The impact of short tandem repeat variation on gene expression

Stephanie Feupe Fotsing, Jonathan Margoliash, Catherine Wang, Shubham Saini, Richard Yanicky, Sharona Shleizer-Burko, Alon Goren and Melissa Gymrek

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 use 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 genome-wide association study 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 should serve as a valuable resource for future studies of complex traits.

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. Most of these have focused on biallelic SNPs or short indels. Yet multiple studies dissecting genome-wide association study (GWAS) loci have found repetitive and structural variants to be the underlying causal variants, highlighting the need to consider additional variant classes beyond SNPs.

Short tandem repeats, consisting of consecutively repeated units of 1–6 base pairs (bp), represent a large source of genetic variation. STR mutation rates are orders of magnitude higher than those of SNPs and short indels, and each individual is estimated to harbor around 100 de novo mutations in STRs. Expansions at several dozen STRs have been known for decades to cause mendelian disorders, 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. Due to bioinformatics challenges of analyzing repetitive regions, many STRs are often filtered from genome-wide studies. However, increasing evidence supports a widespread role of common variation at STRs in complex traits, such as gene expression.

STRs may regulate gene expression through a variety of mechanisms. For example, the CCG repeat implicated in fragile X syndrome was shown to disrupt DNA methylation, altering expression of FMR1. Yeast studies have demonstrated that homopolymer repeats act as nucleosome positioning signals with downstream regulatory effects. Dinucleotide repeats may alter affinity of nearby DNA-binding sites. Furthermore, certain STR repeat units may form noncanonical DNA and RNA secondary structures such as G-quadruplexes, R-loops and Z-DNA.

We previously identified more than 2,000 STRs for which the number of repeats was associated with the expression of nearby genes, 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 (4–6×) whole-genome sequencing data performed using short reads (50–100 bp), which are unable to span many STRs. As a result, STR genotype calls exhibited poor quality with less than 50% genotyping accuracy. Additionally, the study used a single cell type (lymphoblastoid cell lines) with potentially limited relevance to most complex traits.

While our study and others demonstrated that eSTRs explain a sizable portion (10–15%) of the 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) 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 dataset for which both high-coverage WGS and...
RNA-sequencing of multiple tissues were available (Fig. 1a). We used HipSTR\textsuperscript{2} 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 kilobases (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 and Supplementary Figs. 1–3). Analysis was restricted to 17 tissues where we had data for at least 100 samples (Supplementary Table 1 and 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 and 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 lymphoblastoid cell lines 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\textsuperscript{30–35} 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\textsuperscript{32}, 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 and 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 and Supplementary Data 2). Expected gene annotations are more strongly enriched (CAVIAR score $> 0.3$) with each gene under our model (Methods and Fig. 1a). On average across tissues, 12.2\% of eSTRs had the highest causality scores of all variants tested.

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 also provide results recomputed using a range of score thresholds in the Supplementary Information. 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 differs from that of STRs overall (Fig. 2a,b and 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 and near DNsase I hypersensitive (HS) sites are more likely to be FM-eSTRs (Fig. 2c,d and Extended Data Fig. 6). FM-eSTRs are strongly enriched at 5' UTRs (odds ratio (OR) = 5.0; Fisher's two-sided $P = 4.9 \times 10^{-12}$), 3' UTRs (OR = 2.78; $P = 5.85 \times 10^{-18}$) and within 3 kb of transcription start sites (OR = 3.39; $P = 3.94 \times 10^{-20}$). 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 DNsase-seq) in a variety of cell and tissue types within 500 bp of FM-eSTRs (Extended Data Fig. 7). As expected from previous studies\textsuperscript{14,31}, 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 DNsase-seq read count compared to non-eSTRs (Mann–Whitney $U$-test two-sided $P = 3.9 \times 10^{-37}$ in GM12878; Extended Data Fig. 7d–f). DNsase I 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\textsuperscript{45}, we hypothesize that a subset of homopolymer FM-eSTRs may act by shifting nucleosome positions and thus modulating the accessibility of adjacent sites.

Next, we compared the sequence characteristics of FM-eSTRs with all STRs. We find that the total lengths of FM-eSTRs are significantly higher (Mann–Whitney $U$-test 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 and 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\textsuperscript{46} or RNA\textsuperscript{47}, 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.
Multitissue 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 genotypes obtained from deep WGS for 652 individuals from the GTEx Project. Supplementary Table 1 gives the number of FM-eSTRs identified in each tissue. Rows and columns were clustered using hierarchical clustering (Methods).
Fig. 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 DNase I HS sites. Plots are centered at ENCODE DNase I 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 DNase I HS clusters. Values were smoothed using a sliding average of each four consecutive bins (a–d). e, Repeat unit enrichment at FM-eSTRs. The x axis shows all repeat units for which there are at least three FM-eSTRs across all tissues. The y axis center values denote the log2 OR comparing FM-eSTRs to all STRs. Error bars represent ±1 s.e.m. Asterisks denote repeat units that are significantly enriched or depleted in FM-eSTRs (based on two-sided Fisher's 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: example eSTRs are shown from skeletal muscle (f); esophagus mucosa (g) and transformed fibroblasts (h). 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 nontemplate (dashed) strands. The x axis shows STR lengths relative to hg19 (bp). Gene diagrams are not drawn to scale.
and increased expression of nearby genes (Supplementary Note, Fig. 2f–h and Supplementary Figs. 7–10).

We next examined effect-size biases in FM-eSTR associations. Overall, FM-eSTRs are equally likely to show positive versus negative correlations between repeat length and gene expression (Supplementary Fig. 11; binomial two-sided \(P = 0.94\)). We additionally observe that FM-eSTRs with repeat units of the form \((A,C,G,T)\) 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 (binomial two-sided \(P = 0.0015\)). These T-rich FM-eSTRs tend to have more positive effect sizes, with the most notable differences for AC versus GT repeats. These patterns are observed in transcribed regions across multiple distinct repeat types (A/T, AC/GT, AAC/GTT, AAAC/GGTT) 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-eSTRs (Supplementary Figs. 12 and 13). These include two clusters of FM-eSTRs present across many tissues (clusters 2 and 8) as well as several more tissue-specific clusters (for example, 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 that 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 the 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 megabase (Mb) of GWAS hits (Methods and 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 a CAVIAR score >0.5, 403 and 26 are in moderate and strong 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: (1) strong similarity in association statistics across variants for both the GWAS trait and expression of a particular gene, indicating the signals may be colocalized, that is, driven by, the same causal variant; and (2) strong evidence that the FM-eSTR causes variation in expression of that gene (Fig. 3a). Colocalization 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 of individuals, or more, are publicly available (Methods).

For each trait, we identified FM-eSTRs within 1 Mb of published GWAS signals from Supplementary Data 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 colocalized. 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 colocalized. We used coloc to test a total of 276 gene \(\times\) 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) colocalization posterior probability between the target gene and the trait >50%, meaning colocalization of the eQTL and GWAS signals is the most probable model (Extended Data Figs. 9 and 10). Out of the 62 FM-eSTRs colocalized with GWAS signals, 40 have CAVIAR scores >0.5. Results of all colocalization tests are provided in Supplementary Table 6.

A top example is an FM-eSTR for RFT1, a gene encoding an enzyme involved in the N-glycosylation of proteins\(^{51}\), which has 97.8% colocalization probability with a GWAS signal for height (Fig. 3b,c). The lead SNP in the NHGRI catalog (rs2336725:C\(\rightarrow\)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\(^{42}\) 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 \((P = 0.00328; \beta = 0.010; n = 6,393,\) where \(\beta\) is the effect size), although with a slightly weaker \(P\) value compared to the top SNP (Fig. 3e). Notably, even in the case where the STR is the causal variant, power is likely to be reduced due to the lower quality of the 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 due to the lower quality of the 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.

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 characterized the top fine-mapped eSTRs. The 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 noncanonical 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 with biallelic SNPs, to drive phenotypic variation along a spectrum of multiple alleles. In multiple examples, eSTRs show a linear trend between
Fig. 3 | FM-eSTRs colocalize 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, defined as in Fig. 1c. c, Summary statistics for RFT1 expression and height. The $-\log_{10} P$ values of association between each variant and RFT1 expression are shown in the middle panel. The $-\log_{10} P$ values of association for each variant with height are shown in the bottom panel. 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.m. g, Reporter assay testing repeat number versus 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.
repeat length and expression across a range of repeat numbers, a signal that cannot be easily explained by tagging nearby biallelic variants. Notably, our analysis is based only on signals that could be detected by standard SNP-based GWAS, which are underpowered to detect underlying multiallelic associations from STRs. 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. (1) While we applied stringent fine-mapping approaches to find eSTRs whose signals are probably not explained by nearby SNPs in LD, some signals could plausibly be explained by other variant classes, such as structural variants or Alu elements 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. (2) Our study was limited to tissues available from GTEx with sufficient sample sizes. 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 STRs. (3) 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 SNP-based analyses. To facilitate these and other studies, all summary-level eSTR data are publicly available at http://webstr.ucsd.edu/.
Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0521-9.

Received: 13 February 2019; Accepted: 25 September 2019; Published online: 1 November 2019

References
1. GTEx Consortium Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017).  2. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 510, 506–511 (2013).  3. Grünewald, T. G. P. et al. Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat. Genet. 47, 1073–1078 (2015).  4. Song, J. H. T., Lowe, C. B. & Kingsley, D. M. Characterization of a human-specific tandem repeat associated with bipolar disorder and schizophrenia. Am. J. Hum. Genet. 103, 421–430 (2018).  5. Boettger, L. M. et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat. Genet. 48, 359–366 (2016).  6. Leffler, E. M. A. resistance to malaria through structural variation of red blood cell invasion receptors. Science 356, eaam6393 (2017).  7. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. Nature 530, 177–183 (2016).  8. Sun, J. X. et al. A direct characterization of human mutation based on population summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0521-9.

Online content
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
24. Lin, Y., Dent, S. Y., Wilson, J. H., Wells, R. D. & Napierala, M. R loops
44. 8. Sun, J. X. et al. A direct characterization of human mutation based on population summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0521-9.

Online content
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2019
Methods

Dataset and preprocessing. Next-generation sequencing data were obtained from GTEx via dbGaP under phs000424.v7.p2. This included high-coverage (30x) Illumina 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 VCF files containing SNP genotype calls.

STRs were genotyped using HipSTR27 v.0.5, which returns the maximum likelihood diplodid STR allele sequences for each sample based on aligned reads as input. Samples were genotyped separately with nondefault parameters, namely 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—hg19.genomicSuperDups table); penta- and hexameric STRs containing homopolymer runs of at least five or six nucleotides, respectively, in the hg19 reference genome, since we previously found these STRs to have high error rates due to indels in homopolymer regions28; and STRs whose frequencies did not meet the percentage of homozygous versus 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.

For each tissue, we performed a gene-level RPMK 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 v19 (Ensembl 74) annotation.

Before 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 ten principal components (PCs) 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 Project32 samples genotyped using Omni 2.5 SNP genotyping arrays (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapet2_scaffolds/ld_chip_scaffolds/). Analysis was restricted to biallelic SNPs present in the Omni 2.5 data and resulting loci were LD pruned using plink v1.90b3.44 with option —indep 50 5 2. PCA on resulting SNP genotypes was performed using smartpcax,35,36,38,130500. To control for technical variation in expression, we applied PEER factor correction60. Based on an analysis of number of PEER factors versus 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 other factors reported previously for GTEx samples (Supplementary Fig. 3), such as ischemic time.

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

$$Y = \beta X + e$$

where $X$ denotes STR genotypes, $Y$ denotes expression values adjusted for the covariates described above, $\beta$ denotes the effect size and $e$ 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 module41 (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 genotype values, we removed samples with genotypes observed in fewer than three 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$^{10}$) 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 $fdr correction function in the statsmodels.multitest module to obtain a $q$ value for the best eSTR for each gene. FDR analysis was performed separately for each tissue.

Expressions SNPs (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 three samples and removed SNPs with fewer than three unique genotypes remaining after filtering.

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 as 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: $Y \approx eSNP$ (SNP-only model) versus $Y \approx eSNP + eSTR$ (SNP + STR model) using the anova_lm function in the Python statsmodels.api.stats.model. $Q$ values were obtained using the $fdr correction function in the statsmodels.stats.statscalculations module. $Q$ values were obtained using the 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$^{36}$ v.2.2 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 the 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 $-f 1$ 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 before running CAVIAR.

Multitissue eSTR analysis. We used an R implementation of mash$^{47}$ (mashR) v0.2.21 to compute posterior estimates of eSTR effect sizes and standard errors across tissues (https://stephenslab.github.io/mash/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 five 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. 14, 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. To give 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 phyloseq package (https://dawgberger.github.io/phyloseq/articles/phyloseq_fisherexact.21). Overlapping eSTRs with each annotation was performed using the intersectBed tool from the BEDTools suite. The distance between each STR and the center of the nearest RNA-seq or Dnase I HS cluster was restricted to RNA-seq clusters annotated in at least 20 cell types. The distance between each STR and the center of the nearest Dnase I HS cluster was computed using the closestBed tool from the BEDTools suite.

Analysis of DNA-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/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz). Analysis was restricted to Dnase I HS clusters annotated in at least 20 cell types. The distance between each STR and the center of the nearest Dnase I HS cluster was computed using the closestBed tool from the BEDTools suite.
Genotype data were available for 7,196, 6,100, 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 RH allele frequency at least 10% and with genotype frequencies expected under Hardy-Weinberg equilibrium (P > 1 × 10^{-5}). We performed LD pruning using the plink option --indep 50 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 supplementary PCA using only the identified European samples and used the top ten 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 ancestry PCs. Residuals were standardized 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 (ref. 14) 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 chr1:33022501–3266470.

The RFT1 FM-eSTR was imputed into the imputed SNP genotypes using Beagle 5 (ref. 15) with option gb=true and using our SNP–STR reference haplotype panel for the previously estimated 97% accuracy. We ran LD clustering 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 6,145 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 (chr3:3128363–3128413) plus approximately 170-bp genomic context on either side (RFT1_0rpt, RFT1_5rpt, RFT1_10rpt in Supplementary Table 7) were ordered as gBlocks from Integrated DNA Technologies. Each construct additionally contained homology arms for cloning into pGL4.27 (below). All constructs were amplified using Taq polymerase and sequenced on top of a 97% accuracy. We performed Illumina HiSeq sequencing of eight random samples for each construct.

Plasmids were transfected into the human embryonic kidney 293 cell line (HEK293T; ATCC CRL-3216) and grown in DMEM media (Gibco, catalog no. 10566-016), supplemented with 10% fetal bovine serum (Gibco, catalog no. 10090-136) and penicillin, 100 μg ml^-1 amphotericin B and 0.25 μg ml^-1 antibiotic B (Gibco, Antibiotic-Antimycotic, catalog no. 15240062). Cells were maintained at 37°C in 5% CO2 incubator. HEK293T cells (2 × 10^5) were plated onto each well of a 25 μg ml^-1 poly-D-lysine (EDM Millipore, catalog no. A-003-E)-coated 24-well plate, the day before transfection. On the day of the transfection, medium was changed to Opti-MEM. We conducted cotransfection experiments to test expression of each construct. Empty pGL4.27 vector (100 ng) (Promega, pha1E451) or 100 ng of each one of the pGL4.27 derivatives, was mixed with 5 μg of the reference plasmid pGL4.73 (Promega, catalog no. E0911), harboring an SV40 promoter upstream of Renilla luciferase, and added to the cells in the presence of Lipofectamine 2000 (Invitrogen, catalog no. L3000015). Three individual using PCR, the region from genomic DNA for sample NA12878 with 12 copies of AC (NIGMS Human Genetic Repository, Coriell) using the renilla plasmid was linearized using EcoRV (New England Biolabs, catalog no. R3195) and the region from genomic DNA for sample.

Capture Hi-C interactions (Extended Data Fig. 10) were visualized using the 3D Genome Browser. The visualization depictions interactions profiled in GM12878 (ref. 16) 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/MedicalBioclinic), c3 (Health/Medical/Medical-Biogenic Studies Only-No Insurance Companies) and c4 (Health/Medical/Medical-Biogenic Studies Only). Height data were available for samples in cohorts c1 (phs000888.v1.phs004680.v1.p1.c1), c3 (phs000888.v1.phs004680.v1.p1.c3), and c4 (phs000888.v1.phs004680.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.

**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 the 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 based on the elbow method. Using different values of K produced similar groups. We tested for nonuniform distributions of FM-eSTR repeat units across clusters using a chi-squared test implemented in the scipy.stats chi2_contingency function.
**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability**
All eSTR summary statistics are available for download on WebSTR [http://webstr.ucsd.edu/downloads](http://webstr.ucsd.edu/downloads).

**Code availability**
Code for performing analyses and generating figures is available at [http://github.com/gymreklab/gtex-estrs-paper](http://github.com/gymreklab/gtex-estrs-paper).

**References**
55. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
56. The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
57. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
58. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
59. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
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**, 500–507 (2012).
61. Seabold, S. P. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 57–61 (SCIPY, 2010).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
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**, 576–589 (2010).
64. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3466–3475 (2003).
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).
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).
67. Howie, B. N., 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).
68. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).

**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.

**Author contributions**
S.F. performed all eSTR and SNP mapping, helped to perform downstream analyses and helped to draft the manuscript. J.M. performed multitissue analysis using mashR and helped to revise the manuscript. C.W. optimized and performed the reporter assay. S.S. participated in the 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 declare no competing interests.

**Additional information**
Extended data is available for this paper at [https://doi.org/10.1038/s41588-019-0521-9](https://doi.org/10.1038/s41588-019-0521-9).
Supplementary information is available for this paper at [https://doi.org/10.1038/s41588-019-0521-9](https://doi.org/10.1038/s41588-019-0521-9).
Correspondence and requests for materials should be addressed to A.G. or M.G.
Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints).
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 log2 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. Top plots compare repeat units of the form C,G 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 -log10 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 -log10 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.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
  - Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

| Data collection | No software was used for data collection. |
| Data analysis   | HipSTR v0.5, bedtools v2.28.0, plink v1.90b3.44, Smartpca v1.3050, PEER (https://github.com/PMBio/peer), Python statsmodels package v0.8.0, CAVIAR v2.2, mashR v0.2.21, Python seaborn package v0.9.0, Python scipcy package v1.2.1, Homer v4.10, Python sklearn package 0.20.3, Mfold v3.6, Custom code for additional analyses is available at https://github.com/gymreklab/gtx-estrs-paper |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | We used 652 samples with available whole genome sequencing from GTEx. This sample size was based on the sample size in the data available from GTEx (phs000424.v7.p2). |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | No samples were excluded                                                                                                                       |
| Replication | We compared eSTR summary statistics to previously published results. Replication was measured using Pearson correlation between effect sizes previously reported and those reported in this study (computed for each tissue, mean Pearson r=0.45). |
| Randomization | Our study primarily relies on association testing. Analyses where randomization would typically be used (e.g. trials) were not performed here.                                                                 |
| Blinding | Our study primarily relies on association testing. Analyses where blinding would typically be used (e.g. trials) were not performed here.                                                                 |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a Involved in the study | n/a Involved in the study |
| ☒ ☐ Antibodies | ☒ ☐ ChiP-seq |
| ☒ ☒ Eukaryotic cell lines | ☒ ☐ Flow cytometry |
| ☒ ☐ Palaeontology | ☒ ☐ MRI-based neuroimaging |
| ☒ ☐ Animals and other organisms | |
| ☒ ☐ Human research participants | |
| ☒ ☐ Clinical data | |

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s) | HEK293T cells from ATCC
Authentication | Cells were purchased from ATCC. We did no additional authentication.
Mycoplasma contamination | Mycoplasma was last tested using PCR in October 2018. Experiments were performed November 2018. All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines
(See ICLAC register) | No cell lines used are listed in the ICLAC register. Experiments were performed in HEK293T cells (ATCC CRL-3216)