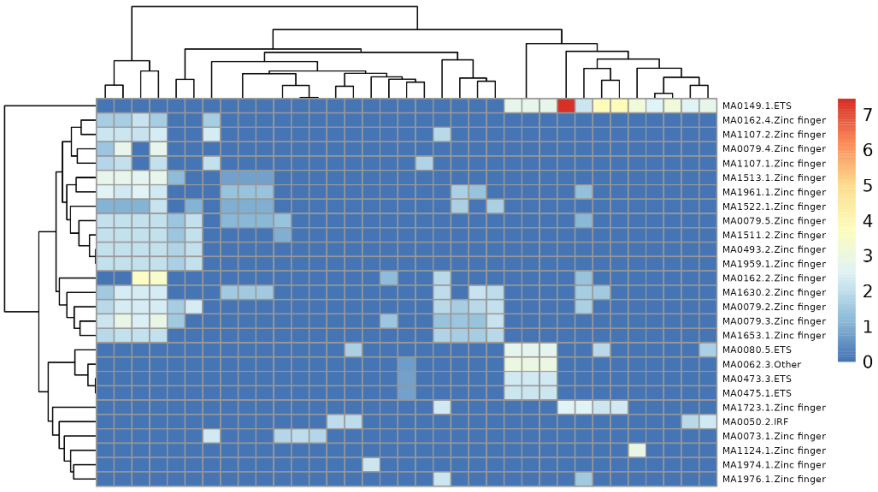


Figure S5. Predicted binding transcription factors for eMotifs in the promoter region, Related to Figure 2.

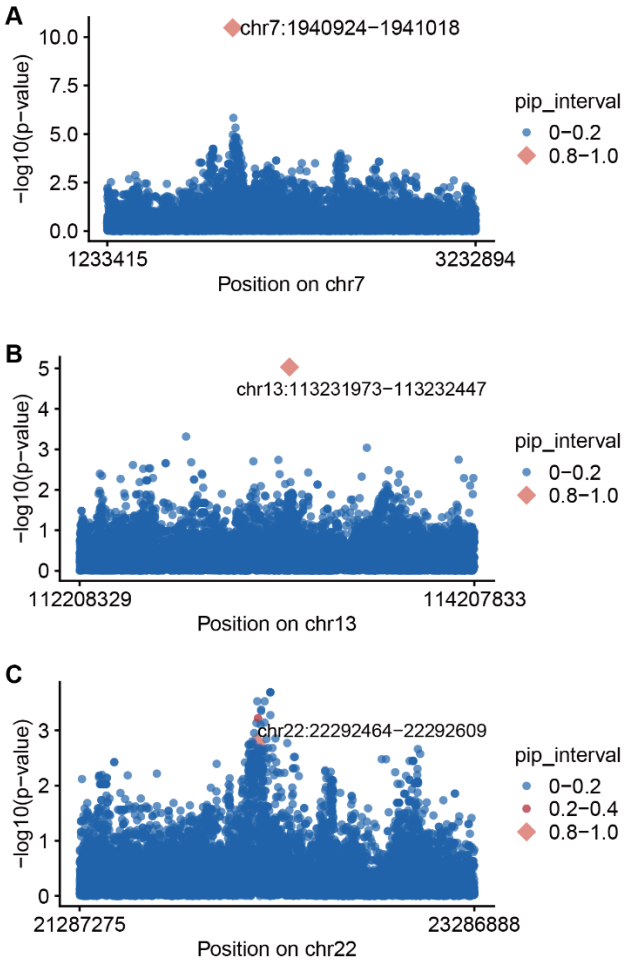


Figure S6. *P*-values of three fine-mapped eVNTRs and nearby SNP eQTLs, Related to Figure 3.

Dot plot of the eVNTR in the locus of chr7:1940924 (**A**), chr13:113231973 (**B**) and chr22:22292464 (**C**). Blue dots indicate the *P*-values of eVNTRs and nearby SNP eQTLs. Shapes represent the pip intervals for each VNTR or SNPs. *P*-values were calculated by two-sided t-test in the linear model.

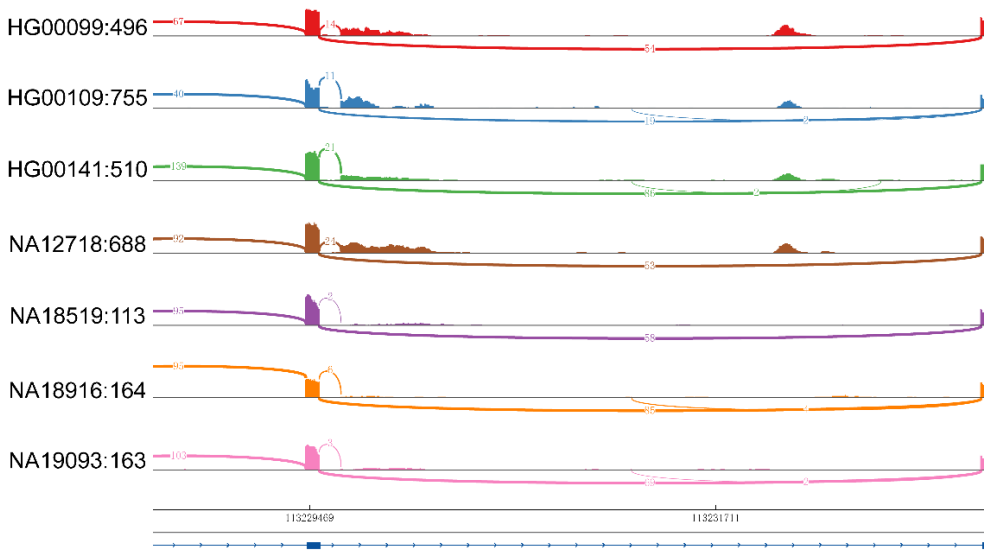


Figure S7. Sashimi plot of the eVNTR in the locus of chr13:113231973 with alternative splicing events, Related to Figure 3.

The x-axis indicates genomic locations, and the y-axis indicates transcription intensity. Each panel represents each sample in the Geuvadis project. SampleID: eVNTR length in chr13:113231973 was shown on the left.

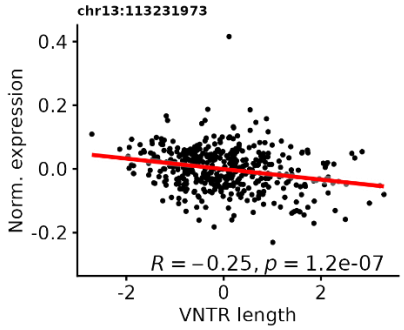


Figure S8. Correlations between the length of the eVNTR in chr13:113231973 and the expression of *CUL4A*, Related to Figure 3.

The red line indicates the best fit under simple linear regression.

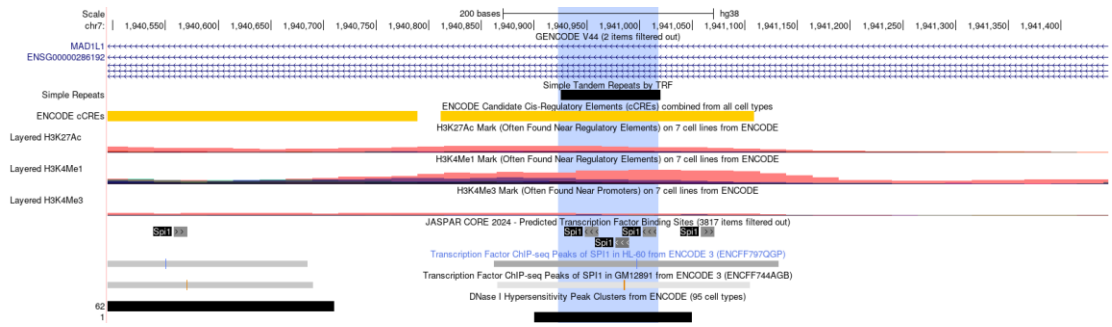


Figure S9. UCSC Genome Browser view[2] of VNTR *MAD1L1*, Related to Figure 3.

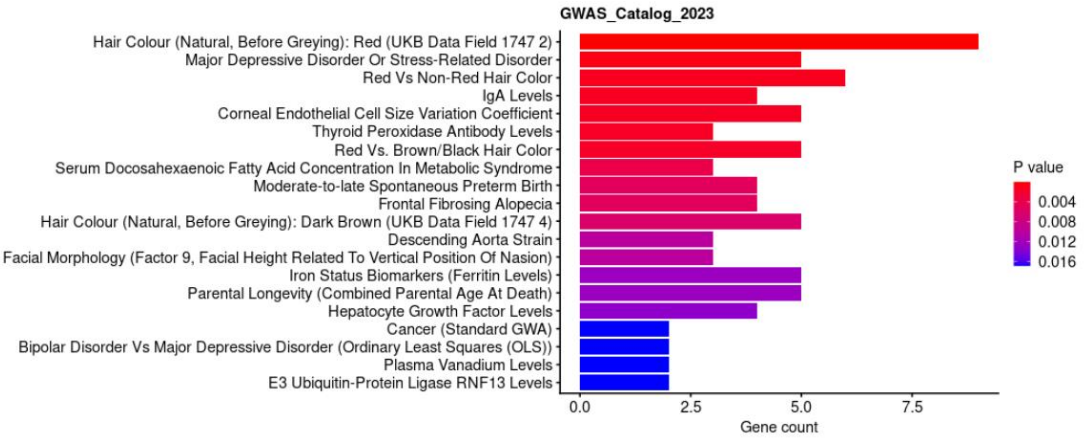


Figure S10. GWAS Catalog trait enrichments of exonic VNTRs, Related to Figure 4.

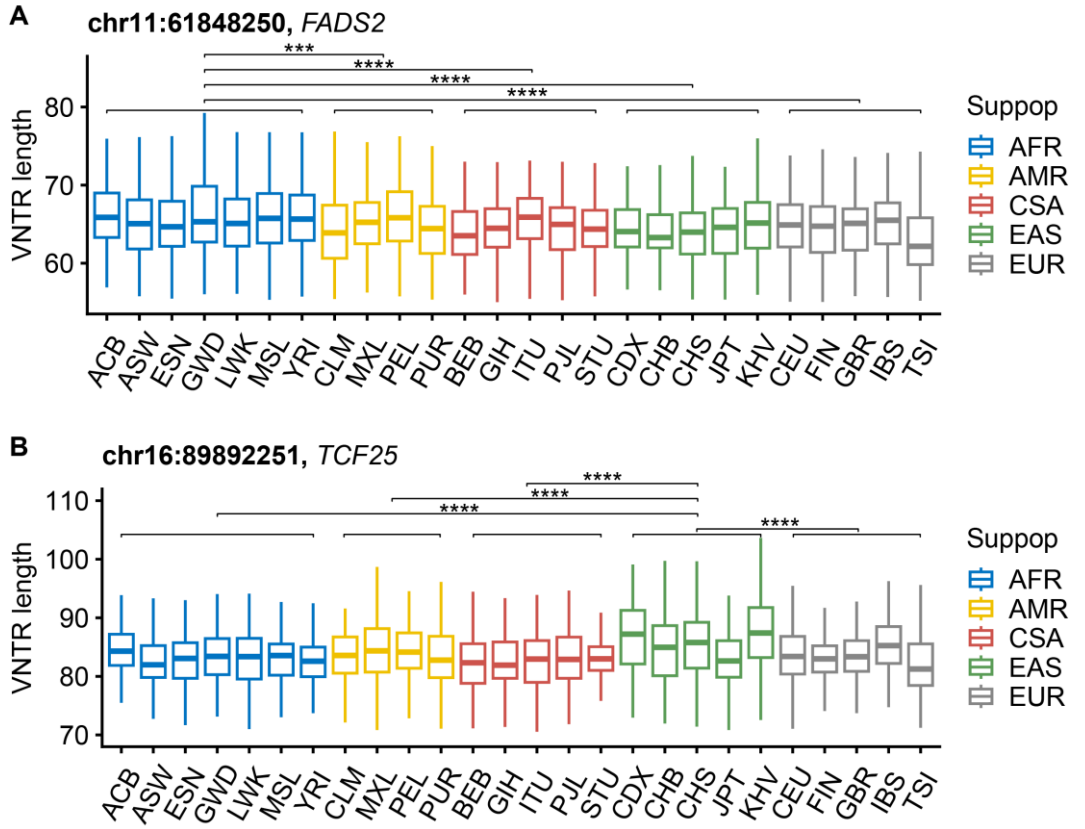


Figure S11. Distributions of lengths for two exonic VNTRs significantly differences across populations related to IgA levels (A) and hair color (B), Related to Figure 4.

P-values were computed using the two-sided Wilcoxon rank sum test. ***, $P < 0.001$; ****, $P < 0.0001$.

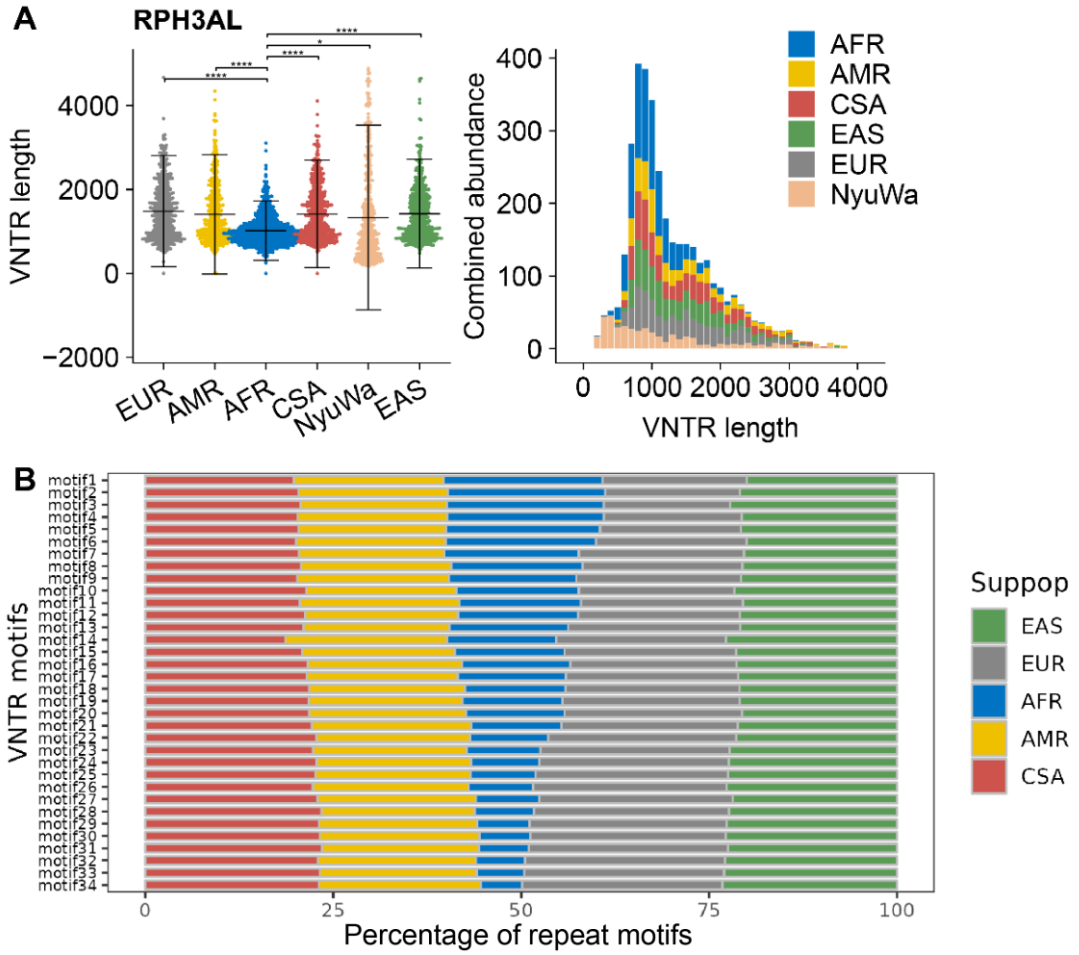


Figure S12. Distributions of lengths and motif frequency for human-specific expansion VNTR *RPH3AL* differed across superpopulations, Related to Figure 5.

(A) Distribution of VNTR length for VNTR *RPH3AL* in NyuWa and five superpopulations. P-values were computed using the two-sided Wilcoxon rank sum test. *, $P < 0.05$; ****, $P < 0.0001$.

(B) Cumulative frequency of repeat motifs in VNTR *RPH3AL* across superpopulations. Repeat motifs are ordered by decreasing abundance in the African.

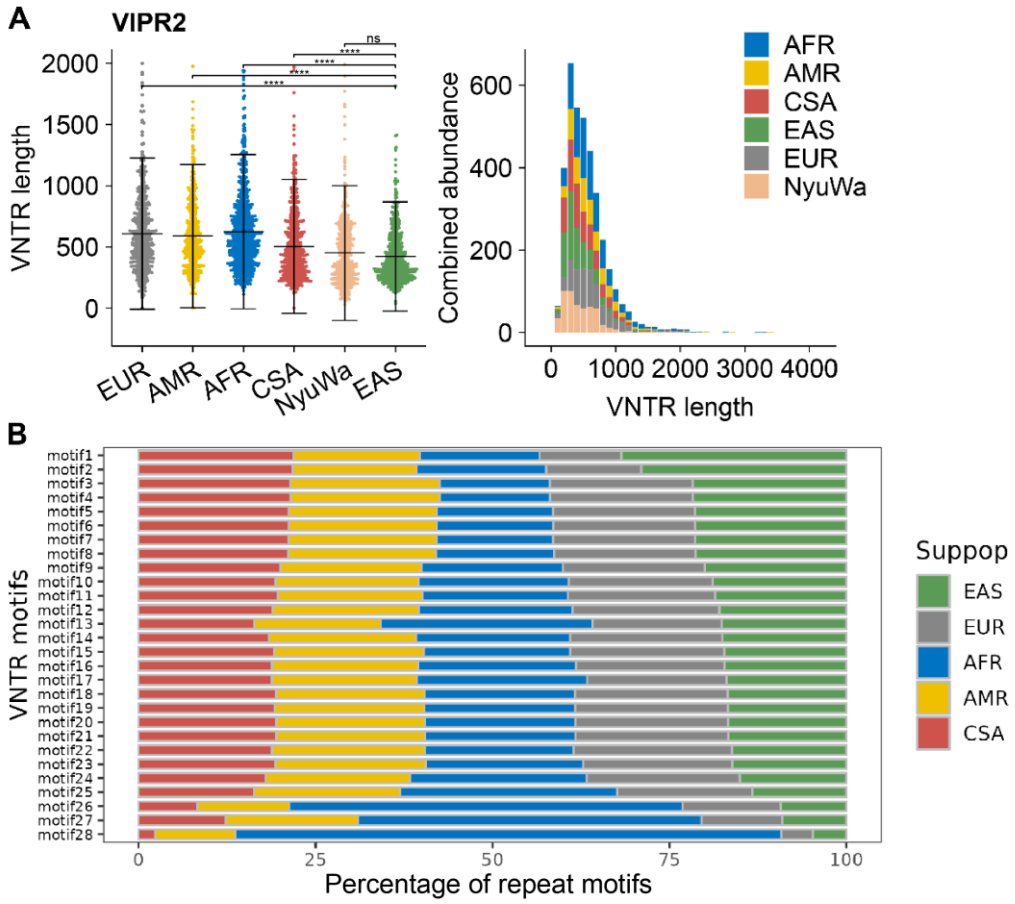


Figure S13 Distributions of lengths and motif frequency for human-specific expansion VNTR *VIPR2* differed across superpopulations, Related to Figure 5.

(A) Distribution of VNTR length for VNTR *VIPR2* in NyuWa and five superpopulations. P-values were computed using the two-sided Wilcoxon rank sum test. ns, $P > 0.05$; ****, $P < 0.0001$. **(B)** Cumulative frequency of repeat motifs in VNTR *VIPR2* across superpopulations. Repeat motifs are ordered by decreasing abundance in the African.

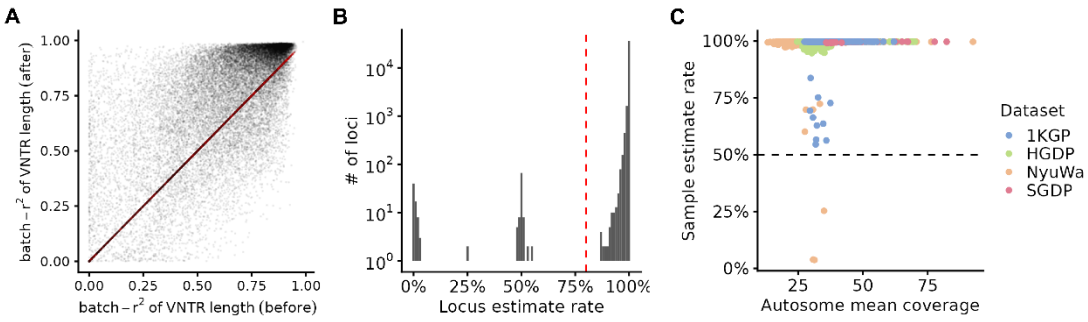


Figure S14. Quality control of VNTR length, Related to STAR Methods.

(A) Bias correction on the VNTR length for 38,685 loci. The red line indicates no improvement after bias correction. **(B)** Distribution of VNTR length estimate rate for 38,685 loci. The red dashed line indicates the estimate rate of 80%. **(C)** The VNTR length estimate rate for 8,225 samples. The black dashed line indicates the estimate rate of 50%. Colors represent diverse datasets used in this study.

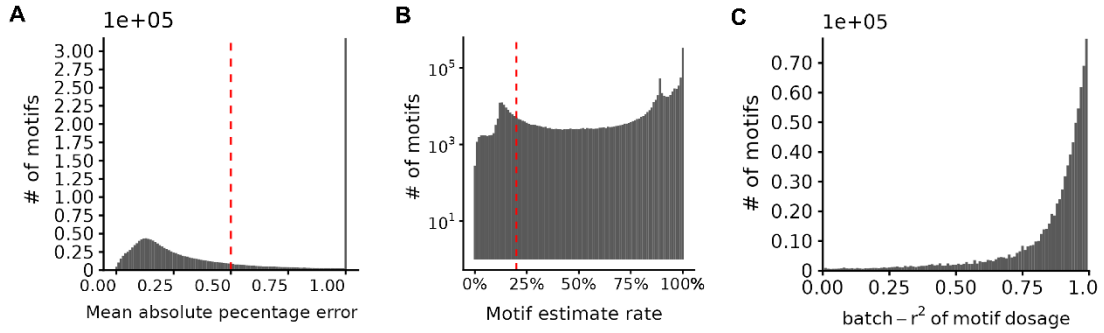


Figure S15 Quality control of VNTR motif dosage, Related to STAR Methods.

(A) Distribution of mean absolute percentage error (MAPE) of motifs. Total 1,703,470 motifs are shown in the plot, excluding motifs with MAPE greater than 1. The red dashed line indicates the MAPE of 0.5. **(B)** Distribution of motif dosage estimate rate. The red dashed line indicates the estimate rate of 20%. **(C)** Distribution of motif batch-r^2.

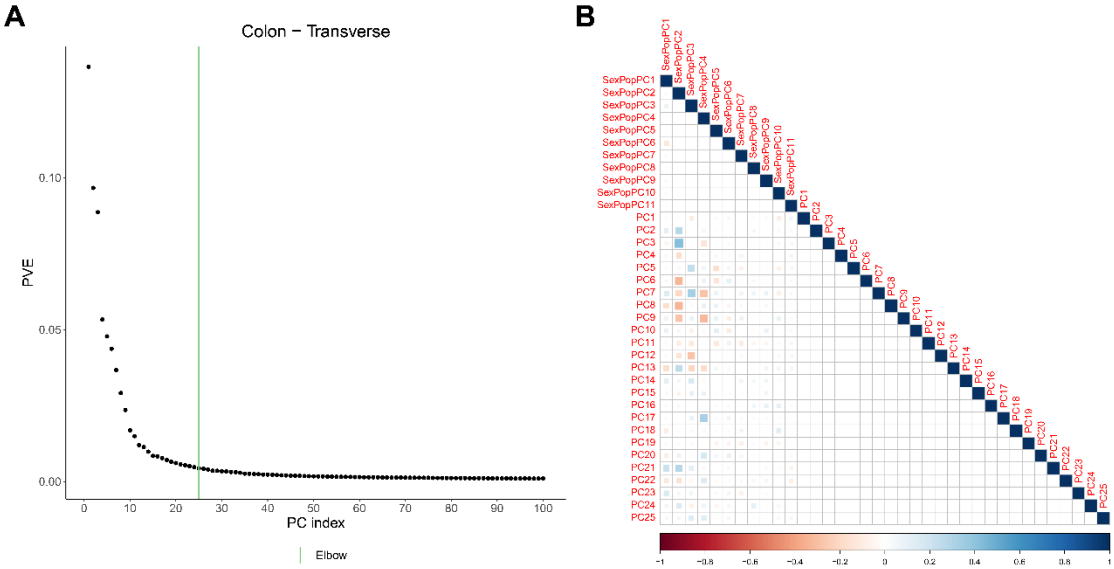


Figure S16 Detecting hidden covariates with PCA for gene expression, Related to STAR Methods.

(A) Proportion of variance explained of PCs. The green line indicates the PC index 25. **(B)** Correlations of the top 25 hidden factors for gene expression data.

References

1. Lu, T.Y., Smaruj, P.N., Fudenberg, G., Mancuso, N., and Chaisson, M.J.P. (2023). The motif composition of variable number tandem repeats impacts gene expression. Genome research *33*, 511-524. 10.1101/gr.276768.122.
2. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Research *12*, 996-1006. 10.1101/gr.229102.

Table S1. Fourteen fine-mapped eMotifs corresponding to seventeen eMotif-eGene pairs identified in this study, Related to Figure 3.

VNTR locus	Motif	eGene Ensembl ID	Posterior inclusion probability (PIP)
chr1_819912_823661	CAGGTAGTGTAGATAGCGTGG	ENSG00000225880	0.92
chr2_241774609_241774747	CGCTGCTCCCCGGCTGCTCCC	ENSG00000180902	0.98
chr3_195710216_195711492	GATCTCAATTAACTCACTCAC	ENSG00000242086	1
chr6_278651_278845	TCCAGGCTCTCCCCACAGCCTTCCACCA	ENSG00000112679	1
chr7_1940924_1941018	CCTCCTCTTCCTCTCCCAGGCCTCA	ENSG00000002822	1
chr10_67854352_67854524	GTGTGTGTGTGTGTGTGTATGTATATATGTGTATATA	ENSG00000096717	1
chr12_132148889_132150763	GAGCGAGGCGGCGGCACTCAC	ENSG00000184967	1
chr15_34454180_34454239	ATAATATATGTTTATATATATA	ENSG00000215252	1
chr21_42476948_42477299	CCCCGGGCTGTGGGACAGAGGG	ENSG00000160188	1
chr21_44208704_44209144	ACAAATGCCCACCTGCACCCGTGTCCACTGGCACAAATGCC	ENSG00000160223	1
chr21_44208704_44209144	ACAAATGCCCACCTGCACCCGTGTCCACTGGCACAAATGCC	ENSG00000232124	0.98
chr21_44928811_44929048	ACCCTGGATGCCTGTGGGCTGCCTTCCTCACC	ENSG00000227039	1
chr22_22578433_22578685	ATATATATGAAACATATATATG	ENSG00000275004	1
chr22_22875305_22875558	ATATATATACTATACACATATG	ENSG00000211676	1
chr22_22875305_22875558	ATATATATACTATACACATATG	ENSG00000211677	1
chr22_22875305_22875558	GTGTGTATATATATACATATGTGTATA	ENSG00000211678	1
chr22_22875305_22875558	GTGTGTATATATATACATATGTGTATA	ENSG00000211679	1

Table S3. Pearson correlation coefficients between VNTR lengths estimated from long-read sequencing and short-read sequencing, Related to STAR Methods.

Locus	**Pearson correlation coefficients (r^2)**
chr2:231472900-231473034	0.832417464
chr5:110610185-110610271	0.944082065
chr6:170614112-170614287	0.775630177
chr12:57678308-57678361	0.60250742
chr14:70424352-70425355	0.967093601