Title

The impact of short tandem repeat variation on gene expression.

Journal

Nature genetics, 51(11)

ISSN

1061-4036

Authors

Fotsing, Stephanie Feupe
Margoliash, Jonathan
Wang, Catherine
et al.

Publication Date

2019-11-01

DOI

10.1038/s41588-019-0521-9

Peer reviewed

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2020 May 01.

Published in final edited form as:

Nat Genet. 2019 November ; 51(11): 1652–1659. doi:10.1038/s41588-019-0521-9.

The impact of short tandem repeat variation on gene expression

Stephanie Feupe Fotsing[1,2,+], **Jonathan Margoliash**[3], **Catherine Wang**[4], **Shubham Saini**[3], **Richard Yanicky**[5], **Sharona Shleizer-Burko**[5], **Alon Goren**[5,#], **Melissa Gymrek**[3,5,#]

[1]Biomedical Informatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA

[2]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA USA

[3]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA USA

[4]Department of Bioengineering, University of California San Diego, La Jolla, CA USA

[5]Department of Medicine, University of California San Diego, La Jolla, CA USA

Abstract

Short tandem repeats (STRs) have been implicated in a variety of complex traits in humans. However, genome-wide studies of the effects of STRs on gene expression thus far have had limited power to detect associations and provide insights into putative mechanisms. Here, we leverage whole genome sequencing and expression data for 17 tissues from the Genotype-Tissue Expression Project to identify more than 28,000 STRs for which repeat number is associated with expression of nearby genes (eSTRs). We employ fine-mapping to quantify the probability that each eSTR is causal and characterize the top 1,400 fine-mapped eSTRs. We identify hundreds of eSTRs linked with published GWAS signals and implicate specific eSTRs in complex traits including height, schizophrenia, inflammatory bowel disease, and intelligence. Overall, our results support the hypothesis that eSTRs contribute to a range of human phenotypes and our data will serve as a valuable resource for future studies of complex traits

[#] Correspondence should be addressed to mgymrek@ucsd.edu or agoren@ucsd.edu.

[+]Present address: La Jolla Institute of Immunology, La Jolla, CA USA

Author Contributions

S.F.F. performed all eSTR and SNP mapping, helped perform downstream analyses and helped draft the manuscript. J.M. performed multi-tissue analysis using mashR and helped revise the manuscript. C.W. optimized and performed the reporter assay. S.S. participated in design of the STR imputation analysis. S.S.-B. lead, designed, and analyzed data from the reporter assay. R.Y. implemented the WebSTR web application. A.G. conceived and planned analyses and validation experiments of regulatory effects of eSTRs and wrote the manuscript. M.G. conceived the study, designed and performed analyses, and wrote the manuscript. All authors have read and approved the final manuscript.

Competing Interests

The authors have no competing interests to declare.

Introduction

Expression quantitative trait loci (eQTL) studies attempt to link genetic variation to gene expression changes as potential molecular intermediates that drive disease and variation in complex traits. Recent studies have identified tens of thousands of eQTLs (genetic variants associated with expression of nearby genes) across multiple human tissue types[1,2]. Most of these have focused on bi-allelic single nucleotide polymorphisms (SNPs) or short indels. Yet multiple studies dissecting genome wide association study (GWAS) loci have found repetitive[3,4] and structural variants[5–7] to be the underlying causal variants, highlighting the need to consider additional variant classes beyond SNPs.

Short tandem repeats (STRs), consisting of consecutively repeated units of 1–6bp, represent a large source of genetic variation. STR mutation rates are orders of magnitude higher than those of SNPs[8] and short indels[9], and each individual is estimated to harbor around 100 *de novo* mutations in STRs[10]. Expansions at several dozen STRs have been known for decades to cause Mendelian disorders[11] including Huntington's Disease and hereditary ataxias. Importantly, these pathogenic STRs represent a small minority of the more than 1.5 million STRs in the human genome[12]. Due to bioinformatics challenges of analyzing repetitive regions, many STRs are often filtered from genome-wide studies[13]. However, increasing evidence supports a widespread role of common variation at STRs in complex traits such as gene expression[14–17].

STRs may regulate gene expression through a variety of mechanisms[18]. For example, the CCG repeat implicated in Fragile X Syndrome was shown to disrupt DNA methylation, altering expression of *FMR1*[19]. Yeast studies have demonstrated that homopolymer repeats act as nucleosome positioning signals with downstream regulatory effects[20,21]. Dinucleotide repeats may alter affinity of nearby DNA binding sites[22]. Furthermore, certain STR repeat units may form non-canonical DNA and RNA secondary structures such as G-quadruplexes[23], R-loops[24], and Z-DNA[25].

We previously identified more than 2,000 STRs for which the number of repeats were associated with the expression of nearby genes[14], termed expression STRs (eSTRs). However, the quality of the datasets available for that study reduced our power to detect associations and prevented accurate fine-mapping of individual signals. STR genotypes were based on low coverage (4–6x) whole genome sequencing data performed using short reads (50–100bp) which are unable to span many STRs. As a result, STR genotype calls exhibited poor quality with less than 50% genotyping accuracy[12]. Additionally, the study used a single cell-type (lymphoblastoid cell lines; LCLs) with potentially limited relevance to most complex traits[26]. While our study and others[14,16] demonstrated that eSTRs explain a significant portion (10–15%) of the *cis* heritability of gene expression, the resulting eSTR catalogs were not powered to robustly implicate eSTRs over other nearby variants.

Here, we leverage deep whole genome sequencing (WGS) and gene expression data collected by the Genotype-Tissue Expression Project (GTEx)[1] to identify more than 28,000 eSTRs in 17 tissues. We employ fine-mapping to quantify the probability of causality of each eSTR and characterize the top 1,400 (top 5%) fine-mapped eSTRs. We additionally

identify hundreds of eSTRs that are in strong linkage disequilibrium (LD) with published GWAS signals and implicate specific eSTRs in height, schizophrenia, inflammatory bowel disease, and intelligence. To further validate our findings, we demonstrate evidence of a causal link between height and an eSTR for the gene *RFT1* and use a reporter assay to experimentally validate an effect of this STR on expression. Finally, our eSTR catalog is publicly available as a resource for future studies of complex traits.

Results

Profiling expression STRs across 17 human tissues

We performed a genome-wide analysis to identify associations between the number of repeats at each STR and expression of nearby genes (expression STRs, or "eSTRs", which we use to refer to a unique STR by gene association). We focused on 652 individuals from the GTEx[1] dataset for which both high coverage WGS and RNA-sequencing of multiple tissues were available (Fig. 1a). We used HipSTR[27] to genotype STRs in each sample. After filtering low quality calls (Methods), 175,226 STRs remained for downstream analysis. To identify eSTRs, for each gene and for each STR within 100 kb of that gene, we performed a linear regression between the average length of the STR in each person and normalized expression of the gene, controlling for sex, population structure, and technical covariates (Methods, Supplementary Figs. 1–3). Analysis was restricted to 17 tissues where we had data for at least 100 samples (Supplementary Table 1, Methods) and to genes with median Reads Per Kilobase of transcript, per Million mapped reads (RPKM) greater than 0. Altogether, we performed an average of 262,593 STR-gene tests across 15,840 protein-coding genes per tissue.

Using this approach, we identified 28,375 unique eSTRs associated with 12,494 genes in at least one tissue at a gene-level false discovery rate (FDR) of 10% (Fig. 1b, Supplementary Table 1, Supplementary Data 1). The number of eSTRs detected per tissue correlated with sample size as expected (Pearson $r = 0.75$; $p = 0.00059$; $n = 17$), with the smallest number of eSTRs detected in the two brain tissues presumably due to their low sample sizes (Extended Data Fig. 1). eSTR effect sizes previously measured in LCLs were significantly correlated with effect sizes in all GTEx tissues ($p < 0.01$ for all tissues, mean Pearson $r = 0.45$). We additionally examined previously reported eSTRs[28–35] that were mostly identified using *in vitro* constructs. Six of eight examples were significant eSTRs in GTEx ($p < 0.01$) in at least one tissue analyzed (Supplementary Table 2).

eSTRs identified above could potentially be explained by their tagging nearby causal variants. To prioritize potentially causal eSTRs we employed CAVIAR[36], a statistical fine-mapping framework. CAVIAR models the relationship between LD-structure and association statistics of local variants to quantify the posterior probability of causality for each variant (which we refer to as the CAVIAR score). We used CAVIAR to fine-map eSTRs against all SNPs nominally associated ($p < 0.05$) with each gene under our model (Methods, Fig. 1a). On average across tissues, 12.2% of eSTRs had the highest causality scores of all variants tested.

We ranked eSTRs by their best CAVIAR score across tissues and chose the top 5% for downstream analysis (1,420 unique eSTRs with best CAVIAR score >0.3). We hereby refer to these as fine-mapped eSTRs (FM-eSTRs) (Supplementary Table 1, Supplementary Data 2). Expected gene annotations are more strongly enriched in this subset compared to the entire set (Extended Data Fig. 2), and stricter thresholds reduced power to detect eSTR-enriched features described below. Of FM-eSTRs in each tissue, on average 78% explained gene expression variation beyond that explained by the best SNP (ANOVA $q < 0.1$). Furthermore, on average each FM-eSTR had CAVIAR score 0.41 higher (41% higher posterior probability) than the top-scoring SNP (Supplementary Fig. 4). Multiple STRs with known disease implications[35,37–40] were captured by this list (Fig. 1c). In many cases, FM-eSTRs show clear relationships between the number of repeats and gene expression across a wide range of repeat lengths (Extended Data Fig. 3).

To minimize power differences across tissues and enable cross-tissue comparisons of eSTR effects, we applied multivariate adaptive shrinkage (mash[41]) (Fig. 1a). Mash takes the per-tissue effect sizes and standard errors computed above as input and recomputes posterior estimates for each while considering cross-tissue effect size correlations. We compared FM-eSTR mash effect sizes across all pairs of tissues (Fig. 1d) and recovered previously observed relationships[41]. For example, tissues with similar origins (*e.g.*, Adipose-Visceral/Adipose-Subcutaneous) are highly concordant, whereas Whole Blood effects are less correlated with other tissues. These tissue sharing patterns are similar to those obtained using unadjusted effect sizes of single-tissue eSTRs (Supplementary Fig. 5). We further examined tissue sharing of FM-eSTRs by counting for each FM-eSTR the number of tissues for which mash computed a posterior Z-score with absolute value >4. Most eSTRs are either shared across all tissues analyzed or are shared by only a small number of tissues (Extended Data Fig. 4), again similar to previously reported SNP analyses in this cohort[1].

FM-eSTRs demonstrate unique genomic characteristics

We next sought to characterize properties of STRs that might provide insights into their biological function. We reasoned that genomic characteristics that distinguish FM-eSTRs from all analyzed STRs would support the hypothesis that a subset of them are acting as causal variants. While results below are presented for FM-eSTRs as defined above (CAVIAR score >0.3), we additionally provide results recomputed using a range of score thresholds in the Supplementary Material These results show that the major characteristics of FM-eSTRs identified below are robust to the precise threshold used.

We first considered whether the localization of FM-eSTRs differ from that of STRs overall (Fig. 2a–b, Extended Data Fig. 5). Overall, the majority of FM-eSTRs occur in intronic or intergenic regions, and only 11 FM-eSTRs fall in coding exons (Supplementary Table 3). However, compared to all STRs, those closest to transcription start sites (TSSs) and near DNAseI hypersensitive (HS) sites are more likely to be FM-eSTRs (Fig. 2c–d, Extended Data Fig. 6). FM-eSTRs are strongly enriched at 5' UTRs (odd's ratio (OR) = 5.0; Fisher's two-sided $p = 4.9 \times 10^{-13}$), 3' UTRs (OR = 2.78; $p = 5.85 \times 10^{-10}$), and within 3 kb of transcription start sites (OR = 3.39; $p = 3.94 \times 10^{-70}$). These enrichments are considerably

stronger for FM-eSTRs compared to all eSTRs (Supplementary Table 4), suggesting as expected that FM-eSTRs are more likely to be causal.

We next examined nucleosome occupancy in the lymphoblastoid cell line GM12878 and DNA accessibility (measured by DNAseI-seq) in a variety of cell and tissue types within 500 bp of FM-eSTRs (Extended Data Fig. 7). As expected from previous studies[42], regions near homopolymer repeats are strongly nucleosome-depleted. STRs with other repeat lengths also show distinct patterns of nucleosome positioning (Extended Data Fig. 7a–c). Nucleosome occupancy is broadly similar for FM-eSTRs compared to all STRs. Yet FM-eSTRs are generally located in regions with higher DNAseI-seq read count compared to non-eSTRs (Mann-Whitney two-sided $p = 3.9 \times 10^{-37}$ in GM12878; Extended Data Fig. 7d–f). DNAseI HS signal around homopolymer FM-eSTRs shows a periodic pattern in multiple cell and tissue types with peaks located at multiples of 147 bp upstream and downstream from the STR (Extended Data Fig. 7d). Given that 147 bp is the length of DNA typically wrapped around a single nucleosome[42], we hypothesize that a subset of homopolymer FM-eSTRs may act by shifting nucleosome positions and thus modulating accessibility of adjacent sites.

Next, we compared the sequence characteristics of FM-eSTRs to all STRs. We find that the total lengths of FM-eSTRs are significantly higher (Mann-Whitney two-sided $p = 0.00032$ and $p = 2.4 \times 10^{-10}$ when comparing total repeat number and total length in bp, respectively, based on the sequence present in hg19). We tested FM-eSTRs combined across all tissues for enrichment of each canonical STR repeat unit (defined lexicographically, see Methods) and found that FM-eSTRs are most strongly enriched for repeats with GC-rich repeat units (Fig. 2e, Supplementary Table 5, Supplementary Fig. 6). For example, the canonical repeat units CCCCGG, CCCCCG, and CCG are 22, 13, and 7-fold enriched in FM-eSTRs compared to all STRs, respectively. During transcription, these GC-rich repeat units have been shown to form highly stable secondary structures such as G4 quadruplexes in single-stranded DNA[43] or RNA[44] that may be involved in regulation of gene expression. We found that in general higher repeat numbers at GC-rich eSTRs are associated with greater DNA or RNA stability and increased expression of nearby genes (Supplementary Note, Fig. 2f–h, Supplementary Fig. 7–10).

We next examined effect size biases in FM-eSTR associations. Overall, FM-eSTRs are equally likely to show positive vs. negative correlations between repeat length and gene expression (Supplementary Fig. 11; two-sided binomial $p = 0.94$). We additionally observe that FM-eSTRs with repeat units of the form (A_nC/G_nT) show strand-specific effects when in or near transcribed regions. Transcribed FM-eSTRs are more likely to have the T-rich version of the repeat unit on the template strand (two-sided binomial $p = 0.0015$). These T-rich FM-eSTRs tend to have more positive effect sizes, with the most notable differences for AC vs. GT repeats. These patterns are observed in transcribed regions across multiple distinct repeat types (A/T, AC/GT, AAC/GTT, AAAC/GGGT) but are not present in intergenic regions (Extended Data Fig. 8).

Finally, we wondered whether eSTRs might exhibit distinct characteristics in different tissues. We clustered tissue-specific Z-scores (absolute value) for each FM-eSTR calculated jointly across tissues by mash (Methods) to identify eight categories of FM-eSTR

(Supplementary Fig. 12– 13). These include two clusters of FM-eSTRs present across many tissues (Clusters 2 and 8) as well as several more tissue-specific clusters (*e.g.,* Thyroid for Cluster 1). Notably, clusters do not necessarily imply tissue specificity, but rather enrich for FM-eSTRs with particularly strong effects in one or more tissues (Supplementary Fig. 13). Clusters show similar repeat unit enrichment to all FM-eSTRs and do not exhibit distinct enriched repeat units (Supplementary Fig. 14). Similar results were achieved using different numbers of clusters. Overall, our results suggest the majority of eSTRs act by global mechanisms and do not implicate tissue-specific characteristics of FM-eSTRs. However, low numbers of tissue-specific effects limit power to detect differences.

eSTRs are potential drivers of published GWAS signals

We wondered whether our eSTR catalog could identify STRs affecting complex traits in humans. We first leveraged the NHGRI/EBI GWAS catalog[45] to identify FM-eSTRs that are nearby and in LD with published GWAS signals. Overall, 1,380 unique FM-eSTRs are within 1 Mb of GWAS hits (Methods, Supplementary Data 3). Of these, 847 are in moderate LD ($r^2 > 0.1$) and 65 are in strong LD ($r^2 > 0.8$) with the lead SNP. When considering a more stringent set of FM-eSTRs with CAVIAR score >0.5, 403 and 26 are in strong and moderate LD with a GWAS hit, respectively.

We next sought to determine whether specific published GWAS signals could be driven by changes in expression due to an underlying but previously unobserved FM-eSTR. We reasoned that such loci would exhibit the following properties: (***i***) strong similarity in association statistics across variants for both the GWAS trait and expression of a particular gene, indicating the signals may be co-localized, *i.e.*, driven by the same causal variant; and (***ii***) strong evidence that the FM-eSTR causes variation in expression of that gene (Fig. 3a). Co-localization analysis requires high-resolution summary statistic data. Thus, we focused on several example complex traits (height[46], schizophrenia[47], inflammatory bowel disease (IBD)[48], and intelligence[49]) for which detailed summary statistics computed on cohorts of tens of thousands or more individuals are publicly available (Methods).

For each trait, we identified FM-eSTRs within 1 Mb of published GWAS signals from Supplementary Dataset 3. We then used coloc[50] to compute the probability that the FM-eSTR signals we derived from GTEx and the GWAS signals derived from other cohorts are co-localized. The coloc tool compares association statistics at each SNP in a region for expression and the trait of interest and returns a posterior probability that the signals are co-localized. We used coloc to test a total of 276 gene×trait pairs (138, 45, 29, and 64 for height, intelligence, IBD, and schizophrenia, respectively). In total, we identified 62 GWAS loci with (1) an FM-eSTR in at least moderate LD ($r^2 > 0.1$) with a nearby SNP for that trait in the GWAS catalog and (2) co-localization posterior probability between the target gene and the trait >50%, meaning co-localization of the eQTL and GWAS signals is the most probable model (Extended Data Fig. 9–10). Out of the 62 FM-eSTRs co-localized with GWAS signals, 40 have CAVIAR scores >0.5. Results of all co-localization tests are provided in Supplementary Table 6.

A top example is an FM-eSTR for *RFT1*, a gene encoding enzyme involved in the N-glycosylation of proteins[51], that has 97.8% co-localization probability with a GWAS signal

for height (Fig. 3b–c). The lead SNP in the NHGRI catalog (rs2336725:C>T) is in high LD ($r^2 = 0.85$) with an AC repeat that is a significant eSTR in 15 tissues. This STR falls in a cluster of transcription factor and chromatin regulator binding regions identified by ENCODE near the 3' end of the gene (Fig. 3d) and exhibits a positive correlation with expression.

To more directly test for association between this FM-eSTR and height, we used our recently developed STR-SNP reference haplotype panel[52] to impute STR genotypes into available GWAS data. We focused on the eMERGE cohort (Methods) for which imputed genotype array data and height measurements are available. We tested for association between height and SNPs as well as for height and AC repeat number after excluding samples with low STR imputation quality (Methods). Imputed AC repeat number is significantly associated with height in the eMERGE cohort ($p = 0.00328$; beta = 0.010; $n = 6,393$), although with a slightly weaker p-value compared to the top SNP (Fig. 3e). Notably, even in the case that the STR is the causal variant, power is likely reduced due to the lower quality of imputed STR genotypes. Notably, AC repeat number shows a strong positive relationship with height across a range of repeat lengths (Fig. 3f), similar to the relationship between repeat number and *RFT1* expression.

To further investigate whether the FM-eSTR for *RFT1* could be a causal driver of gene expression variation, we devised a dual reporter assay in HEK293T cells to test for an effect of the number of repeats on gene expression (0, 5, 10, or 12 repeats plus approximately 170 bp of genomic sequence context on either side) (Supplementary Table 7, Methods). We observed a positive linear relationship between the number of AC repeats and reporter expression as predicted (Fig. 3g) (Pearson $r = 0.97$; $p = 0.013$). Furthermore, all pairs of constructs with consecutive repeat numbers showed significantly different expression (one-sided t-test $p < 0.01$) with the exception of 10 vs. 12 repeats. Overall, these results further support the hypothesis that eSTRs may act as causal drivers of gene expression.

Discussion

Here we present the most comprehensive resource of eSTRs to date, which reveals more than 28,000 associations between the number of repeats at STRs and expression of nearby genes across 17 tissues. We performed fine-mapping to quantify the probability that each eSTR causally effects gene expression and characterize top fine-mapped eSTRs. eSTRs analyzed here consist of a large spectrum of repeat classes with a variety of repeat unit lengths and sequences. Based on the diverse characteristics of eSTRs, we hypothesize that different repeat classes work by distinct regulatory effects (Fig. 4). While we explored several potential mechanisms, including nucleosome positioning and the formation of non-canonical DNA or RNA secondary structures, our results do not rule out other potential mechanisms.

We leveraged our resource to provide evidence that FM-eSTRs may drive a subset of published GWAS associations for a variety of complex traits. STRs have a unique ability compared to bi-allelic SNPs to drive phenotypic variation along a spectrum of multiple alleles. In multiple examples, eSTRs show a linear trend between repeat length and

expression across a range of repeat numbers, a signal that cannot be easily explained by tagging nearby bi-allelic variants. Notably, our analysis is based only on signals that could be detected by standard SNP-based GWAS, which are underpowered to detect underlying multi-allelic associations from STRs[52]. Further work to directly test for associations between STRs and phenotypes may reveal a widespread role for repeat number variation in complex traits.

Our study faced several limitations: (***i***) While we applied stringent fine-mapping approaches to find eSTRs whose signals are likely not explained by nearby SNPs in LD, some signals could plausibly be explained by other variant classes such as structural variants[53] or *Alu* elements[54] that were not considered. Furthermore, our fine-mapping procedure may be vulnerable to false negatives for STRs in strong or perfect LD with nearby SNPs or false positives due to noise present with small sample sizes. (***ii***) Our study was limited to tissues available from GTEx with sufficient sample sizes. While this greatly expanded on the single tissue used in our previous eSTR analysis, some tissues such as brain were not well represented. Further, due to overwhelming sharing of eSTRs across tissues, we were unable to identify tissue-specific characteristics of eSTRs. (***iii***) Despite strong evidence that the FM-eSTRs for *RFT1* and other genes may drive published GWAS signals, we have not definitively proved causality. Additional work is needed to validate effects on expression and evaluate the impact of these STRs in trait-relevant cell types.

Altogether, our eSTR catalog provides a valuable resource for studying the role of STRs in complex traits. Example applications of this resource include: further analysis of the genetic architecture of gene expression by quantifying the contribution of different variant classes, genome-wide analyses to confirm or refute hypotheses about eSTR mechanisms, and integration of eSTRs into GWAS fine-mapping to identify candidate variants not identified by SNP-based analyses. To facilitate these and other studies, all summary-level eSTR data are publicly available at http://webstr.ucsd.edu.

Online Methods

Dataset and preprocessing

Next-generation sequencing data was obtained from the Genotype-Tissue Expression (GTEx) through dbGaP under phs000424.v7.p2. This included high coverage (30x) Illumina whole genome sequencing (WGS) data and expression data from 652 unrelated individuals (Supplementary Fig. 1). The WGS cohort consisted of 561 individuals with reported European ancestry, 75 of African ancestry, and 8, 3, and 5 of Asian, Amerindian, and Unknown ancestry, respectively. For each sample, we downloaded BAM files containing read alignments to the hg19 reference genome and VCFs containing SNP genotype calls.

STRs were genotyped using HipSTR[27] v0.5, which returns the maximum likelihood diploid STR allele sequences for each sample based on aligned reads as input. Samples were genotyped separately with non-default parameters --min-reads 5 and --def-stutter-model. VCFs were filtered using the filter_vcf.py script available from HipSTR using recommended settings for high coverage data (--min-call-qual 0.9, --max-call-flank-indel 0.15, and --max-call-stutter 0.15). VCFs were merged across all samples and further filtered to exclude STRs

meeting the following criteria: call rate <80%; STRs overlapping segmental duplications (UCSC Genome Browser[55] hg19.genomicSuperDups table); penta- and hexamer STRs containing homopolymer runs of at least 5 or 6 nucleotides, respectively in the hg19 reference genome, since we previously found these STRs to have high error rates due to indels in homopolymer regions[52]; and STRs whose frequencies did not meet the percentage of homozygous vs. heterozygous calls expected under Hardy-Weinberg Equilibrium (binomial two-sided $p < 0.05$). Additionally, to restrict to polymorphic STRs we filtered STRs with heterozygosity <0.1. Altogether, 175,226 STRs remained for downstream analysis.

We additionally obtained gene-level RPKM values for each tissue from dbGaP project phs000424.v7.p2. We focused on 15 tissues with at least 200 samples, and included two brain tissues with slightly more than 100 samples available (Supplementary Table 1). Genes with median RPKM of 0 were excluded and expression values for remaining genes were quantile normalized separately per tissue to a standard normal distribution. Analysis was restricted to protein-coding genes based on GENCODE version 19 (Ensembl 74) annotation.

Prior to downstream analyses, expression values were adjusted separately for each tissue to control for sex, population structure, and technical variation in expression as covariates. For population structure, we used the top 10 principal components resulting from performing principal components analysis (PCA) on the matrix of SNP genotypes from each sample. PCA was performed jointly on GTEx samples and 1000 Genomes Project[56] samples genotyped using Omni 2.5 SNP genotyping arrays (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapeit2_scaffolds/hd_chip_scaffolds/). Analysis was restricted to bi-allelic SNPs present in the Omni 2.5 data and resulting loci were LD-pruned using plink[57] v1.90b3.44 with option --indep 50 5 2. PCA on resulting SNP genotypes was performed using smartpca[58,59] v13050. To control for technical variation in expression, we applied PEER factor correction[60]. Based on an analysis of number of PEER factors vs. number of eSTRs identified per tissue (Supplementary Fig. 2), we determined an optimal number of N/10 PEER factors as covariates for each tissue, where N is the sample size. PEER factors were correlated with covariates reported previously for GTEx samples (Supplementary Fig. 3) such as ischemic time.

eSTR and eSNP identification

For each STR within 100kb of a gene, we performed a linear regression between STR lengths and adjusted expression values:

$$Y' = \beta X + \epsilon$$

Where X denotes STR genotypes, Y' denotes expression values adjusted for the covariates described above, β denotes the effect size, and ε is the error term. A separate regression analysis was performed for each STR-gene pair in each tissue. For STR genotypes, we used the average repeat length of the two alleles for each individual, where repeat length was computed as a length difference from the hg19 reference, with 0 representing the reference allele. Linear regressions were performed using the OLS function from the Python

statsmodels.api module[61] (https://www.statsmodels.org, v0.8.0), which returns estimated regression coefficients computed using ordinary least squares and two-sided p-values for each regression coefficient testing the null hypothesis $\beta = 0$ computed from t-statistics for each coefficient. As a control, for each STR-gene pair we performed a permutation analysis in which sample identifiers were shuffled.

Samples with missing genotypes or expression values were removed from each regression analysis. To reduce the effect of outlier STR genotypes, we removed samples with genotypes observed in fewer than 3 samples. If after filtering samples there were fewer than three unique genotypes, the STR was excluded from analysis. Adjusted expression values and STR genotypes for remaining samples were then Z-scaled to have mean 0 and variance 1 before performing each regression. This step forces resulting effect sizes to be between −1 and 1.

We used a gene-level FDR threshold (described previously[14]) of 10% to identify significant STR-gene pairs. We assume most genes have at most a single causal eSTR. For each gene, we determined the STR association with the strongest P-value. This P-value was adjusted using a Bonferroni correction for the number of STRs tested per gene to give a P-value for observing a single eSTR association for each gene. We then used the list of adjusted P-values (one per gene) as input to the fdrcorrection function in the statsmodels.stats.multitest module to obtain a q-value for the best eSTR for each gene. FDR analysis was performed separately for each tissue.

eSNPs were identified using the same model covariates and normalization procedures but using SNP dosages (0, 1, or 2) rather than STR lengths. Similar to the STR analysis, we removed samples with genotypes occurring in fewer than 3 samples and removed SNPs with fewer than 3 unique genotypes remaining after filtering. On average, we tested 17 STRs and 533 SNPs per gene.

Fine-mapping eSTRs

We used model comparison as an orthogonal validation to CAVIAR findings to determine whether the best eSTR for each gene explained variation in gene expression beyond a model consisting of the best eSNP. For each gene with an eSTR we determined the eSNP with the strongest p-value. We then compared two linear models: Y'~eSNP (SNP-only model) vs. Y'~eSNP+eSTR (SNP+STR model) using the anova_lm function in the python statsmodels.api.stats module. Q-values were obtained using the fdrcorrection function in the statsmodels.stats.multitest module. On average across tissues, 17.4% of eSTRs tested improved the model over the best eSNP for the target gene (10% FDR). When restricting to FM-eSTRs, 78% improved the model (10% FDR).

We used CAVIAR[36] v2.2 to further fine-map eSTR signals against all nominally significant eSNPs ($p < 0.05$) within 100 kb of each gene. On average, 121 SNPs per gene passed this threshold and were included in CAVIAR analysis. Pairwise-LD between the eSTR and eSNPs was estimated using the Pearson correlation between SNP dosages (0, 1, or 2) and STR genotypes (average of the two STR allele lengths) across all samples. CAVIAR was run with parameters -f 1 -c 2 to model up to two independent causal variants per locus. In some

cases, initial association statistics for SNPs and STRs might have been computed using different sets of samples if some were filtered due to outlier genotypes. To provide a fair comparison between eSTRs and eSNPs, for each CAVIAR analysis we recomputed Z-scores for eSTRs and eSNPs using the same set of samples prior to running CAVIAR.

Multi-tissue eSTR analysis

We used an R implementation of mash[41] (mashR) v0.2.21 to compute posterior estimates of eSTR effect sizes and standard errors across tissues (https://stephenslab.github.io/mashr/articles/intro_mash_dd.html). Briefly, mashR takes as input effect sizes and standard error measurements per-tissue, learns various covariance matrices of effect sizes between tissues, and outputs posterior estimates of effect sizes and standard errors accounting for global patterns of effect size sharing. We used all eSTRs with a nominal p-value of $<1 \times 10^{-5}$ in at least one tissue as a set of strong signals to compute covariance matrices. eSTRs that were not analyzed in all tissues were excluded from this step. We included "canonical" covariance matrices (identity matrix and matrices representing condition-specific effects) and matrices learned by extreme deconvolution initialized using PCA with 5 components as suggested by mashR documentation. After learning covariance matrices, we applied mashR to estimate posterior effect sizes and standard errors for each eSTR in each tissue. For eSTRs that were filtered from one or more tissues in the initial regression analysis, we set input effect sizes to 0 and standard errors to 10 in those tissues to reflect high uncertainty in effect size estimates at those eSTRs. For Fig. 1d, rows and columns of the effect size correlation matrix were clustered using default parameters from the clustermap function in the Python seaborn library (https://seaborn.pydata.org/, v0.9.0).

Canonical repeat units

For each STR, we defined the canonical repeat unit as the lexicographically first repeat unit when considering all rotations and strand orientations of the repeat sequence. For example, the canonical repeat unit for the repeat sequence CAGCAGCAGCAG would be AGC.

Enrichment analyses

Enrichment analyses were performed using a two-sided Fisher's exact test as implemented in the fisher_exact function of the python package scipy.stats (https://docs.scipy.org/doc/scipy/reference/stats.html, v1.2.1). Overlapping STRs with each annotation was performed using the intersectBed tool of the BEDTools[62] suite v2.28.0. Genomic annotations were obtained by downloading custom tables using the UCSC Genome Browser[55] table browser tool to select either coding regions, introns, 5'UTRs, or 3'UTRs. An STR could be assigned to more than one category in the case of overlapping transcripts. STRs not assigned to one of those categories were labeled as intergenic. ENCODE DNAseI HS clusters were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz). Analysis was restricted to DNAseI HS clusters annotated in at least 20 cell types. The distance between each STR and the center of the nearest DNAseI HS cluster was computed using the closestBed tool from the BEDTools suite.

Analysis of DNAseI-seq, ChIP-seq, and Nucleosome occupancy

Genome-wide nucleosome occupancy signal in GM12878 was downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhNsome/wgEncodeSydhNsomeGm12878Sig.bigWig). ChIP-seq reads for RNAPII and DNAseI-seq reads were downloaded from the ENCODE Project website (https://www.encodeproject.org) (Accessions GM12878 RNAPII: ENCFF000OBB, heart RNAPII: ENCFF643EGO, lung RNAPII: ENCSR033NHF, tibial nerve RNAPII: ENCFF750HDH, human embryonic stem cells RNAPII: ENCFF526YGE; GM12878 DNAseI: ENCFF775ZJX, fat DNAseI: ENCFF880CAD, tibial nerve DNAseI: ENCFF226ZCG, skin DNAseI: ENCFF238BRB). Histograms of aggregate read densities and heatmaps for individual STR regions were generated using the annoatePeaks.pl tool of Homer[63] v4.10. For nucleosome occupancy and DNAseI analyses on all STRs, we used parameters -size 1000 -hist 1. For analysis of GC-rich repeats in promoters, we used parameters -size 10000 -hist 5.

Characterization of tissue-specific eSTRs

We clustered FM-eSTRs based on Z-scores computed by mash for each eSTR in each tissue. We first created a tissue by FM-eSTR matrix of the absolute value of the Z-scores. We then Z-normalized Z-scores for each FM-eSTR to have mean 0 and variance 1. We used the KMeans class from the Python sklearn.cluster module to perform K-means clustering with K=8 (https://scikit-learn.org/stable/, v0.20.3). The number of clusters was chosen by visualizing the sum of squared distances from centroids for values of K ranging from 1 to 20 and choosing a value of K based on the “elbow method”. Using different values of K produced similar groups. We tested for non-uniform distributions of FM-eSTR repeat units across clusters using a chi-squared test implemented in the scipy.stats chi2_contingency function.

Analysis of DNA and RNA secondary structure

For each STR, we extracted the repeat plus 50 bp flanking sequencing from the hg19 reference genome. We additionally created sequences containing each common allele for each STR. Common alleles were defined as those seen at least 5 times in a previously generated deep catalog of STR variation in 1,916 samples[52]. For each sequence and its reverse complement, we ran mfold[64] v3.6 on the DNA and corresponding RNA sequences with mfold arguments NA=DNA and NA=RNA, respectively, and otherwise default parameters to estimate the free energy of each single-stranded sequence. Mann-Whitney tests were performed using the mannwhitneyu function of the scipy.stats python package.

Co-localization of FM-eSTRs with published GWAS signals

Published GWAS associations were obtained from the NHGRI/EBI GWAS catalog available from the UCSC Genome Browser Table Browser (table hg19.gwasCatalog) downloaded on July 24, 2019. Height GWAS summary statistics were downloaded from the GIANT Consortium website (https://portals.broadinstitute.org/collaboration/giant/images/0/0f/Meta-analysis_Locke_et_al%2BUKBiobank_2018.txt.gz). Schizophrenia GWAS summary statistics were downloaded from the Psychiatric Genomics Consortium website

(Schizophrenia GWAS summary statistics, https://www.med.unc.edu/pgc/results-and-downloads). IBD summary statistics were downloaded from the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) website. We used the file EUR.IBD.gwas_info03_filtered.assoc with summary statistics in Europeans (https://ftp://ftp.sanger.ac.uk/pub/consortia/ibdgenetics/iibdgc-trans-ancestry-filtered-summary-stats.tgz). Intelligence summary statistics were downloaded from the Complex Trait Genomics lab website (https://ctg.cncr.nl/documents/p1651/SavageJansen_IntMeta_sumstats.zip). LD between STRs and SNPs was computed by taking the squared Pearson correlation between STR lengths and SNP dosages in GTEx samples for each STR-SNP pair. STR genotypes seen less than 3 times were filtered from LD calculations.

Co-localization analysis of eQTL and GWAS signals was performed using the coloc.abf function of the coloc[50] package. For all traits, dataset 1 was specified as type="quant" and consisted of SNP effect sizes and their variances as input. We specified sdY=1 since expression was quantile normalized to a standard normal distribution. Dataset 2 was specified differently for height and schizophrenia to reflect quantitative vs. case-control analyses. For height and intelligence, we specified type="quant" and used effect sizes and their variances as input. We additionally specified minor allele frequencies listed in the published summary statistics file and the total sample size of N = 695,647 and N = 269,720 for height and intelligence, respectively. For schizophrenia and IBD, we specified type="CC" and used effect sizes and their variances as input. We additionally specified the fraction of cases as 33%.

Capture Hi-C interactions (Extended Data Fig. 10) were visualized using the 3D Genome Browser[65]. The visualization depicts interactions profiled in GM12878[66] and only shows interactions overlapping the STR of interest.

Association analysis in the eMERGE cohort

We obtained SNP genotype array data and imputed genotypes from dbGaP accessions phs000360.v3.p1 and phs000888.v1.p1 from consent groups c1 (Health/Medical/Biomedical), c3 (Health/Medical/Biomedical - Genetic Studies Only - No Insurance Companies), and c4 (Health/Medical/Biomedical - Genetic Studies Only). Height data was available for samples in cohorts c1 (phs000888.v1.pht004680.v1.p1.c1), c3 (phs000888.v1.pht004680.v1.p1.c3), and c4 (phs000888.v1.pht004680.v1.p1.c4). We removed samples without age information listed. If height was collected at multiple times for the same sample, we used the first data point listed.

Genotype data was available for 7,190, 6100, and 3,755 samples from the c1, c3, and c4 cohorts respectively (dbGaP study phs000360.v3.p1). We performed PCA on the genotypes to infer ancestry of each individual. We used plink to restrict to SNPs with minor allele frequency at least 10% and with genotype frequencies expected under Hardy-Weinberg Equilibrium ($p > 1 \times 10^{-4}$). We performed LD pruning using the plink option --indep 50 5 1.5 and used pruned SNPs as input to PCA analysis. We visualized the top two PCs and identified a cluster of 14,147 individuals overlapping samples with annotated European ancestry. We performed a separate PCA using only the identified European samples and used the top 10 PCs as covariates in association tests.

A total of 11,587 individuals with inferred European ancestry had both imputed SNP genotypes and height and age data available. Samples originated from cohorts at Marshfield Clinic, Group Health Cooperative, Northwestern University, Vanderbilt University, and the Mayo Clinic. We adjusted height values by regressing on top 10 ancestry PCs, age, and cohort. Residuals were inverse normalized to a standard normal distribution. Adjustment was performed separately for males and females.

Imputed genotypes (from dbGaP study phs000888.v1.p1) were converted from IMPUTE2[67] to plink's binary format using plink, which marks calls with uncertainty >0.1 (score<0.9) as missing. SNP associations were performed using plink with imputed genotypes as input and with the "linear" option with analysis restricted to the region chr3:53022501–53264470.

The *RFT1* FM-eSTR was imputed into the imputed SNP genotypes using Beagle 5[68] with option gp=true and using our SNP-STR reference haplotype panel[52]. We previously estimated imputation concordance of 97% at this STR in a separate European cohort. Samples with imputed genotype probabilities of less than 0.9 were removed from the STR analysis. We additionally restricted analysis to STR genotypes present in at least 100 samples to minimize the effect of outlier genotypes. We regressed STR genotype (defined above as the average of an individual's two repeat lengths) on residualized height values for the remaining 6,393 samples using the Python statsmodels.regression.linear_model.OLS function (https://www.statsmodels.org).

Dual luciferase reporter assay

Constructs for 0, 5, or 10 copies of AC at the FM-eSTR for *RFT1* (chr3:53128363–53128413 plus approximately 170bp genomic context on either side (RFT1_0rpt, RFT1_5rpt, RFT1_10rpt in Supplementary Table 7) were ordered as gBlocks from Integrated DNA technologies (IDT). Each construct additionally contained homology arms for cloning into pGL4.27 (below). We additionally PCR amplified the region from genomic DNA for sample NA12878 with 12 copies of AC (NIGMS Human Genetic Repository, Coriell) using PrimeSTAR max DNA Polymerase (Clontech R045B) and primers RFT1eSTR_F and RFT1eSTR_R (Supplementary Table 7) which included the same homology arms.

Constructs were cloned into plasmid pGL4.27 (Promega, E8451), which contains the firefly luciferase coding sequence and a minimal promoter. The plasmid was linearized using EcoRV (New England Biolabs, R3195) and purified from agarose gel (Zymo Research, D4001). Constructs were cloned into the linearized vector using In-Fusion (Clontech, 638910). Sanger sequencing of isolated clones for each plasmid validated expected repeat numbers in each construct.

Plasmids were transfected into the human embryonic kidney 293 cell line (HEK293T; ATCC CRL-3216) and grown in DMEM media (Gibco, 10566–016), supplemented with 10% fetal bovine serum (Gibco, 10438–026), 2 mM glutamine (Gibco, A2916801), 100 units/mL of penicillin, 100 μg/mL of streptomycin, and 0.25 μg/mL Amphotericin B (Anti-Anti Gibco, 15240062). Cells were maintained at 37°C in a 5% CO_2 incubator. 2×10^5 HEK293T cells were plated onto each well of a 25 ug/ml poly-D lysine (EMD Millipore, A-003-E) coated

24-well plate, the day prior to transfection. On the day of the transfection medium was changed to Opti-MEM. We conducted co-transfection experiments to test expression of each construct. 100ng of the empty pGL4.27 vector (Promega, E8451) or 100 ng of each one of the pGL4.27 derivatives, were mixed with 5ng of the reference plasmid, pGL4.73 (Promega, E6911), harboring SV40 promoter upstream of Renilla luciferase, and added to the cells in the presence of Lipofectamine™ 3000 (Invitrogen, L3000015), according to the manufacturer's instructions. Cells were incubated for 24 h at 37°C, washed once with phosphate-buffered saline, and then incubated in fresh completed medium for an additional 24 h.

48 hours after transfection the HEK293T cells were washed 3 times with PBS and lysed in 100μl of Passive Lysis Buffer (Promega, E1910). Firefly luciferase and Renilla luciferase activities were measured in 10 μl of HEK293T cell lysate using the Dual-Luciferase Reporter assay system (Promega, E1910) in a Veritas™ Microplate Luminometer. Relative activity was defined as the ratio of firefly luciferase activity to Renilla luciferase activity. For each plasmid, transfection and the expression assay were done in triplicates using three wells of cultured cells that were independently transfected (biological repeats), and three individually prepared aliquots of each transfection reaction (technical repeats). Values from each technical replicate were averaged to get one ratio for each biological repeat.

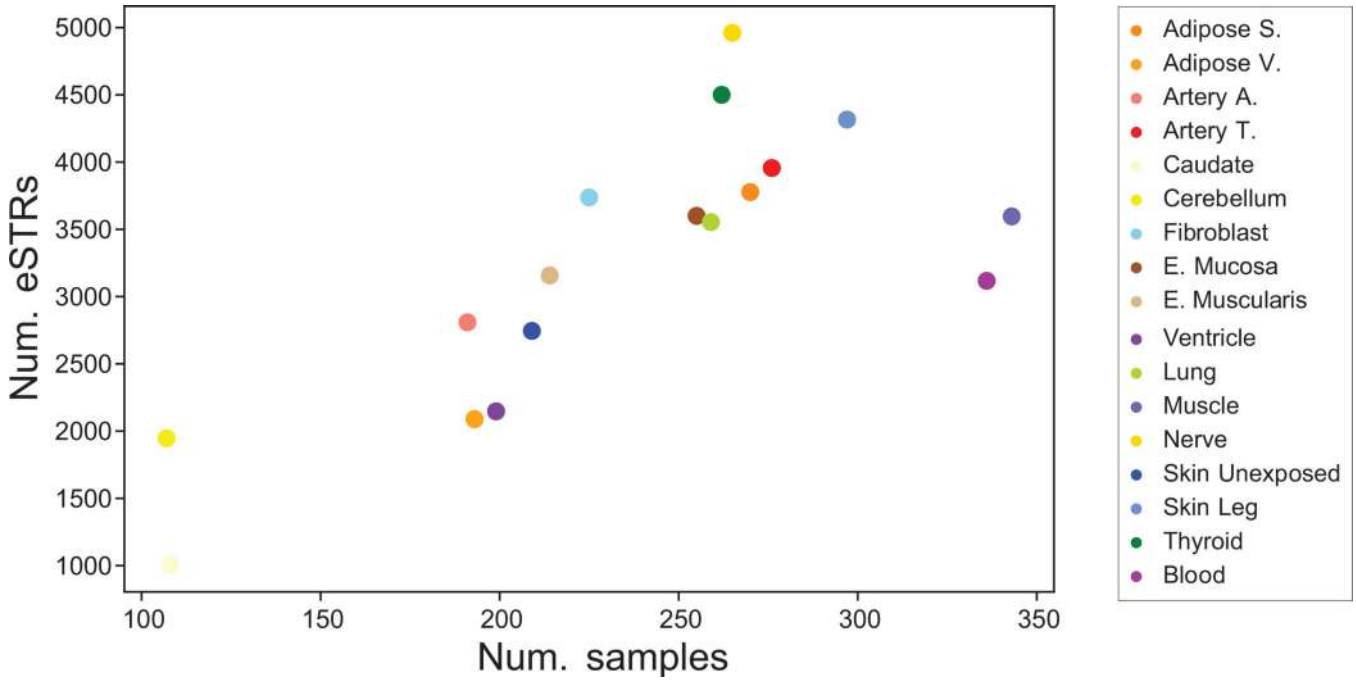
Data Availability

All eSTR summary statistics are available for download on WebSTR http://webstr.ucsd.edu/downloads.

Code Availability

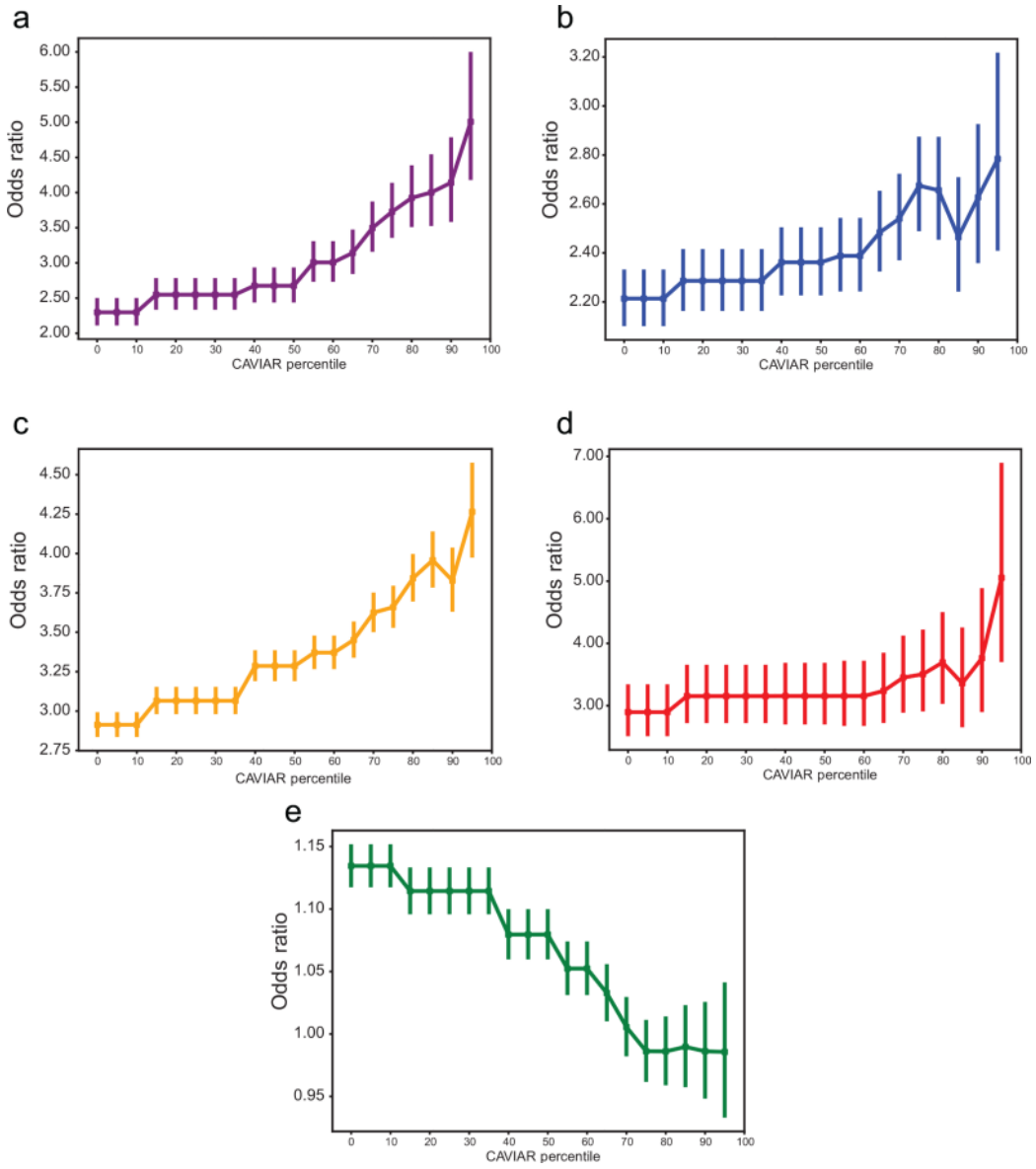
Code for performing analyses and generating figures is available at http://github.com/gymreklab/gtex-estrs-paper.

Extended Data



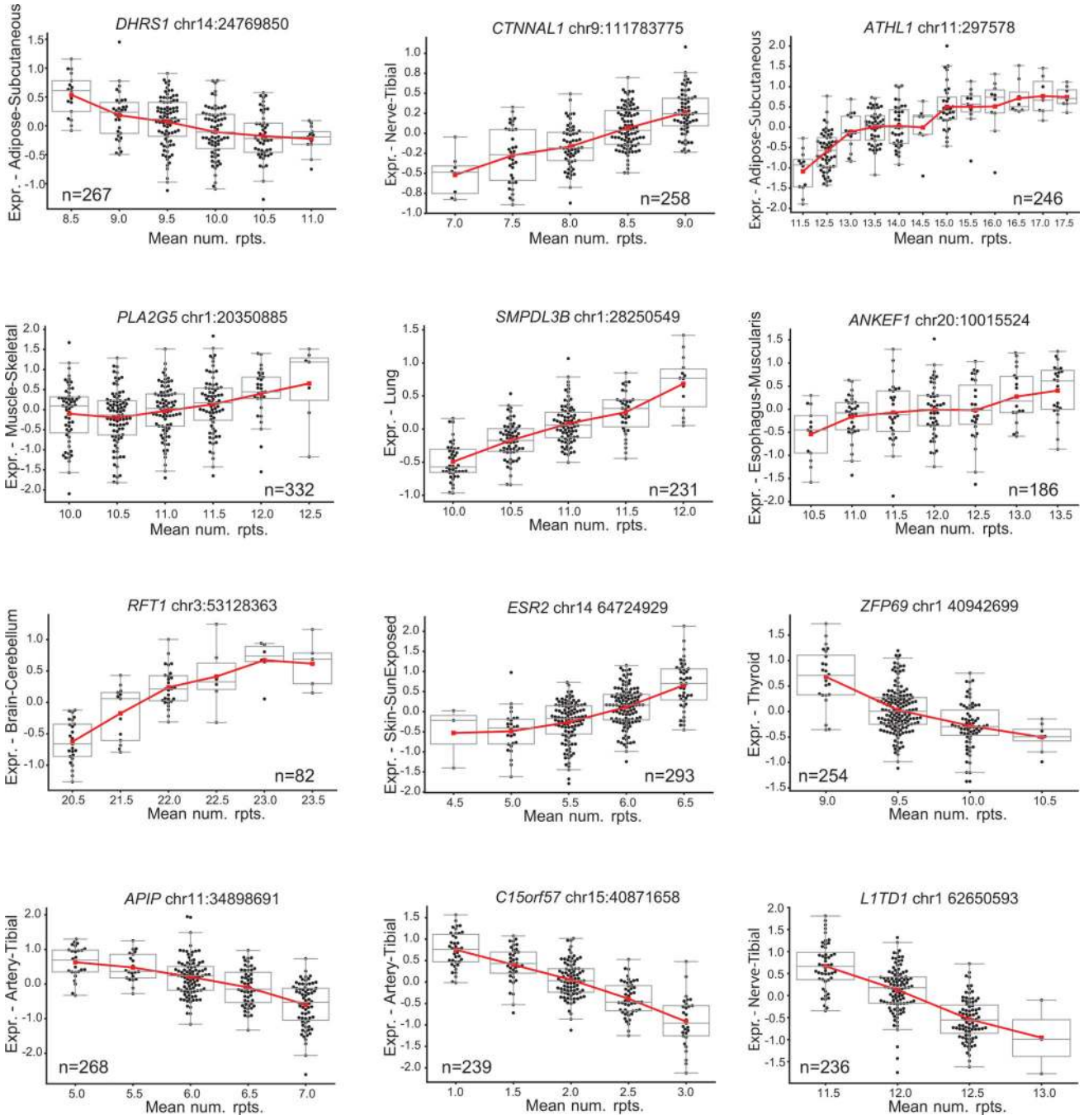
Extended Data Fig. 1: Relationship between sample size and number of eSTRs detected

The x-axis shows the number of samples per tissue. The y-axis shows the number of eSTRs (gene-level FDR<10%) detected in each tissue. Each dot represents a single tissue, using the same colors as shown in Fig. 1 in the main text (see box on the right). Notably, although whole blood and skeletal muscle had the highest number of samples, we identified fewer eSTRs in those tissues than in others with lower sample sizes. This is concordant with previous results for SNPs in the GTEx cohort and may reflect higher cell-type heterogeneity in these tissue samples.



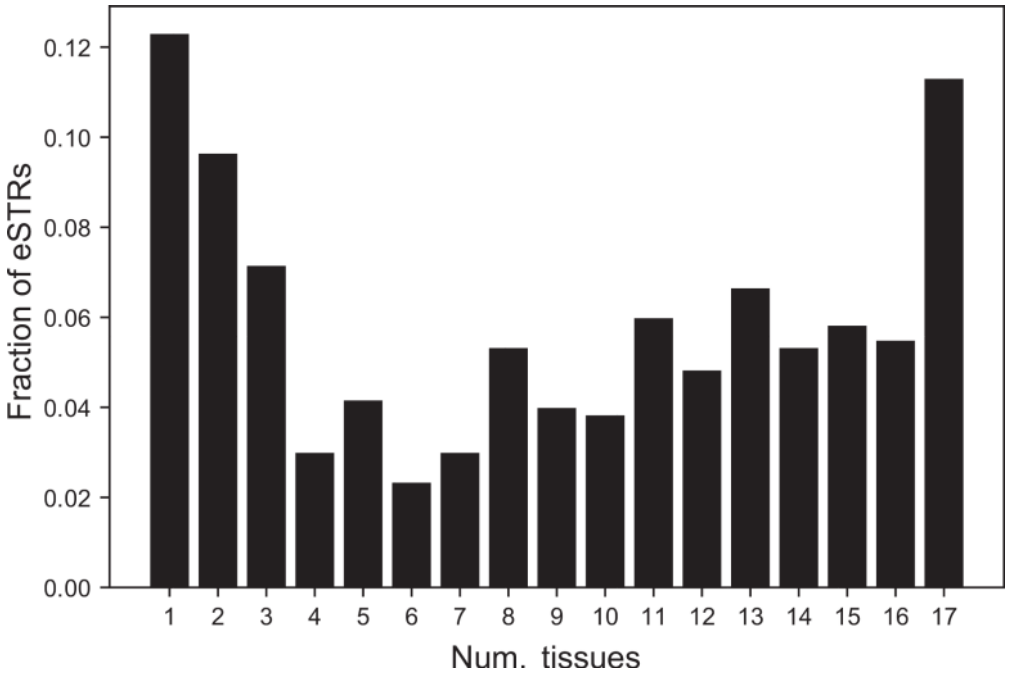
Extended Data Fig. 2: Enrichment of genomic annotations as a function of CAVIAR threshold

The x-axis represents CAVIAR thresholds in terms of the percentile (percentage of all 28,375 eSTRs excluded by those thresholds). The y-axis represents the odds ratio for enrichment in eSTRs above each percentile threshold in each of these categories: **a.** 5'UTRs (purple); **b.** 3'UTRs (blue); **c.** promoters (orange; TSS +/− 3kb); **d.** Coding regions (red) and **e.** Introns (green). The y-axis center values denote the $\log_2$ odds ratios comparing eSTRs passing each threshold to all STRs. Error bars represent +/−1 s.e.



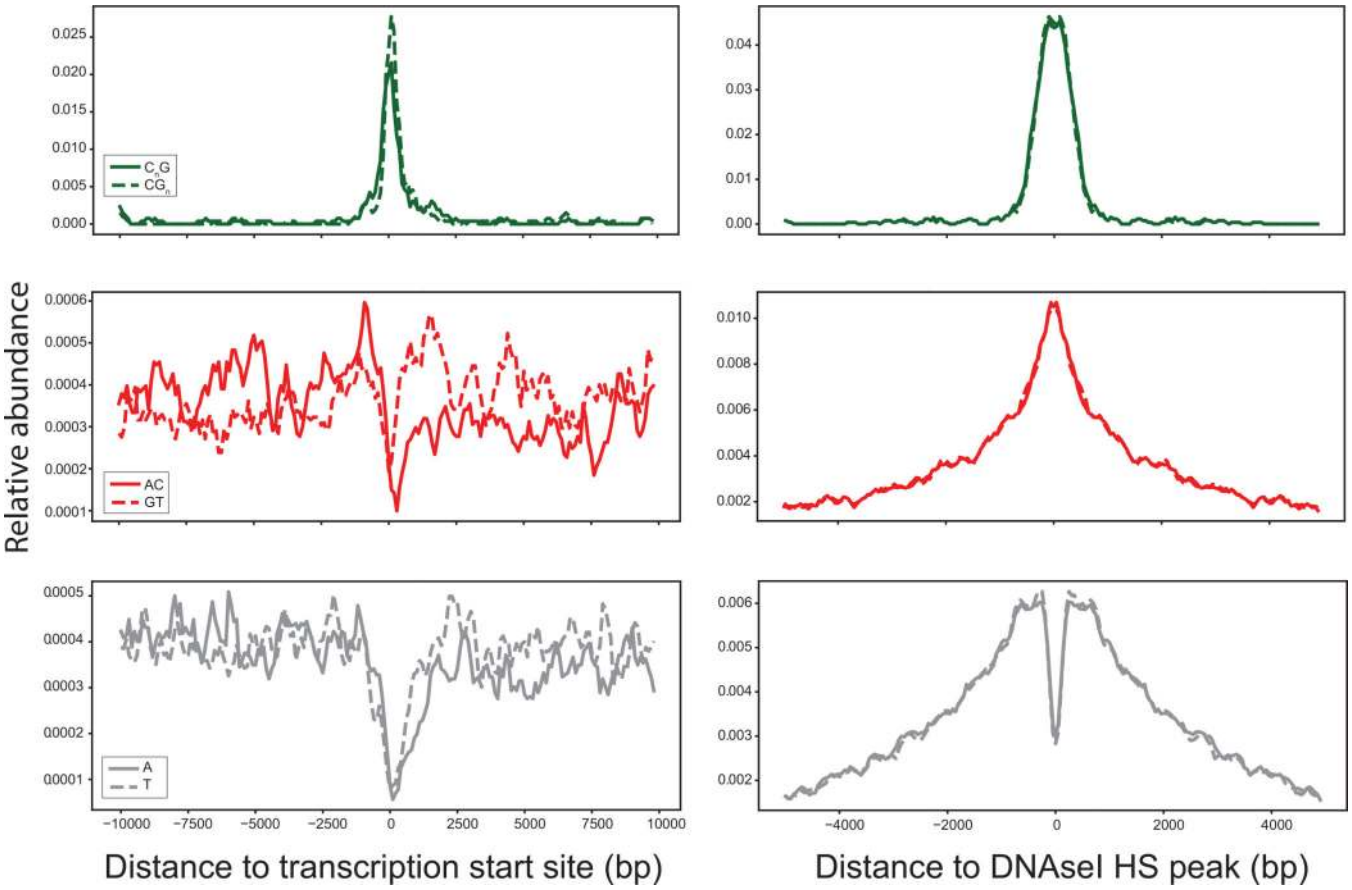
Extended Data Fig. 3: Example multi-allelic FM-eSTRs

For each plot, the x-axis represents the mean number of repeats in each individual and the y-axis represents normalized expression in the tissue for which the eSTR was most significant. Boxplots summarize the distribution of expression values for each genotype. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1–1.5*IQR (bottom) and Q3+1.5*IQR (top), where IQR gives the interquartile range (Q3-Q1). The red line shows the mean expression for each x-axis value.



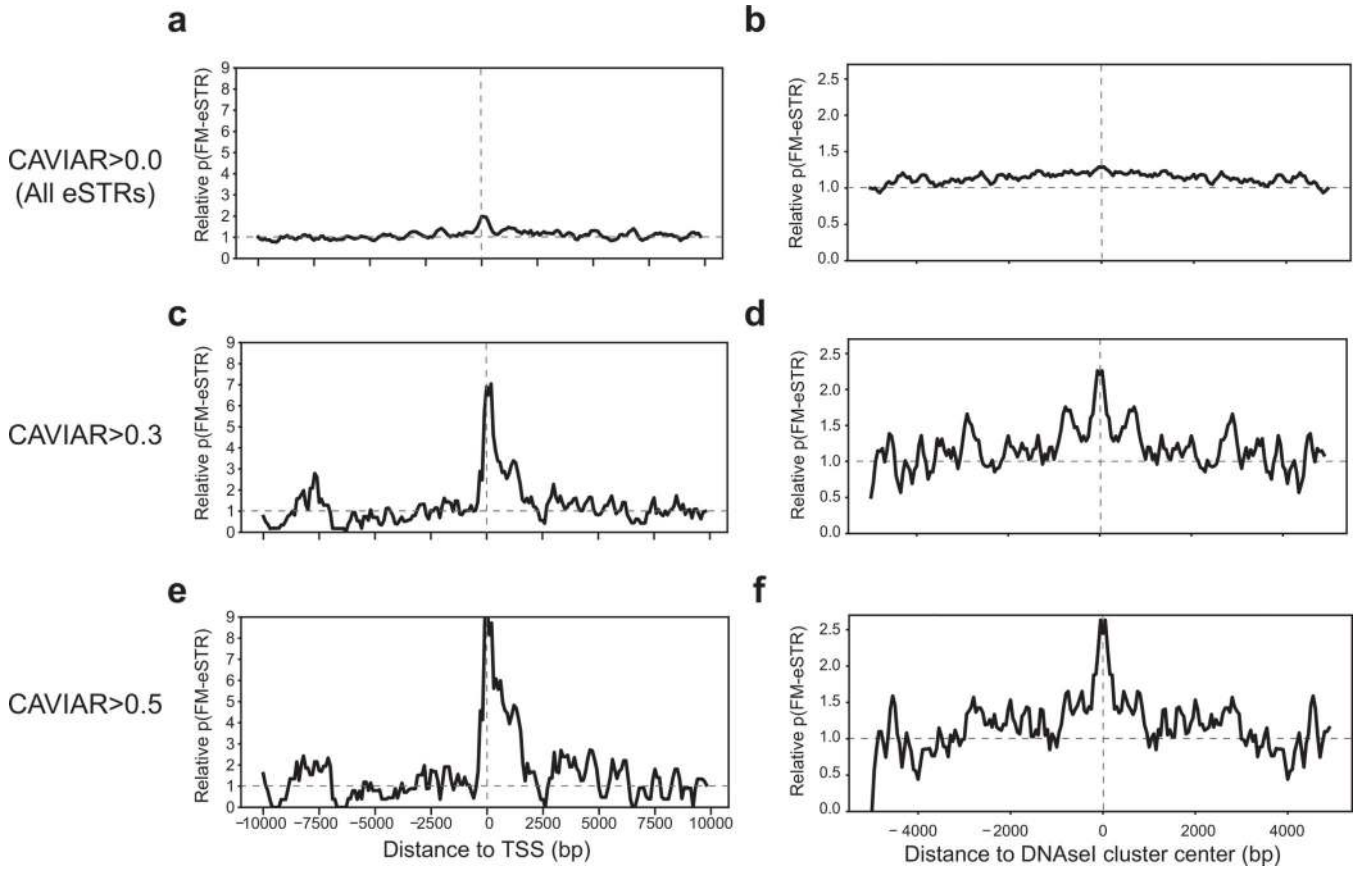
Extended Data Fig. 4: Sharing of eSTRs across tissues

The x-axis represents the number of tissues that share a given eSTR (absolute value of mashR Z-score >4). The y-axis represents the number of eSTRs shared across a given number of tissues.



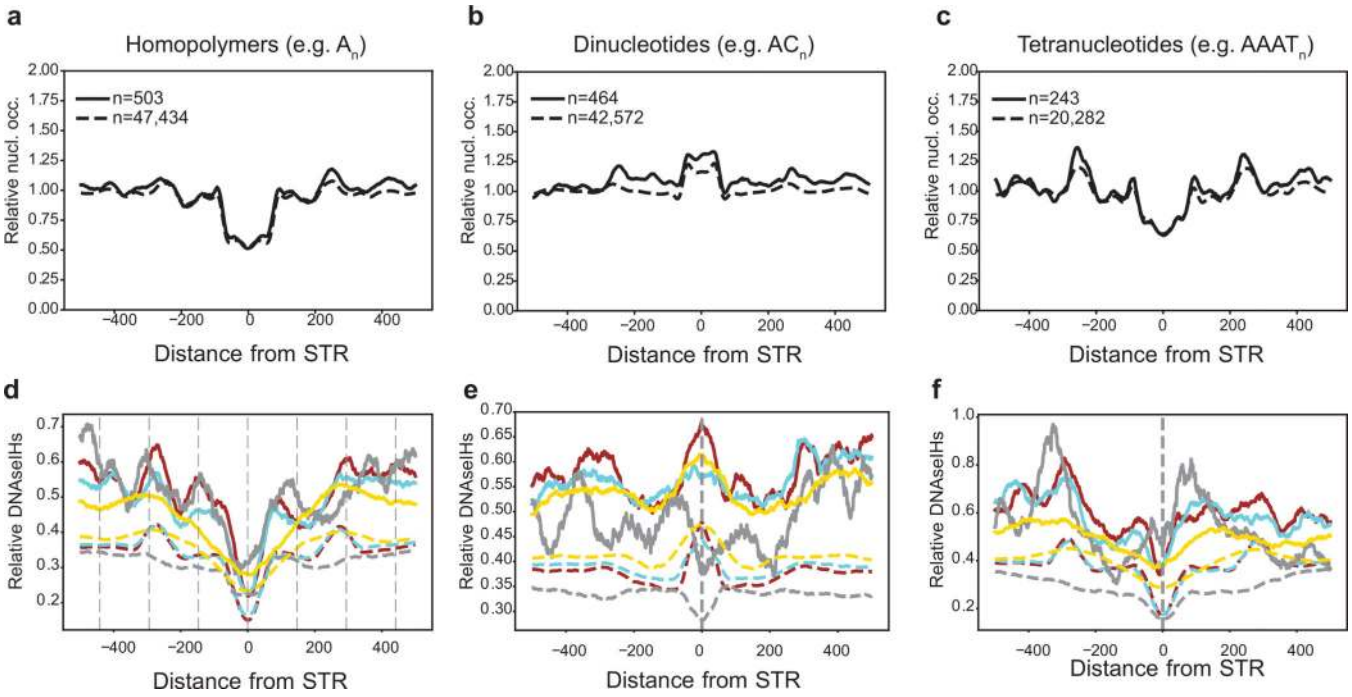
Extended Data Fig. 5: Localization of all STRs around putative regulatory regions

Left and right plots show localization around transcription start sites and DNAseI HS clusters, respectively. The y-axis denotes the fraction of STRs of each type in each bin. For promoters, the x-axis is divided into 100bp bins. For DNAseI HS sites, the x-axis is divided into 50bp bins. In each plot, values were smoothed by taking a sliding average of each four consecutive bins. Only STR-gene pairs included in our analysis are considered. Each plot compares localization of the two possible sequences of a given repeat unit on the coding strand. *i.e.* top plots compare repeat units of the form C_nG vs. their reverse complement on the opposite strand, middle plots compare AC vs. GT repeats, and bottom plots compare A vs. T repeats. The strand of each STR was determined based on the coding strand of each target gene.



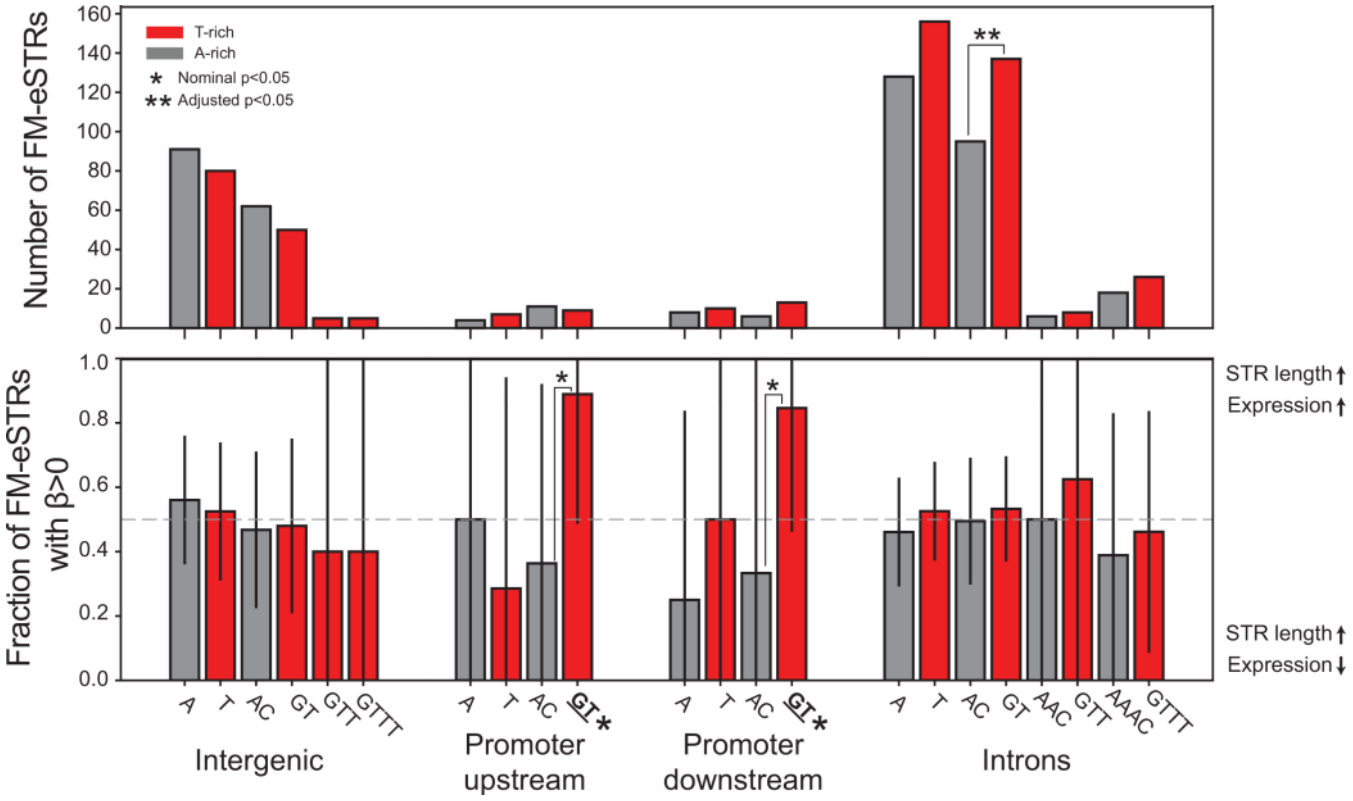
Extended Data Fig. 6: Relative probability of eSTRs around TSSs and DNAseI HS sites for a range of CAVIAR scores

Plots are shown for FM-eSTRs defined using multiple CAVIAR thresholds (0, corresponding to all eSTRs, 0.3, as used in the main text, or 0.5). **a., c.,** and **e.** show the relative probability of an STR to be an FM-eSTR around TSSs. The black lines represent the probability of an STR in each bin to be an FM-eSTR. Values were scaled relative to the genome-wide average. **b., d.,** and **f.** show the relative probability of an STR to be an FM-eSTR around DNAseI HS clusters. Values were smoothed by taking a sliding average of each four consecutive bins.



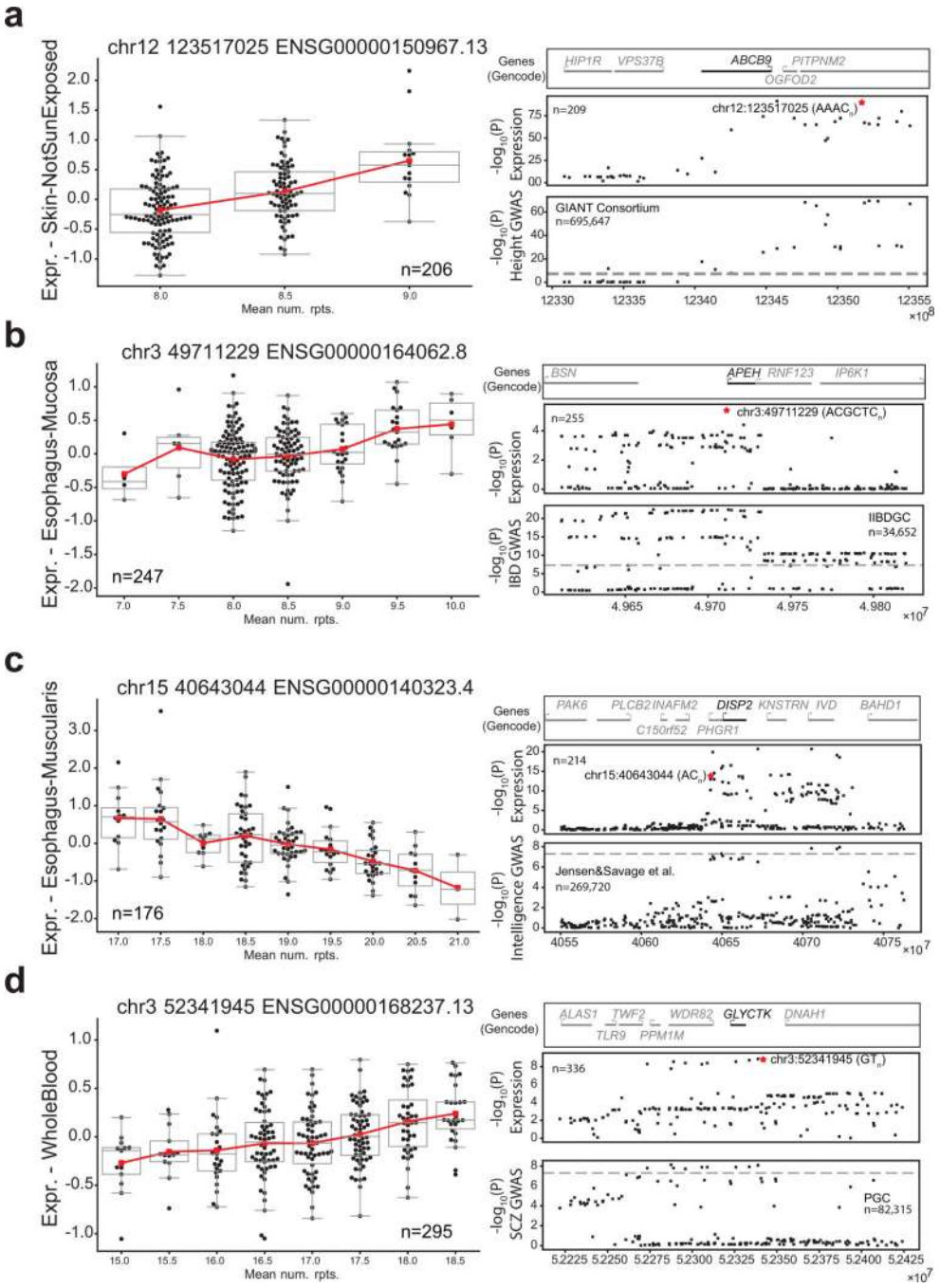
Extended Data Fig. 7: Nucleosome occupancy and DNAseI hypersensitivity show distinct patterns around eSTRs

a-c. Nucleosome density around STRs with different repeat unit lengths. Nucleosome density in GM12878 in 5bp windows is averaged across all STRs analyzed (dashed) and FM-eSTRs (solid) relative to the center of the STR. **b. DNAseI HS density around STRs with different repeat unit lengths.** The number of DNAseI HS reads in GM12878 (gray), fat (red), tibial nerve (yellow), and skin (cyan) is averaged across all STRs in each category. Solid lines show FM-eSTRs. Dashed lines show all STRs. Left=homopolymers, middle=dinucleotides, right=tetranucleotides. Other repeat unit lengths were excluded since they have low numbers of FM-eSTRs (see Fig. 4a). Dashed vertical lines in **(d)** show the STR position +/− 147bp.



Extended Data Fig. 8: Strand-biased characteristics of FM-eSTRs

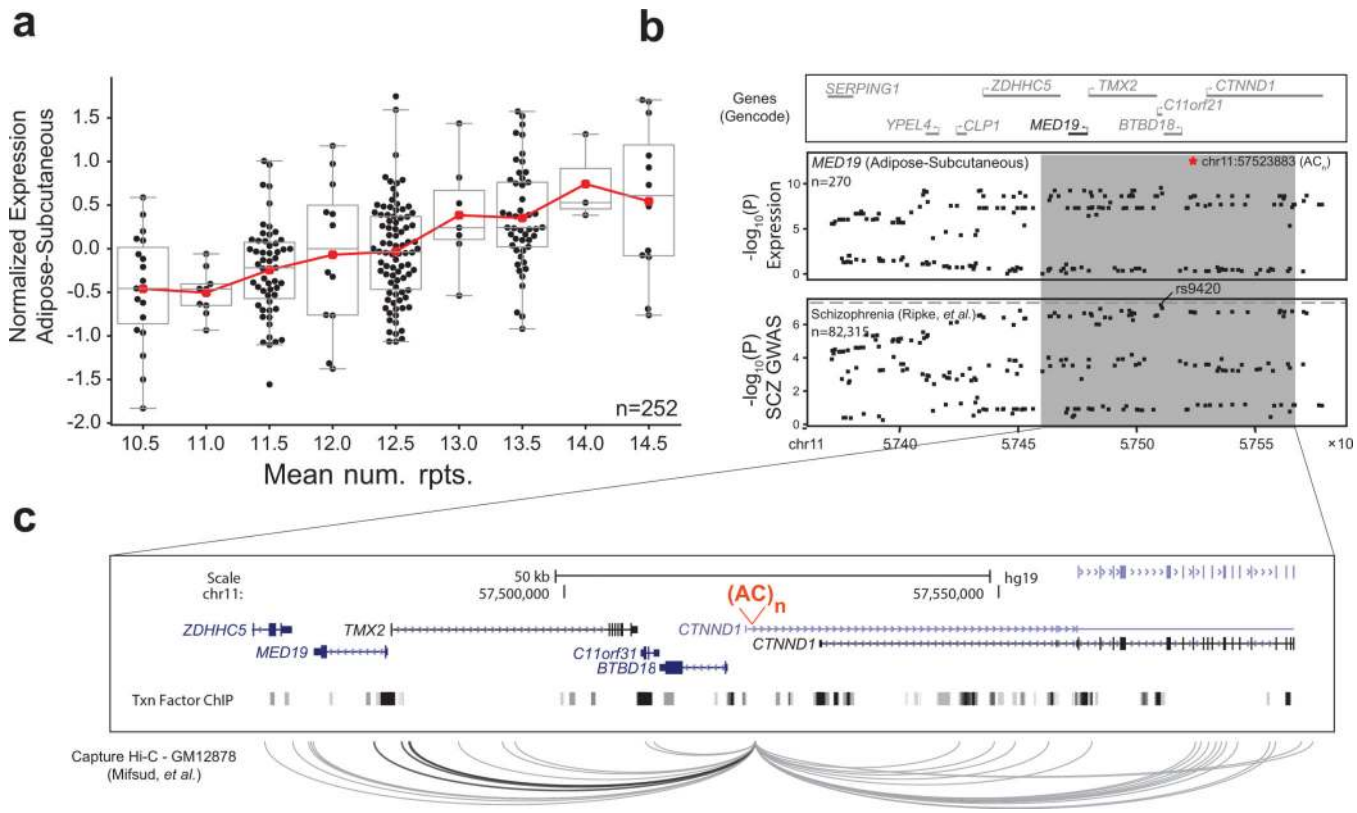
Top panel: the y-axis shows the number of FM-eSTRs with each repeat unit on the template strand. Bottom panel: the y-axis shows the percentage of FM-eSTRs with each repeat unit on the template strand that have positive effect sizes. Gray bars denote A-rich repeat units (A/AC/AAC/AAAC) and red bars denote T-rich repeat units (T/GT/GTT/GTTT). Single asterisks denote repeat units nominally enriched or depleted (two-sided binomial $p<0.05$). Double asterisks denote repeat units significantly enriched after controlling for multiple hypothesis testing (Bonferroni adjusted $p<0.05$). Asterisks above brackets show significant differences between repeat unit pairs. Asterisks on x-axis labels denote departure from the 50% positive effect sizes expected by chance. Error bars give 95% confidence intervals.



Extended Data Fig. 9: Example GWAS signals co-localized with FM-eSTRs

Left: For each plot, the x-axis represents the mean number of repeats in each individual and the y-axis represents normalized expression in the tissue with the most significant eSTR signal at each locus. Boxplots summarize the distribution of expression values for each genotype. Box plots are as defined in Fig. 1c. The red line shows the mean expression for each x-axis value. Right: Top panels give genes in each region. The target gene for the eQTL associations is shown in black. Middle panels give the -log$_{10}$ p-values of association of the effect-size between each SNP (black points) and the expression of the target gene. The FM-

eSTR is denoted by a red star. Bottom panels give the -$\log_{10}$ p-values of association between each SNP and the trait based on published GWAS summary statistics. P-values are two-sided and are based on t-statistics computed for effect sizes (β) (see Methods). Dashed gray horizontal lines give the genome-wide significance threshold of 5E-8.



Extended Data Fig. 10: Example GWAS signal for schizophrenia potentially driven by an eSTR for *MED19*

a. eSTR association for *MED19*. The x-axis shows STR genotypes at an AC repeat (chr11:57523883) as the mean number of repeats in each individual and the y-axis shows normalized *MED19* expression in subcutaneous adipose. Each point represents a single individual. Red lines show the mean expression for each x-axis value. Boxplots are as defined in Fig. 1c. **b. Summary statistics for *MED19* expression and schizophrenia.** The top panel shows genes in the region around *MED19*. The middle panel shows the -$\log_{10}$ p-values of association between each variant and *MED19* expression in subcutaneous adipose tissue in the GTEx cohort. The FM-eSTR is denoted by a red star. The bottom panel shows the -$\log_{10}$ p-values of association for each variant with schizophrenia reported by the Psychiatric Genomics Consortium. The dashed gray horizontal line shows genome-wide significance threshold of 5E-8. **c. Detailed view of the *MED19* locus.** A UCSC genome browser screenshot is shown for the region in the gray box in **(b)**. The FM-eSTR is shown in red. The bottom track shows transcription factor (TF) and chromatin regulator binding sites profiled by ENCODE. The bottom panel shows long-range interactions reported by Mifsud, *et al.* using Capture Hi-C on GM12878. Interactions shown in black include *MED19*. Interactions to loci outside of the window depicted are not shown.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Research reported in this publication was supported in part by the Office Of The Director, National Institutes of Health under Award Number DP5OD024577 (M.G.). We thank V. Bafna, E. Mendenhall, J. Gleeson, and Y. Liu for helpful comments. See the Supplementary Note for additional acknowledgements.

References for Main Text

1. Consortium, G. et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017). [PubMed: 29022597]
2. Lappalainen T et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–11 (2013). [PubMed: 24037378]
3. Grünewald TGP et al. Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat. Genet. 47, 1073–1078 (2015). [PubMed: 26214589]
4. Song JHT, Lowe CB & Kingsley DM Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. Am J Hum Genet 103, 421–430 (2018). [PubMed: 30100087]
5. Boettger LM et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat Genet 48, 359–66 (2016). [PubMed: 26901066]
6. Leffler EM et al. Resistance to malaria through structural variation of red blood cell invasion receptors. Science 356(2017).
7. Sekar A et al. Schizophrenia risk from complex variation of complement component 4. Nature 530, 177–183 (2016). [PubMed: 26814963]
8. Sun JX et al. A direct characterization of human mutation based on microsatellites. Nat. Genet 44, 1161–1165 (2012). [PubMed: 22922873]
9. Lynch M Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A 107, 961–8 (2010). [PubMed: 20080596]
10. Willems T et al. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. Am. J. Hum. Genet 98, 919–933 (2016). [PubMed: 27126583]
11. Mirkin SM Expandable DNA repeats and human disease. Nature 447, 932–940 (2007). [PubMed: 17581576]
12. Willems T et al. The landscape of human STR variation. Genome Res. 24, 1894–1904 (2014). [PubMed: 25135957]
13. Li H Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. arXiv [q-bio.GN] (2014).
14. Gymrek M et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 48, 22–9 (2016). [PubMed: 26642241]
15. Nasrallah MP et al. Differential effects of a polyalanine tract expansion in Arx on neural development and gene expression. Hum Mol Genet 21, 1090–8 (2012). [PubMed: 22108177]
16. Quilez J et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. 44, 3750–3762 (2016). [PubMed: 27060133]
17. Vinces MD, Legendre M, Caldara M, Hagihara M & Verstrepen KJ Unstable tandem repeats in promoters confer transcriptional evolvability. Science 324, 1213–1216 (2009). [PubMed: 19478187]
18. Gemayel R, Vinces MD, Legendre M & Verstrepen KJ Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44, 445–77 (2010). [PubMed: 20809801]
19. Liu XS et al. Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. Cell 172, 979–992 e6 (2018). [PubMed: 29456084]
20. Raveh-Sadka T et al. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat Genet 44, 743–50 (2012). [PubMed: 22634752]

21. Suter B, Schnappauf G & Thoma F Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res 28, 4083–9 (2000). [PubMed: 11058103]
22. Afek A, Schipper JL, Horton J, Gordan R & Lukatsky DB Protein-DNA binding in the absence of specific base-pair recognition. Proc Natl Acad Sci U S A 111, 17140–5 (2014). [PubMed: 25313048]
23. Conlon EG et al. The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. Elife 5(2016).
24. Lin Y, Dent SY, Wilson JH, Wells RD & Napierala M R loops stimulate genetic instability of CTG.CAG repeats. Proc Natl Acad Sci U S A 107, 692–7 (2010). [PubMed: 20080737]
25. Rothenburg S, Koch-Nolte F, Rich A & Haag F A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. Proc. Natl. Acad. Sci. U. S. A 98, 8985–8990 (2001). [PubMed: 11447254]
26. Min JL et al. The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. PLoS One 6, e22070 (2011). [PubMed: 21789213]
27. Willems T et al. Genome-wide profiling of heritable and de novo STR variations. Nat. Methods (2017).
28. Borel C et al. Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. Hum. Mutat. 33, 1302–1309 (2012). [PubMed: 22573514]
29. Contente A, Dittmer A, Koch MC, Roth J & Dobbelstein M A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat. Genet 30, 315–320 (2002). [PubMed: 11919562]
30. Gebhardt F, Zänker KS & Brandt B Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. J. Biol. Chem 274, 13176–13180 (1999). [PubMed: 10224073]
31. Johnson AD et al. Genome-wide association meta-analysis for total serum bilirubin levels. Hum Mol Genet 18, 2700–10 (2009). [PubMed: 19414484]
32. Matsuzono K et al. Antisense Oligonucleotides Reduce RNA Foci in Spinocerebellar Ataxia 36 Patient iPSCs. Mol Ther Nucleic Acids 8, 211–219 (2017). [PubMed: 28918022]
33. Saha A et al. Functional IFNG polymorphism in intron 1 in association with an increased risk to promote sporadic breast cancer. Immunogenetics 57, 165–71 (2005). [PubMed: 15900487]
34. Shimajiri S et al. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. FEBS Lett. 455, 70–74 (1999). [PubMed: 10428474]
35. Vikman S et al. Functional analysis of 5-lipoxygenase promoter repeat variants. Hum Mol Genet 18, 4521–9 (2009). [PubMed: 19717473]
36. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B & Eskin E Identifying causal variants at loci with multiple signals of association. Genetics 198, 497–508 (2014). [PubMed: 25104515]
37. Kobayashi H et al. Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am J Hum Genet 89, 121–30 (2011). [PubMed: 21683323]
38. Lalioti MD et al. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. Nature 386, 847–51 (1997). [PubMed: 9126745]
39. Mougey E et al. ALOX5 polymorphism associates with increased leukotriene production and reduced lung function and asthma control in children with poorly controlled asthma. Clin Exp Allergy 43, 512–20 (2013). [PubMed: 23600541]
40. Stephensen CB et al. ALOX5 gene variants affect eicosanoid production and response to fish oil supplementation. J Lipid Res 52, 991–1003 (2011). [PubMed: 21296957]
41. Urbut SM, Wang G, Carbonetto P & Stephens M Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet 51, 187–195 (2019). [PubMed: 30478440]
42. Jiang C & Pugh BF Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10, 161–72 (2009). [PubMed: 19204718]

43. Bochman ML, Paeschke K & Zakian VA DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet 13, 770–80 (2012). [PubMed: 23032257]
44. Ciesiolka A, Jazurek M, Drazkowska K & Krzyzosiak WJ Structural Characteristics of Simple RNA Repeats Associated with Disease and their Deleterious Protein Interactions. Front Cell Neurosci 11, 97 (2017). [PubMed: 28442996]
45. MacArthur J et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45, D896–D901 (2017). [PubMed: 27899670]
46. Yengo L et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. bioRxiv (2018).
47. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427 (2014). [PubMed: 25056061]
48. Liu JZ et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 47, 979–986 (2015). [PubMed: 26192919]
49. Savage JE et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet 50, 912–919 (2018). [PubMed: 29942086]
50. Guo H et al. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. Hum Mol Genet 24, 3305–13 (2015). [PubMed: 25743184]
51. Haeuptle MA et al. Human RFT1 deficiency leads to a disorder of N-linked glycosylation. Am J Hum Genet 82, 600–6 (2008). [PubMed: 18313027]
52. Saini S, Mitra I, Mousavi N, Fotsing SF & Gymrek M A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat. Commun. 9, 4397 (2018). [PubMed: 30353011]
53. Chiang C et al. The impact of structural variation on human gene expression. Nat Genet 49, 692–699 (2017). [PubMed: 28369037]
54. Hasler J & Strub K Alu elements as regulators of gene expression. Nucleic Acids Res 34, 5491–7 (2006). [PubMed: 17020921]

Methods-only references

55. Kent WJ et al. The human genome browser at UCSC. Genome Res. 12, 996–1006 (2002). [PubMed: 12045153]
56. Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]
57. Purcell S et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559–75 (2007). [PubMed: 17701901]
58. Patterson N, Price AL & Reich D Population structure and eigenanalysis. PLoS Genet 2, e190 (2006). [PubMed: 17194218]
59. Price AL et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–9 (2006). [PubMed: 16862161]
60. Stegle O, Parts L, Piipari M, Winn J & Durbin R Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 7, 500–7 (2012). [PubMed: 22343431]
61. Seabold SP, Josef. Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference (2010).
62. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–2 (2010). [PubMed: 20110278]
63. Heinz S et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–89 (2010). [PubMed: 20513432]
64. Zuker M Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406–15 (2003). [PubMed: 12824337]

65. Wang Y et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol 19, 151 (2018). [PubMed: 30286773]

66. Mifsud B et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 47, 598–606 (2015). [PubMed: 25938943]

67. Howie BN, Donnelly P & Marchini J A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009). [PubMed: 19543373]

68. Browning BL, Zhou Y & Browning SR A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet 103, 338–348 (2018). [PubMed: 30100085]

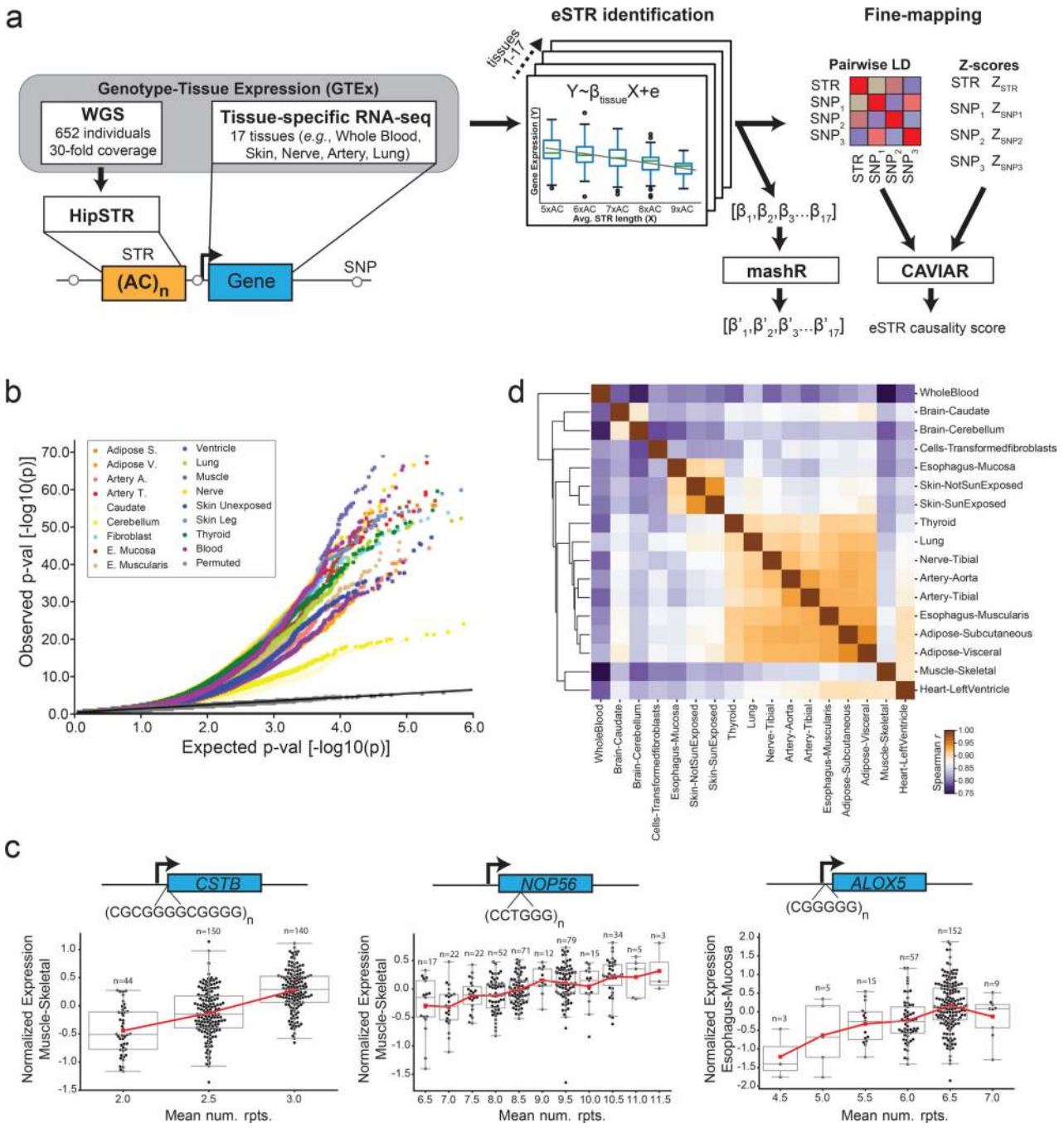


Figure 1: Multi-tissue identification of eSTRs.

(a) Schematic of eSTR discovery pipeline. We analyzed eSTRs using RNA-seq from 17 tissues and STR genotypes obtained from deep WGS for 652 individuals from the GTEx Project.

(**b) eSTR association results.** The quantile-quantile plot compares observed p-values for each STR-gene test vs. the expected uniform distribution for each tissue. Gray dots denote permutation controls (n = 336). Supplementary Table 1 gives the number of tests performed in each tissue.

(c) Example eSTRs previously implicated in disease. Example FM-eSTRs previously implicated in myoclonus epilepsy (left), spinocerebellar ataxia 36 (middle), and reduced lung function and cardiovascular disease (right) are shown. Black points represent single individuals. For each plot, the x-axis represents the mean number of repeats in each individual and the y-axis represents normalized expression in a representative tissue. Boxplots summarize the distribution of expression values. Horizontal lines show median values, boxes span from the 25th percentile (Q1) to the 75th percentile (Q3). Whiskers extend to Q1–1.5*IQR (bottom) and Q3+1.5*IQR (top), where IQR gives the interquartile range (Q3-Q1). The red line shows the mean expression for each x-axis value. Gene diagrams not drawn to scale.

(d) eSTR correlations across tissues. Each cell shows the Spearman correlation between mashR FM-eSTR effect sizes for each pair of tissues. Only eSTRs with CAVIAR score >0.3 (FM-eSTRs) in one of the two tissues were included in each correlation. Supplementary Table 1 gives the number of FM-eSTRs identified in each tissue. Rows and columns were clustered using hierarchical clustering (Methods).

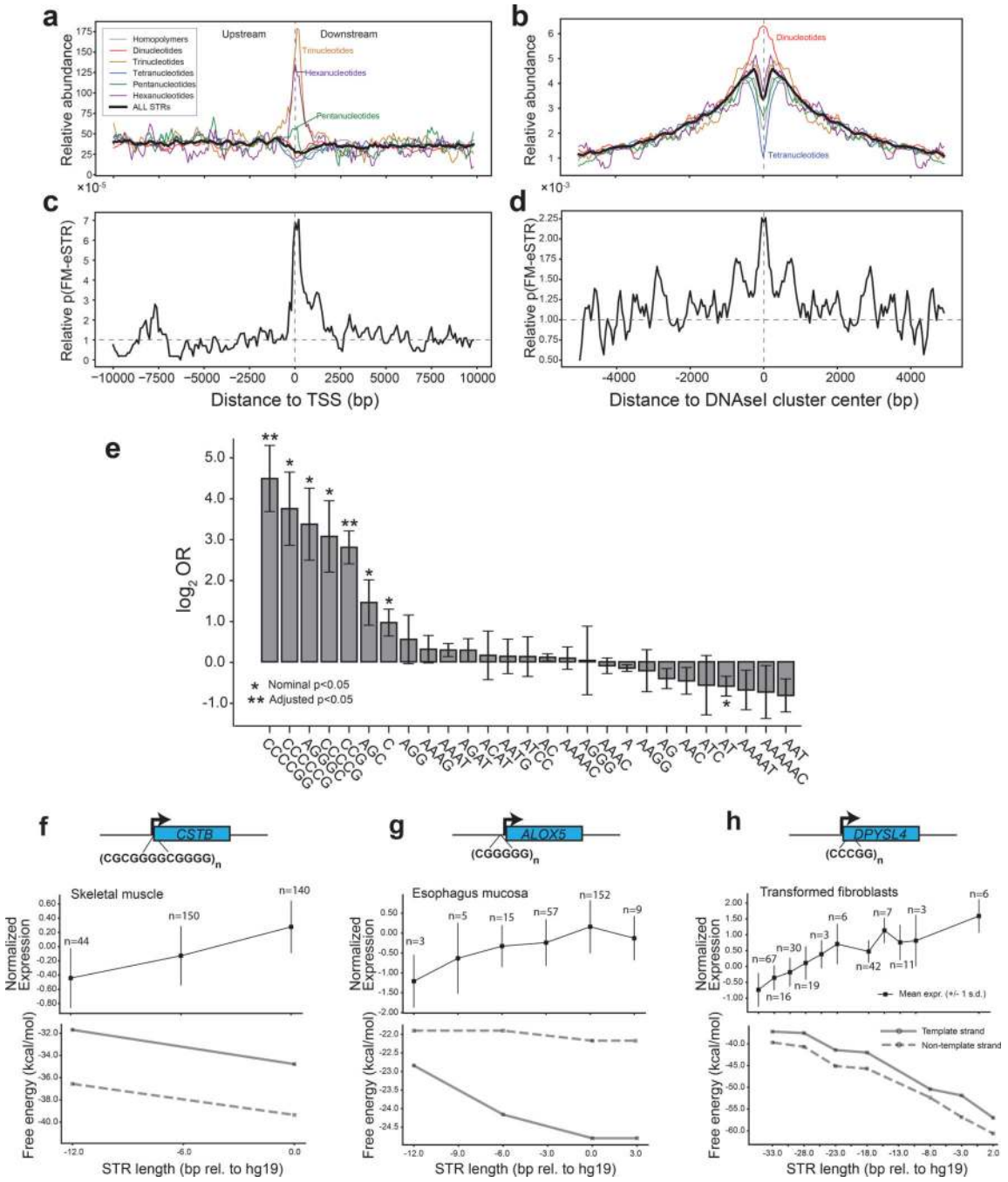


Figure 2: Characterization of FM-eSTRs

(a) Density of all STRs around transcription start sites (TSS). The y-axis shows the fraction of STRs with each repeat unit type located in each 100 bp bin around the TSS.

(**b) Density of all STRs around DNAseI hypersensitive sites.** Plots are centered at ENCODE DNAseI HS clusters and represent the fraction of STRs with each repeat unit type located in each 50 bp bin.

(**c) Relative probability to be an FM-eSTR around TSSs**.

(**d) Relative probability to be an FM-eSTR around DNAseI HS clusters.** For **a-d,** values were smoothed using a sliding average of each four consecutive bins.

(**e) Repeat unit enrichment at FM-eSTRs**. The x-axis shows all repeat units for which there are at least 3 FM-eSTRs across all tissues. The y-axis center values denote the $\log_2$ odds ratios comparing FM-eSTRs to all STRs. Error bars represent ± 1 s.e. Asterisks denote repeat units that are significantly enriched or depleted in FM-eSTRs (based on two-sided Fisher exact p-value). Per repeat unit sample sizes and Fisher exact statistics are provided in Supplementary Table 5.

(**f-h) Example GC-rich FM-eSTRs in promoters predicted to modulate secondary structure**. Top plots show mean expression across all individuals with each mean STR length. Vertical bars represent ± 1 s.d. Bottom plots show the free energy computed for each allele based on template (solid) and non-template (dashed) strands. The x-axis shows STR lengths relative to hg19 (bp). Gene diagrams are not drawn to scale.

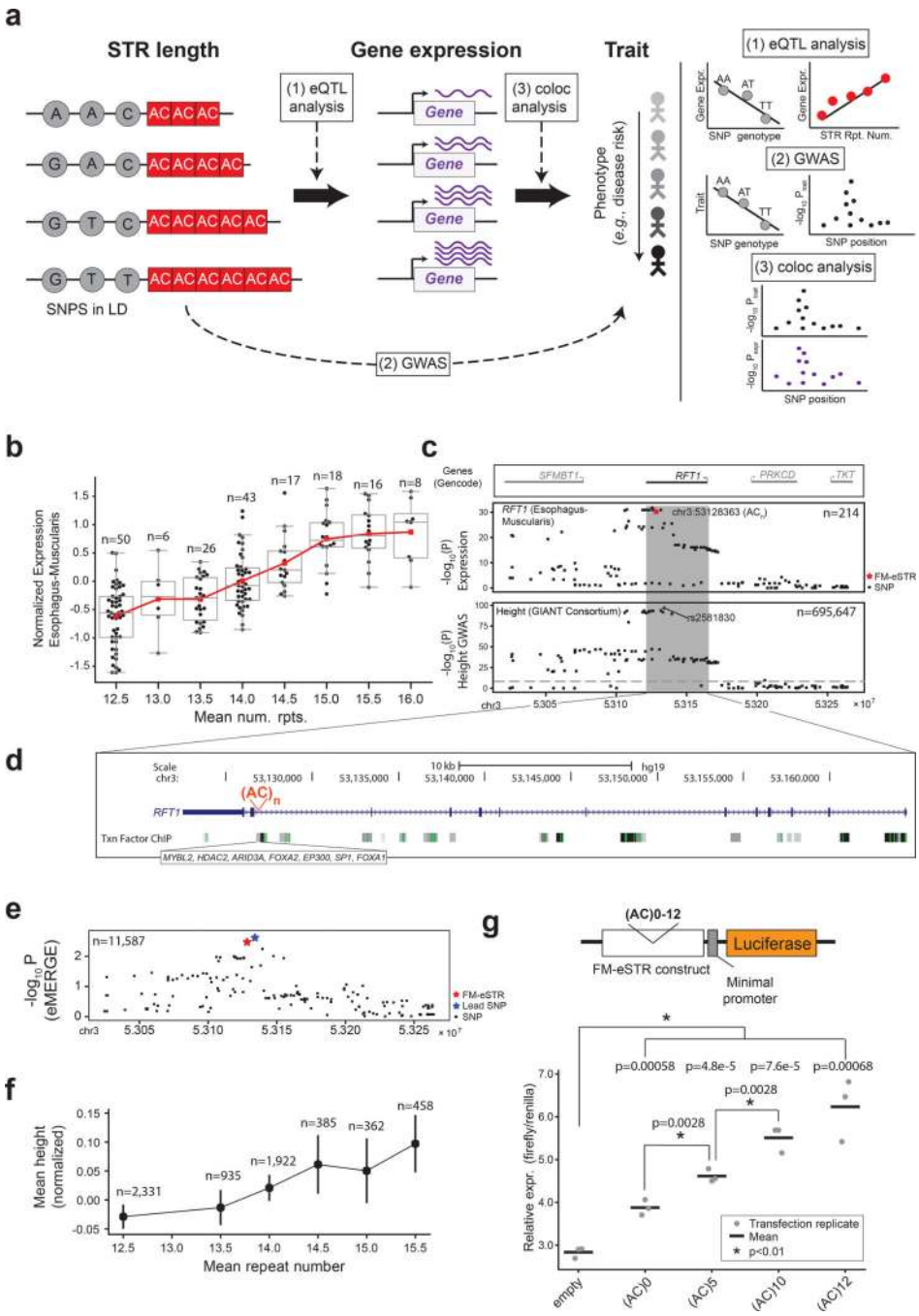


Figure 3: FM-eSTRs co-localize with GWAS signals.

(a) Overview of analyses to identify FM-eSTRs involved in complex traits. We assumed a model where variation in STR repeat number alters gene expression, which in turn affects the value of a particular complex trait.

(b) eSTR association for *RFT1*. The x-axis shows STR genotype as the mean number of AC repeats and the y-axis gives normalized *RFT1* expression. Boxplots defined as in Fig. 1c.

(c) Summary statistics for *RFT1* expression and height. The middle panel shows the -$\log_{10}$ p-values of association between each variant and *RFT1* expression. The bottom panel

shows the -$\log_{10}$ p-values of association for each variant with height. Black dots=SNPs; red star=FM-eSTR; gray dashed line=genome-wide significance threshold.
(d) Genomic view of the *RFT1* locus.
(e) eSTR and SNP associations with height in the eMERGE cohort. The y-axis denotes association p-values for each variant. Black dots=SNPs; red star=imputed FM-eSTR; blue star=top eMERGE SNP.
(f) Imputed *RFT1* repeat number is correlated with height. The x-axis shows the mean number of AC repeats. The y-axis shows the mean normalized height for all samples included in the analysis with a given genotype. Error bars show ± 1 s.e.
(g) Reporter assay testing repeat number vs. expression. A variable number of AC repeats plus genomic context were introduced upstream of a reporter gene. Gray dots show the value for each of n=3 transfections, each averaged across three technical replicates. Black lines show the mean across the three transfections.

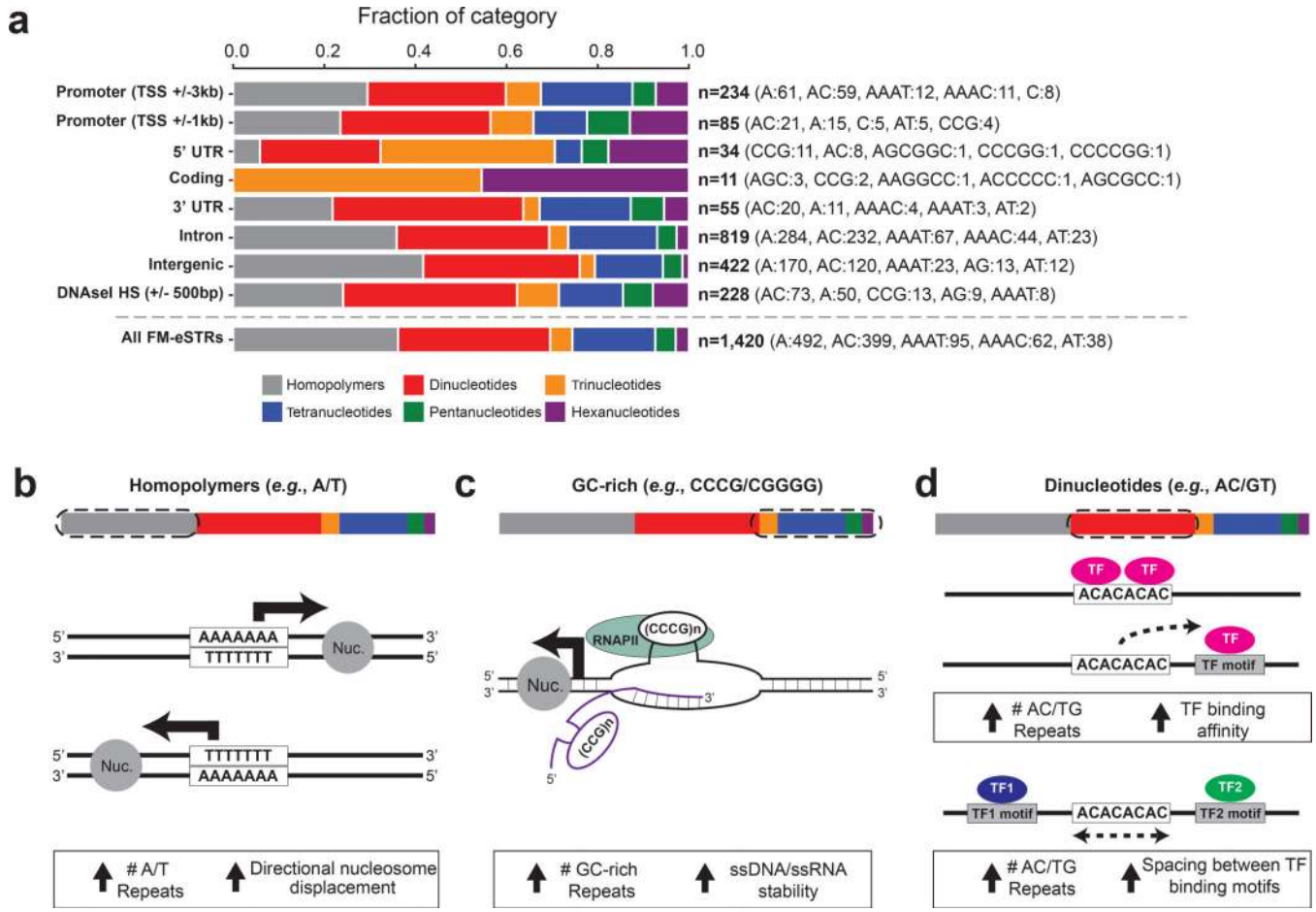


Figure 4: Summary of FM-eSTRs classes and potential regulatory mechanisms

(a) Distribution of FM-eSTR classes across genomic annotations. Each bar shows the fraction of FM-eSTRs falling in each annotation consisting of homopolymer (gray), dinucleotide (red), trinucleotide (orange), tetranucleotide (blue), pentanucleotide (green) or hexanucleotide (purple) repeats. The total number of FM-eSTRs and the top five most common repeat units in each category are shown on the right. Note, FM-eSTRs may be counted in more than one category.

(b) Homopolymer A/T STRs are predicted to modulate nucleosome positioning. Homopolymer repeats are depleted of nucleosomes (gray circles) and may modulate expression changes in nearby genes through altering nucleosome positioning.

(c) GC-rich STRs form DNA and RNA secondary structures during transcription. Highly stable secondary structures such as G4 quadruplexes may act by expelling nucleosomes (gray circle) or stabilizing RNAPII (light green circle). These structures may form in DNA (black) or RNA (purple) The stability of the structure can depend on the number of repeats.

(d) Dinucleotide STRs can alter transcription factor binding. Dinucleotides are prevalent in putative enhancer regions. They may potentially alter transcription factor binding by forming binding sites themselves (top), changing affinity of nearby binding sites (middle), or modulating spacing between nearby binding sites (bottom).

For **(b)-(d)**, text and arrows in the white boxes provide a summary of the predicted eSTR mechanism depicted in each panel.