nature genetics

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41588-019-0521-9

In the format provided by the authors and unedited.

The impact of short tandem repeat variation on gene expression

Stephanie Feupe Fotsing[1,2,6], Jonathan Margoliash[3,4], Catherine Wang[5], Shubham Saini[3], Richard Yanicky[4], Sharona Shleizer-Burko[4], Alon Goren[4*] and Melissa Gymrek[3,4*]

[1]Biomedical Informatics and Systems Biology, University of California San Diego, La Jolla, CA, USA. [2]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA. [3]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [4]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [5]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [6]Present address: La Jolla Institute of Immunology, La Jolla, CA, USA. *e-mail: agoren@ucsd.edu; mgymrek@ucsd.edu

Supplementary Note

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

FM-eSTRs are most strongly enriched for repeats with high GC content (*e.g.*, canonical repeat units CCG, CCCCG, CCCCCG, AGGGC) (**Fig. 2e**, **Supplementary Table 5**) which are found almost exclusively in promoter regions (**Extended Data Fig. 5**). Given their strong enrichment, we decided to further explore potential biological mechanisms by which this subset of FM-eSTRs might be working. These GC-rich repeat units have been shown to form highly stable secondary structures during transcription such as G4 quadruplexes in single-stranded DNA[1] or RNA[2] that may regulate gene expression. 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 properties of two classes of GC-rich FM-eSTRs: (*i*) those following the standard G4 motif ($G_3N_{1-7}G_3N_{1-7}G_3N_{1-7}G_3$)[3] and (*ii*) repeats with canonical repeat unit CCG which does not meet the standard G4 definition. Notably, the majority of CCG FM-eSTRs (79%) occur in 5' UTRs compared to only 11% for G4 repeats. We observed that both classes of GC-rich repeats are associated with higher RNAPII (**Supplementary Fig. 7**) and lower nucleosome occupancy (**Supplementary Fig. 8**) compared to all STRs. This relationship with RNAPII was observed across a diverse range of cell and tissue types.

To evaluate whether GC-rich repeats could be modulating DNA or RNA secondary structure, we used mfold[4] to calculate the free energy of each STR and 50bp of its surrounding context 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. We then computed the correlation between the number of repeats and free energy at each STR region. Overall, both G4 and CGG STRs have lower mean free energy (greater stability) and more negative correlations between repeat number and free energy compared to all STRs (**Supplementary Fig. 9**; adjusted Mann-Whitney [MW] one-sided $p<0.05$). Compared to all STRs, FM-eSTRs tend to have lower free energy and more negative correlations with repeat number (MW $p<0.05$ in all categories except for CGG STRs). Notably, both metrics (mean free energy and correlation of repeat number vs. free energy) are significantly correlated with the total sequence length of the STR in all cases (Pearson correlation $p<0.01$). FM-eSTRs tend to be longer than STRs overall (see main text), which may partially explain the secondary structure trends observed.

Based on previous observations[5], we predicted that a higher number of repeat units at GC-rich eSTRs would result in greater DNA or RNA stability and in turn would increase expression of nearby genes. Three example FM-eSTRs following this trend are shown in **Fig. 2f-h**. We tested whether FM-eSTRs were biased toward negative vs. positive effect sizes. As described in the main text, overall FM-eSTRs show no bias in effect direction. However, when considering only repeats in promoter regions (TSS +/- 3kb), 59% of FM-eSTRs have positive effect sizes, significantly more than the 50% expected by chance (binomial two-sided p=0.04; n=137). This effect was stronger when considering only G4 FM-eSTRs (87% positive effect sizes; p=0.0074; n=15) but not significant for CCG FM-eSTRs (62% positive; n=13; p=0.58). This direction of effect bias was consistent across a range of CAVIAR thresholds used to define FM-eSTRs (**Supplementary Fig. 10**). Altogether, these results support a model in which higher repeat numbers at GC-rich eSTRs in promoter regions stabilize DNA secondary structures which promote transcription. Lastly, the contradictory results for CCG STRs may indicate that those repeats could act by distinct mechanisms compared to G4 STRs, but also may be due in part to limited power from a smaller sample size.

Acknowledgements

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.

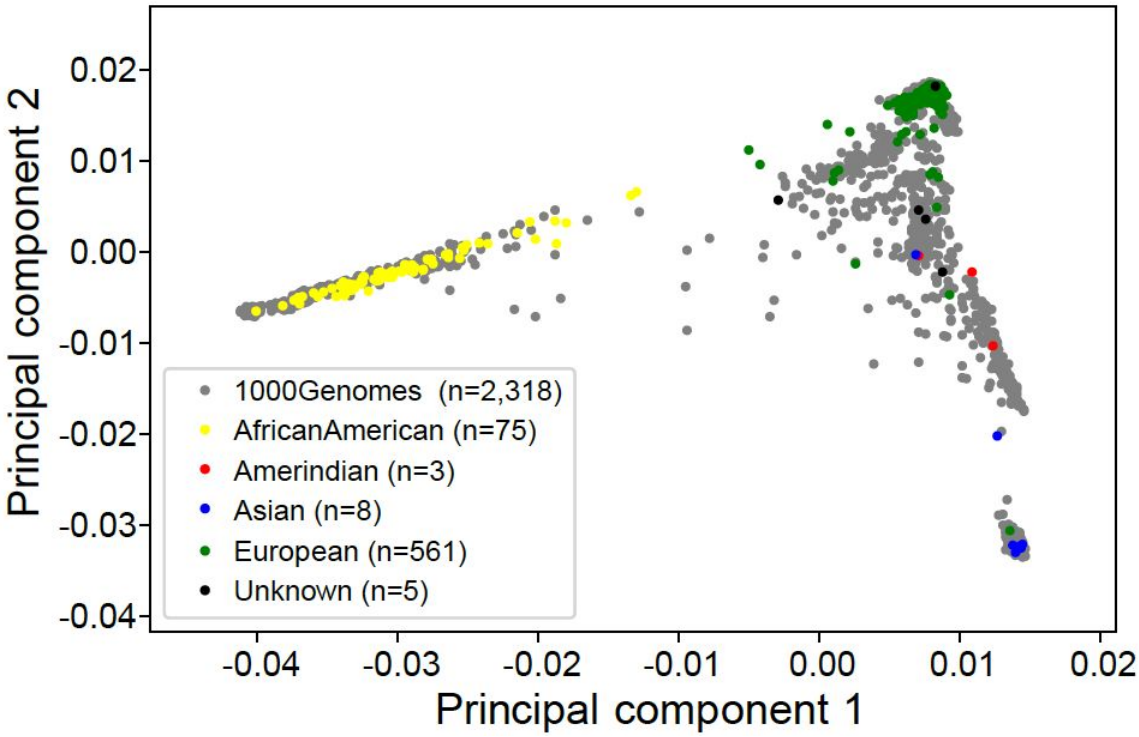
eMERGE Project - Vanderbilt University

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

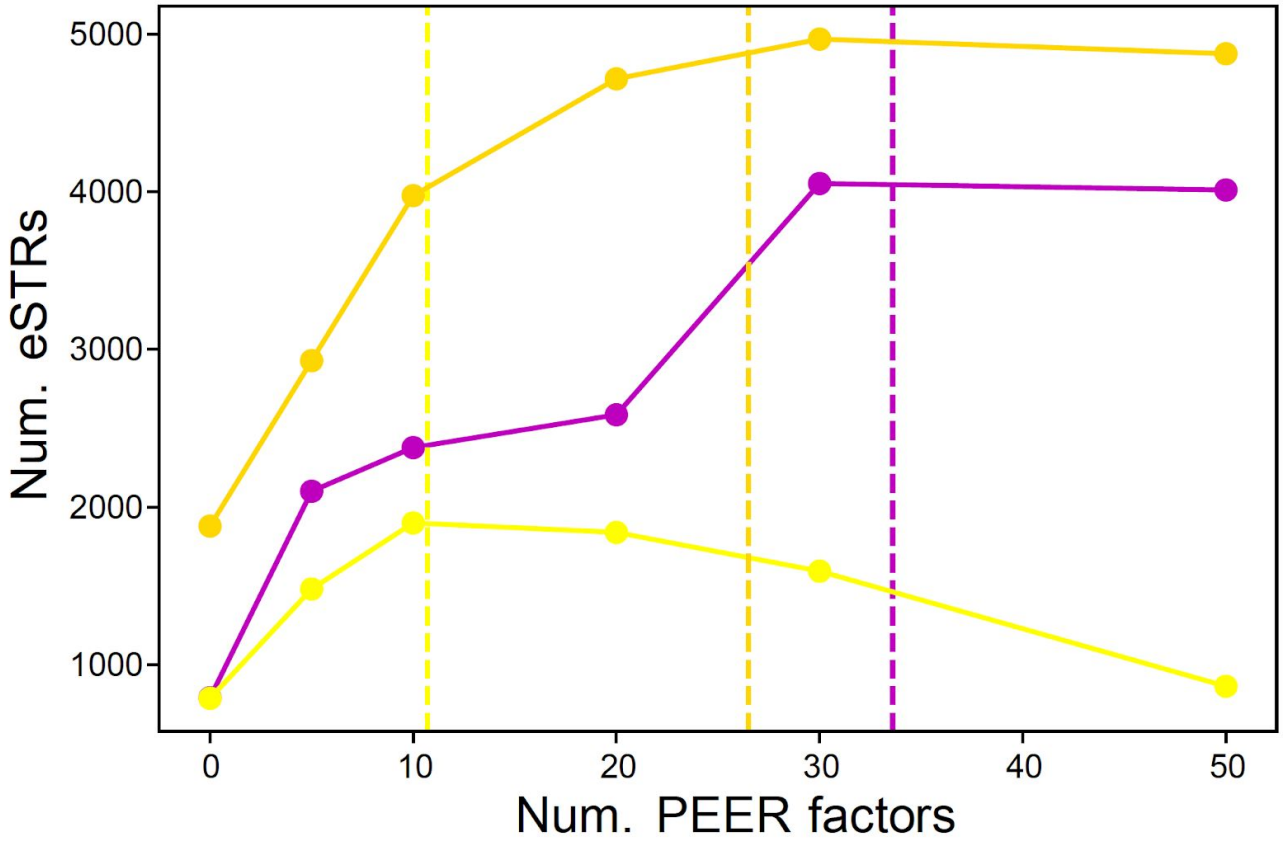
Funding support for Alzheimer's Disease Patient Registry (ADPR) and Adult Changes in Thought (ACT) study was provided by a U01 from the National Institute on Aging (Eric B. Larson, PI, U01AG006781). A gift from the 3M Corporation was used to expand the ACT cohort. DNA aliquots sufficient for GWAS from 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.

Supplementary Figure 1: Analysis of GTEx population structure



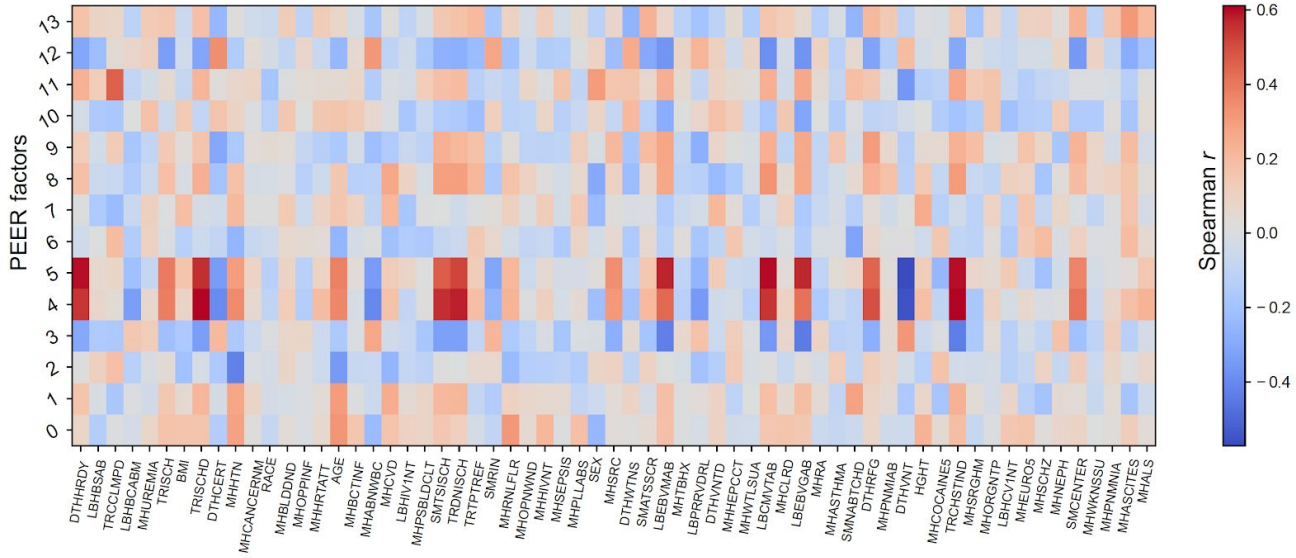
Principal component analysis was performed using SNP genotypes from the GTEx and 1000 Genomes cohorts. Samples from the 1000 Genomes project are shown in gray and GTEx samples are shown as colored dots based on ethnicity provided for each sample (yellow=African American; red=Amerindian; blue=Asian; green=European, black=Unknown).

Supplementary Figure 2: Effect of varying numbers of PEER factors on power to detect eSTRs



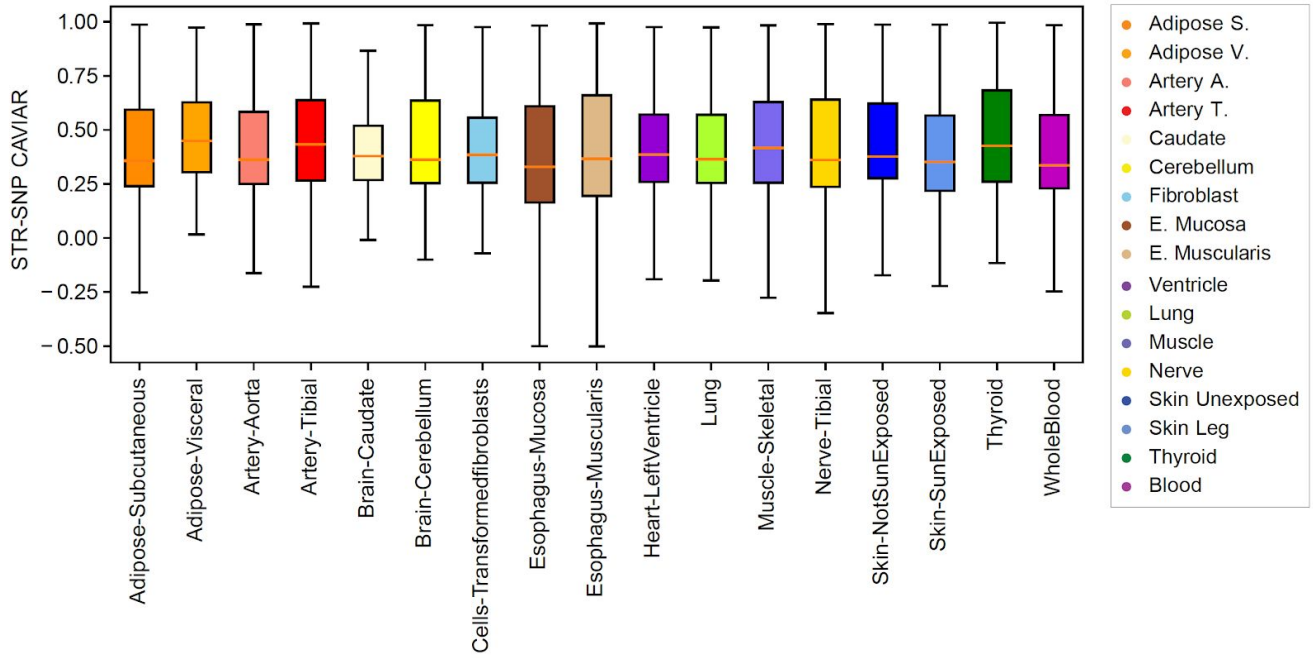
eSTRs (gene-level FDR 10%) were computed for each tissue after adjusting for a number of PEER factors ranging from 0 to 50. The numbers of STRxgene tests performed for each tissue are reported in **Supplementary Table 1**. The x-axis shows the number of PEER factors adjusted for. The y-axis shows the number of significant eSTRs at gene-level FDR of 10% (see **Methods**). Dashed vertical lines show the number of PEER factors equal to N/10, where N is the number of samples analyzed for each tissue. Purple=whole blood, yellow=Brain-Cerebellum, gold=Nerve-Tibial.

Supplementary Figure 3: Correlation of sample metadata with PEER factors



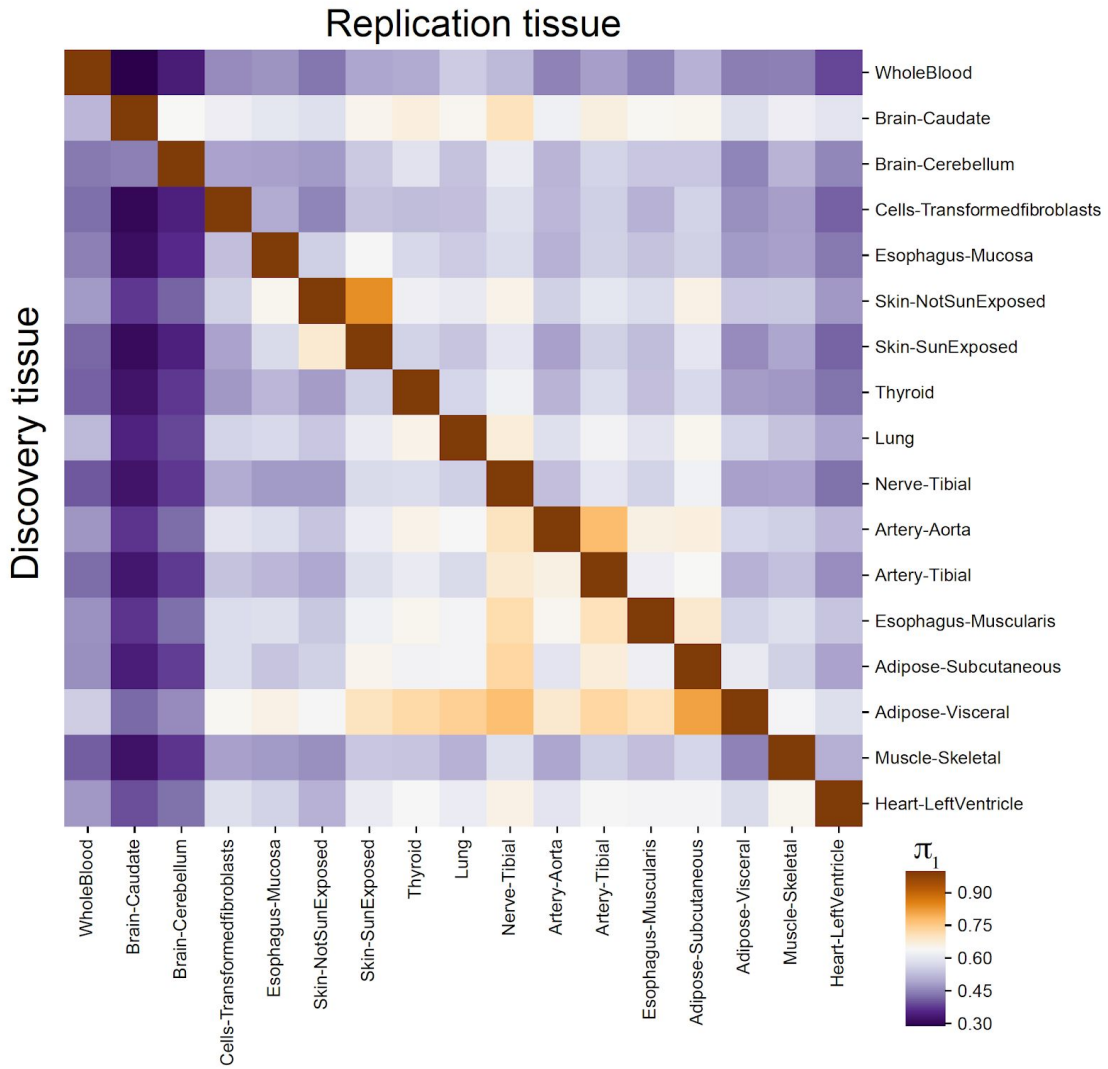
Each cell in the matrix shows the Spearman correlation of each PEER factor with data processing covariates. The x-axis represents each variable as defined for the GTEx cohort in dbGaP study phs000424.v7.p2. For example, covariates most strongly associated with PEER factors included DTHHRDY (Hardy scale for death classification) and TRISCHD (ischemic time). The y-axis represents factors obtained from PEER analysis of gene expression from Adipose-subcutaneous tissue (n=270 samples). Similar correlations were observed for other tissues.

Supplementary Figure 4: Difference in CAVIAR score between the top eSTR and top eSNP for each gene



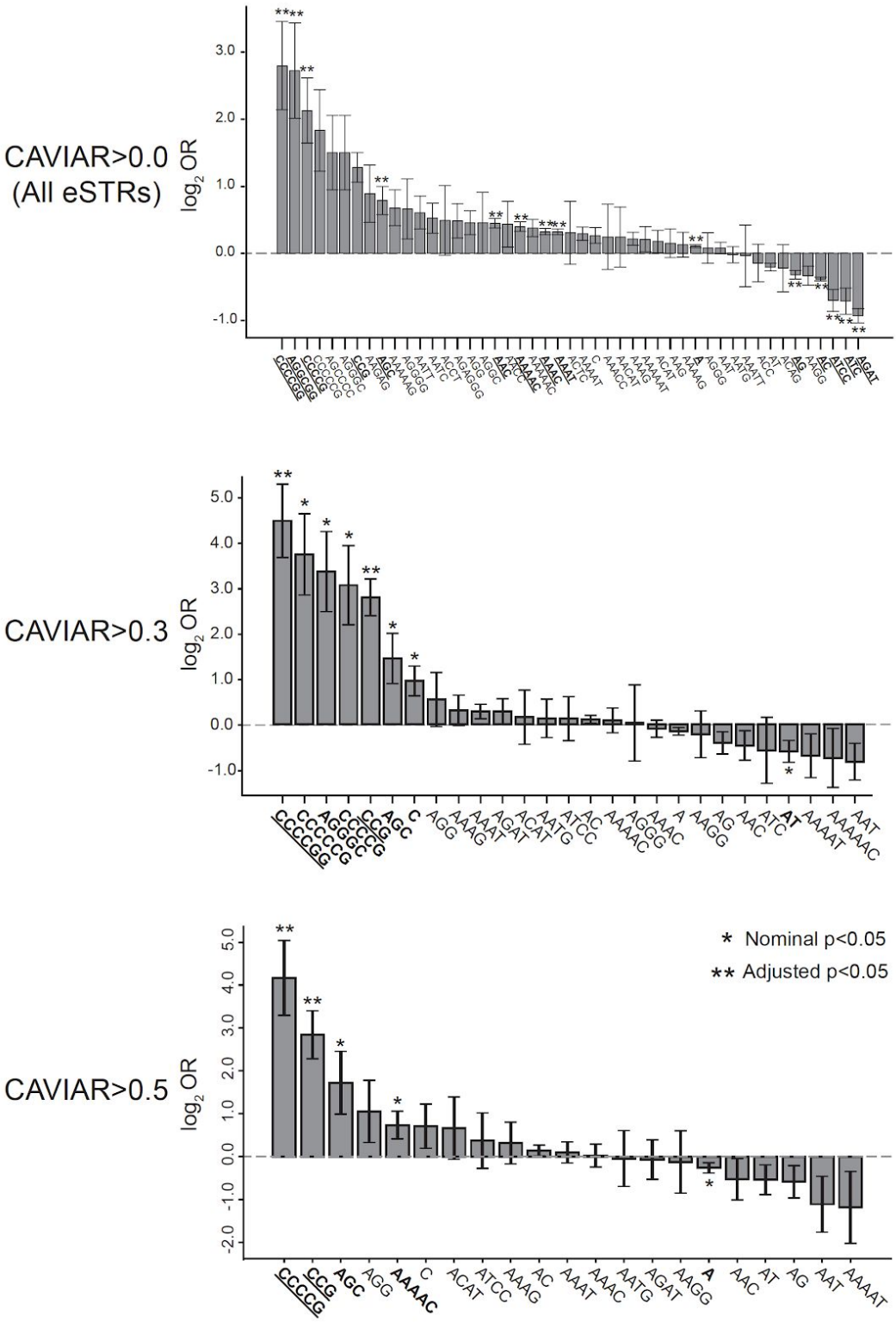
For each tissue, the boxplot shows the distribution of differences between the CAVIAR posterior score for the best STR and the best SNP for each gene. Data is only shown for genes with FM-eSTRs in each tissue. The numbers of FM-eSTRs identified in each tissue are shown in **Supplementary Table 1**. The colors of each box correspond to the different tissues (see legend on the right). Box plots are defined as in **Fig 1c.**

Supplementary Figure 5: Pairwise sharing of effect sizes across tissues



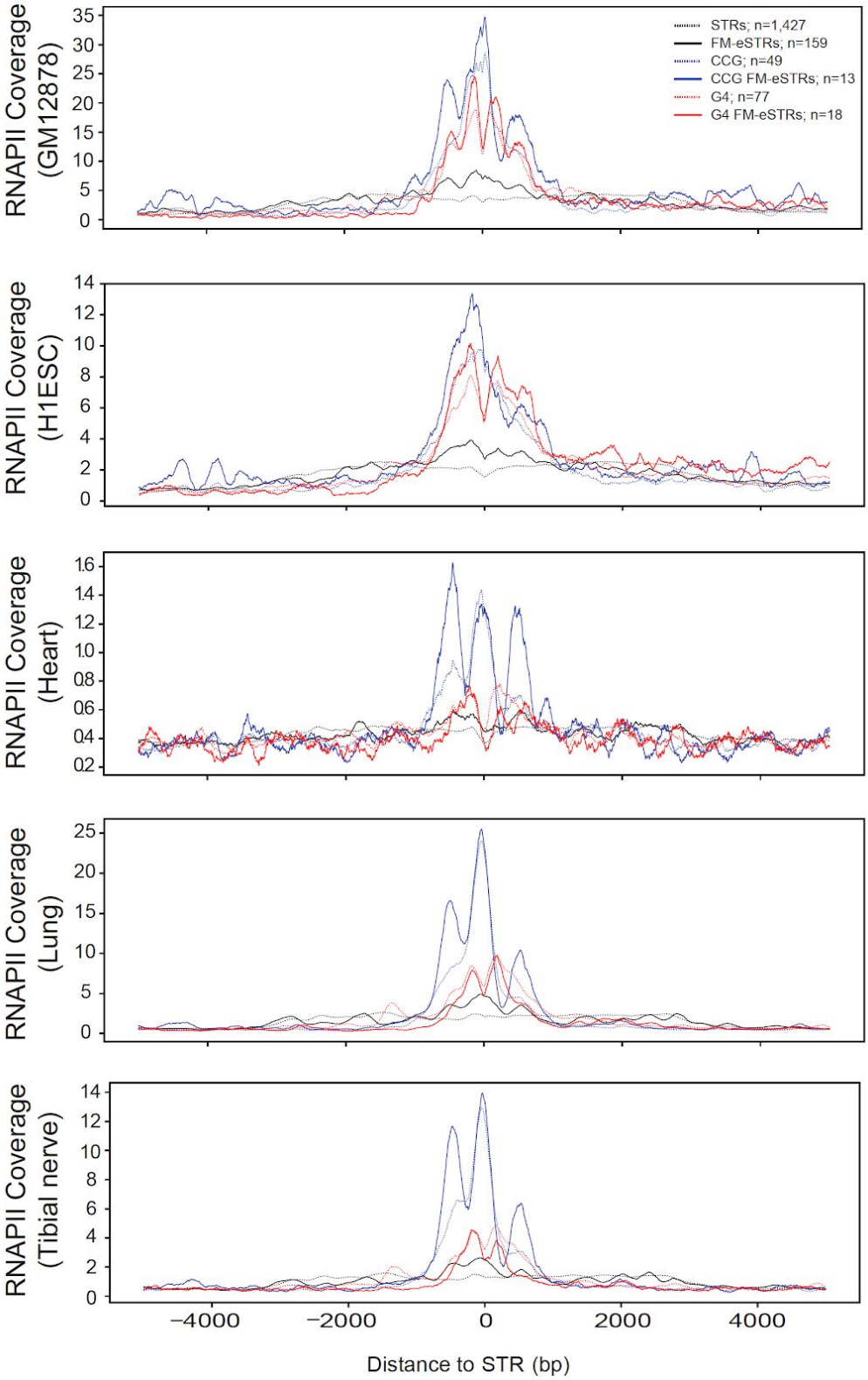
For each discovery tissue (rows), all eSTRs (gene-level FDR<10%) were tested for association in each other (replication) tissue (columns). The value in each cell gives the percent of eSTRs that were replicated (π_1). An eSTR was considered to be replicated if the eSTR measured effect size (β as defined in **Methods**) was nominally significantly different from 0 in the replication tissue (two-sided p<0.05). The numbers of eSTRs in each tissue are shown in **Supplementary Table 1**.

Supplementary Figure 6: Repeat unit enrichment at FM-eSTRs across all tissues for various CAVIAR thresholds.



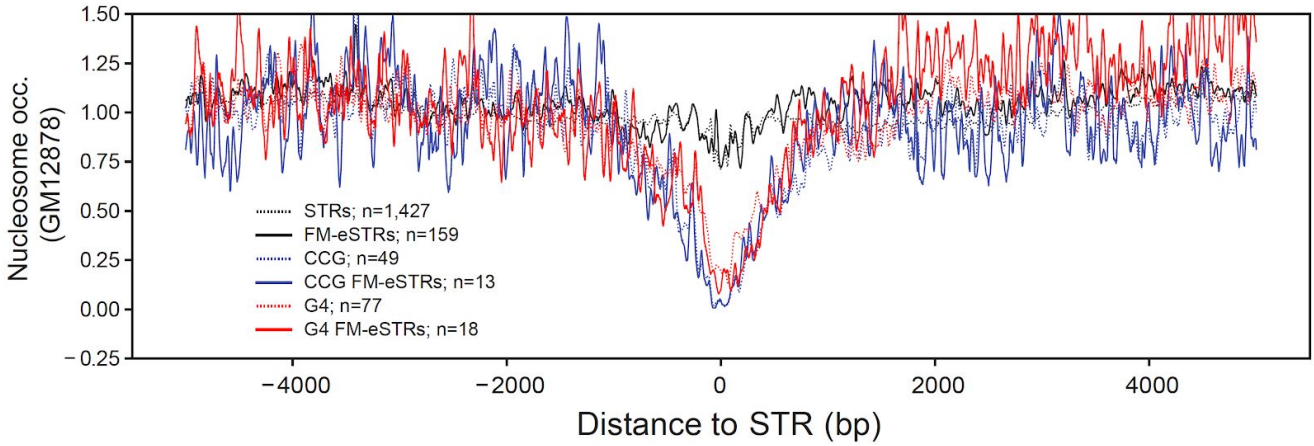
Repeat unit enrichment is shown for FM-eSTRs defined using multiple thresholds for CAVIAR scores (>0, corresponding to all eSTRs, >0.3, as used in the main text, or >0.5). The x-axis shows all repeat units for which there are at least 3 FM-eSTRs across all tissues. The y-axis denotes the $\log_2$ odds ratios (OR) from performing a Fisher's exact test comparing FM-eSTRs to all STRs. Center values are the $\log_2$ odds ratios, error bars represent +/-1 s.e. Single asterisks and bolded text denote repeat units nominally enriched or depleted (two-sided Fisher exact test $p<0.05$). Double asterisks and bolded-underlined text denote repeat units significantly enriched after controlling for the number of repeat units tested (Bonferroni adjusted $p<0.05$). Nominally significant repeat units are not denoted in the top plot due to difficulty in visualizing them for the large number of repeat units shown. The top plot only includes repeat units for which there are at least 10 FM-eSTRs across all tissues. Per repeat unit sample sizes and Fisher exact statistics are provided in **Supplementary Table 5**.

Supplementary Figure 7: Density of RNAPII localization around STRs



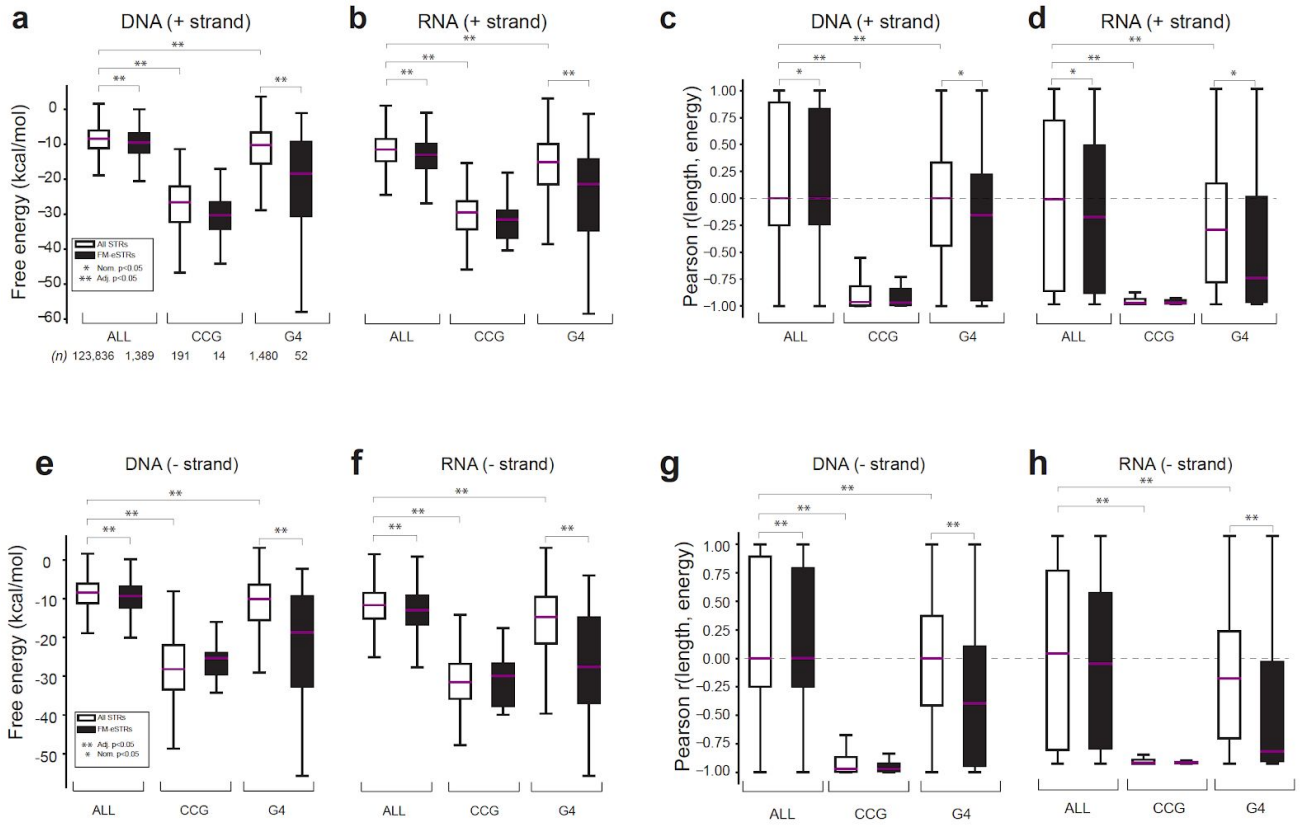
The y-axis shows the average number of ChIP-seq reads for RNA Polymerase II in 5bp bins centered at STRs within 3kb of any TSS. Black lines denote all STRs, blue lines denote CCG STRs, and red lines denote STRs matching the canonical G4 motif. Dashed lines represent all STRs of each class and solid lines represent FM-eSTRs. Plots show read counts in different cell and tissue types. From top to bottom: GM12878, human embryonic stem cells, heart, lung, and tibial nerve.

Supplementary Figure 8: Nucleosome occupancy around STRs.



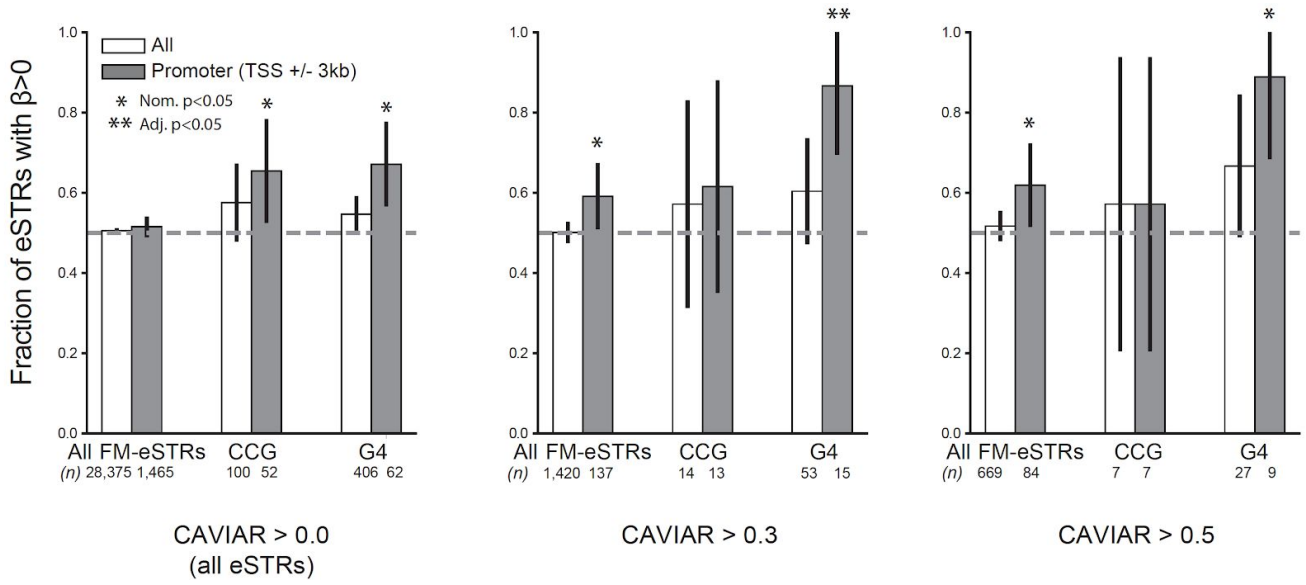
The y-axis denotes the average nucleosome occupancy in 5bp bins centered at STRs in GM12878. Black lines denote all STRs found within 3kb of any TSS, blue lines denote CCG STRs, and red lines denote STRs matching the canonical G4 motif. Dashed lines represent all STRs of each class and solid lines represent FM-eSTRs. Only STRs within 3kb of a TSS are included.

Supplementary Figure 9: GC-rich eSTRs are predicted to modulate DNA secondary structure.



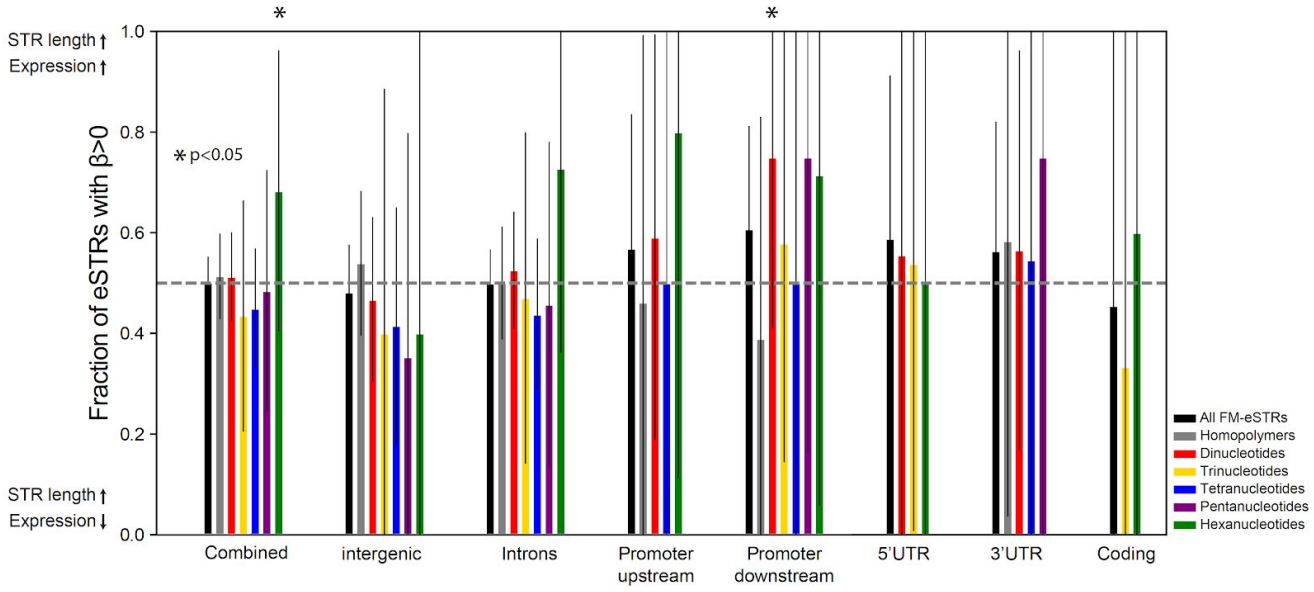
(**a-b, e-f) Free energy of STR regions.** Boxplots denote the distribution of free energy for each STR +/- 50bp of context sequence, computed as the average across all alleles at each STR. (**c-d, g-h) Pearson correlation between STR length and free energy.** Correlations were computed separately for each STR, and plots show the distribution of correlation coefficients across all STRs. The dashed horizontal line denotes 0 correlation as expected by chance. **(a)** and **(c)** show results computed using the template strand for DNA. **(b)** and **(d)** show results computed using the template strand for RNA. **(e)** and **(g)** show results computed using the non-template strand for DNA. **(f)** and **(h)** show results computed using the non-template strand for RNA. Nominally significant (Mann Whitney one-sided p<0.05) differences between distributions are denoted with a single asterisk. Differences significant after controlling for multiple hypotheses are denoted with double asterisks. For each category (free energy and Pearson correlation), we used a Bonferroni correction to control for 20 total comparisons: comparing all vs. FM-eSTRs separately in each category, comparing CCG vs. all STRs, and comparing G4 vs. all STRs, in four conditions (DNA +/- and RNA +/-). Box plots as in **Fig. 1c**. The numbers of FM-eSTRs in each distribution are shown beneath panel **a.** Numbers are identical for **b-h**.

Supplementary Figure 10: Bias in the direction of GC-rich eSTR effect sizes



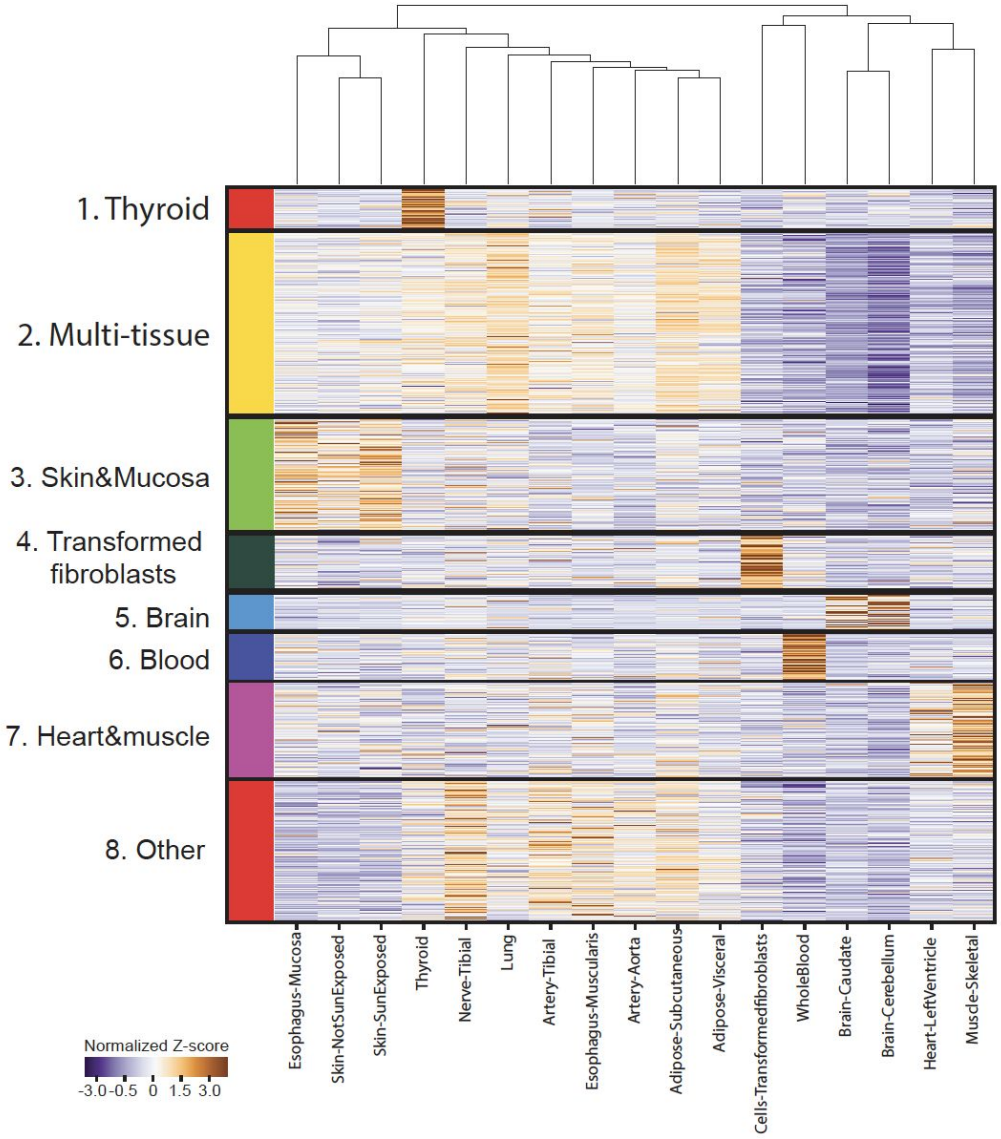
The y-axis shows the percentage of FM-eSTRs in each category with positive effect sizes, meaning a positive correlation between STR length and expression. White bars denote all STRs in each category. Gray bars denote STRs falling within 3kb of the TSS of the gene whose expression it is correlated with. Error bars give 95% confidence intervals. Results are shown for multiple thresholds for CAVIAR scores (>0, corresponding to all eSTRs, >0.3, as used in the main text, or >0.5).

Supplementary Figure 11: Bias in the direction of eSTR effect sizes



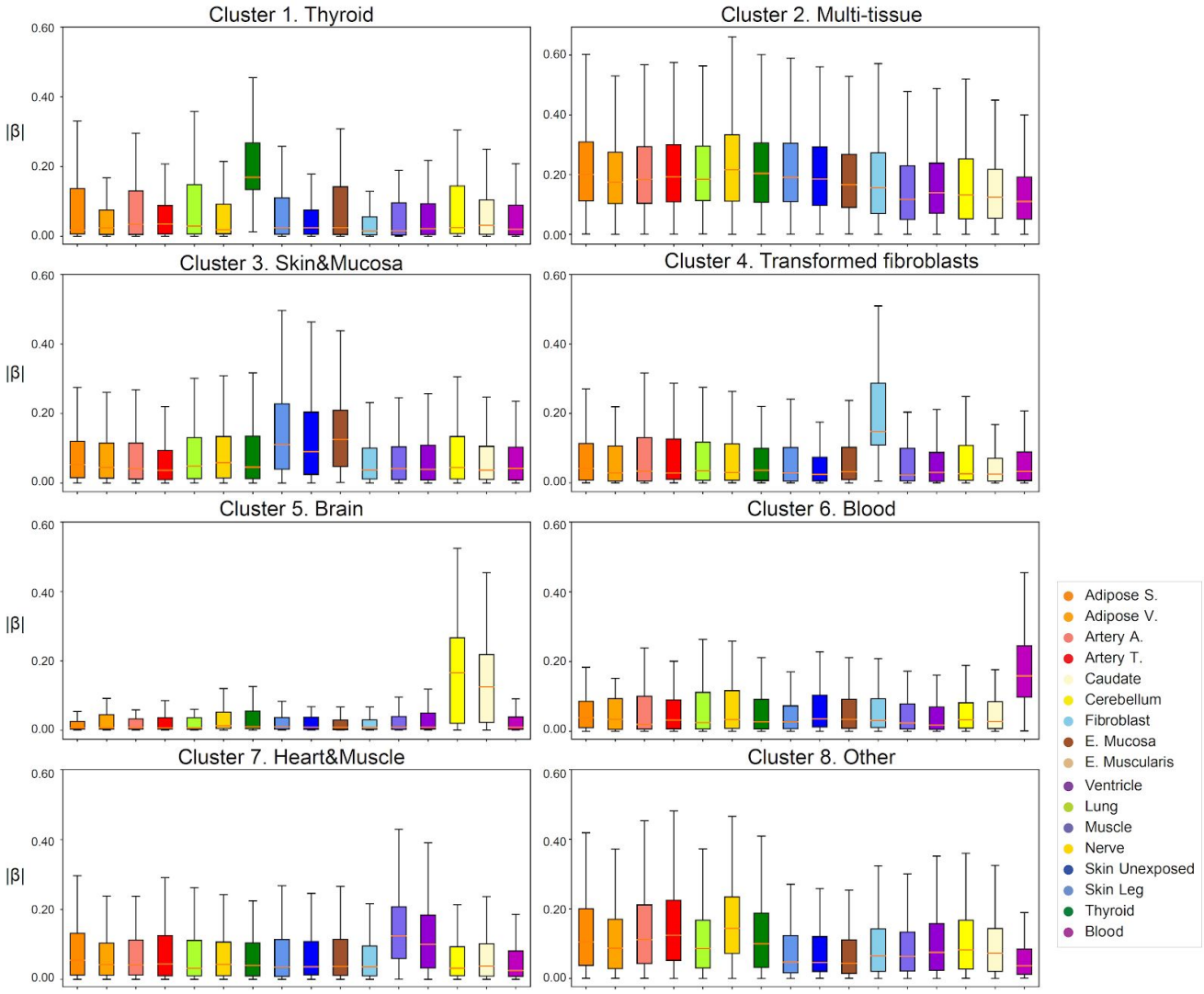
The y-axis shows the percentage of FM-eSTRs (n=1,420) in each category with positive effect sizes, meaning a positive correlation between STR length and expression. Colored bars represent different repeat unit lengths (black=all FM-eSTRs; gray=homopolymers; red=dinucleotides; gold=trinucleotides; blue=tetranucleotides; purple=pentanucleotides; green=hexanucleotides). Error bars show 95% confidence intervals. Asterisks denote categories that are nominally significant (binomial two-sided $p<0.05$) for having significantly more or less positive effect sizes than expected by chance (50%). No category was significant after accounting for multiple hypothesis testing. The number of FM-eSTRs in each category and breakdown by repeat unit length are shown in **Fig. 4a**.

Supplementary Figure 12: Characterization of tissue-specific FM-eSTRs



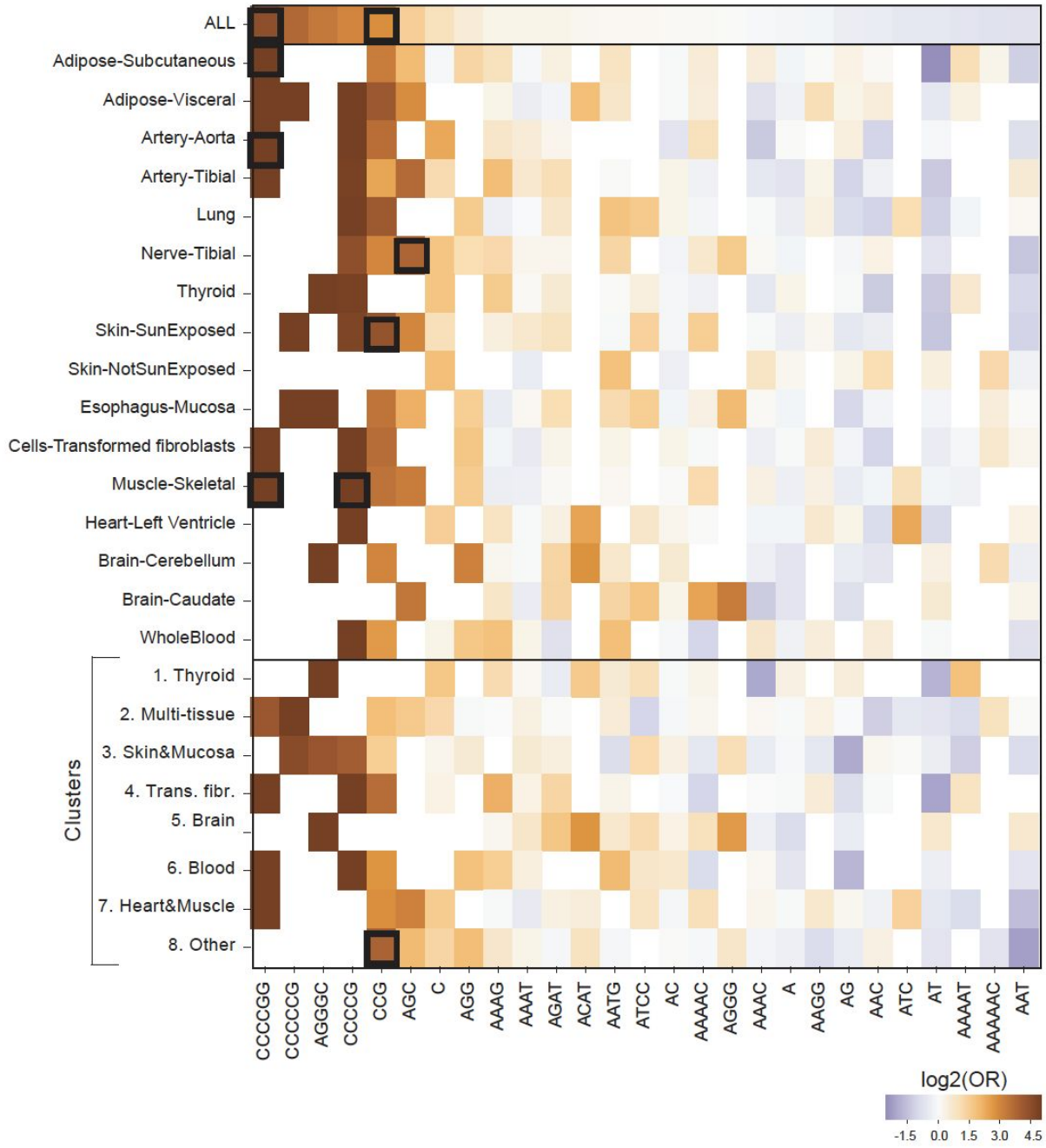
FM-eSTRs were clustered by absolute Z-scores computed by mash using K-means (**Methods**). The heatmap shows absolute values of Z-scores in each tissue, Z-normalized by row. (Number of genes in each cluster: Cluster 1=86, Cluster 2=360, Cluster 3=227; Cluster 4=106, Cluster 5=79, Cluster 6=98, Cluster 7=186, Cluster 8=278).

Supplementary Figure 13: Characterization of tissue-specific eSTR clusters



Each panel shows the distribution of the absolute value of posterior effect sizes computed by mash in each tissue for the set of FM-eSTRs in each cluster (see Supplementary **Fig. 12** above). 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. (Number of genes in each cluster: Cluster 1=86, Cluster 2=360, Cluster 3=227; Cluster 4=106, Cluster 5=79, Cluster 6=98, Cluster 7=186, Cluster 8=278).

Supplementary Figure 14: eSTR repeat unit enrichment



We evaluated repeat unit enrichment in multiple FM-eSTR groups: all FM-eSTRs combined across tissues (similar to **Fig. 2e**), FM-eSTRs identified per-tissue (see **Supplementary Table 1** for the total number of FM-eSTRs in each tissue), and FM-eSTRs belonging to each cluster (see **Supplementary Fig. 12** for the total number of FM-eSTRs in each cluster). For each group of FM-eSTRs, the heatmap shows the $\log_2$ of the odds ratio computed using a Fisher's Exact test (scipy.stats.fisher_exact). Columns are sorted from highest to lowest enrichment in all FM-eSTRs.

Bold boxes indicate enrichments statistically significant (adjusted $p<0.05$, adjusted separately per row for the number of repeat units tested).

Supplementary References

1. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. Nat Rev Genet 13, 770-780, doi:10.1038/nrg3296 (2012).
2. Sawaya, S. et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One 8, e54710, doi:10.1371/journal.pone.0054710 (2013).
3. Todd, A. K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res 33, 2901-2907, doi:10.1093/nar/gki553 (2005).
4. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406-3415, doi:10.1093/nar/gkg595 (2003).
5. Hansel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. Nat Genet 48, 1267-1272, doi:10.1038/ng.3662 (2016).