Multi-tissue analysis reveals short tandem repeats as ubiquitous regulators of gene expression and complex traits

Stephanie Feupe Fotsing[1,2], Catherine Wang[3], Shubham Saini[4], Sharona Shleizer-Burko[5], Alon Goren[5,#] & Melissa Gymrek[4,5,6,#]

[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 Bioengineering, University of California San Diego, La Jolla, CA USA
[4] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA USA
[5] Department of Medicine, University of California San Diego, La Jolla, CA USA
[6] Lead contact

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

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 elucidate the underlying biological mechanisms. Here, we leverage whole genome sequencing and expression data for 17 tissues from GTEx to identify STRs whose repeat lengths are associated with expression of nearby genes (eSTRs). Our analysis reveals more than 3,000 high-confidence eSTRs, which are enriched in known or predicted regulatory regions. We show eSTRs may act through a variety of mechanisms. We further identify hundreds of eSTRs that potentially drive published GWAS signals and implicate specific eSTRs in height and schizophrenia. Overall, our results demonstrate that eSTRs potentially contribute to a range of human phenotypes. We expect that our comprehensive eSTR catalog will serve as a valuable resource for future studies of complex traits.

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with complex traits (Visscher et al., 2017), but determining the causal variants, target genes, and biological mechanisms responsible for each signal has proven challenging. The vast majority of GWAS signals lie in non-coding regions (Edwards et al., 2013) which are difficult to interpret. Expression quantitative trait loci (eQTL) studies attempt to link regulatory genetic variation to gene expression changes as a potential molecular intermediate that drives the biological aberrations leading to disease (Lappalainen et al., 2013). Indeed, recent studies have utilized eQTL catalogs to pinpoint causal genes and relevant tissues for a variety of traits (Gamazon et al., 2015; Gusev et al., 2016; Zhu et al., 2016).

An additional major challenge in interpreting GWAS is that lead variants are rarely causal themselves, but rather tag a set of candidate variants in linkage disequilibrium (LD). A variety of statistical fine-mapping techniques have been developed to identify the most likely causal variant (Hormozdiari et al., 2014; Kichaev et al., 2014; Pickrell, 2014) taking into account factors such as summary association statistics, LD information, and functional annotations. However, these methods have been limited by focusing almost exclusively on bi-allelic single nucleotide polymorphisms (SNPs) or short indels. On the other hand, many recent studies to dissect GWAS loci have found complex repetitive (Grünewald et al., 2015; Song et al., 2018) or structural variation (Boettger et al., 2016; Leffler et al., 2017; Sekar et al., 2016) to be the underlying causal variants, highlighting the need to consider additional variant classes.

Short tandem repeats (STRs), consisting of repeated motifs of 1-6bp in tandem, represent a large source of genetic variation. STRs exhibit mutation rates that are orders of magnitude higher than SNPs (Sun et al., 2012) or short indels (Lynch, 2010) and each individual is estimated to harbor around 100 *de novo* mutations in STRs (Willems et al., 2016). Expansions at several dozen STRs have been known for decades to cause Mendelian disorders (Mirkin, 2007) 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 (Willems et al., 2014). Due to bioinformatics challenges of analyzing repetitive regions, most of these STRs are often filtered from genome-wide studies (Li, 2014). However, increasing

evidence supports a widespread role of common variation at STRs in complex traits such as gene expression (Hannan, 2010, 2018; Press et al., 2014).

STRs may regulate gene expression through a variety of mechanisms (Gemayel et al., 2010). For example, the CCG repeat implicated in Fragile X Syndrome was shown to disrupt DNA methylation, altering expression of *FMR1* (Liu et al., 2018). Yeast studies have demonstrated that homopolymer repeats act as nucleosome positioning signals with downstream regulatory effects (Raveh-Sadka et al., 2012; Suter et al., 2000). Dinucleotide repeats may alter affinity of nearby DNA binding sites (Afek et al., 2014). Furthermore, certain STR motifs have been shown to form non-canonical DNA and RNA secondary structures such as G-quadruplexes (Conlon et al., 2016), R-loops (Lin et al., 2010), and Z-DNA (Rothenburg et al., 2001).

We previously performed a genome-wide analysis to identify more than 2,000 STRs for which the number of repeats were associated with expression of nearby genes (Gymrek et al., 2016), termed expression STRs (eSTRs). However, the quality of the datasets available for this study reduced power to detect associations and prevented accurate fine-mapping of individual eSTRs. First, STR genotypes were based on low coverage (4-6x) whole genome sequencing data performed using short reads (50-100bp). As a result, individual STR genotype calls exhibited poor quality with less than 50% genotyping accuracy (Willems et al., 2014). Second, it was based on a single cell-type (lymphoblastoid cell lines; LCLs) with potentially limited relevance to most complex traits (Min et al., 2011). While our and other studies (Gymrek et al., 2016; Quilez et al., 2016) demonstrated that eSTRs explain a significant portion of the *cis* heritability of gene expression, the resulting eSTR catalogs were not powered to causally implicate eSTRs over other nearby variants.

Here, we leverage deep whole genome sequencing (WGS) and expression data collected by the Genotype-Tissue Expression Project (GTEx) (GTEx Consortium, 2015) to map eSTRs in 17 tissues. Our analysis reveals more than 25,000 unique eSTRs. Fine-mapping shows that more than 3,000 of these eSTRs explain variation in gene expression beyond that explained by nearby SNPs. Our results support a role of a subset of eSTRs in regulating nucleosome positioning and modulating DNA or RNA secondary structure. We additionally identify hundreds of eSTRs that may drive published GWAS signals and implicate specific eSTRs in height and schizophrenia. Further, we employ available GWAS data to demonstrate evidence of a causal

link between an eSTR for *RFT1* and height and use reporter assays to experimentally validate the effect of this STR on expression. Finally, the complete catalog of eSTRs is publicly available and will likely be a valuable 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 in 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 samples included in the Genotype Tissue Expression (GTEx) (GTEx Consortium, 2015) dataset for which both high coverage whole genome sequencing (WGS) and RNA-sequencing of multiple tissues were available. 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. We used HipSTR (Willems et al., 2017) to genotype STRs in each sample. Resulting genotypes were subjected to stringent filtering to remove low quality calls (**Methods**). After filtering, 175,226 STRs remained for downstream analysis. To identify eSTRs, we performed a linear regression between average STR length and normalized gene expression for each individual at each STR within 100kb of a gene, controlling for sex, population structure, and technical covariates (**Methods, Figures S1, S2**). Analysis was restricted to 17 tissues where we had data for at least 100 samples (**Figure 1A, Table S1, Methods**) and to genes with median RPKM greater than 0. As a control, for each STR-gene pair we performed a permutation analysis in which sample identifiers were shuffled. Altogether, we performed an average of 278,521 STR-gene tests across 16,065 genes per tissue.

Using this approach, we identified 25,561 unique eSTRs associated with 11,810 genes in at least one tissue at a gene-level FDR of 10% (**Methods**). Of these, 8,417 (32.5%) were shared by two or more tissues and 469 were shared by 10 or more tissues (**Figure S3**). P-values from the permuted controls followed a uniform distribution as expected under the null hypothesis of no association (**Figure 1B**). The number of eSTRs detected per tissue correlated with sample size as expected (Pearson r=0.76; p=0.00043; n=17), with the smallest number of eSTRs detected in the two brain tissues due to their low sample sizes (107 and 108 samples and 1,551 and 675 eSTRs for cerebellum and caudate, respectively; **Table S1**). Effect sizes were strongly

correlated across tissues (**Figure 1C**), with related tissues, such as subcutaneous adipose vs. visceral adipose showing the strongest concordance.

We evaluated these eSTRs by comparing to eSTRs we previously generated using an orthogonal cohort (Gymrek et al., 2016) that included expression data from LCLs. Effect sizes from LCLs were significantly correlated with effect sizes in all GTEx tissues ($p<0.01$ for all tissues, mean Pearson $r=0.55$), conveying the robustness of eSTR signals across cohorts and tissues. Notably, pairwise comparisons of eSTR effects between tissues within the GTEx cohort showed markedly higher concordance (mean pairwise Pearson $r=0.85$), likely as a result of being generated from the same set of individuals, increased power to accurately estimate effect sizes, and study-specific variation in RNA-seq protocols. We additionally examined previously reported eSTRs from single-gene studies. Most of these were originally tested using *in vitro* constructs in cell lines rather than primary tissues, and thus may not recapitulate *in vivo* conditions. Still, six of eight examples had significant eSTRs (nominal $p<0.01$) in at least one tissue analyzed (**Table S2**).

eSTRs identified above could potentially be explained by tagging nearby causal variants such as single nucleotide polymorphisms (SNPs). To identify potentially causal eSTRs we employed two fine-mapping approaches. First, for comparison to our previous study, we used ANOVA to determine whether eSTRs explained additional variability in expression of the target gene beyond the best eSNP (**Methods**). On average across tissues, 27.9% of eSTRs improved the model (10% FDR), consistent with previous results (23%) obtained in LCLs (Gymrek et al., 2016). To determine whether the eSTR could be tagging additional SNPs besides the best eSNP, we employed CAVIAR (Hormozdiari et al., 2014), a statistical fine-mapping framework for identifying causal variants. CAVIAR models the relationship between local LD-structure and association scores 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 the top 100 associated SNPs for each gene (**Methods, Figure 1A**). On average across tissues, 14.3% of eSTRs had the highest causality scores of all variants tested. We focused on the more stringent CAVIAR causality scores for downstream analyses.

We identified a group of 3,474 unique high-confidence eSTRs with CAVIAR scores of at least 0.1 across 3,046 genes (**Table S1, Dataset S1**) corresponding to 14% of all eSTRs (**Figure**

S4). We hereby refer to these STRs as fine-mapped eSTRs (FM-eSTRs), with the caveat that further validation will be needed to definitively determine these to be causally related to gene expression. We were encouraged to observe that multiple STRs with known disease implications were captured by this list. For example, a CG-rich repeat upstream of *CSTB* which was previously implicated in myoclonus epilepsy (Lalioti et al., 1997) and as a causal eQTL for *CSTB (Borel et al., 2012)* was identified as an FM-eSTR in 11 tissues (**Figure 1D left,** CAVIAR score >0.9 in seven tissues). In many cases, top FM-eSTRs show clear linear trends between the number of repeats and gene expression across a wide range of repeat lengths. For example, a highly polymorphic intronic CCTGGG repeat in *NOP56* implicated in spinocerebellar ataxia 36 (Kobayashi et al., 2011) was identified as an FM-eSTR in three tissues (**Figure 1D middle**). In another example, a multi-allelic CGGGGG repeat in the promoter of *ALOX5* was identified as an FM-eSTR in esophagus mucosa (CAVIAR score=0.38) (**Figure 1D**, right). This repeat was previously shown to regulate *ALOX5* expression in leukocytes (Vikman et al., 2009a) and is associated with reduced lung function (Mougey et al., 2013; Vikman et al., 2009b) and cardiovascular disease (Stephensen et al., 2011).

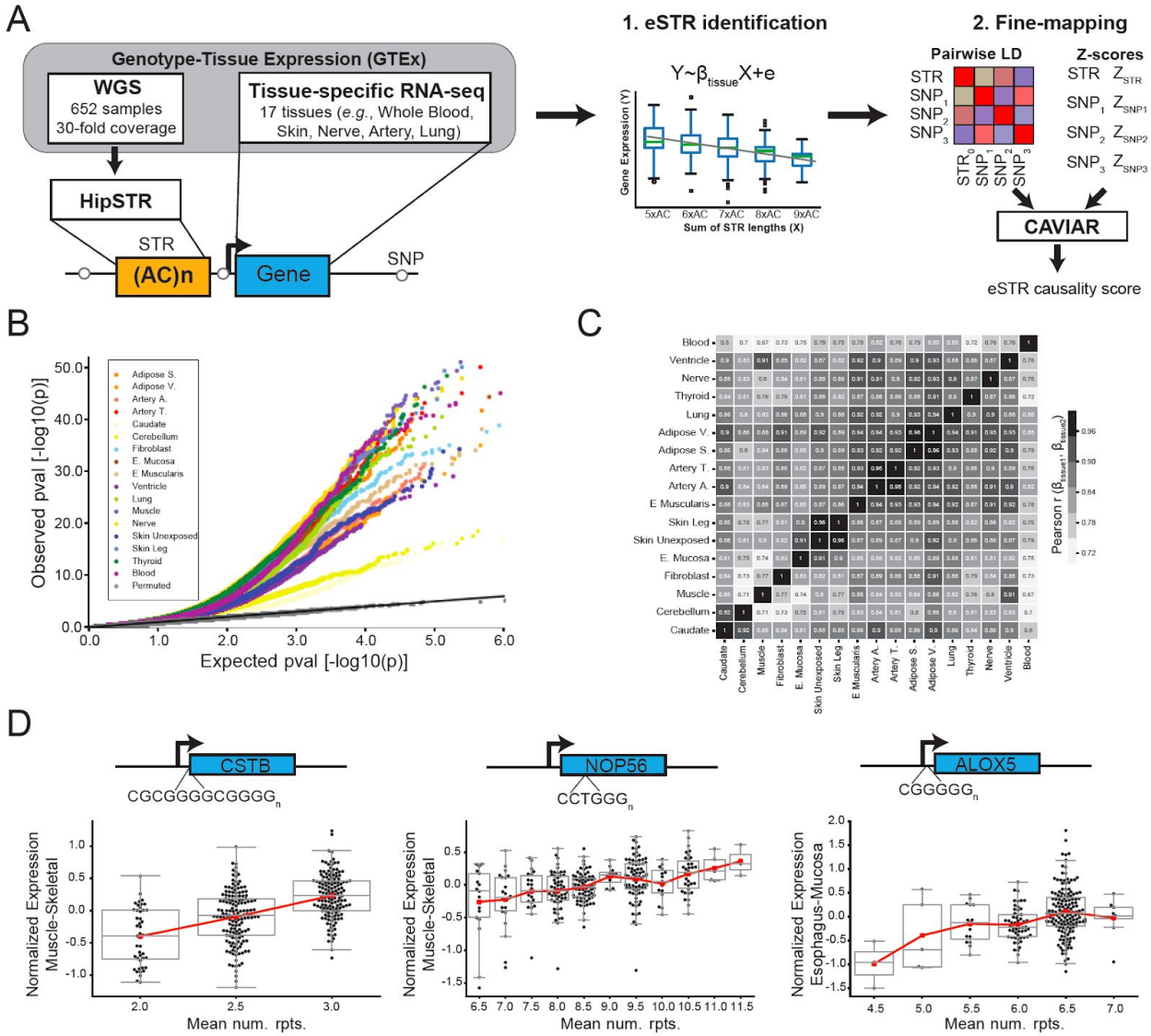

Figure 1: Multi-tissue identification of eSTRs.
(A) Schematic of eSTR discovery pipeline. We analyzed RNA-seq for 17 tissues and STR genotypes obtained from deep WGS for 652 individuals from the GTEx Project. For each STR within 100kb of a gene, we tested for association between length of the STR and expression of the gene in each tissue. For each gene, CAVIAR was used to fine-map the effects of eSTRs vs. the top 100 *cis* SNPs on gene expression. CAVIAR takes as input pairwise variant LD and effect sizes (Z-scores) and outputs a posterior probability of causality for each variant.
(**B) eSTR association results.** The quantile-quantile plot compares observed p-values for each STR-gene test vs. the expected uniform distribution. Gray dots denote permutation controls. The black line shows the diagonal. Colored dots show observed p-values for each tissue.
(C) eSTR effect sizes are correlated across tissues and studies. Each cell in the matrix shows the Pearson correlation between eSTR effect sizes (β) for each pair of tissues.
(D) Example fine-mapped eSTRs (FM-eSTRs). Each black point represents a single individual. 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. 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 are not drawn to scale.
See also Figures S1-S4, Tables S1-S2, and Dataset S1.

eSTRs are strongly enriched in promoters and putative enhancer regions

We next sought to characterize properties of STRs that might provide insights into their biological function. We first examined the distribution of all 175,226 STRs that were included in our analysis in regulatory regions by visualizing the localization of STRs around transcription start sites (TSSs) and DNAseI hypersensitive (DNAseI HS) sites identified by the ENCODE Project (ENCODE Project Consortium, 2012) (**Methods**). In general, STRs are depleted directly at TSSs (**Figure 2A**). However, certain classes of STR motifs, including GC-rich motifs and trinucleotides are highly enriched in promoter regions (**Figure S5**, **Figure 2A**). Notably, strand-specific localization patterns in promoters, but not enhancers, showed striking differences for several STR motifs, with A- and T-rich repeats most prevalent on the template strand upstream and downstream of the TSS, respectively (**Figure S5**) similar to previous observations (Aporntewan et al., 2013; Sawaya et al., 2013). STRs, particularly dinucleotides, are highly prevalent at DNAseI HS sites, consistent with a similar finding in *Drosophila melanogaster* (Yáñez-Cuna et al., 2014) which identified dinucleotides as a hallmark of enhancer sequences (**Figure 2B**).

We next considered whether the localization of eSTRs differed from that of STRs overall. Compared to all STRs, eSTRs are strongly enriched at 5' UTRs (OR=3.3; Fisher's two-sided p=1.9e-37), 3' UTRs (OR=3.1; Fisher's two-sided p=3.1e-77), and within 3kb of transcription start sites (OR=3.2; Fisher's two-sided p<1e-300) (**Table S3**). Further, STRs closest to TSSs, near DNAseI HS sites, and falling within binding sites for transcription factor (TF) and chromatin regulators profiled by ENCODE (**Table S4**) were most likely to act as eSTRs (**Figure 2C-D**). These enrichments strengthen substantially when considering only FM-eSTRs (**Figure 2C-D**), suggesting that CAVIAR scores accurately identify the eSTRs most likely to be causal variants. Notably, similar to other published eQTLs (GTEx Consortium, 2015), the majority of eSTRs are located in non-coding regions (**Table S3**), with only 16 FM-eSTRs falling in coding exons (**Table S5)**.

We next examined the sequence characteristics of eSTRs compared to all STRs. For each STR, we defined the canonical motif as the lexicographically first motif when considering all

rotations and strand orientations of the repeat sequence. For example, the canonical motif for the repeat sequence CAGCAGCAGCAG would be AGC. Using this approach, we tested eSTRs combined across all tissues for enrichment of each canonical motif. eSTRs are strongly enriched for repeats with GC-rich motifs across all tissues (**Figure 2E, Table S6, Methods**). These enrichments strengthen when restricting to FM-eSTRs. For example, the canonical motifs CCCGG, CCCCG, and CCCCCG are 9.0, 6.0, and 6.5-fold, enriched in eSTRs compared to all STRs, respectively. When considering only FM-eSTRs compared to all STRs, these motifs are 15.5, 14.6, and 10.3-fold enriched, respectively (Fisher's exact test two-sided $p<10^{-3}$ in all cases). Additionally, several A-rich motifs were enriched in FM-eSTRs compared to all STRs (*e.g.*, AAAT – Fisher exact test odds ratio=1.33, two-sided p=8.5e-5; AAAAAG – Fisher exact test odds ratio=2.8, two-sided p=2.75e-3). Few motifs show tissue-specific enrichment (**Figure 2F**) suggesting that most eSTRs act via global mechanisms common across multiple cell types. Notably, the majority of FM-eSTRs consist of homopolymer (40%) and dinucleotide (31%) motifs, although these motifs were not enriched compared to STRs overall.

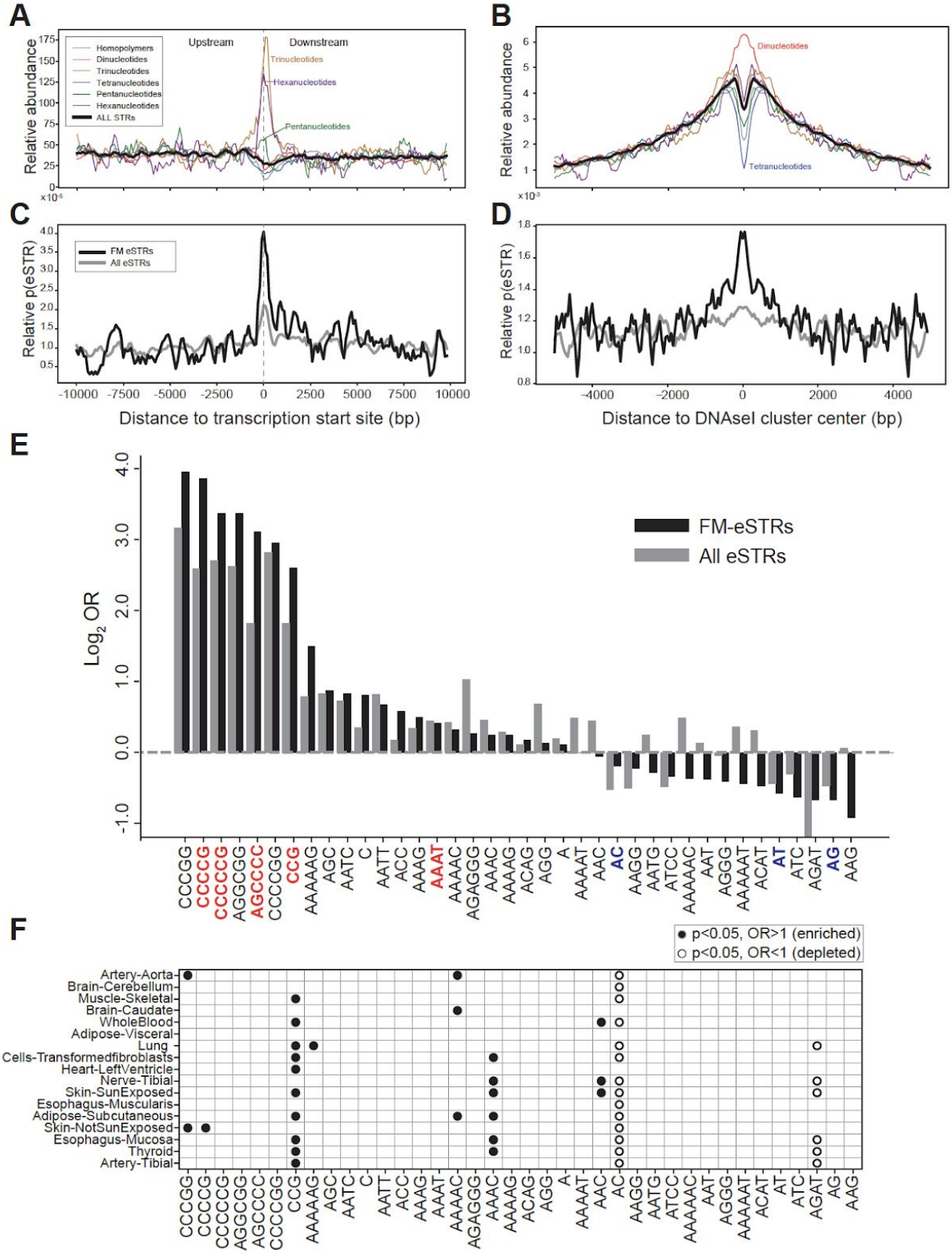
A
Relative abundance
Upstream
Downstream
Homopolymers
Dinucleotides
Trinucleotides
Tetranucleotides
Pentanucleotides
Hexanucleotides
ALL STRs
Trinucleotides
Hexanucleotides
Pentanucleotides
B
Relative abundance
Dinucleotides
Tetranucleotides
C
Relative p(eSTR)
FM eSTRs
All eSTRs
Distance to transcription start site (bp)
D
Relative p(eSTR)
Distance to DNAseI cluster center (bp)
E
Log2 OR
FM-eSTRs
All eSTRs
F
p<0.05, OR>1 (enriched)
p<0.05, OR<1 (depleted)
Artery-Aorta
Brain-Cerebellum
Muscle-Skeletal
Brain-Caudate
WholeBlood
Adipose-Visceral
Lung
Cells-Transformedfibroblasts
Heart-LeftVentricle
Nerve-Tibial
Skin-SunExposed
Esophagus-Muscularis
Adipose-Subcutaneous
Skin-NotSunExposed
Esophagus-Mucosa
Thyroid
Artery-Tibial

Figure 2: Characterization of eSTRs in predicted regulatory regions.
(A) Density of all STRs around transcription start sites. The y-axis shows the relative number of STRs in each 100bp bin around the TSS. Negative x-axis numbers denote upstream regions and positive numbers denote downstream regions.
(**B) Density of all STRs around DNAseI HS sites.** Plots are centered at ENCODE DNAseI HS clusters and represent the relative number of STRs in each 50bp bin. For (**A)** and (**B)** the black line denotes all STRs and colored lines denote repeats with different motif lengths (gray=homopolymers, red=dinucleotides, gold=trinucleotides, blue=tetranucleotides, green=pentanucleotides, purple=hexanucleotides).
(**C) Relative probability to be an eSTR around TSSs**. The gray and black lines represent the probability of an STR in each bin to be an eSTR or FM-eSTR, respectively. Each line was scaled relative to the genome-wide average.
(**D) Relative probability to be an eSTR around DNAseI HS sites.** Axes are similar to those in **(C)** except centered around DNAseI HS sites. For **A-D,** values were smoothed by taking a sliding average of each four consecutive bins.
(**E) Motif enrichment at eSTRs across all tissues**. The x-axis shows all motifs for which there are at least 3 FM-eSTRs across all tissues. The y-axis denotes the $\log_2$ odds ratios from performing a Fisher's exact test comparing eSTRs to all STRs (gray bars) or all FM-eSTRs to all STRs (black bars). Bolded motifs indicate those that are enriched (red) or depleted (blue) across all FM-eSTRs (Bonferroni corrected $p<0.05$).
(**F) TIssue-specific motif enrichment at FM-eSTRs.** Dots in each cell of the matrix denote motifs that are enriched (closed circles) or depleted (open circles) in FM-eSTRs in each tissue (Bonferroni corrected $p<0.05$). The order of motifs is the same as in **(E)**.
See also Figure S5 and Tables S3-S6.

Homopolymer eSTRs are associated with low nucleosome occupancy

Given the high occurrence of homopolymer A/T (also referred to as poly(dA/dT)) FM-eSTRs, we further examined their potential role in regulating gene expression. These repeats have previously been shown to act as modulators of nucleosome positioning both in humans and yeast (Rhodes, 1979; Struhl, 1985), and are strongly nucleosome-depleted (**Figure 3A**). We thus wondered whether homopolymer repeat variation may be affecting DNA accessibility of nearby regions. Indeed, homopolymer FM-eSTRs are generally located in regions of increased DNAseI HS sensitivity in the lymphoblastoid cell line GM12878 (**Figure 3B**). Furthermore, DNAseI-seq read counts within 500bp of FM-eSTRs are significantly higher than when considering all A/T STRs genome-wide (Kolmogorov-Smirnov two-sided $p=4.47e-8$). DNAseI-HS signal around A/T FM-eSTRs shows a periodic pattern, with notable peaks located at multiples of 147bp upstream or downstream from the STR (**Figure 3B,C**). Given that 147bp is the length of DNA typically wrapped around a single nucleosome (Jiang and Pugh, 2009), we hypothesize that while A/T STRs themselves show a depletion of DNAseI HS signals, they may act by shifting nucleosome positions and thus modulating openness of adjacent hypersensitive sites.

Previous studies in yeast have shown that A/T repeats in promoter regions alter transcriptional activity via strand-specific effects on nucleosome positioning (Krietenstein et al., 2016; Raveh-Sadka et al., 2012). We thus wondered whether A/T repeats in promoters would show different characteristics based on which motif was on the template strand. Overall, we found that FM-eSTRs are equally likely to show positive vs. negative effect sizes (where positive effects denote higher expression with increasing repeat number and vice versa) (binomial two-sided p=0.33; 50.9% positive effect sizes; n=3,474). Similarly, 50.8% of all A/T FM-eSTRs have positive effect sizes (binomial two-sided p=0.58; n=1,337). On the other hand, A/T FM-eSTRs directly downstream of TSSs (within 3kb) show strand-specific directional biases (**Figure 3D**). Downstream repeats with T on the template strand are significantly more likely to show positive effect sizes (71.4% positive effects; binomial two-sided p=0.036; n=28), whereas downstream A repeats show a suggestive opposite bias toward negative effects (23.1% positive effects; binomial two-sided p=0.092; n=13). Overall, homopolymer FM-eSTRs downstream of the TSS and within gene bodies are more likely to have T on the template strand (**Figure 3E**) and show positive effect sizes (**Figure 3F**). Altogether, these directional and strand-specific biases are consistent with a previous finding that A/T promoter repeats shift nucleosomes in a strand-specific manner (Krietenstein et al., 2016).

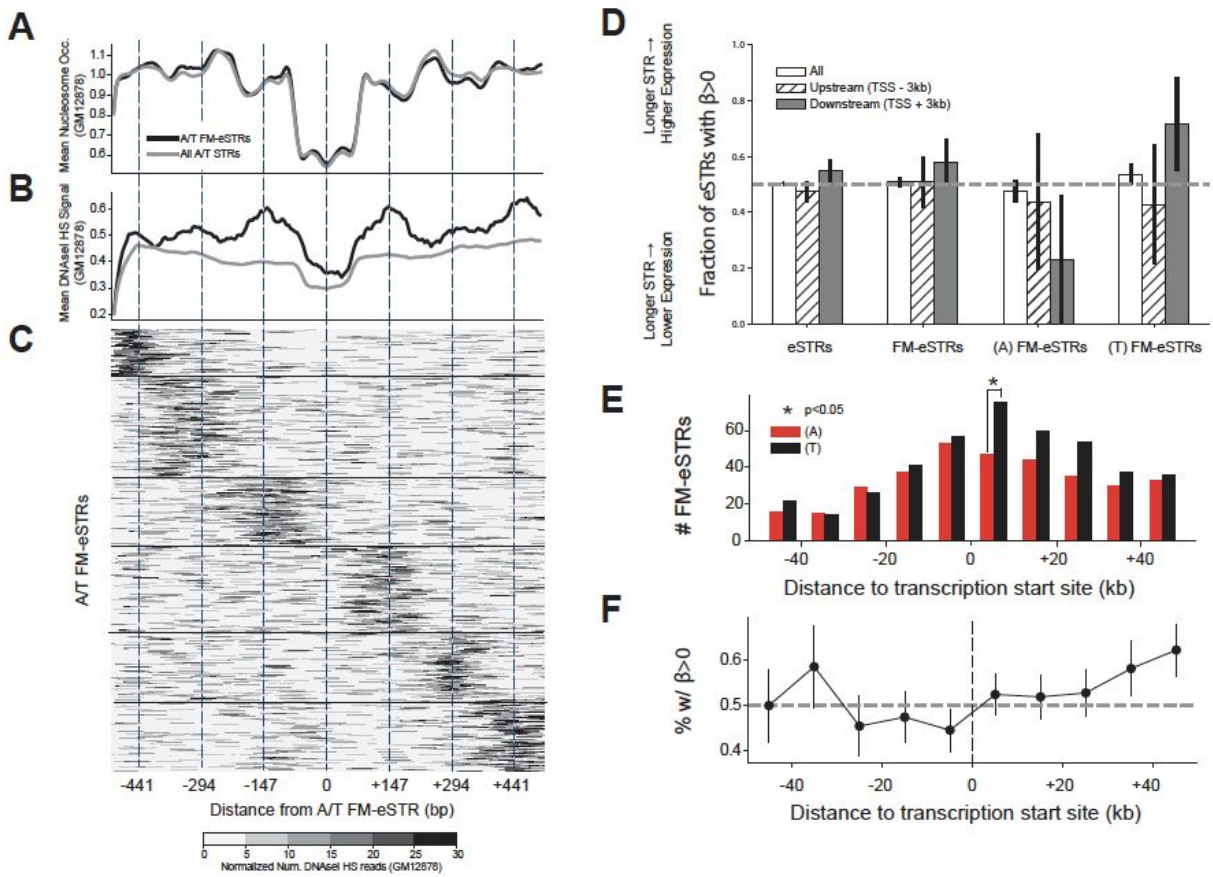

Figure 3: Homopolymer A/T FM-eSTRs are associated with nucleosome depletion.
(A) Nucleosome density around A/T STRs. Nucleosome density in the GM12878 cell line in 5bp windows is averaged across all A/T STRs (gray) and A/T FM-eSTRs (black) relative to the center of the STR.
(**B) DNAseI HS density around A/T STRs.** The number of DNAseI HS reads in GM12878 in 5bp windows is averaged across all A/T STRs (gray) and A/T FM-eSTRs (black) relative to the STR center.
(**C) DNAseI HS signal for individual A/T FM-eSTRs.** Only FM-eSTRs with at least 50 reads within the window of +/- 500bp are shown. Rows were normalized such that the counts in each row sum to 1 and the color bar represents the normalized number of reads in each cell of the matrix. Loci were clustered using K-means with K=6 and clusters were sorted based on the location of maximum DNAseI HS signal. Horizontal black lines demarcate different clusters. For **A-C**, dashed vertical gray lines denote the STR location (0bp) and positions that are +/- 147bp from the center.
(D) Bias in the direction of eSTR effect sizes. The y-axis shows the percentage of eSTRs in each category with positive effect sizes, meaning a positive correlation between STR length and expression. White bars denote all STRs in each category, hatched bars denote STRs in each category that fall within 3kb upstream of TSSs, and black bars denote STRs in each category falling within 3kb downstream of TSSs.
(**E) Number of FM-eSTRs relative to TSSs.** Each bar represents a 10kb bin. Red=A STRs and black=T STRs relative to the template strand.
(**F) Effect size direction bias relative to TSSs.** Dashed gray line gives the 50% expected by chance. Each dot represents a 10kb bin. Vertical bars give +/- 1 s.d.

GC-rich eSTRs are predicted to modulate DNA and RNA secondary structure

eSTRs were most strongly enriched for repeats with high GC content (*e.g.*, canonical motifs CCG, CCCCG, CCCCCG, AGCCCC) (**Table S6**) which are found almost exclusively in promoter and 5'UTR regions (Sawaya et al., 2013) (**Figure S5**). These GC-rich motifs have been shown to form highly stable secondary structures during transcription such as G4 quadruplexes in single-stranded DNA (Bochman et al., 2012) or RNA (Ciesiolka et al., 2017). G4 quadruplexes may regulate gene expression through a variety of mechanisms, including acting as nucleosome exclusion signals (Halder et al., 2009; Hänsel-Hertsch et al., 2016), promoting RNA polymerase II (RNAPII) pausing (Szlachta et al., 2018), and generating R-loops, a DNA:RNA hybrid structure formed by hybridization of nascent RNA to the template DNA strand (Duquette et al., 2004).

We hypothesized that the effects of GC-rich eSTRs may be in part due to formation of non-canonical nucleic acid secondary structures that modulate DNA or RNA stability as a function of repeat number. We considered two classes of eSTRs: those following the standard G4 motif ($G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$) (Todd et al., 2005) (**Table S7**; referred to below as G4 repeats) and repeats with canonical motif CCG, which may form hairpins or other secondary structures in single-stranded RNA (Kiliszek et al., 2012) or DNA (Moore et al., 1999), but do not meet the standard G4 definition. We observed that both classes of GC-rich repeats are associated with higher RNAPII (**Figure 4A)** and lower nucleosome occupancy (**Figure 4B)** in GM12878 compared to all STRs. This difference is more pronounced when considering only FM-eSTRs. Notably, the majority of CCG repeats (78%) occur in 5' UTRs compared to only 6% for G4 repeats.

To evaluate whether GC-rich repeats could be modulating DNA or RNA secondary structure, we used mfold (Zuker, 2003) to calculate the free energy of each STR and 50bp of its surrounding context, which we refer to as the STR region, in single stranded DNA or RNA. We considered all common allele lengths (number of repeats) observed at each STR (**Methods**) and computed energies for both the template and non-template strands. For template-strand DNA sequences, both classes of GC-rich STRs are predicted to have lower free energy (greater stability) compared to all STRs in promoter regions (Kolomogorov-Smirnov [KS] two-sided p=3.5e-54 and p=8.3e-140 for G4 and CCG repeats, respectively; **Figure 4C**). Overall, FM-eSTR regions exhibit significantly lower free energy compared to all STR regions (1.1 fold decrease, KS

two-sided p=4.6e-17), with the strongest effect when restricting to G4 repeats (1.6 fold decrease; p=6.3e-8).

We next tested whether STR variation affects the predicted DNA stability by computing the correlation between the number of repeats and free energy at each STR region. Overall, FM-eSTRs are about equally likely to increase or decrease free energy (47% with negative effects, defined as Pearson r<0). On the other hand, 70% of G4 repeats and 100% of CCG FM-eSTRs show negative correlations between free energy and number of repeats (**Figure 4D)**, significantly more than the 50% expected by chance (binomial two-sided p=5.3e-4 and p=4.8e-7 for G4 repeats and CGG, respectively)**.** Finally, the magnitude of the change in free energy across common allele lengths is by far the strongest for GC-rich FM-eSTRs (**Figure 4E**). Similar trends were observed when considering structures formed on the non-template DNA strand or in transcribed RNA (**Figure S6**). Overall, these results suggest that higher repeat numbers at GC-rich STRs result in more stable secondary structures formed in both DNA and RNA during transcription.

Finally, we hypothesized that changes in DNA stability resulting from variation in repeat numbers at GC-rich STRs might be impacting the expression of nearby genes. Based on previous observations (Hänsel-Hertsch et al., 2016), we predicted that higher repeat numbers at GC-rich eSTRs would result in increased expression of nearby genes. To that end, we tested whether eSTRs were biased toward negative vs. positive effect sizes. As described above, FM-eSTRs overall show no bias in effect direction. In line with this result, the subset of FM-eSTRs with CCG motifs also do not show a significant directional bias (binomial two-sided p=0.40; 60.1% positive effects; n=23). However, 67% of FM-eSTRs with G4 repeats have positive effects, significantly more than the 50% expected by chance (binomial two-sided p=0.0017; n=93) (**Figure 4F**). The effect direction bias is most pronounced for FM-eSTRs with G4 repeats within 3kb downstream of TSSs, for which 17/18 (94%) have positive effect sizes (binomial two-sided p=0.00014). In many cases changes in expression levels across allele lengths follows an inverse relationship with free energy levels (**Figure 3G-I**).

Altogether, these results support a model in which higher repeat numbers at GC-rich eSTRs in promoter regions form more stable DNA secondary structures which promote transcription. The contradictory results for CCG STRs may indicate that those repeats could act by distinct

mechanisms compared to G4-forming eSTRs, potentially through forming RNA secondary structures given their high prevalence in 5'UTRs.

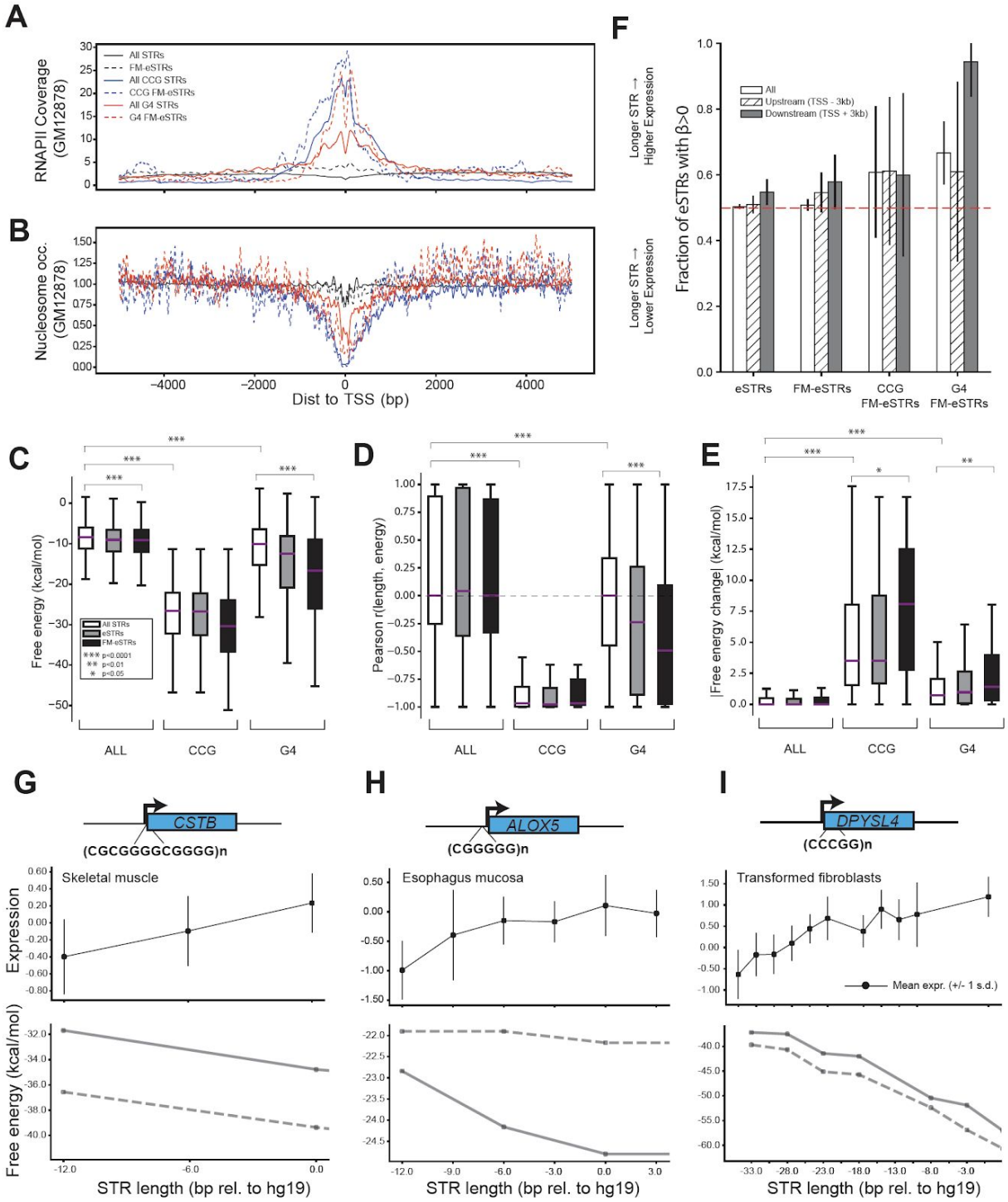

Figure 4: GC-rich eSTRs are predicted to modulate DNA secondary structure.
(A) Density of RNAPII localization around STRs. The y-axis denotes the average number of ChIP-seq reads for RNA Polymerase II in GM12878 in 5bp bins centered at STRs.
(**B) Nucleosome occupancy around STRs.** The y-axis denotes the average nucleosome occupancy in 5bp bins centered at STRs in GM12878. For **A.** and **B.,** black lines denote all STRs found within 5kb of TSSs, blue lines denote CCG STRs, and red lines denote STRs matching the canonical G4 motif. Solid lines represent all STRs of each class and dashed lines represent FM-eSTRs.
(**C) Free energy of STR regions.** Boxplots denote the distribution of free energy for each STR +/- 50bp of context sequence, computed as the average across all alleles at each STR.
(**D) Pearson correlation between STR length and free energy.** Correlations were computed separately for each STR, and plots show the distribution of correlation coefficients across all STRs. The dashed horizontal line denotes 0 correlation as expected by chance.
(**E) Magnitude of free energy change across alleles.** For each STR the absolute value of the difference in energy between the shortest and longest allele was computed. Plots show the distribution of these values across all STRs. For **C-E**, purple lines show medians and 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 represents the interquartile range (Q3-Q1). White boxes show all STRs in each category, gray boxes show eSTRs, and black boxes show FM-eSTRs. Upper brackets denote significant differences for all STRs (white) across categories or for significant differences within each category between all STRs (white) and FM-eSTRs (black).
(**F) Bias in the direction of eSTR effect sizes.** The y-axis shows the percentage of eSTRs in each category with positive effect sizes, meaning a positive correlation between STR length and expression. White bars denote all STRs in each category, hatched bars denote STRs in each category that fall within 3kb upstream of TSSs, and black bars denote STRs in each category falling within 3kb downstream of TSSs.
(**G-I) Examples of G4 FM-eSTRs in promoter regions predicted to modulate secondary structure**. For each example, top plots show the mean expression across all samples with each mean STR length. Vertical bars represent +/- 1 s.d. Bottom plots show the free energy computed by mfold for each number of repeats for the STR. Solid and dashed gray lines show energies for alleles on the template and non-template strands, respectively. The x-axis shows STR lengths relative to the hg19 reference genome in bp.
See also Figure S6 and Table S7.

eSTRs are potential drivers of published GWAS signals

Finally, we wondered whether our eSTR catalog could identify STRs affecting complex traits in humans. The vast majority of GWAS studies have been performed using SNP genotyping arrays that do not directly assay STRs. Notably, many highly polymorphic STRs are not well tagged by SNPs (Saini et al., 2018) and thus many underlying associations between repeat number and phenotypes are unlikely to be detected by standard SNP-based GWAS. On the other hand, we reasoned that GWAS may detect a subset of signals driven by causal eSTRs if they are in high LD with common SNPs, although with attenuated effects compared to testing the STR directly for association.

To that end, we leveraged the NHGRI/EBI GWAS catalog (MacArthur et al., 2017) to identify FM-eSTRs that are nearby and in LD with published GWAS signals. Overall, 2,559 FM-eSTRs are within 50kb of GWAS hits (**Methods, Dataset S2**). Of these, 1,341 are in moderate LD (r^2>0.1) and 180 are in strong LD (r^2>0.8) with the lead SNP. For 15 loci in at least moderate LD, the lead GWAS variant is within the STR itself (**Table S8**). In one example, the lead variant (rs10709981 -/A) at a locus associated with basophil percentage of granulocytes (Astle et al., 2016) falls within a homopolymer repeat positively associated (nominal p<0.01) with expression of *SLC36A1* in 15 tissues (**Figure S7A-B**). In another example, the lead variant (rs369552432; -/CGGCGGCGG) at a locus associated with hemoglobin (Astle et al., 2016) falls within a multi-allelic 5'UTR trinucleotide CCG repeat negatively associated with expression of *VLDLR* in whole blood (**Figure S7C-D**). Notably, in many cases, despite being multi-allelic in the GTEx cohort, the variant is listed as a bi-allelic indel in dbSNP (Sherry, 2001).

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 are co-localized, *i.e.*, driven by the same causal variant; and (***ii***) strong evidence that the FM-eSTR is causally associated with variation in expression of that gene (**Figure 5A**). Co-localization analysis requires high-resolution summary statistic data. Thus, we focused on two complex traits: height (Yengo et al., 2018) and schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) for which detailed summary statistics computed on cohorts of tens of thousands or more individuals are publicly available.

Additionally, the large sample sizes for these traits maximize power to detect associations driven by STRs even when genotyped SNPs may only be in weak LD with the number of repeats.

To this end, we first used coloc (Guo et al., 2015) to identify GWAS signals for each trait co-localized with expression of a particular gene. The coloc method 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. In parallel, we identified associations for each trait where the lead SNP is within 50kb and in significant LD ($p<0.05$) with an FM-eSTR and the FM-eSTR is the most probable causal variant for expression identified by CAVIAR.

Using this approach, we identified a total of 10 and 7 putative causal FM-eSTRs for schizophrenia and height, respectively (**Table S9**). For schizophrenia, this list includes a GWAS signal (lead SNP rs9420) with 90.1% co-localization probability with expression of *MED19,* a component of the Mediator complex and regulator of neuronal gene expression (Ding et al., 2009). The lead SNP is in LD (r^2=0.68) with an AC repeat that is an FM-eSTR for *MED19* (CAVIAR score=0.47) that is a nominally significant eSTR in 14 tissues. The STR shows a linear positive association with expression across a range of repeat numbers (**Figure 5B,C**). This region overlaps a DNAseI HS cluster identified by ENCODE and is located between two CCCTC-binding factor (CTCF) binding sites. The FM-eSTR is located in an intron of *CTNND1* approximately 43kb upstream of *MED19* (**Figure 5D**). Capture Hi-C (Mifsud et al., 2015) revealed a hotspot of long-range interactions between the region and many other loci, including the promoter region of *MED19*. Notably, a recently developed burden test taking into account promoter-enhancer identified a significant association between *MED19* expression and schizophrenia risk (Wu and Pan, 2018) supporting *MED19* as a causal gene at this locus.

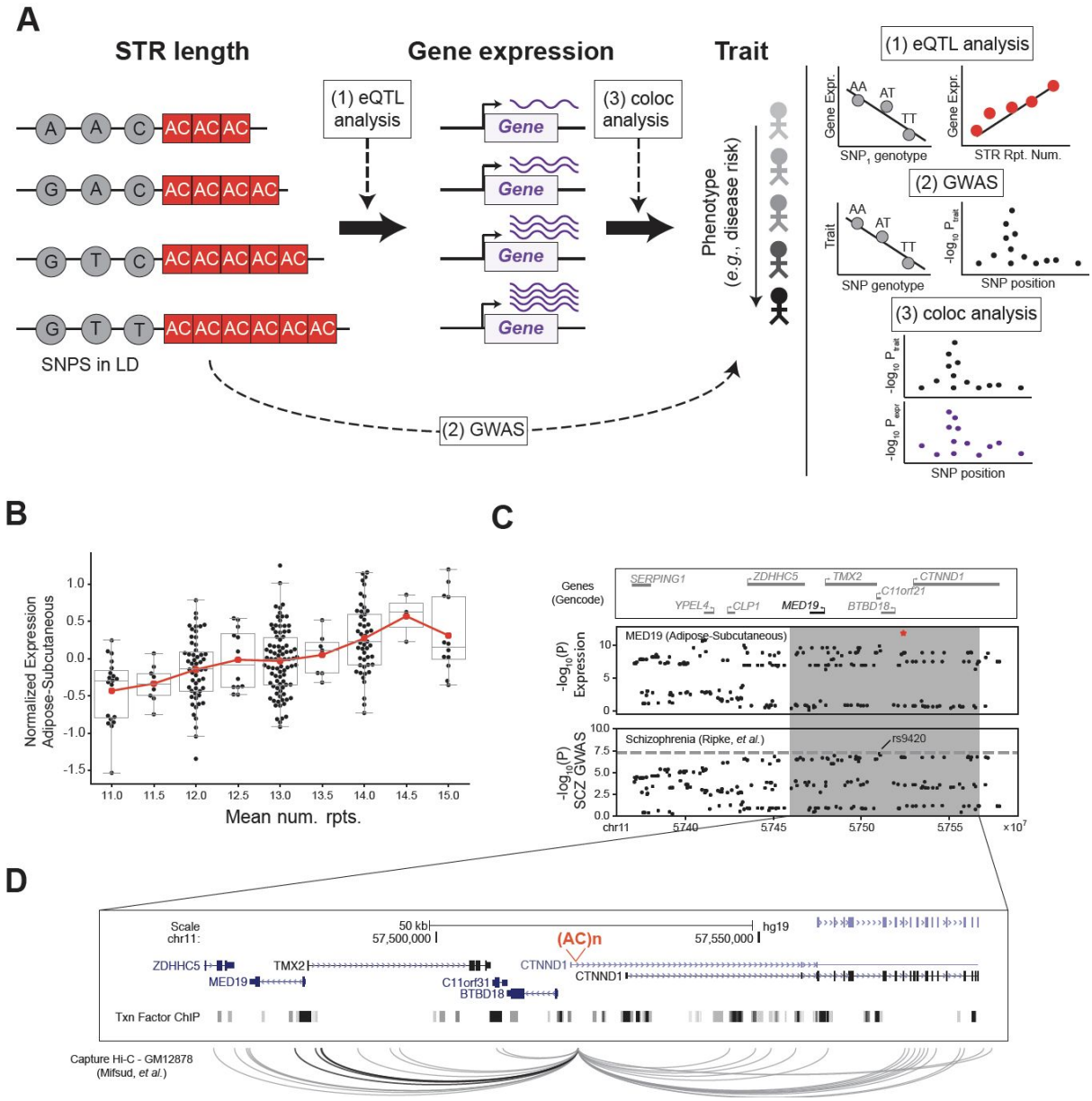

Figure 5: FM-eSTRs co-localize with published GWAS signals for schizophrenia.
(A) Overview of analyses to identify FM-eSTRs involved in complex traits. In cases where FM-eSTRs may drive complex phenotypes, we assumed a model where variation in STR repeat number (red; left) alters gene expression (purple; middle), which in turn affects the value of a particular complex trait (right). Even when the STR is the causal variant, nearby SNPs (gray; left) in LD may be associated with both gene expression through eQTL analysis and the trait of interest through GWAS. Black arrows indicate assumed causal relationships. Dashed arrows indicate analyses (1-3) used to test each relationship. (1) eQTL analysis tests for associations between SNP genotype or STR repeat number and gene expression. (2) GWAS tests for associations between SNP genotype and trait value. (3) coloc analysis tests whether the association signals for expression and the trait are driven by the same underlying causal variant.
(B) 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 summarizing the expression distribution for each genotype are as described in **Figure 1D**.
(C) 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 from Ripke, *et al.*. The dashed gray horizontal line shows genome-wide significance threshold of 5E-8.
(D) Detailed view of the *MED19* locus. A UCSC genome browser screenshot is shown for the region in the gray box in **(C)**. The FM-eSTR is shown in red [(AC)n]. 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.
See also Figure S7, Tables S8-S9, and Dataset S2.

The top example in our height analysis identified a GWAS signal (lead SNP rs2336725) with 99.0% co-localization probability with *RFT1*, an enzyme involved in the pathway of N-glycosylation of proteins (Haeuptle et al., 2008) (**Figure 6A,B**). The lead SNP is in high LD (r^2=0.85) with an AC repeat that is an FM-eSTR for *RFT1* in aortic artery (CAVIAR score=0.70) and heart tissue (CAVIAR score=0.39) and is a nominally significant eSTR in 17 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 (**Figure 6C**) and exhibits a positive correlation with expression across a range of repeat numbers.

To more directly test for association between this FM-eSTR and height, we used our recently developed STR-SNP reference haplotype panel (Saini et al., 2018) 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 adjusted height values for age, population structure, sex, and cohort. We restricted analysis to individuals for which all covariates were available and with inferred European ancestry to maximize imputation performance. In total our analysis consisted of 11,588 samples. We tested for association between height and SNPs as well as for AC repeat number after excluding samples with low imputation quality (**Methods**). Overall SNP associations were qualitatively similar to those reported in the much larger cohort (700,000 samples) used to obtain summary statistics above (Yengo et al., 2018). The top associated SNP is the same in both studies (rs2581830, p=0.0024; n=11,687 in eMERGE), validating our analysis (**Figure 6D**). 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. 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. Encouragingly, AC repeat number shows a strong positive relationship with height across a range of repeat lengths (**Figure 6E**), 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 to test for an effect of the number of repeats on gene expression. We synthesized constructs consisting of varying numbers of AC repeats (0, 5, 10) with approximately 170bp of genomic sequence context on either side (**Table S10**). We additionally obtained a construct with 12 repeats from genomic DNA, since repeats of length 12 or longer could not be commercially synthesized (**Methods**). We cloned these constructs, as well as a cyclic AMP responsive element (CRE) as a positive control, into a plasmid upstream of a minimal promoter for a firefly luciferase reporter (**Methods**). We transfected each plasmid into the human cell line HEK293T and measured normalized expression from three biological repeats and three technical repeats for each construct. We observed a strong positive linear relationship between the number of AC repeats and reporter expression as predicted (**Figure 6F**) (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 provide further evidence that eSTRs identified by our genome-wide analysis may act as causal drivers of gene expression.

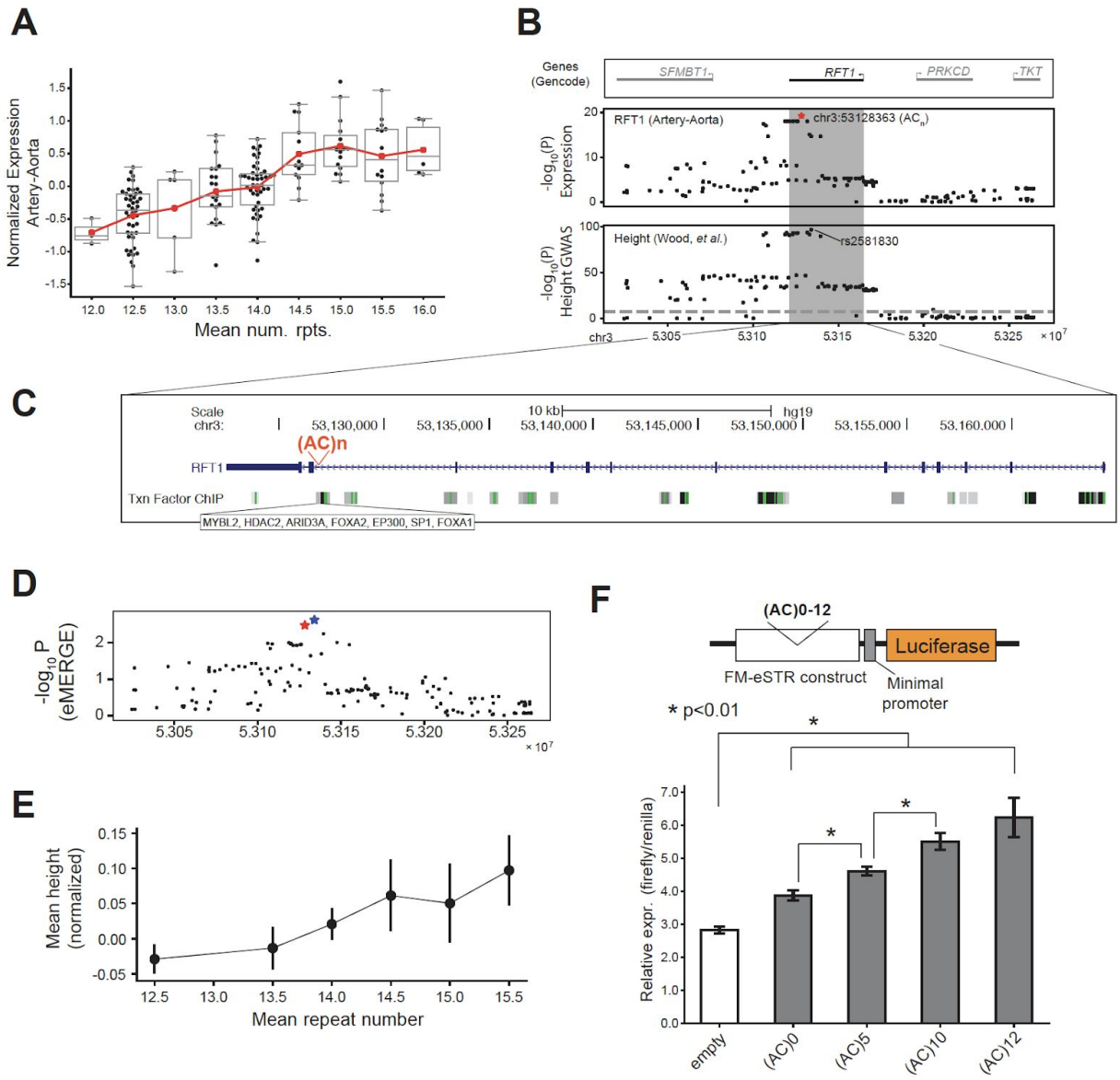

Figure 6: Evidence of a role for an FM-eSTR for *RFT1* in height.
(A) eSTR association for *RFT1*. The x-axis shows STR dosage at an AC repeat (chr3:53128363) as the mean number of repeats and the y-axis gives normalized *RFT1* expression in aortic artery. Each point represents a single individual. Boxplots summarize expression distributions for each genotype as described in **Figure 1D**. Red lines show the mean expression for each x-axis value.
(B) Summary statistics for *RFT1* expression and height. The top panel shows genes in the region around *RFT1*. The middle panel shows the -log$_{10}$ p-values of association between each variant and *RFT1* expression in aortic artery. The FM-eSTR is denoted by a red star. The bottom panel shows the -log$_{10}$ p-values of association for each variant with height based on available summary statistics (Yengo et al., 2018). The dashed gray horizontal line shows genome-wide significance threshold.
(C) Detailed view of the *RFT1* locus. A UCSC genome browser (Kent et al., 2002) 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) binding clusters profiled by ENCODE.
(D) eSTR and SNP associations with height at the *RFT1* locus in the eMERGE cohort. The x-axis shows the same genomic region as in **(B).** The y-axis denotes association p-values for each variant in the

subset of eMERGE cohort samples analyzed here. Black dots represent SNPs. The blue star denotes the top variant in the region identified by Yengo, *et al.* (rs2581830), which was also the top SNP in eMERGE. The red star represents the imputed FM-eSTR.

(E) Imputed *RFT1* repeat number is positively 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. Vertical black lines show +/- 1 s.d.

(F) Reporter assay shows expected positive trend between FM-eSTR repeat number and luciferase expression. Top: schematic of the experimental design (not to scale). A variable number of AC repeats plus surrounding genomic context (hg19 chr3:53128201-53128577) were introduced upstream of a minimal promoter driving reporter expression. Bottom: white bar shows results from the unmodified plasmid (empty). Gray bars show expression results for constructs with each number of repeats (0, 5, 10, and 12). Reporter expression results normalized to a Renilla control. Error bars show +/- 1 s.d. Asterisks represent significant differences between conditions (one-sided t-test $p<0.01$).

See also Tables S8-S9.

Discussion

A widespread role for eSTRs in regulating gene expression

Here we presented the most comprehensive resource of eSTRs to date, which reveals more than 3,000 high confidence associations across 17 tissues between the number of repeats at STRs and expression of nearby genes. The depth and breadth of this catalog allowed us to explore potential eSTR mechanisms on a genome-wide scale and to identify specific eSTRs likely driving complex traits in humans.

Until recently, STRs have routinely been missing from population-level genomic analyses due to computational and experimental challenges in genotyping these repetitive loci. Furthermore, because STRs are often highly multi-allelic, they do not fit easily into common association testing frameworks built primarily for analysis of bi-allelic SNPs and indels. Now, a combination of improved computational methods, the reduced cost and improved quality of sequencing as well as consortium led efforts to generate deep population-wide datasets, enable highly accurate genotyping of complex variants such as STRs from short reads (Willems et al., 2017).

The recent explosion in the number of whole genome sequencing datasets available and improved quality of eQTL catalogs demonstrate the impact of these technological advancements. In our initial study (Gymrek et al., 2016), we used low-coverage WGS from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015) to identify eSTRs in a single cell type (lymphoblastoid cell lines) using RNA-seq generated by the gEUVADIS Project (Lappalainen et al., 2013). While this study revealed a widespread role for STRs in regulating gene expression, it was based on shallow sequencing data generated using sequencing technologies that have since been greatly improved. As a result, we had low power to distinguish individual eSTRs that were likely to be causal and thus could not explore eSTR mechanisms in detail. The GTEX Project (GTEx Consortium, 2015) includes deep (30x) WGS data with 100bp paired end reads for more than 650 individuals together with matched RNA-seq data for dozens of tissues. A recent study (Chiang et al., 2017) demonstrated the power of the GTEx dataset by performing joint eQTL mapping across SNPs, indels, and structural variants (SVs) to identify SVs likely to be causally associated with gene expression. However, repetitive variants were not included in this study.

eSTRs regulate expression through diverse mechanisms

STRs have long been hypothesized to affect transcription through a variety of regulatory processes. A unique feature of our resource is the ability to examine the impact of eSTRs on these mechanisms as a function of repeat number across individuals in a population. While we explored two specific eSTR mechanisms here, our results suggest that they likely work via a variety of mechanisms to impact gene expression. Overall, eSTRs are most strongly enriched at promoter and putative enhancers, but thousands of the ones we identified occur outside of these regions, including in UTRs, introns, and intergenic regions. Moreover, eSTRs analyzed here consist of a large spectrum of repeat classes with a variety of repeat motif lengths and sequences, ranging from homopolymers to hexanucleotide repeats (**Figure 7A**). It is probable that each type induces distinct regulatory effects.

Homopolymer A/T (also referred to as poly(dA/dT)) repeats make up more than one third of FM-eSTRs we identified. Our results agree with previous reports demonstrating that these STRs act as nucleosome positioning signals and are strongly nucleosome depleted (Segal and Widom, 2009). Intriguingly, we find that homopolymer eSTRs tend to be located near DNAseI HS sites at a distance that is a multiple of 147bp, the length of DNA wrapped around nucleosomes (Jiang and Pugh, 2009), further suggesting homopolymer eSTRs may act primarily through modulating nucleosome positioning. Furthermore, homopolymer eSTRs in promoter regions show a directional bias according to the motif on the template strand and localization relative to the TSS, consistent with a previous finding that A/T repeats shift nucleosomes in a strand-specific manner (Krietenstein et al., 2016). Overall, our results support a model in which longer homopolymer tracts more strongly push nucleosomes in a direction dependent on the strand orientation of the repeat (**Figure 7B**).

Another highly prevalent type of FM-eSTR consisted of repeats with GC-rich motifs in close proximity to TSSs. An association between promoter GC content and gene expression has been firmly established (Deaton and Bird, 2011; Saxonov et al., 2006). Multiple mechanisms driving this relationship have been proposed, including reduced DNA methylation at CpG islands (Bergman and Cedar, 2013) and recruitment of the the Polycomb group complex (Mendenhall et al., 2010). Our analysis suggests an additional mechanism by which GC-rich repeats form non-canonical DNA and RNA secondary structures such as G4-quadruplexes that can impact

transcription (Bochman et al., 2012; Hänsel-Hertsch et al., 2016) in a repeat number-dependent manner (**Figure 7C**).

We found that longer repeat numbers at these eSTRs are predicted to have strong stabilizing effects on both single-stranded DNA and RNA. In almost all cases, G4-forming repeats in non-coding regions near TSSs showed a positive effect between number of repeats and DNA stability, and corresponding positive effect on gene expression, consistent with previous studies showing that enhanced G4 formation on single-stranded DNA drives transcription activity (Hänsel-Hertsch et al., 2016). On the other hand, CCG repeats, found primarily in 5'UTRs, show no such directional bias on expression suggesting they act by an alternative mechanism, potentially related to formation of R-loops or other secondary structures involving mRNA (**Figure 7C**).

Notably, our results do not rule out other hypothesized mechanisms for these eSTRs. For instance, homopolymers may modulate spacing between regulatory elements (Panne et al., 2007; Swanson et al., 2010), alter affinity of nearby TF binding sites (Afek et al., 2014), or induce non-canonical secondary structures. Strand-specific effects ranging tens of kb downstream of TSSs fall primarily in introns and may affect RNA splicing rather than nucleosome positioning, although the effect size direction bias would not be easily explained by this phenomenon. Additionally, a subset of GC-rich repeats have previously been suggested to act by forming SP1 binding sites (Drazen et al., 1999) or altering DNA methylation (Liu et al., 2018). Furthermore, homopolymers and GC-rich repeats make up only a subset of FM-eSTRs. Other repeat classes have been suggested to work by different mechanisms including altering affinity or spacing of DNA- or RNA- protein binding sites (Balestrieri et al., 2018; Goodwin and Swanson, 2014) (**Figure 7D**), regulating alternative splicing (Hefferon et al., 2004), and forming transcription-inducing Z-DNA structures (Rothenburg et al., 2001).

eSTRs are potential drivers of complex traits

We leveraged our resource to demonstrate that FM-eSTRs are likely driving published GWAS associations for a variety of complex traits. For schizophrenia, we detected a dinucleotide repeat which CAVIAR identified as the most probable causal variant for expression of *MED19*. Co-localization analysis shows *MED19* and the schizophrenia signal in that region are likely driven by the same variant, suggesting the eSTR is causal for the GWAS signal. *MED19* is part

of the mediator complex, which has previously been shown to play a role in schizophrenia and neurodevelopmental phenotypes (Snijders Blok et al., 2018; Spaeth et al., 2011). Furthermore, a recently developed burden test taking into account promoter-enhancer identified a significant association between *MED19* expression and schizophrenia risk (Wu and Pan, 2018).

In a similar analysis, we found another dinucleotide repeat acting as an FM-eSTR for *RFT1* is co-localized with a GWAS signal for height. We imputed the STR into an independent cohort and found a positive association between repeat number and height. We further validated that this FM-eSTR has a repeat number-dependent impact on expression using a reporter assay. This repeat falls within a cluster of binding sites for transcription factors and chromatin regulators annotated by ENCODE. We hypothesize that this STR could either be forming a binding site itself or could act by modulating spacing between the numerous nearby binding motifs on either side (**Figure 7D**).

Altogether, our complex trait analysis demonstrates that STRs may represent an important class of variation that is largely missed by current GWAS. STRs have a unique ability compared to bi-allelic variants such as SNPs or small indels to drive phenotypic variation along a spectrum of multiple alleles, each with different numbers of repeats. In both the *MED19* and *RFT1* examples, the eSTR shows 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. Importantly, the cases identified here likely represent a minority of eSTRs driving complex traits. Our analysis is based only on GWAS signals that could be detected by standard SNP-based GWAS, which are underpowered to detect underlying multi-allelic signals from STRs (Saini et al., 2018). Further work to directly test for associations between STRs and phenotypes is likely to reveal a widespread role for repeat number variation in complex traits.

Challenges in detection of eSTRs

Our study faced several limitations. First, eSTR discovery was limited to linear associations between repeat number and expression. Our analysis did not consider non-linear effects or the effect of sequence imperfections, such as SNPs or small indels, within the STR sequence itself. Second, 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 (Chiang et al., 2017) or *Alu* elements (Häsler and

Strub, 2006) that were not considered. Third, 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 and had low power for eSTR detection. This additionally made tissue-specific analyses difficult due to variation in power across different tissues. Finally, while our results strongly suggest several eSTR mechanisms including nucleosome positioning and formation of non-canonical DNA secondary structures, further work will be needed to experimentally validate both of these in relevant cell types.

In summary, our eSTR catalog provides a valuable resource both for obtaining deeper insights into both known and unexplored biological roles of eSTRs in regulating gene expression and for identifying potential causal variants underlying a variety of complex traits.

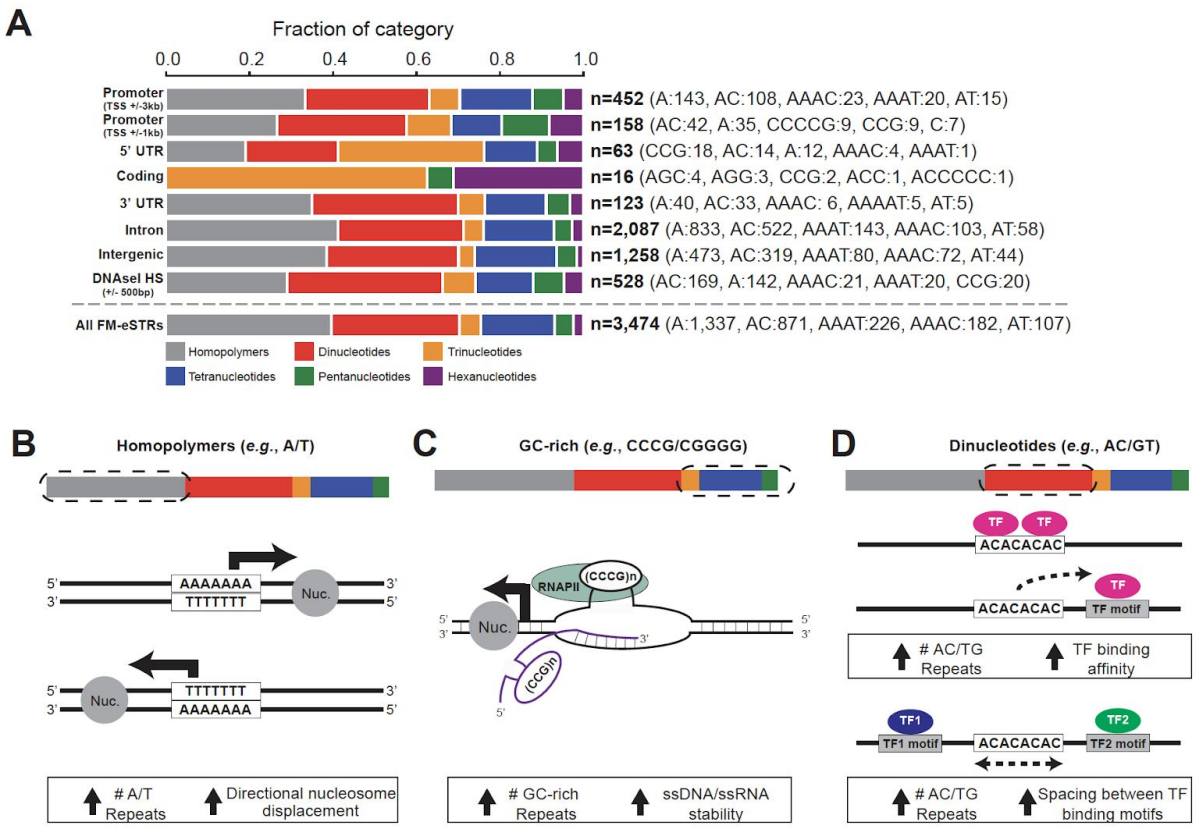

Figure 7: Summary of FM-eSTRs classes and potential regulatory mechanisms
(A) Distribution of FM-eSTR classes across genomic annotations. Each bar shows the fraction of FM-eSTRs falling in each annotation consisting of homopolymer (gray), dinucleotide (red), trinucleotide (orange), tetranucleotide (blue), pentanucleotide (green) or hexanucleotide (purple) repeats. The total number of FM-eSTRs and the top five most common motifs 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 expel nucleosomes in a direction dependent on the strand of the motif (indicated by the thick black arrows). The strength of the expulsion may depend on the number of repeats.
(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.

Materials and 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 samples with diverse ethnic backgrounds (**Figure S1**). 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 (Willems et al., 2017), 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-cal-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 (Kent et al., 2002) hg19.genomicSuperDups table); penta- and hexamer STRs containing homopolymer runs of at least 5 or 6 nucleotides, respectively in the hg19 reference genome; and STRs whose frequencies did not meet the percentage of homozygous vs. heterozygous calls based on 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 (**Table S1**). Genes with median expression level of 0 were excluded and expression values for remaining genes were quantile normalized to a standard normal distribution.

eSTR and eSNP identification

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

$$Y = \beta X + C + \varepsilon$$

Where X denotes STR genotypes, Y denotes expression values, β denotes the effect size, C denotes various covariates, and ε is the error term. A separate regression analysis was performed for each tissue. We used "STR dosage", defined as the sum of repeat lengths of the two alleles for each sample, to define STR genotypes. All repeat lengths are reported as length difference from the hg19 reference, with 0 representing the reference allele. STR dosages were scaled to have mean 0 and variance 1.

We included 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 (1000 Genomes Project Consortium et al., 2015) samples genotyped using Omni 2.5 genotyping arrays (see **URLs**). Analysis was restricted to bi-allelic SNPs present in the Omni 2.5 data and resulting loci were LD-pruned using plink (Purcell et al., 2007) with option --indep 50 5 2. PCA on resulting SNP genotypes was performed using smartpca (Patterson et al., 2006; Price et al., 2006).

To control for technical variation in expression, we applied PEER factor correction (Stegle et al., 2012). As suggested for the PEER method, we used N/4 PEER factors as covariates for each tissue, where N is the sample size. PEER factors altogether accounted for more than 75% of variation in gene expression and were correlated with covariates reported previously for GTEx samples (**Figure S2**). We used a gene-level FDR threshold (described previously (Gymrek et al., 2016)) of 10% to identify significant STR-gene pairs. eSNPs were identified using the same model and covariates, but using SNP dosages (0, 1, or 2) rather than STR dosages.

Fine-mapping eSTRs

We used model comparison 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 qvalue package in R (see **URLs**).

We used CAVIAR (Hormozdiari et al., 2014) v1.0 to further fine-map eSTR signals against the top 100 eSNPs within 100kb of each gene. Pairwise-LD between the eSTR and eSNPs was estimated using the Pearson correlation between SNP dosages (0, 1, or 2) and STR dosages (sum of the two repeat allele lengths) across all samples. CAVIAR was run with parameters -f 1 -c 1.

Enrichment analyses

Enrichment analysis were performed using a two-sided Fisher's exact test as implemented in the fisher_exact function of the python package scipy.stats (see **URLs**). Overlapping STRs with each annotation was performed using the intersectBed tool of the BEDTools (Quinlan and Hall, 2010) suite.

Genomic annotations were obtained by downloading custom tables using the UCSC Genome Browser (Kent et al., 2002) table browser tool to select either coding regions, introns, 5'UTRs, 3'UTRs, or upstream regions. For **Figure S4** and **Table S3**, each STR was assigned to a single annotation using the following order of preference: coding, 5'UTR, 3'UTR, promoter (within 3kb of the TSS), or intron. STRs not assigned to one of those categories were labeled as intergenic.

ENCODE DNAseI HS clusters were downloaded from the UCSC Genome Browser (see **URLs**). Analysis was restricted to clusters seen in at least 20 cell types. ENCODE TF binding sites were downloaded from the UCSC Genome Browser (see **URLs**). The distance between each STR and the center of the nearest DNAseI HS cluster was computed using the closestBed tool from the BEDTools suite (Quinlan and Hall, 2010).

Analysis of nucleosome occupancy

DNAseI-seq reads (accession ENCFF775ZJX) and ChIP-seq reads for RNAPII (accession ENCFF865BUP) generated using the lymphoblastoid cell line GM12878 were downloaded from the ENCODE Project website (see **URLs**). Genome-wide nucleosome occupancy signal in GM12878 was downloaded from the UCSC Genome Browser (see **URLs**). Histograms of aggregate read densities and heatmaps for individual STR regions were generated using the annoatePeaks.pl tool of Homer (Heinz et al., 2010).

Analysis of DNA and RNA secondary structure

For each STR, we extracted the repeat plus 50bp flanking sequencing from the hg19 reference genome. We additionally created sequences containing each STR allele for each locus seen at least 5 times in a previously generated catalog STR variation 1,916 samples (Saini et al., 2018). For each sequence and its reverse complement, we ran mfold (Zuker, 2003) 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. Kolmogorov-Smirnov tests were performed using the ks_2samp function of the scipy.stats python package (see **URLs**).

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 September 18, 2018. Height GWAS summary statistics were downloaded from the GIANT Consortium website (see **URLs**). Schizophrenia GWAS summary statistics were downloaded from the Psychiatric Genomics Consortium website (see **URLs**).

LD between STRs and SNPs was computed by taking the squared Pearson correlation between STR dosages 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 (Guo et al., 2015) package. For all traits, dataset 1 was specified as type="quant" and consisted of eSNP 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, we specified type="quant" and used height 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. For schizophrenia, we specified type="CC" and used schizophrenia effect sizes and their variances as input. We additionally specified the fraction of cases (33%) and minor allele frequencies listed in the published summary statistics.

Capture Hi-C interactions were visualized using the 3D Genome Browser (Wang et al., 2017). The visualization depicts interactions profiled in GM12878 (Mifsud et al., 2015) 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 (Purcell et al., 2007) to restrict to SNPs with minor allele frequency at least 10% and with genotype frequencies expected under Hardy-Weinberg Equilibrium ($p>1e-4$). 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 (Howie et al., 2009) 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 (Browning et al., 2018) with option gp=true and using our SNP-STR reference haplotype panel (Saini et al., 2018). 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 dosage (defined above as the sum 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 (see **URLs**).

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 **Table S10**) 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 (**Table S10**) which included the same homology arms.

We cloned constructs 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 hr at 37°C, washed once with phosphate-buffered saline, and then incubated in fresh completed medium for an additional 24 hr.

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 10µl 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. Values shown in **Figure 6F** represent the mean and standard deviation across the three biological replicates for each construct.

URLs

1000 Genomes phased Omni2.5 SNP data, ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/shapeit2_scaffolds/hd_chip_scaffolds/

Q-value package, http://github.com/jdstorey/qvalue

Python scipy.stats package. https://docs.scipy.org/doc/scipy/reference/stats.html

ENCODE DNAseI HS clusters, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegDnaseClustered/wgEncodeRegDnaseClusteredV3.bed.gz

ENCODE transcription factor binding clusters, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeRegTfbsClustered/wgEncodeRegTfbsClusteredV3.bed.gz

DNAseI-seq in GM12878, https://www.encodeproject.org/files/ENCFF775ZJX/@@download/ENCFF775ZJX.bam

Nucleosome occupancy in GM12878, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeSydhNsome/wgEncodeSydhNsomeGm12878Sig.bigWig

RNAPII ChIP-seq in GM12878, https://www.encodeproject.org/files/ENCFF865BUP/@@download/ENCFF865BUP.bam

Height GWAS summary statistics https://portals.broadinstitute.org/collaboration/giant/images/0/0f/Meta-analysis_Locke_et_al%2BUKBiobank_2018.txt.gz

Schizophrenia GWAS summary statistics, https://www.med.unc.edu/pgc/results-and-downloads

Python statsmodels package. https://www.statsmodels.org

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. was supported in part by NIH/NIMH grant R01 MH113715. We thank Vineet Bafna, Eric Mendenhall, Joseph Gleeson, and Yu-Ting Liu for helpful comments.

GTEx Project

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), GTEx Project March 5, 2014 version Page 5 of 8 Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded

through a contract (HHSN268201000029C) to the The Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951,MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819),Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v7.p2.

eMERGE Project - Mayo Clinic

Samples and associated genotype and phenotype data used in this study were provided by the Mayo Clinic. Funding support for the Mayo Clinic was provided through a cooperative agreement with the National Human Genome Research Institute (NHGRI), Grant #: UOIHG004599; and by grant HL75794 from the National Heart Lung and Blood Institute (NHLBI). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Marshfield Clinic

Funding support for the Personalized Medicine Research Project (PMRP) was provided through a cooperative agreement (U01HG004608) with the National Human Genome Research Institute (NHGRI), with additional funding from the National Institute for General Medical Sciences (NIGMS) The samples used for PMRP analyses were obtained with funding from Marshfield Clinic, Health Resources Service Administration Office of Rural Health Policy grant number D1A RH00025, and Wisconsin Department of Commerce Technology Development Fund contract number TDF FYO10718. Funding support for genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). The datasets used for analyses

described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Northwestern University

Samples and data used in this study were provided by the NUgene Project (www.nugene.org). Funding support for the NUgene Project was provided by the Northwestern University's Center for Genetic Medicine, Northwestern University, and Northwestern Memorial Hospital. Assistance with phenotype harmonization was provided by the eMERGE Coordinating Center (Grant number U01HG04603). This study was funded through the NIH, NHGRI eMERGE Network (U01HG004609). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Vanderbilt University

Funding support for the Vanderbilt Genome-Electronic Records (VGER) project was provided through a cooperative agreement (U01HG004603) with the National Human Genome Research Institute (NHGRI) with additional funding from the National Institute of General Medical Sciences (NIGMS). The dataset and samples used for the VGER analyses were obtained from Vanderbilt University Medical Center's BioVU, which is supported by institutional funding and by the Vanderbilt CTSA grant UL1RR024975 from NCRR/NIH. Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

eMERGE Project - Group Health Cooperative

Funding support for Alzheimer's Disease Patient Registry (ADPR) and Adult Changes in Thought (ACT) study was provided by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). A gift from the 3M Corporation was used to expand the ACT cohort. DNA aliquots sufficient for GWAS from ADPR Probable AD cases, who had been enrolled in Genetic Differences in Alzheimer's Cases and Controls (Walter Kukull, PI, R01 AG007584) and obtained under that grant, were made available to eMERGE without charge. Funding support for

genotyping, which was performed at Johns Hopkins University, was provided by the NIH (U01HG004438). Genome-wide association analyses were supported through a Cooperative Agreement from the National Human Genome Research Institute, U01HG004610 (Eric B. Larson, PI). The datasets used for analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000360.v3.p1 and phs000888.v1.p1.

Author Contributions

S.F.F. performed all eSTR and eSNP mapping, helped perform downstream analyses and helped draft 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. 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 drafted the initial manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors have no competing financial interests to disclose.

References

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Afek, A., Schipper, J.L., Horton, J., Gordân, R., and Lukatsky, D.B. (2014). Protein-DNA binding in the absence of specific base-pair recognition. Proc. Natl. Acad. Sci. U. S. A. *111*, 17140–17145.

Aporntewan, C., Pin-on, P., Chaiyaratana, N., Pongpanich, M., Boonyaratanakornkit, V., and Mutirangura, A. (2013). Upstream mononucleotide A-repeats play a cis-regulatory role in mammals through the DICER1 and Ago proteins. Nucleic Acids Res. *41*, 8872–8885.

Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. Cell *167*, 1415–1429.e19.

Balestrieri, C., Alfarano, G., Milan, M., Tosi, V., Prosperini, E., Nicoli, P., Palamidessi, A., Scita, G., Diaferia, G.R., and Natoli, G. (2018). Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity. Cell *173*, 1150–1164.e14.

Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. Nat. Struct. Mol. Biol. *20*, 274–281.

Bochman, M.L., Paeschke, K., and Zakian, V.A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. *13*, 770–780.

Boettger, L.M., Salem, R.M., Handsaker, R.E., Peloso, G.M., Kathiresan, S., Hirschhorn, J.N., and McCarroll, S.A. (2016). Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat. Genet. *48*, 359–366.

Borel, C., Migliavacca, E., Letourneau, A., Gagnebin, M., Béna, F., Sailani, M.R., Dermitzakis, E.T., Sharp, A.J., and Antonarakis, S.E. (2012). Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. Hum. Mutat. *33*, 1302–1309.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one penny imputed genome from next generation reference panels.

Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., GTEx Consortium, et al. (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.

Ciesiolka, A., Jazurek, M., Drazkowska, K., and Krzyzosiak, W.J. (2017). Structural Characteristics of Simple RNA Repeats Associated with Disease and their Deleterious Protein Interactions. Front. Cell. Neurosci. *11*, 97.

Conlon, E.G., Lu, L., Sharma, A., Yamazaki, T., Tang, T., Shneider, N.A., and Manley, J.L. (2016). The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and

sequesters hnRNP H to disrupt splicing in ALS brains. Elife *5*.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Ding, N., Tomomori-Sato, C., Sato, S., Conaway, R.C., Conaway, J.W., and Boyer, T.G. (2009). MED19 and MED26 are synergistic functional targets of the RE1 silencing transcription factor in epigenetic silencing of neuronal gene expression. J. Biol. Chem. *284*, 2648–2656.

Drazen, J.M., Yandava, C.N., Dubé, L., Szczerback, N., Hippensteel, R., Pillari, A., Israel, E., Schork, N., Silverman, E.S., Katz, D.A., et al. (1999). Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. Nat. Genet. *22*, 168–170.

Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., and Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev. *18*, 1618–1629.

Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. *44*, 445–477.

Goodwin, M., and Swanson, M.S. (2014). RNA-binding protein misregulation in microsatellite expansion disorders. Adv. Exp. Med. Biol. *825*, 353–388.

Grünewald, T.G.P., Bernard, V., Gilardi-Hebenstreit, P., Raynal, V., Surdez, D., Aynaud, M.-M., Mirabeau, O., Cidre-Aranaz, F., Tirode, F., Zaidi, S., et al. (2015). Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene EGR2 via a GGAA microsatellite. Nat. Genet. *47*, 1073–1078.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J.A., and Wallace, C. (2015). 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*, 3305–3313.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale

transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., et al. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29.

Haeuptle, M.A., Pujol, F.M., Neupert, C., Winchester, B., Kastaniotis, A.J., Aebi, M., and Hennet, T. (2008). Human RFT1 deficiency leads to a disorder of N-linked glycosylation. Am. J. Hum. Genet. *82*, 600–606.

Halder, K., Halder, R., and Chowdhury, S. (2009). Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals. Mol. Biosyst. *5*, 1703–1712.

Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for “missing heritability.” Trends Genet. *26*, 59–65.

Hannan, A.J. (2018). Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet. *19*, 286–298.

Hänsel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., et al. (2016). G-quadruplex structures mark human regulatory chromatin. Nat. Genet. *48*, 1267–1272.

Häsler, J., and Strub, K. (2006). Alu elements as regulators of gene expression. Nucleic Acids Res. *34*, 5491–5497.

Hefferon, T.W., Groman, J.D., Yurk, C.E., and Cutting, G.R. (2004). A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing. Proc. Natl. Acad. Sci. U. S. A. *101*, 3504–3509.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. *10*, 161–172.

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

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. PLoS Genet. *10*, e1004722.

Kiliszek, A., Kierzek, R., Krzyzosiak, W.J., and Rypniewski, W. (2012). Crystallographic characterization of CCG repeats. Nucleic Acids Res. *40*, 8155–8162.

Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y., Habu, T., Liu, W., Okuda, H., and Koizumi, A. (2011). 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*, 121–130.

Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C.L., Pugh, B.F., and Korber, P. (2016). Genomic Nucleosome Organization Reconstituted with Pure Proteins. Cell *167*, 709–721.e12.

Lalioti, M.D., Scott, H.S., Buresi, C., Rossier, C., Bottani, A., Morris, M.A., Malafosse, A., and Antonarakis, S.E. (1997). Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. Nature *386*, 847–851.

Lappalainen, T., Sammeth, M., Friedländer, M.R., ’t Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Leffler, E.M., Band, G., Busby, G.B.J., Kivinen, K., Le, Q.S., Clarke, G.M., Bojang, K.A., Conway, D.J., Jallow, M., Sisay-Joof, F., et al. (2017). Resistance to malaria through structural variation of red blood cell invasion receptors. Science *356*.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics *30*, 2843–2851.

Lin, Y., Dent, S.Y.R., Wilson, J.H., Wells, R.D., and Napierala, M. (2010). R loops stimulate genetic instability of CTG.CAG repeats. Proc. Natl. Acad. Sci. U. S. A. *107*, 692–697.

Liu, X.S., Wu, H., Krzisch, M., Wu, X., Graef, J., Muffat, J., Hnisz, D., Li, C.H., Yuan, B., Xu, C., et al. (2018). Rescue of Fragile X Syndrome Neurons by DNA Methylation Editing of the FMR1 Gene. Cell *172*, 979–992.e6.

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. U. S. A. *107*, 961–968.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Min, J.L., Taylor, J.M., Richards, J.B., Watts, T., Pettersson, F.H., Broxholme, J., Ahmadi, K.R.,

Surdulescu, G.L., Lowy, E., Gieger, C., et al. (2011). The use of genome-wide eQTL associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. PLoS One *6*, e22070.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.

Moore, H., Greenwell, P.W., Liu, C.P., Arnheim, N., and Petes, T.D. (1999). Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. U. S. A. *96*, 1504–1509.

Mougey, E., Lang, J.E., Allayee, H., Teague, W.G., Dozor, A.J., Wise, R.A., and Lima, J.J. (2013). ALOX5 polymorphism associates with increased leukotriene production and reduced lung function and asthma control in children with poorly controlled asthma. Clin. Exp. Allergy *43*, 512–520.

Panne, D., Maniatis, T., and Harrison, S.C. (2007). An atomic model of the interferon-beta enhanceosome. Cell *129*, 1111–1123.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.

Press, M.O., Carlson, K.D., and Queitsch, C. (2014). The overdue promise of short tandem repeat variation for heritability. Trends Genet. *30*, 504–512.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R.S., Mittelman, D., and Sharp, A.J. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. *44*, 3750–3762.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Raveh-Sadka, T., Levo, M., Shabi, U., Shany, B., Keren, L., Lotan-Pompan, M., Zeevi, D., Sharon, E., Weinberger, A., and Segal, E. (2012). Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. Nat. Genet. *44*, 743–750.

Rhodes, D. (1979). Nucleosome cores reconstituted from poly (dA-dT) and the octamer of histones. Nucleic Acids Res. *6*, 1805–1816.

Rothenburg, S., Koch-Nolte, F., Rich, A., and Haag, F. (2001). 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.

Saini, S., Mitra, I., Mousavi, N., Fotsing, S.F., and Gymrek, M. (2018). A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat. Commun. *9*, 4397.

Sawaya, S., Bagshaw, A., Buschiazzo, E., Kumar, P., Chowdhury, S., Black, M.A., and Gemmell, N. (2013). Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One *8*, e54710.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. U. S. A. *103*, 1412–1417.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

Segal, E., and Widom, J. (2009). Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr. Opin. Struct. Biol. *19*, 65–71.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183.

Sherry, S.T. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Snijders Blok, L., Hiatt, S.M., Bowling, K.M., Prokop, J.W., Engel, K.L., Cochran, J.N., Bebin, E.M., Bijlsma, E.K., Ruivenkamp, C.A.L., Terhal, P., et al. (2018). De novo mutations in MED13, a component of the Mediator complex, are associated with a novel neurodevelopmental disorder. Hum. Genet.

Song, J.H.T., Lowe, C.B., and Kingsley, D.M. (2018). Characterization of a Human-Specific Tandem Repeat Associated with Bipolar Disorder and Schizophrenia. Am. J. Hum. Genet. *103*, 421–430.

Spaeth, J.M., Kim, N.H., and Boyer, T.G. (2011). Mediator and human disease. Semin. Cell Dev. Biol. *22*, 776–787.

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. *7*, 500–507.

Stephensen, C.B., Armstrong, P., Newman, J.W., Pedersen, T.L., Legault, J., Schuster, G.U., Kelley, D., Vikman, S., Hartiala, J., Nassir, R., et al. (2011). ALOX5 gene variants affect eicosanoid production and response to fish oil supplementation. J. Lipid Res. *52*, 991–1003.

Struhl, K. (1985). Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc. Natl. Acad. Sci. U. S. A. *82*, 8419–8423.

Sun, J.X., Helgason, A., Masson, G., Ebenesersdóttir, S.S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., et al. (2012). A direct characterization of human mutation

based on microsatellites. Nat. Genet. *44*, 1161–1165.

Suter, B., Schnappauf, G., and Thoma, F. (2000). Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res. *28*, 4083–4089.

Swanson, C.I., Evans, N.C., and Barolo, S. (2010). Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev. Cell *18*, 359–370.

Szlachta, K., Thys, R.G., Atkin, N.D., Pierce, L.C.T., Bekiranov, S., and Wang, Y.-H. (2018). Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. Genome Biol. *19*, 89.

Todd, A.K., Johnston, M., and Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res. *33*, 2901–2907.

Vikman, S., Brena, R.M., Armstrong, P., Hartiala, J., Stephensen, C.B., and Allayee, H. (2009a). Functional analysis of 5-lipoxygenase promoter repeat variants. Hum. Mol. Genet. *18*, 4521–4529.

Vikman, S., Brena, R.M., Armstrong, P., Hartiala, J., Stephensen, C.B., and Allayee, H. (2009b). Functional analysis of 5-lipoxygenase promoter repeat variants. Hum. Mol. Genet. *18*, 4521–4529.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.

Wang, Y., Zhang, B., Zhang, L., An, L., Xu, J., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., Hardison, R., et al. (2017). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions.

Willems, T., Gymrek, M., Highnam, G., 1000 Genomes Project Consortium, Mittelman, D., and Erlich, Y. (2014). The landscape of human STR variation. Genome Res. *24*, 1894–1904.

Willems, T., Gymrek, M., Poznik, G.D., Tyler-Smith, C., 1000 Genomes Project Chromosome Y Group, and Erlich, Y. (2016). Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. Am. J. Hum. Genet. *98*, 919–933.

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods.

Wu, C., and Pan, W. (2018). Integration of Enhancer-Promoter Interactions with GWAS Summary Results Identifies Novel Schizophrenia-Associated Genes and Pathways. Genetics *209*, 699–709.

Yáñez-Cuna, J.O., Arnold, C.D., Stampfel, G., Boryń, L.M., Gerlach, D., Rath, M., and Stark, A. (2014). Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. *24*, 1147–1156.

Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M.,

Hirschhorn, J., Yang, J., Visscher, P.M., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. *31*, 3406–3415.