Abundant contribution of short tandem repeats to gene expression variation in humans

Melissa Gymrek[1,2,3,4], Thomas Willems[1,4,5], Haoyang Zeng[6], Barak Markus[1], Mark J. Daly[3,7], Alkes L. Price[3,8], Jonathan Pritchard[9,10], and Yaniv Erlich[1,4,11,*]

[1] Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts, USA

[2] Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts, USA.

[3] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

[4] New York Genome Center, New York, NY, USA.

[5] Computational and Systems Biology Program, MIT, Cambridge, MA 02139, USA

[6] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

[7] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA

[8] Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.

[9] Department of Genetics and Biology, Stanford University, Stanford, California, USA.

[10] Howard Hughes Medical Institute, Chevy Chase, Maryland, USA.

[11] Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA.

[*]To whom correspondence should be addressed (yaniv@cs.columbia.edu)

Abstract

Expression quantitative trait loci (eQTLs) are a key tool to dissect cellular processes mediating complex diseases. However, little is known about the role of repetitive elements as eQTLs. We report a genome-wide survey of the contribution of Short Tandem Repeats (STRs), one of the most polymorphic and abundant repeat classes, to gene expression in humans. Our survey identified 2,060 significant expression STRs (eSTRs). These eSTRs were replicable in orthogonal populations and expression assays. We used variance partitioning to disentangle the contribution of eSTRs from linked SNPs and indels and found that eSTRs contribute 10%-15% of the *cis*-heritability mediated by all common variants. Functional genomic analyses showed that eSTRs are enriched in conserved regions, co-localize with regulatory elements, and are predicted to modulate histone modifications. Our results show that eSTRs provide a novel set of regulatory variants and highlight the contribution of repeats to the genetic architecture of quantitative human traits.

Introduction

In recent years, there has been tremendous progress in identifying genetic variants that affect expression of nearby genes, termed *cis* expression quantitative trait loci (*cis*-eQTLs). Multiple studies have shown that disease-associated variants often overlap *cis*-eQTLs in the affected tissue[1,2]. These observations suggest that understanding the genetic architecture of the transcriptome may provide insights into the cellular-level mediators underlying complex traits[3-5]. So far, eQTL-mapping studies have mainly focused on SNPs and to a lesser extent on bi-allelic indels and CNVs as determinants of gene expression[6-10]. However, these variants do not account for all of the heritability of gene expression attributable to *cis*-regulatory elements as measured by twin studies, leaving on average about 20-30% unexplained[7,11]. It has been speculated that such heritability gaps could indicate the involvement of repetitive elements that are not well tagged by common SNPs[12,13].

To augment the repertoire of eQTL classes, we focused on Short Tandem Repeats (STRs), one of the most polymorphic and abundant type of repetitive elements in the human genome[14,15]. These loci consist of periodic DNA motifs of 2-6bp spanning a median length of around 25bp. There are about 700,000 STR loci covering almost 1% of the human genome. Their repetitive structure induces DNA-polymerase slippage events that add or delete repeat units, creating mutation rates that are orders of magnitude higher than those of most other variant types[14,16]. Over 40 Mendelian disorders, such as Huntington's Disease, are attributed to STR mutations, most of which are caused by large expansions of trinucleotide coding repeats[17]. However, trinucleotide coding STRs are only a minute fraction of all genomic STRs. The majority consist of di- and tetranucleotide motifs, which are overrepresented in promoter and regulatory regions[18].

Multiple lines of evidence support the potential role of STRs in regulating gene expression. *In vitro* studies have shown that STR variations can modulate the binding of transcription factors[19,20], change the distance between promoter elements[21,22], alter splicing efficiency[23,24], and induce irregular DNA structures that may modulate transcription[25]. Recent computational work showed that dinucleotide STRs are a hallmark of enhancer elements in *Drosophila*[26]. *In vivo* experiments have reported specific examples of STR variations that control gene expression across a wide range of taxa, including *Haemophilus influenza*[27], *Saccharomyces cerevisiae*[28], *Arabidopsis thaliana*[29], and vole[30]. In humans, several dozen candidate-gene studies used reporter assay experiments to show that STR variations modulate gene expression[19,31-35] and

alternative splicing[23,36,37]. However, there has been no systematic evaluation of the contribution of STRs to gene expression in humans.

To that end, we conducted a genome-wide analysis of STRs that affect expression of nearby genes, termed expression STRs (eSTRs), in lymphoblastoid cell lines (LCLs), a central *ex-vivo* model for eQTL studies. This well-studied model permitted the integration of whole genome sequencing data, expression profiles from RNA-sequencing and arrays, and functional genomics data. We tested for association in close to 190,000 STR×gene pairs and found over 2,000 significant eSTRs. Using a multitude of statistical genetic and functional genomics analyses, we show that hundreds of these eSTRs are predicted to be functional, uncovering a new class of genetic variants that modulate gene expression.

Results

Initial genome-wide discovery of eSTRs

The initial genome-wide discovery of potential eSTRs relied on finding associations between STR length and expression of nearby genes. We focused on 311 European individuals whose LCL expression profiles were measured using RNA-sequencing by the gEUVADIS[8] project and whose whole genomes were sequenced by the 1000 Genomes Project[38]. The STR genotypes were obtained in our previous study[39] in which we created a catalog of STR variation as part of the 1000 Genomes Project using lobSTR, a specialized algorithm for profiling STR variations from high throughput sequencing data[40]. Briefly, lobSTR identifies reads with repetitive sequences that are flanked by non-repetitive segments. It then aligns the non-repetitive regions to the genome using the STR motif to narrow the search, thereby overcoming the gapped alignment problem and conferring alignment specificity. Finally, lobSTR aggregates aligned reads and employs a model of STR-specific sequencing errors to report the maximum likelihood genotype at each locus. lobSTR recovered most (r^2=0.71) of the additive genetic variance of STR loci in the 1000 Genomes datasets based on large-scale validation using 5,000 STR genotype calls obtained by capillary electrophoresis, the gold standard for STR genotyping[39]. The majority of genotype errors were from dropout of one allele at heterozygote sites due to low sequencing coverage. We simulated the performance of STR associations using lobSTR calls compared to the capillary calls. This process showed that STR genotype errors reduce the power to detect eSTRs by 30-50% but importantly do not create spurious associations (**Supplementary Note** and **Supplementary Fig. 1**).

To detect eSTR associations, we regressed gene expression on STR dosage, defined as the sum of the two STR allele lengths in each individual. We opted to use this measure based on previous findings that reported a linear trend between STR length and gene expression[19,32,34] or disease phenotypes[41,42]. As covariates, we included sex, population structure, and other technical parameters (**Fig. 1a** and **Supplementary Methods**). We employed this process on 15,000 coding genes whose expression profiles were detected in the RNA-sequencing data. For each gene, we considered all polymorphic STR variations that passed our quality criteria (**Methods**) within 100kb of the transcription start and end sites of the gene transcripts as annotated by Ensembl[43]. On average, 13 STR loci were tested for each gene (**Supplementary Fig. 2**), yielding a total of 190,016 STR×gene tests.

Our analysis identified 2,060 unique protein-coding genes with a significant eSTR (gene level FDR≤5%) (**Fig. 1b**, **Supplementary Table 1**). The majority of these were di- and tetra-nucleotide STRs (**Supplementary Tables 2**, **3**). Only 13 eSTRs fall in coding exons but eSTRs were nonetheless strongly enriched in 5'UTRs ($p=1.0\times10^{-8}$), 3'UTRs ($p=1.7\times10^{-9}$) and regions near genes ($p<10^{-28}$) compared to all STRs analyzed (**Supplementary Table 4**). We repeated the association tests with two negative control conditions by regressing expression on (i) STR dosages permuted between samples and (ii) STR dosages from randomly chosen unlinked loci (**Fig. 1b, Supplementary Fig. 3**). Both negative controls produced uniform p-value distributions expected under the null hypothesis. This provides support for the absence of spurious associations due to inflation of the test statistic or the presence of uncorrected population structure.

The initial discovery set of eSTRs was largely reproducible in an independent set of individuals using an orthogonal expression assay technology. We obtained an additional set of over 200 individuals whose genomes were also sequenced as part of the 1000 Genomes Project and whose LCL expression profiles were measured by Illumina expression array[44]. These individuals belong to cohorts with African, Asian, European, and Mexican ancestry, enabling testing of the associations in a largely distinct set of populations. The Illumina expression array allowed testing 882 eSTRs out of the 2,060 identified above. The association signals of 734 of the 882 (83%) tested eSTRs showed the same direction of effect in both datasets (sign test $p=2.7\times10^{-94}$) and the effect sizes were strongly correlated ($R=0.73$, $p=1.4\times10^{-149}$) (**Fig. 1c**), despite only moderate reproducibility of expression profiles across platforms (**Supplementary Note** and **Supplementary Fig. 4**). Overall, these results show that eSTR association signals are robust and reproducible across populations and expression assay technologies.

Gymrek et al - Figure 1

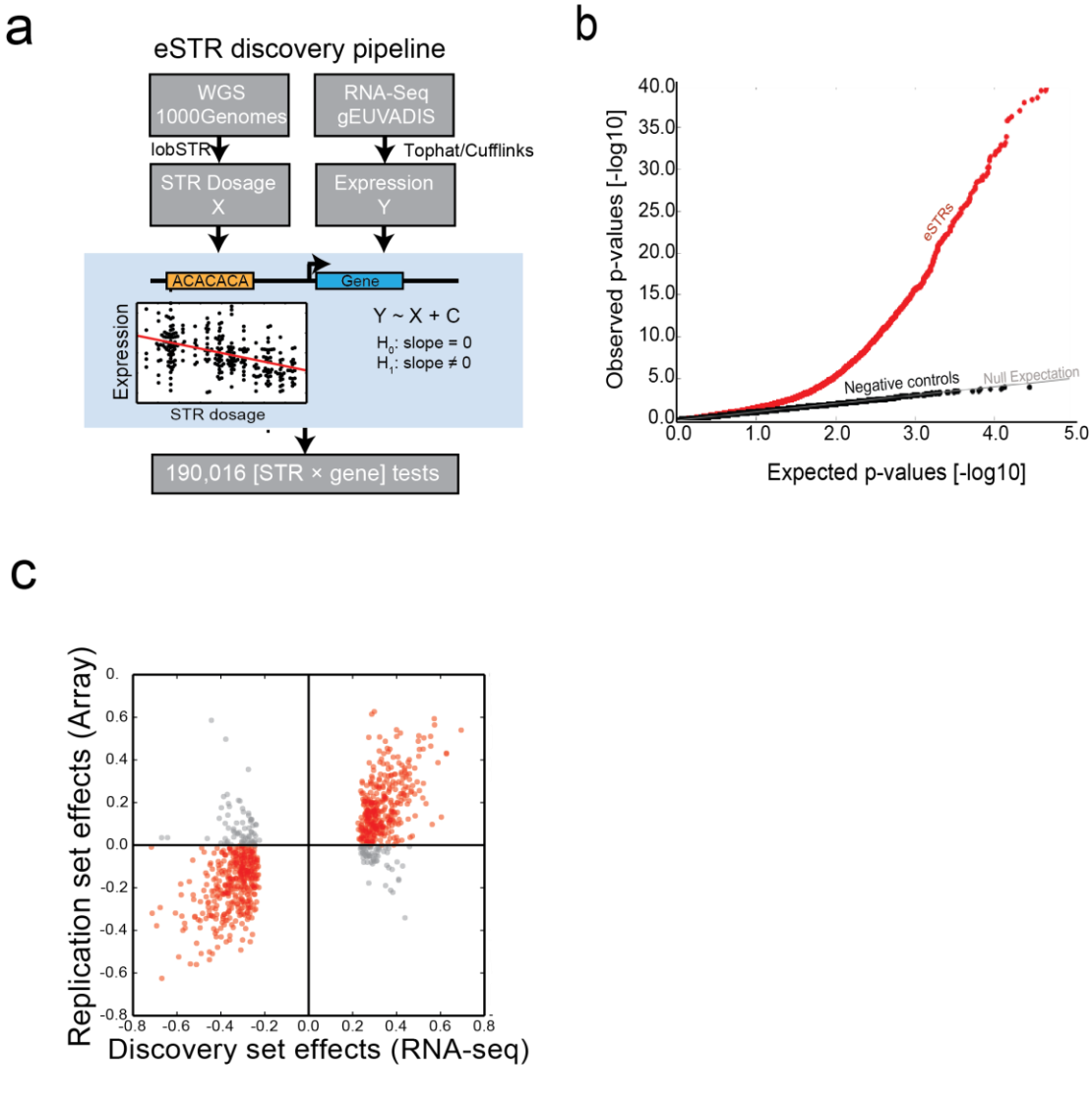


Figure 1: eSTR discovery and replication. (a) eSTR discovery pipeline. An association test using linear regression was performed between STR dosage and expression level for every STR within 100kb of a gene **(b)** Quantile-quantile plot showing results of association tests. The gray line gives the expected p-value distribution under the null hypothesis of no association. Black dots give p-values for permuted controls. Red dots give the results of the observed association tests **(c)** Comparison of eSTR effect sizes as Pearson correlations in the discovery dataset vs. the replication dataset. Red points denote eSTRs whose direction of effect was concordant in both datasets and gray points denote discordant directions.

Partitioning the contribution of eSTR and nearby variants

An important question is whether eSTR association signals stem from causal STR loci or are merely due to tagging SNPs or other variants in linkage disequilibrium (LD). Previous results reported that the average STR-SNP LD is approximately half of the traditional SNP-SNP LD[39,45], but there are known examples of STRs tagging GWAS SNPs[46].

To address this question, we partitioned the relative contributions of eSTRs versus all common (MAF≥1%) bi-allelic SNPs, indels, and structural variants (SV) in the *cis* region using a linear mixed model (LMM) (**Fig. 2a**). Multiple studies have used this approach to measure the total contributions of common variants to the heritability of quantitative traits and to partition the contribution of different classes of variants[47,48]. Taking a similar approach, we included two types of effects for each gene: a random effect (h^2_b) that captures all common bi-allelic loci detected within 100kb of the gene and a fixed effect (h^2_{STR}) that captures the best STR. To test whether other causal variants on the local region could inflate the estimate of the STR contribution, we simulated gene expression with a causal SNP eQTL per gene while preserving the local haplotype structure. In this negative control scenario, the LMM correctly reported a median $h^2_{STR}/h^2_{cis}\approx 0$ across all conditions (**Supplementary Note** and **Supplementary Fig. 5**), where $h^2_{cis}= h^2_b+h^2_{STR}$. This suggests that other causal variants in LD do not inflate the estimator of the relative contribution of STRs. As the LMM is expected to downwardly bias the variance explained in the presence of genotyping errors, the reported h^2_{STR} is likely to be conservative.

The LMM results showed that eSTRs contribute about 12% of the genetic variance attributed to common *cis* polymorphisms. For genes with a significant eSTR, the median h^2_{STR} was 1.80%, whereas the median h^2_b was 12.0% **(Fig. 2b)**, with a median ratio of h^2_{STR}/h^2_{CIS} = 12.3% ($CI_{95\%}$ 11.1%-14.2%; n=1,928) (**Table 1**). We repeated the same analysis for genes with at least moderate (≥5%) *cis*-heritability (**Methods**) regardless of the presence of a significant eSTR in the discovery set. The motivation for this analysis was to avoid potential winner's curse[49] and to obtain a transcriptome-wide perspective on the role of STRs in gene expression (**Fig. 2c**). In this set of genes, eSTRs contribute about 13% ($CI_{95\%}$ 12.2%-13.4%; n=6,272) of the genetic variance attributed to *cis* common polymorphisms. The median h^2_{STR} was 1.45% of the total expression variance, whereas the median h^2_b was 9.10% (**Table 1**). Repeating the analysis while treats STRs as a random effect showed highly similar results (**Supplementary Note**, **Supplementary Table 5, Supplementary Fig. 5-6,**). Taken together, this analysis shows that

STR variations explain a sizeable component of gene expression variation after controlling for all variants that are well tagged by common bi-allelic markers on in the *cis* region.

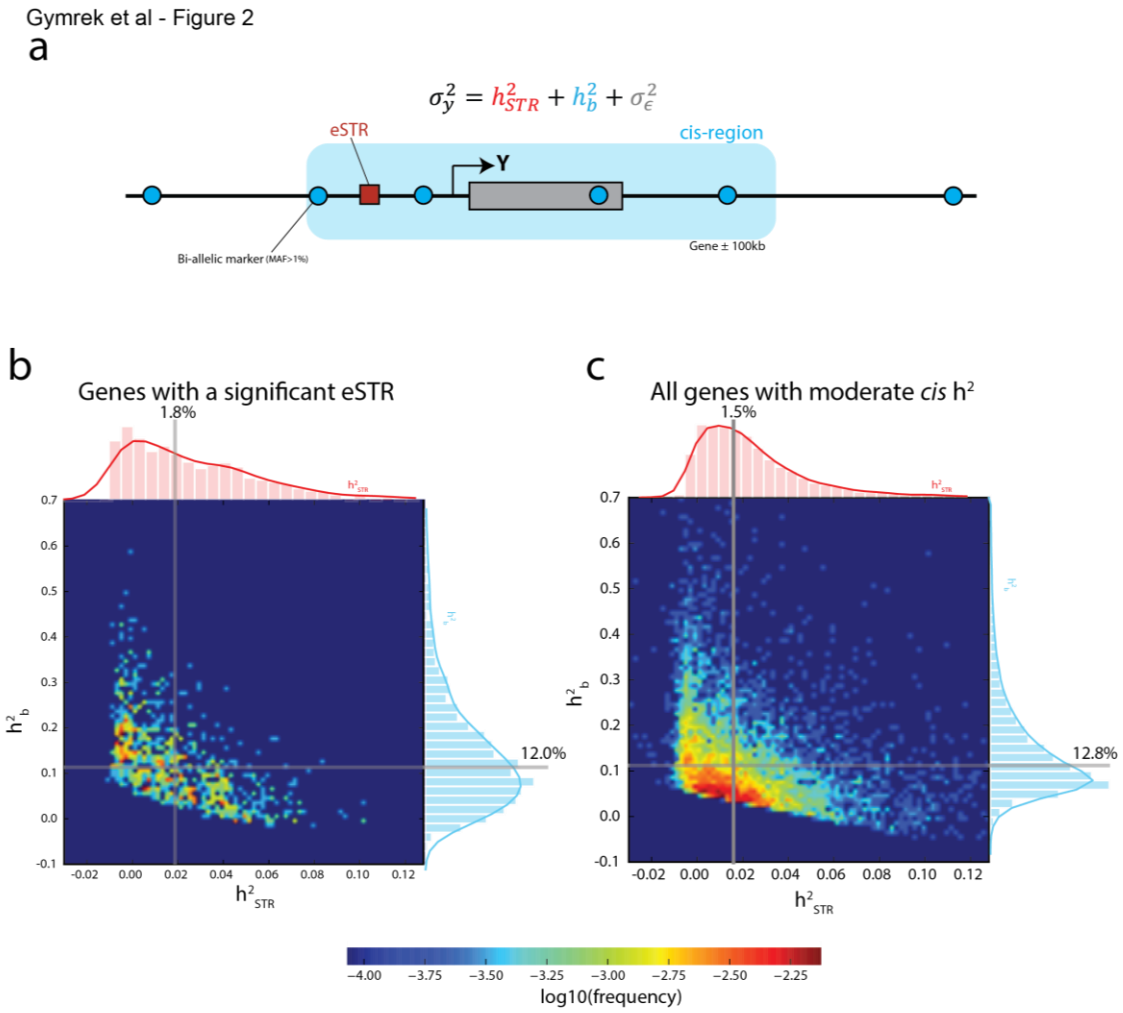


Figure 2: Variance partitioning using linear mixed models (a) The normalized variance of the expression of gene Y was modeled as the contribution of the best eSTR and common bi-allelic markers in the *cis* region (±100kb from the gene boundaries) **(b&c)** Heatmaps show the joint distributions of variance explained by eSTRs and by the *cis* region. Gray lines denote the median variance explained **(b)** Variance partitioning across genes with a significant eSTR in the discovery set and **(c)** variance partitioning across genes with moderate *cis* heritability.

The effect of eSTRs in the context of individual SNP eQTLs

To further assess the contribution of eSTRs in the context of other variants, we also inspected the relationship between eSTRs and individual cis-SNP eQTLs (eSNPs). We performed a traditional eQTL analysis with the whole genome sequencing data of 311 individuals that were part of the discovery set to identify common eSNPs [minor allele frequency (MAF) ≥5%] within 100kb of the gene. This process identified 4,290 genes with an eSNP (gene-level FDR≤5%). We then re-analyzed the eSTR association signals while conditioning on the genotype of the most significant eSNP (**Fig. 3a**). For each eSTR, we ascertained the subset of individuals that were homozygous for the major allele of the best eSNP in the region. If the eSTR simply tags this eSNP, its conditioned effect should be randomly distributed compared to the unconditioned effect. Alternatively, if the eSTR is causal, the direction of the conditioned effect should match the original effect. We conducted this analysis for eSTR loci with at least 25 individuals homozygous for the best eSNP and for which these individuals had at least two unique STR genotypes (1,856 loci). After conditioning on the best eSNP, the direction of effect for 1,395 loci (75%) was identical to that in the original analysis (sign test $p<4.2\times10^{-109}$) and the effect sizes were significantly correlated ($R=0.52$; $p=3.2\times10^{-130}$) (**Fig. 3b**). This further supports the additional role of eSTRs beyond traditional cis-eQTLs.

We also found that hundreds of eSTRs in the discovery set provide additional explanatory value for gene expression beyond the best eSNP. In 23% of genes, the eSTR significantly improved the explained variance of gene expression over considering only the best eSNP according to an ANOVA model comparison (FDR<5%) (**Fig. 3c-e, Methods**). Combined with the 183 genes with an eSTR but no significant eSNP, these results show that at least 30% of the eSTRs identified by our initial scan cannot be explained by mere tagging of the best eSNP. Given the reduced quality of STR compared to SNP genotypes, this analysis is likely to underestimate the true contribution of STRs. Nonetheless, our results show concrete examples for hundreds of associations in which the eSTR increases the variance explained by the best eSNP.

Gymrek et al - Figure 3

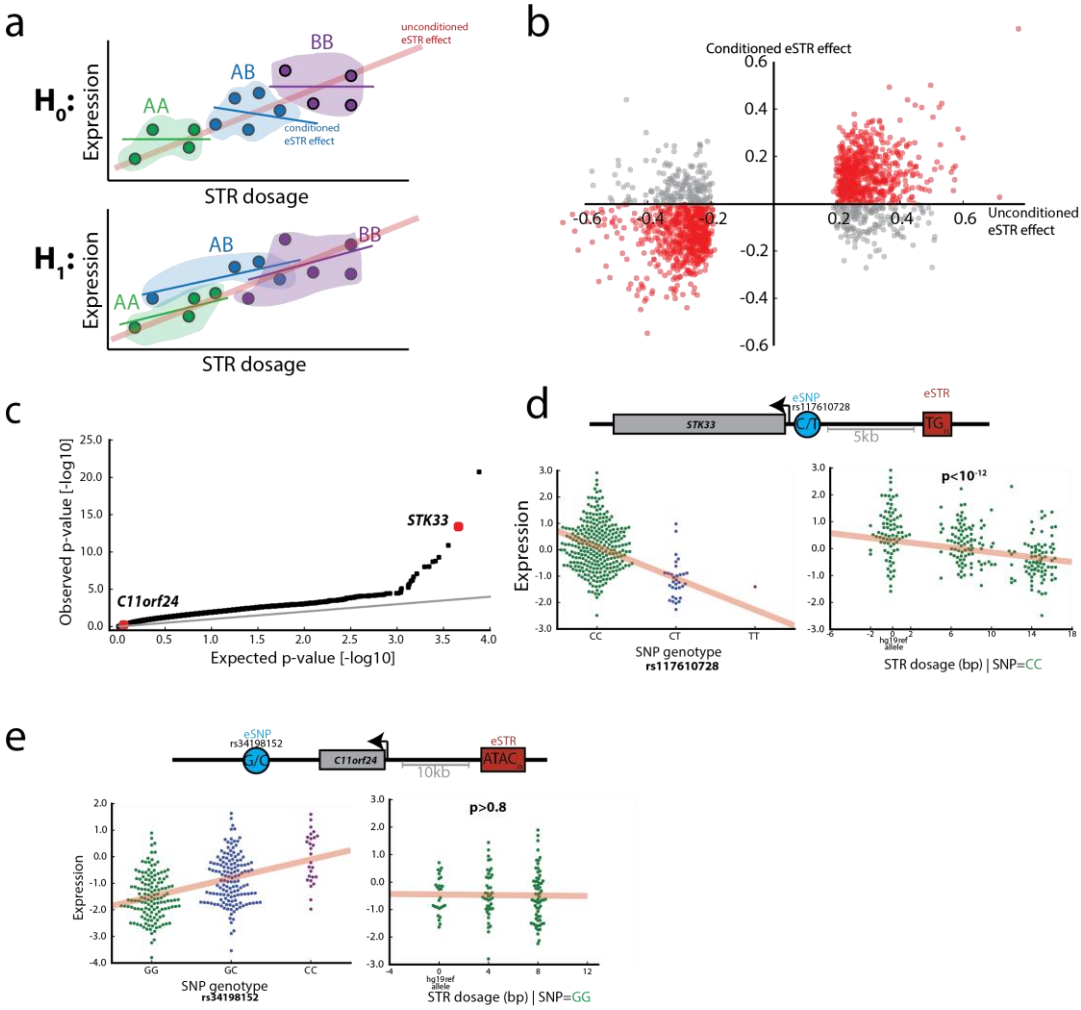


Figure 3: eSTR associations in the context of eSNPs (a) Schematic of the eSTR effect versus the effect conditioned on the best eSNP genotype. Under the null expectation, the original association (red line) comes from mere tagging of eSNPs. Thus, the eSTR effect disappears when restricting to a group of individuals (dots) with the same eSNP genotype (colored patches). Under the alternative hypothesis, the effect is concordant between the original and conditioned associations **(b)** The original eSTR effect versus the conditioned eSTR effect. Red points denote eSTRs whose direction of effect was concordant in both datasets and gray points denote discordant directions **(c)** Quantile-quantile plot of p-values from ANOVA testing of the explanatory value of eSTRs beyond that of eSNPs (**d**) *STK33* is an example of a gene for which the eSTR (red rectangle) has a strong explanatory value beyond the best eSNP (blue circle) based on ANVOA. Indeed, when conditioning on individuals that are homozygous for the "C" eSNP allele (bottom left, green dots), the STR dosage still shows a significant effect (bottom right) **(e)** *C11orf24* is an example of a gene for which the eSTR was part of the discovery set but did not pass the ANOVA threshold. After conditioning on individuals that are homozygous for the "G" eSNP allele (bottom left, green dots), the STR effect is lost (bottom right).

Functional Genomics Supports the Causal Role of eSTRs

To provide further evidence of their causal role, we analyzed eSTRs in the context of functional genomics data. First, we assessed the potential functionality of STR regions by measuring signatures of purifying selection, since previous findings have reported that putatively causal eSNPs are slightly enriched in conserved regions[50]. We inspected the sequence conservation[51] across 46 vertebrates in the sequence upstream and downstream of the eSTRs in our discovery dataset (**Fig. 4a**). To tune the null expectation, we matched each tested eSTR to a random STR that did not reach significance in the association analysis but had a similar distance to the nearest transcription start site (TSS). The average conservation level of a ±500bp window around eSTRs was slightly but significantly higher ($p<0.03$) than for control STRs. Tightening the window size to shorter stretches of ±50bp showed a more significant contrast in the conservation scores of the eSTRs versus the control STRs ($p<0.01$) (**Fig. 4a inset**), indicating that the excess in conservation comes from the vicinity of the eSTR loci. Taken together, these results show that eSTRs discovered by our association pipeline reside in regions exposed to relatively higher purifying selection, further suggesting a functional role.

We also found that eSTRs significantly co-localize with functional elements. eSTRs show the strongest enrichment closest to transcription start sites (**Fig. 4b**) and to a lesser extent near transcription end sites (**Supplementary Fig. 7**), similar to patterns previously observed for eSNPs[50]. We then inspected the co-localization of eSTRs with histone modifications as annotated by the Encode Consortium[6] in LCLs. eSTRs were strongly enriched in peaks of histone modifications associated with regulatory regions (H3K4me1, H3K4me2, H3K4me3, H3K27ac, H3K9ac) and transcribed regions (H3K36me3), and highly depleted in repressed regions (H3K27me3) (**Fig. 4b**). These results match previous patterns of enrichments found for putatively causal eSNPs[50]. To test the significance of these signals, we constructed a null distribution for each histone modification by measuring the co-localization of eSTRs with randomly shifted histone peaks similar to the fine-mapping procedure of Trynka *et al.*[52]. This null distribution controls for co-occurrence of eSTRs and histone peaks due to their proximity to other causal variants. We found eSTR/histone co-localizations were significant (weakest p-value<0.01) after the peak shifting procedure, suggesting that these results stem from the eSTRs themselves. We also performed a peak-shifting analysis using ChromHMM annotations[53] (**Fig. 4c**). The two strongest enrichments for eSTRs were weak-promoters ($p<0.002$) and weak-enhancers ($p<0.004$). Again, this analysis shows overlap of eSTRs with elements that are predicted to regulate gene expression.

Finally, we found that eSTR variations are likely to modulate the occupancy of certain histone marks. For each eSTR, we created a series of DNA sequences reflecting the STR alleles observed among individuals in our dataset (**Fig. 4d**). We used these sequences as an input to the WAVE (Whole-genome regulAtory Variants Evaluation) model[54], which predicts ChIP-sequencing experiments directly from genomic sequences (**Methods**). The output of WAVE showed the predicted effect of STR variations on the occupancy of chromatin marks. We then compared the distribution of the magnitude of effect sizes between eSTRs and a randomly chosen control set of STRs. eSTRs had significantly greater effects than control STRs on the predicted occupancy of all tested histone marks ($p_{H3K4me3}=3.4\times10^{-5}$, $p_{H3K9ac}=5.4\times10^{-8}$, $p_{H3K36me3}=2.1\times10^{-9}$, $p_{H3K27me3}=0.0047$; Mann-Whitney rank test) (**Fig. 4e)**. We also discovered a marginally significant effect on DNAseI (p=0.016) but not on P300 (p=0.94). Importantly, since the input material for this analysis is solely STR variations that are independent of any linked variants, these results provide an orthogonal piece of evidence of the functionality of eSTRs and suggest modulating histone marks as a potential mechanism.

Gymrek et al - Figure 4

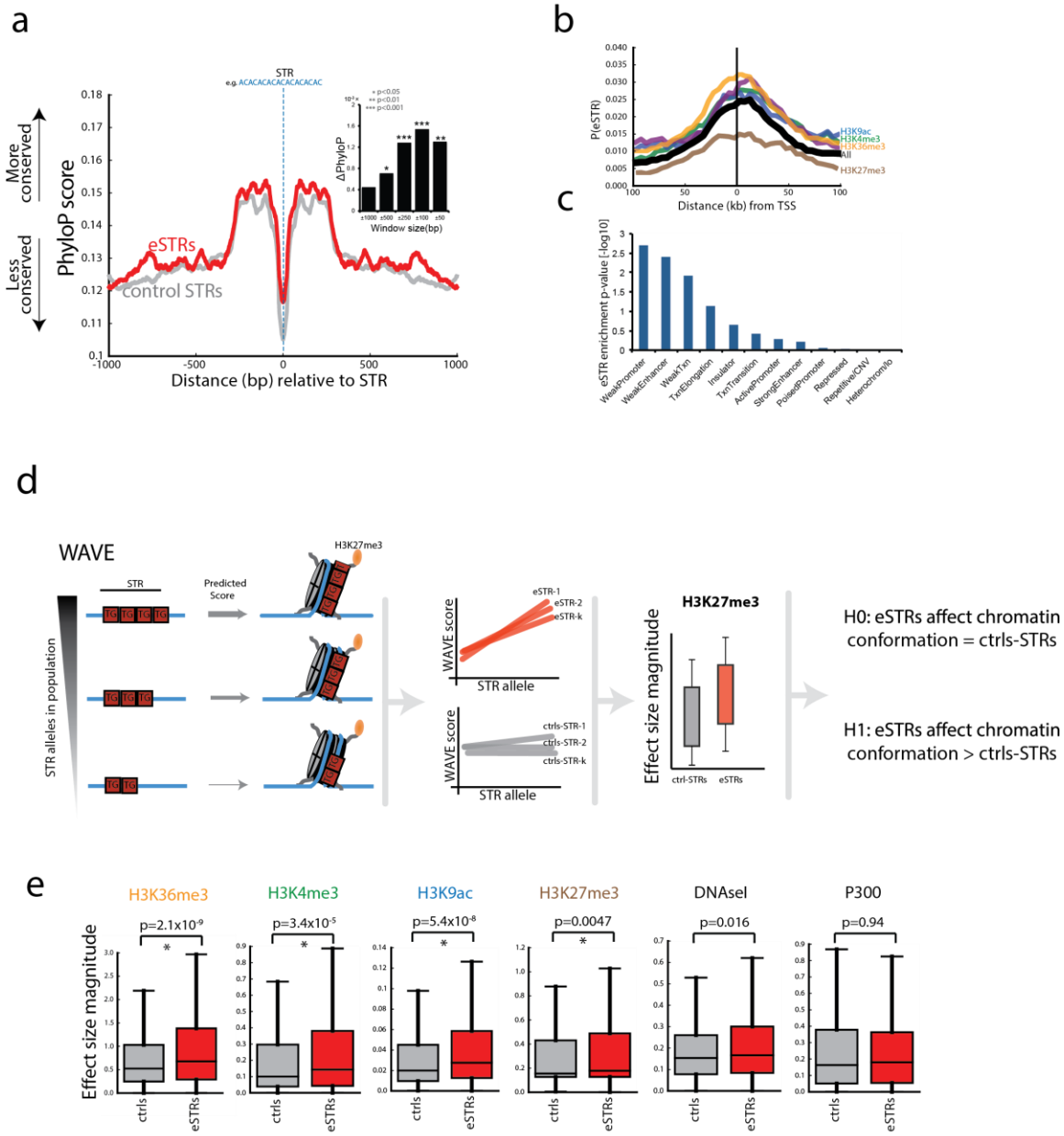


Figure 4: Functional analysis of eSTR loci (a) Median PhyloP conservation score as a function of distance from the STR. Red: eSTR loci, gray: matched control STRs. Inset: the difference in the PhyloP conservation score between eSTRs and matched control STRs as a function of window size around the STR. **(b)** The probability that an STR scores as an eSTR in the discovery set as a function of distance from the transcription start site (TSS). eSTRs show clustering around the TSS (black line). Conditioning on the presence of a histone mark (colored lines) significantly modulated the probability of an STR to be an eSTR **(c)** The enrichment of eSTRs in different chromatin states **(d)** Schematic of the application of WAVE to predict histone modification signatures for different STR alleles. For each eSTR (red) and control STR (gray) we measured the magnitude of the slope between the STR allele and the WAVE score and then tested whether the magnitudes were significantly different between the two sets **(e)** Comparison of the distribution of slope magnitudes for eSTRs (red) and controls (gray). A "*" denotes p-values < 0.01.

Discussion

Our study conducted the first genome-wide characterization of the effect of STR variation on gene expression and identified over 2,000 potential eSTRs. Further statistical analysis showed that eSTRs contribute on average about 10-15% of the *cis*-heritability of gene expression attributed to common (MAF≥1%) polymorphisms and that at least a third of these eSTRs improve the explained heritability beyond the strongest SNP-eQTLs. Functional genomics analyses provide further support for the predicted causal role of eSTRs.

We hypothesize that there are more eSTRs to find in the genome. Variance partitioning across all moderately heritable genes showed that STRs that did not reach significance still account for a sizeable component of gene expression variance. Our analysis also had several technical limitations. First, STRs show higher rates of genotype errors than SNPs, which limited our power to detect eSTRs and likely downwardly biased their estimated contribution in the LMM and ANOVA. In addition, about 10% of STR loci in the genome could not be analyzed because they are too long to be spanned by current sequencing read lengths[39]. Second, based on previous findings in humans[19,32,34], our association tests focused on a linear relationship between STR length and gene expression. However, experimental work in yeast reported that certain loci exhibit non-linear relationships between STR lengths and expression[28], which are unlikely to be captured in our current analysis. Finally, our association pipeline takes into account only the length polymorphisms of STRs and cannot distinguish the effect of sequence variations inside STR alleles with identical lengths (dubbed homoplastic alleles[55]). Addressing these technical complexities would probably require phased STR haplotypes and longer sequence reads that are currently beyond reach for large sample sizes. We envision that recent advancements in sequencing technologies[56] will further expand the catalog of eSTRs.

Previous reports have highlighted the overlap between eQTLs and GWAS hits of human diseases[7,8,11], suggesting a similar role for STRs. These loci, as well as other repetitive elements, have been largely overlooked by GWAS studies due to the technical complexities in genotyping them across a large number of samples. Analyzing the contribution of these loci in complex disease studies will require the availability of large scale whole genome sequencing data or the development of reliable imputation methods from genotyping arrays. Regardless of the technical method, our results suggest that such efforts might reveal exciting biology beyond that observed through the prism of traditional point mutations.

Methods

Genotype datasets

lobSTR genotypes were generated for the phase 1 individuals from the 1000 Genomes Project as described in[39]. Variants from the 1000 Genomes Project phase 1 release were downloaded in VCF format from the project website. HapMap genotypes were used to correct association tests for population structure. Genotypes for 1.3 million SNPs were downloaded for draft release 3 from the HapMap Consortium webpage. SNPs were converted to hg19 coordinates using the liftOver tool and filtered using Plink[57] to contain only the individuals for which both expression array data and STR calls were available. Throughout this manuscript, all coordinates and genomic data are referenced according to hg19.

Expression datasets

RNA-sequencing datasets from 311 HapMap lymphoblastoid cell lines for which STR and SNP genotypes were also available were obtained from the gEUVADIS Consortium. Raw FASTQ files containing paired end 100bp Illumina reads were downloaded from the EBI website. The hg19 Ensembl transcriptome annotation was downloaded as a GTF file from the UCSC Genome Browser[58,59] ensGene table. The RNA-sequencing reads were mapped to the Ensembl transcriptome using Tophat v2.0.7[60] with default parameters. Gene expression levels were quantified using Cufflinks v2.0.2[61] with default parameters and supplied with the GTF file for the Ensembl reference version 71. Genes with median FPKM of 0 were removed, leaving 23,803 genes. We restricted analysis to protein coding genes, giving 15,304 unique Ensembl genes. Expression values were quantile-normalized to a standard normal distribution for each gene.

The replication set consisted of Illumina Human-6 v2 Expression BeadChip data from 730 HapMap lymphoblastoid cell lines from the EBI website. These datasets contain two replicates each for 730 unrelated individuals from 8 HapMap populations (YRI, CEU, CHB, JPT, GIH, MEX, MKK, LWK) and were generated as described by Stranger *et al.*[62]. Background corrected and summarized probeset intensities (by Illumina software) contained values for 7,655 probes. Additionally, probes containing common SNPs were removed[63]. Only probes with a one-to-one correspondence with Ensembl gene identifiers were retained. We removed probes with low concordance across replicates (Spearman correlation ≤ 0.5). In total we obtained 5,388 probes for downstream analysis.

Each probe was quantile-normalized to a standard normal distribution across all individuals separately for each replicate and then averaged across replicates. These values were quantile-normalized to a standard normal distribution for each probe.

eQTL association testing

Expression values were adjusted for individual sex, individual membership, gene expression heterogeneity, and population structure (**Supplementary Methods**). Adjusted expression values were used as input to the eSTR analysis. To restrict to STR loci with high quality calls, we filtered the callset to contain only loci where at least 50 of the 311 samples had a genotype call. To avoid outlier genotypes that could skew the association analysis, we removed any genotypes seen less than three times. If only a single genotype was seen more than three times, the locus was discarded. To increase our power, we further restricted analysis to the most polymorphic loci with

heterozygosity of at least 0.3. This left 80,980 STRs within 100kb of a gene expressed in our LCL dataset.

A linear model was used to test for association between normalized STR dosage and expression for each STR within 100kb of a gene (**Supplementary Methods**). Dosage was defined as the sum of the deviations of the STR allele lengths from the hg19 reference. For example, if the hg19 reference for an STR is 20bp, and the two alleles called are 22bp and 16bp, then the dosage is equal to (22-20)+(16-20) = -4bp. Then, STR genotypes were zscore-normalized to have mean 0 and variance 1. For genes with multiple transcripts we defined the transcribed region as the maximal region spanned by the union of all transcripts. The linear model for each gene is given by:

$$\vec{y}_g = \alpha_g + \beta_{j,g}\,\vec{x}_j + \vec{\epsilon}_{j,g}$$

where $\vec{y}_g = (y_{g,1}, \ldots, y_{g,n})^T$ with $y_{g,i}$ the normalized covariate-corrected expression of gene g in individual i, n is the number of individuals, α_g is the mean expression level of homozygous reference individuals, $\beta_{j,g}$ is the effect of the allelic dosage of STR locus j on gene g, $\vec{x}_j = (x_{j,1}, \ldots, x_{j,n})^T$ with $x_{j,i}$ the normalized allelic dosage of STR locus j in the ith individual, and $\vec{\epsilon}_{j,g}$ is a random vector of length n whose entries are drawn from $N(0, \sigma^2_{\epsilon,j,g})$ where $\sigma^2_{\epsilon,j,g}$ is the unexplained variance after regressing locus j on gene g. The association was performed using the OLS function from the Python statsmodels package. For each comparison, we tested H_0: $\beta_{j,g} = 0$ vs. H_1: $\beta_{j,g} \neq 0$ using a standard t-test. We controlled for a gene-level false discovery rate (FDR) of 5% (**Supplementary Methods**).

Partitioning heritability using linear mixed models

For each gene, we used a linear mixed model to partition heritability between the best explanatory STR and other *cis* variants. We used a model of the form:

$$\vec{y}_g = \alpha_g + \beta_{j,g}\,\vec{x}_j + \vec{u}_g + \vec{\epsilon}_{j,g}$$

where:

- $\vec{y}_g$, α_g, $\beta_{j,g}$, $\vec{x}_j$, and $\vec{\epsilon}_{j,g}$ are as described above.
- $\vec{u}_g$ is a length n vector of random effects and $\vec{u}_g \sim MVN(0, \sigma^2_{u_g} K_g)$ with $\sigma^2_{u_g}$ the percent of phenotypic variance explained by *cis* variants for gene g.
- K_g is a standardized $n \times n$ identity by state (IBS) relatedness matrix constructed using all common bi-allelic variants (MAF≥1%) reported by phase 1 of the 1000 Genomes Project within 100kb of gene g. This includes SNPs, indels, and several biallelic structural variants and is constructed as $K_g = \frac{1}{p}\sum_{i=0}^{p} \frac{1}{var(\vec{x}_i)} (\vec{x}_i - 1_n mean(\vec{x}_i))(\vec{x}_i - 1_n mean(\vec{x}_i))^T$ where p is the total number of variants considered, $\vec{x}_i$ is a length n vector of genotypes for variant i, and 1_n is a length n vector of ones. Note the mean diagonal element of K_g is equal to 1.

We used the GCTA program[64] to determine the restricted maximum likelihood estimates (REML) of $\beta_{j,g}$ and $\sigma^2_{u_g}$. To get unbiased values of $\sigma^2_{u_g}$, the --reml-no-constrain option was used.

We used the resulting estimates to determine the variance explained by the STR and the *cis* region. We can write the overall phenotypic variance-covariance matrix as:

$$var(\vec{y}_g) = \beta^2_{j,g} var(\vec{x}_j) + \sigma^2_{u_g} K_g + \sigma^2_{\epsilon_{j,g}} I_n$$

where:

- $var(\vec{y}_g)$ is an $n \times n$ expression variance-covariance matrix with diagonal elements equal to 1, since expression values for each gene were normalized to have mean 0 and variance 1.
- I_n is the $n \times n$ identity matrix.

This equation shows the relationship:

$$\sigma^2_p = \sigma^2_{STR} + \sigma^2_{haplotype} + \sigma^2_\epsilon$$

where:

- σ^2_p is the phenotypic variance, which is equal to 1.
- h^2_{STR} is the variance explained by the STR. This is equal to $\beta^2_{j,g} var(\vec{x}_j) = \beta^2_{j,g}$ since the STR genotypes were scaled to have mean 0 and variance 1.
- h^2_b is the variance explained by bi-allelic variants in the *cis* region. This is approximately equal to $\sigma^2_{u_g}$ since the local IBS matrix K_g has a mean diagonal value of 1.

We estimated the percent of phenotypic variance explained by STRs, $\beta^2_{j,g}$, using the unbiased estimator $\hat{h}^2_{STR} = E[\beta^2_{j,g}] = \hat{\beta}^2_{j,g} - SE^2$, where $\hat{\beta}_{j,g}$ is the estimate of $\beta_{j,g}$ returned by GCTA, and SE is the standard error on the estimate, using the fact that $\hat{\beta}_{j,g} \sim N(\beta_{j,g}, SE)$. We estimated the percent of phenotypic variance explained by bi-allelic markers as $\hat{h}^2_b$. Note for this analysis the STR was treated as a fixed effect. We also reran the analysis treating the STR as a random effect, and found very little change in the results (**Supplementary Note**).

Results are reported for all eSTR-containing genes and for all genes with moderate total *cis* heritability, which we define as genes where $h^2_{STR} + h^2_b \geq 0.05$. We used this approach as to our knowledge there are no published results about the *cis*-heritability of expression of individual genes in LCLs from twin studies. We used 10,000 bootstrap samples of each distribution to generate 95% confidence intervals for the medians.

Comparing to the best eSNP

We identified SNP eQTLs using SNPs with MAF ≥ 1% as reported by phase 1 of the 1000 Genomes Project. We used an identical pipeline to our eSTR analysis to identify SNP eQTLs replacing the vector $\vec{x}_j$ with a vector of SNP genotypes (0, 1 or 2 reference alleles) that was z-normalized to have mean 0 and variance 1. To determine whether our eSTR signal was indeed

independent of the best SNP eQTL at each gene, we repeated association tests between STR dosages and expression levels while holding the genotype of the SNP with most significant association to that gene constant. For this, we determined all samples at each gene that were either homozygous reference or homozygous non-reference for the best SNP. For the SNP allele with more homozygous samples, we repeated the eSTR linear regression analysis and determined the sign and magnitude of the slope. We removed any genes for which there were less than 25 samples homozygous for the SNP genotype or for which there was no STR variation after holding the SNP constant, leaving 1,856 genes for analysis. We used a sign test to determine whether the direction of effects before and after conditioning on the best SNP are more concordant than expected by chance.

We used model comparison to determine whether eSTRs can explain additional variation in gene expression beyond that explained by the best eSNP for each gene. For each gene with a significant eSTR and eSNP, we analyzed the ability of two models to explain gene expression:

$$\text{Model 1 (eSNP-only): } \vec{y}_g = \alpha_g + \beta_{eNP,g}\,\vec{x}_{eSNP,g} + \vec{\epsilon}_{j,g}$$

$$\text{Model 2 (joint eSNP+eSTR): } \vec{y}_g = \alpha_g + \beta_{eNP,g}\,\vec{x}_{eSNP,g} + \beta_{eSTR,g}\,\vec{x}_{eSTR,g} + \vec{\epsilon}_{j,g}$$

where α_g is the mean expression value for the reference haplotype, $\vec{y}_g$ is a vector of expression values for gene g, $\beta_{eSNP,g}$ is the effect of the eSNP on gene g, $\beta_{eSTR,g}$ is the effect of the eSTR on gene g, $\vec{x}_{eSNP,g}$ is a vector of genotypes for the best eSNP for gene g, $\vec{x}_{eSTR,g}$ is a vector of genotypes for the best eSTR for gene g, and $\vec{\epsilon}_{j,g}$ gives the residual term. A major caveat is that the eSNP dataset has significantly more power to detect associations than the eSTR dataset due to the lower quality of the STR genotype panel (**Supplementary Note**), and this analysis is therefore likely to underestimate the true contribution of STRs to gene expression. We used ANOVA to test whether the joint model performs significantly better than the SNP-only method. We obtained the ANOVA p-value for each gene and used the qvalue package to determine the FDR.

Conservation analysis

Sequence conservation around STRs was determined using the PhyloP track available from the UCSC Genome Browser. To calculate the significance of the increase in conservation at eSTRs, we compared the mean PhyloP score for each eSTR to that for 1000 random sets of STRs with matched distributions of the distance to the nearest transcription start site. For each STR we determined the mean PhyloP score for a given window size centered on the STR. The p-value given is the percentage of random sets whose mean PhyloP score was greater than the mean of the observed eSTR set.

Enrichment in histone modification peaks

Chromatin state and histone modification peak annotations generated by the Encode Consortium for GM12878 were downloaded from the UCSC Genome Browser. Because variants involved in regulating gene expression are more likely to fall near genes compared to randomly chosen variants, simple enrichment tests of eSTRs vs. randomly chosen control regions may return strong enrichments simply because of their proximity to genes. To account for this, we randomly shifted the location of eSTRs by a distance drawn from the distribution of distances between the best STR and best SNP for each gene. We repeated this process 1,000 times. For each set of permuted eSTR locations, we generated null distributions by determining the percent of STRs overlapping

each annotation. We used these null distributions to calculate empirical p-values for the enrichment of eSTRs in each annotation.

Predicting effects of STR variation on histone modifications

The WAVE method builds on a kmer-based statistical model to predict the signal of ChIP-seq experiments from a DNA sequence context. Briefly, the model considers that each k-mer has a spatial effect on ChIP-seq read counts in a window of [-M, M-1] bp centered at the start of the k-mer. The read count at a given base is then modeled as the log-linear combination of the effects of all k-mers whose effect ranges cover that base, where k ranges from 1 to 8.

For each eSTR in our dataset, we generated sequences representing each observed allele. We filtered STRs with interruptions in the repeat motif, since the sequence for different allele lengths is ambiguous for these loci. For each mark, we used the model to predict the read count for each allele in a window of ±M bp from the STR boundaries, where M was set to 1,000 for all marks except p300, for which M was set to 200. Previous findings of WAVE showed that these values of M give the best correlation between predicted and real ChIP-seq signals using cross validation. For each alternate allele, we generated a score as the sum of differences in read counts from the reference allele at each position in this window. We regressed the number of repeats for each allele on this score and took the absolute value of the slope for each locus. We repeated the analysis on a set of 2,060 randomly chosen negative control loci and used a Mann-Whitney rank test to compare the magnitudes of slopes between the eSTR and control sets for each mark.

Acknowledgements

M.G. is supported by the National Defense Science & Engineering Graduate Fellowship. Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by a gift from Andria and Paul Heafy (Y.E), NIJ grant 2014-DN-BX-K089 (Y.E, T.W), and NIH grants 1U01HG007037 (H.Z), R01MH084703(J.P) and R01HG006399 (A.L.P). We thank Tuuli Lappalainen, Alon Goren, Tatsu Hashimoto, and Dina Zielinksi for useful comments and discussions.

Figure Legends

Figure 1: eSTR discovery and replication. (a) eSTR discovery pipeline. An association test using linear regression was performed between STR dosage and expression level for every STR within 100kb of a gene **(b)** Quantile-quantile plot showing results of association tests. The gray line gives the expected p-value distribution under the null hypothesis of no association. Black dots give p-values for permuted controls. Red dots give the results of the observed association tests **(c)** Comparison of eSTR effect sizes as Pearson correlations in the discovery dataset vs. the replication dataset. Red points denote eSTRs whose direction of effect was concordant in both datasets and gray points denote discordant directions.

Figure 2: Variance partitioning using linear mixed models (a) The normalized variance of the expression of gene Y was modeled as the contribution of the best eSTR and common bi-allelic markers in the *cis* region (±100kb from the gene boundaries) (**b**&**c**) Heatmaps show the joint distributions of variance explained by eSTRs and by the *cis* region. Gray lines denote the median variance explained **(b)** Variance partitioning across genes with a significant eSTR in the discovery set and **(c)** variance partitioning across genes with moderate *cis* heritability.

Figure 3: eSTR associations in the context of eSNPs (a) Schematic of the eSTR effect versus the effect conditioned on the best eSNP genotype. Under the null expectation, the original association (red line) comes from mere tagging of eSNPs. Thus, the eSTR effect disappears when restricting to a group of individuals (dots) with the same eSNP genotype (colored patches). Under the alternative hypothesis, the effect is concordant between the original and conditioned associations **(b)** The original eSTR effect versus the conditioned eSTR effect. Red points denote eSTRs whose direction of effect was concordant in both datasets and gray points denote discordant directions **(c)** Quantile-quantile plot of p-values from ANOVA testing of the explanatory value of eSTRs beyond that of eSNPs (**d**) *STK33* is an example of a gene for which the eSTR (red rectangle) has a strong explanatory value beyond the best eSNP (blue circle) based on ANVOA. Indeed, when conditioning on individuals that are homozygous for the "C" eSNP allele (bottom left, green dots), the STR dosage still shows a significant effect (bottom right) **(e)** *C11orf24* is an example of a gene for which the eSTR was part of the discovery set but did not pass the ANOVA threshold. After conditioning on individuals that are homozygous for the "G" eSNP allele (bottom left, green dots), the STR effect is lost (bottom right).

Figure 4: Functional analysis of eSTR loci (a) Median PhyloP conservation score as a function of distance from the STR. Red: eSTR loci, gray: matched control STRs. Inset: the difference in the PhyloP conservation score between eSTRs and matched control STRs as a function of window size around the STR. **(b)** The probability that an STR scores as an eSTR in the discovery set as a function of distance from the transcription start site (TSS). eSTRs show clustering around the TSS (black line). Conditioning on the presence of a histone mark (colored lines) significantly modulated the probability of an STR to be an eSTR **(c)** The enrichment of eSTRs in different chromatin states **(d)** Schematic of the application of WAVE to predict histone modification signatures for different STR alleles. For each eSTR (red) and control STR (gray) we measured the magnitude of the slope between the STR allele and the WAVE score and then tested whether the magnitudes were significantly different between the two sets **(e)** Comparison of the distribution of slope magnitudes for eSTRs (red) and controls (gray). A "*" denotes p-values < 0.01.

Tables

	h^2_b	h^2_{STR}	h^2_{STR}/h^2_{cis}
eSTR genes (n=1,928)	0.1203 (0.1139-0.1259)	0.0180 (0.0166-0.0199)	0.1230 (0.1106-0.1420)
Moderate *cis* h^2 (n=6,272)	0.0910 (0.0884-0.0938)	0.0145 (0.0137-0.0151)	0.1283 (0.1222-0.1346)

Table 1: Heritability of gene expression explained by STRs vs. common bi-allelic variants. Values show the median and 95% confidence interval of the median across all eSTR-containing genes and genes with moderate *cis* heritability (≥5%). h^2_b denotes the variance explained by all common *cis* bi-allelic variants, h^2_{STR} denotes the variance explained by the best STR for each gene, and $h^2_{cis} = h^2_{STR} + h^2_b$.

References and Notes

1. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).
2. Moffatt, M.F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-3 (2007).
3. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895 (2010).
4. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888 (2010).
5. Ward, L.D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095-106 (2012).
6. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
7. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**, 1084-9 (2012).
8. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
9. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
10. Montgomery, S.B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23**, 749-61 (2013).
11. Wright, F.A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46**, 430-7 (2014).
12. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
13. Press, M.O., Carlson, K.D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet* **30**, 504-12 (2014).
14. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nature reviews. Genetics* **5**, 435-45 (2004).
15. Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics* **44**, 445-77 (2010).
16. Weber, J.L. & Wong, C. Mutation of human short tandem repeats. *Hum Mol Genet* **2**, 1123-8 (1993).
17. Mirkin, S.M. Expandable DNA repeats and human disease. *Nature* **447**, 932-40 (2007).
18. Sawaya, S. *et al.* Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One* **8**, e54710 (2013).
19. Contente, A., Dittmer, A., Koch, M.C., Roth, J. & Dobbelstein, M. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nat Genet* **30**, 315-20 (2002).
20. Martin, P., Makepeace, K., Hill, S.A., Hood, D.W. & Moxon, E.R. Microsatellite instability regulates transcription factor binding and gene expression. *Proc Natl Acad Sci U S A* **102**, 3800-4 (2005).

21. Willems, R., Paul, A., van der Heide, H.G., ter Avest, A.R. & Mooi, F.R. Fimbrial phase variation in Bordetella pertussis: a novel mechanism for transcriptional regulation. *EMBO J* **9**, 2803-9 (1990).
22. Yogev, D., Rosengarten, R., Watson-McKown, R. & Wise, K.S. Molecular basis of Mycoplasma surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences. *EMBO J* **10**, 4069-79 (1991).
23. Hefferon, T.W., Groman, J.D., Yurk, C.E. & Cutting, G.R. 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**, 3504-9 (2004).
24. Hui, J. *et al.* Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing. *EMBO J* **24**, 1988-98 (2005).
25. Rothenburg, S., Koch-Nolte, F., Rich, A. & Haag, F. A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc Natl Acad Sci U S A* **98**, 8985-90 (2001).
26. Yanez-Cuna, J.O. *et al.* Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res* (2014).
27. Weiser, J.N., Love, J.M. & Moxon, E.R. The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. *Cell* **59**, 657-65 (1989).
28. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K.J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213-6 (2009).
29. Sureshkumar, S. *et al.* A genetic defect caused by a triplet repeat expansion in Arabidopsis thaliana. *Science* **323**, 1060-3 (2009).
30. Hammock, E.A. & Young, L.J. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**, 1630-4 (2005).
31. Borel, C. *et al.* Tandem repeat sequence variation as causative cis-eQTLs for protein-coding gene expression variation: the case of CSTB. *Hum Mutat* **33**, 1302-9 (2012).
32. Gebhardt, F., Zanker, K.S. & Brandt, B. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* **274**, 13176-80 (1999).
33. Rockman, M.V. & Wray, G.A. Abundant raw material for cis-regulatory evolution in humans. *Molecular biology and evolution* **19**, 1991-2004 (2002).
34. Shimajiri, S. *et al.* Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett* **455**, 70-4 (1999).
35. Warpeha, K.M. *et al.* Genotyping and functional analysis of a polymorphic (CCTTT)(n) repeat of NOS2A in diabetic retinopathy. *FASEB J* **13**, 1825-32 (1999).
36. Hui, J., Stangl, K., Lane, W.S. & Bindereif, A. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat Struct Biol* **10**, 33-7 (2003).
37. Sathasivam, K. *et al.* Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proc Natl Acad Sci U S A* **110**, 2366-70 (2013).
38. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
39. Willems, T. *et al.* The landscape of human STR variation. *Genome Res* (2014).
40. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**, 1154-62 (2012).
41. Duyao, M. *et al.* Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat Genet* **4**, 387-92 (1993).

42. La Spada, A.R. *et al.* Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy. *Nat Genet* **2**, 301-4 (1992).
43. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res* **41**, D48-55 (2013).
44. Stranger, B.E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* **8**, e1002639 (2012).
45. Payseur, B.A., Place, M. & Weber, J.L. Linkage disequilibrium between STRPs and SNPs across the human genome. *Am J Hum Genet* **82**, 1039-50 (2008).
46. Lamina, C. *et al.* A systematic evaluation of short tandem repeats in lipid candidate genes: riding on the SNP-wave. *PLoS One* **9**, e102113 (2014).
47. Gusev, A. *et al.* Regulatory variants explain much more heritability than coding variants across 11 common diseases. *bioRxiv* (2014).
48. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
49. Ioannidis, J.P. Why most discovered true associations are inflated. *Epidemiology* **19**, 640-648 (2008).
50. Gaffney, D.J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol* **13**, R7 (2012).
51. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-21 (2010).
52. Trynka, G. *et al.* Disentangling effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex trait loci. *bioRxiv* (2014).
53. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
54. Zeng, H., Hashimoto, T., Kang, D.D. & Gifford, D.K. Whole Genome Regulatory Variant Evaluation for Transcription Factor Binding. in *bioxRiv* (2015).
55. Weber, J.L. & Broman, K.W. 7 Genotyping for human whole-genome scans: Past, present, and future. *Advances in genetics* **42**, 77-96 (2001).
56. Chaisson, M.J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-11 (2015).
57. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
58. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764-70 (2014).
59. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
60. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11 (2009).
61. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-78 (2012).
62. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
63. Barbosa-Morais, N.L. *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. *Nucleic Acids Res* **38**, e17 (2010).
64. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

Abundant contribution of Short Tandem Repeats to gene expression variation in humans
Supplementary Material

Melissa Gymrek, Thomas Willems, Haoyang Zeng, Barak Markus,
Mark J. Daly, Alkes L. Price, Jonathan Pritchard, and Yaniv Erlich*

April 1, 2015

* To whom correspondence should be addressed

Contents

1 Supplementary Methods

1.1 Controlling for covariates

We controlled for a number of covariates by regressing them out of the expression dataset. The covariate-corrected expression matrix is given by:

$$Y = (1 - H)Y' \tag{1}$$

where Y' is an $n \times m$ matrix of normalized expression values, Y is an $n \times m$ matrix of residualized expression values, n is the number of individuals, m is the number of genes, $H = C(C^T C)^{-1} C^T$ is the hat matrix, and C is an $n \times c$ matrix of c covariates. Specifically, the columns of C consist of the following sub-matrices:

$$C = \left[\begin{array}{c|c|c|c} \vec{c_s} & C_p & C_{exp} & C_{popstruct} \end{array} \right] \tag{2}$$

1. **Individual sex**: this is a binary vector, $\vec{c_s} \in \{0,1\}^{n \times 1}$, where 0 denotes female and 1 male.
2. **Individual population membership**: this is a binary matrix $C_p \in \{0,1\}^{n \times pop-1}$. A "1" in position $C_p(i,j)$ denotes that individual i belongs to population j. Specifically, pop is equal to to 4 for the association tests with the gEUVADIS RNA-seq data.
3. **Gene expression heterogeneity**: Y' is a matrix that consists of all $\vec{y_g}$ as its column vectors, where $\vec{y_g}$ is a vector of expression values for gene g. To reduce variation due to experimental differences or other unidentified confounding factors across expression datasets, the top 10 principal components (PCs) corresponding to the top 10 eigenvectors of $Y'Y'^T$ were included as covariates for both the array and RNA-sequencing datasets. $C_{exp} \in \mathbb{R}^{n \times 10}$ indicates the matrix of the top 10 PCs.
4. **Population structure**: We first preprocessed the HapMap SNP dataset to include SNPs with MAF $> 10\%$. We used Plink [1] for LD-pruning with a pairwise correlation threshold of 0.5, a window size of 50 SNPs, and a step size of 5 SNPs. This left 286,010 SNPs for the RNA-sequencing dataset, which we used to correct for population structure. We used the Tracy-Widom test for population stratification proposed by Patterson, et al. [2] to determine the number of PCs to include as covariates. Let $C_{popstruct} \in \mathbb{R}^{n \times t}$ indicate the matrix of the top t PCs removed, where t=5 for the RNA-sequencing dataset.

Residualized expression values were then used as input to the eQTL analysis.

1.2 Controlling for gene-level FDR

We controlled for a gene-level false discovery rate (FDR) of 5%, assuming that most genes have at most a single causal eSTR. For each gene, we determined the STR association with the best p-value. This p-value was adjusted using a Bonferonni correction for the number of STRs tested per gene to give a p-value for observing a single eSTR association for each gene. Performing separate permutations for each gene was computationally infeasible, and was found to give similar results to a simple Bonferfonni correction on a subset of genes. We then used this list of adjusted p-values as input to the `qvalue` package [3] to determine all genes with qval $\leq 5\%$.

2 Supplementary Notes

2.1 STR genotype error reduces power to detect eSTRs

We performed simulations to evaluate the effect of lobSTR genotype errors on our power to detect eSTR associations.

We used capillary electrophoresis calls from the Marshfield panel [4] as ground truth genotypes and lobSTR calls for the same markers in our catalog as observed genotypes. We filtered for loci with at least 25 calls for comparison. For each gene, we simulated expression values assuming a single causal STR per gene that explains h^2_{STR} percent of expression variance. We performed the analysis for h^2_{STR} equal to 0.01, 0.05, 0.1, 0.3, and 0.5. Expression values were simulated as follows:

$$Y_i = \beta X_i + \epsilon_i \tag{3}$$

where Y_i is the expression level for individual i, X_i is the true STR dosage for individual i, $\beta = \sqrt{h^2_{STR}}$ is the effect size of the STR, and $\epsilon_i \sim N(0, 1 - h^2_{STR})$ is the residual term for individual i.

We performed association analysis regressing $\vec{Y}$ on both $\vec{X}$ and $\vec{X}'$, where $\vec{X}'$ are the observed STR dosages, and tested whether β was significantly different than 0 in each case ($p<0.01$). We found that genotype errors limit our power to detect eSTRs (**Supplementary Fig. 1a**) and cause us to underestimate the true variance explained by STRs (**Supplementary Fig. 1b**) but do not introduce spurious eSTR signals.

2.2 Comparing expression across array and RNA-sequencing datasets

To determine the reproducibility of expression profiling across platforms, we compared gene expression for the 122 individuals profiled by both array and RNA-sequencing. For each platform, we obtained a 122 $\times$ 4,627 matrix Y^{Array} and Y^{RNAseq}, where $Y^{Array}_{(i,g)}$ and $Y^{RNAseq}_{(i,g)}$ give the expression of gene g in individual i on the expression array and the RNA sequencing, respectively, before quantile normalization.

We measured the reproducibility of expression profiles inside subjects by calculating the Spearman rank correlation for each pair of row vectors $Y^{Array}_{(i,.)}$ and $Y^{RNAseq}_{(i,.)}$ for $i \in \{1..122\}$ (**Supplementary Fig. 4a**). The average Spearman correlation was 0.71. A previous study by Maroni et al. [5] measured technical reliability of RNA-seq versus array data with independent datasets. Importantly, they reported an average Spearman correlation of 0.73 for reproducibility of expression profiles inside subjects. This result provides additional support to the technical validity of our expression analysis pipeline.

eQTL replication requires that relative differences between subjects are reproducible across experiments. We compared the order of individuals at each gene as reported by the array and the RNA-sequencing data by measuring the Spearman rank correlation of the column vectors $Y^{Array}_{(.,g)}$ and $Y^{RNAseq}_{(.,g)}$ for $g \in \{1..4,627\}$ (**Supplementary Fig. 4b**). The concordance of rank-order of individuals across platforms was moderate (average Spearman rank correlation 0.22), which implies only moderate power to replicate QTLs across the two platforms. Choy et al. performed a similar analysis with biological replicates of LCLs in two expression arrays independent from our study [6]. They also reported Spearman rank correlations of 0.25-0.3 for relative differences of expression between subjects, in agreement with our analysis.

2.3 Partitioning heritability on simulated datasets

The best STR can often exhibit high collinearity with other *cis* variants. To rule out the possibility that the LMM could be incorrectly partitioning variance to the STR in the case of tagging another causal variant nearby, we performed simulations in which there was a single causal SNP eQTL per gene. For each gene, we simulated expression values using the following process:

1. Choose the best SNP from the eQTL analysis on real data as the causal variant. Let this eQTL explain σ^2 percent of expression variance.
2. Simulate expression values as $y_i = \beta x_i + \epsilon_i$ where y_i is the simulated expression value for individual i, x_i is the SNP genotype for individual i, $\beta = \sqrt{\sigma^2}$, and $\epsilon \sim N(0, 1 - \sigma^2)$.
3. Run the LMM analysis as described above to determine h^2_{STR} and h^2_b.

Notably, this procedure simulates the causal SNP based on the SNP-eQTL analysis, rendering the test more realistic. The simulation was repeated for values of σ^2 equal to 0, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 for each gene on chromosome 18. We performed this analysis for both the cases of treating the STR as a fixed and a random effect.

We observed that in both models, h^2_b was very close to the simulated value of σ^2, as expected. Importantly, the median value for h^2_{STR} was negative for the fixed effects case and 0 for the random effects case across all simulations. The mean values were close to 0 in most realistic values of SNP-eQTL effects and slightly biased (< 0.005) upwards in the case of very strong SNP-eQTLs (**Supplementary Fig. 5**). The median ratio of h^2_{STR} to $h^2_{STR} + h^2_b$ was $< 0.1\%$ for the fixed effects case and exactly 0 for the random effects case for all simulations. These findings suggest that our LMM analysis reflects an accurate partitioning of variance even in the presence of strong SNP-eQTLs.

To further validate that our estimators of h^2_{STR} are not inflated, we also ran the fixed effects LMM analysis on random pairs of eSTRs and local bi-allelic mutations from chromosome 2 and gene expression profiles from chromosome 1. This generated a null distribution for h^2_{STR} in the case of no association. In this negative control condition, h^2_{STR} was distributed symmetrically around 0 with mean 7×10^{-4} and median -0.002, demonstrating that the estimator is unbiased.

2.4 Treating STRs as random vs. fixed effects

In our LMM analysis to partition heritability between STRs and other *cis* variants, we treated the best STR for each gene as a fixed effect. We repeated this analysis treating the STR as a random effect to determine whether this choice significantly affects our results. We used a model of the form:

$$\vec{y}_g = \alpha_g + \vec{v}_g + \vec{u}_g + \vec{\epsilon}_{j,g} \tag{4}$$

where:

- $\vec{v}_g$ is a length n vector of random effects for the best STR
- $\vec{v}_g \sim MVN_n(0, \sigma^2_{v_g} S_g)$ with $\sigma^2_{v_g}$ the percent of phenotypic variance explained by the best STR for gene g
- S_g is a standardized IBS relatedness matrix constructed using the best STR. It was constructed as:

$$S_g = \frac{1}{var(\vec{x})}(\vec{x} - 1_n\overline{\vec{x}})(\vec{x} - 1_n\overline{\vec{x}})^T \tag{5}$$

where $\vec{x}$ is a length n vector consisting of genotypes for the best STR.

- All other variables are as described in the Online Methods.

We used the GCTA program [7] to determine the REML estimates of $\sigma^2_{u_g}$ and $\sigma^2_{v_g}$. GCTA encountered numerical problems using the `--reml-no-constrain` option, likely due to the small sample size for each gene and strong correlation between the STR and bi-allelic variance components. Therefore, estimates were constrained to be between 0 and 1 and are biased to be greater than 0.

The overall phenotypic variance-covariance matrix is:

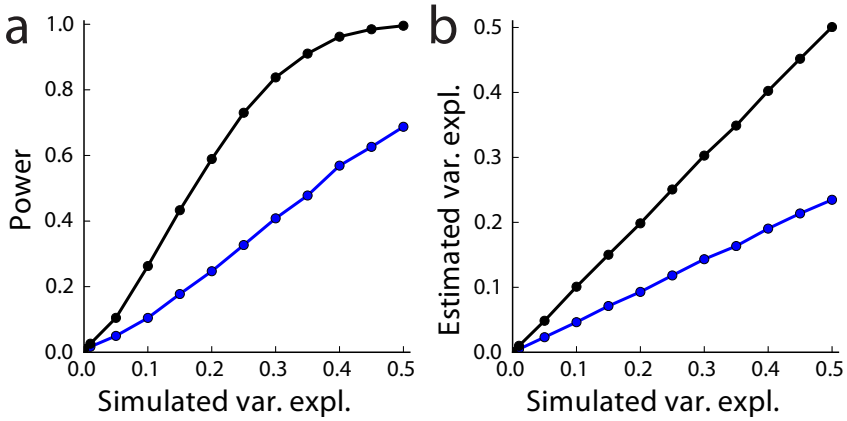
$$var(\vec{y}_g) = \sigma^2_{v_g} S_g + \sigma^2_{u_g} K_g + \sigma^2_{\epsilon_{j,g}} I_n \tag{6}$$

with $\sigma^2_{v_g}$ giving the percent of phenotypic variance explained by the best STR (h^2_{STR}) and $\sigma^2_{u_g}$ giving the percent explained by other *cis* bi-allelic mutations (h^2_b).

Estimates of the variance explained by STRs and by *cis* bi-allelic mutations using this model are consistent with those obtained by treating STRs as a fixed effect (**Table 1** and **Supplementary Table 5**). Because the random effects estimates are constrained to be between 0 and 1, the random effects model tended to partition variance all to a single variance component, but overall distributions of h^2_{STR} and h^2_b were similar to the fixed effects case (**Fig. 3a,b** and **Supplementary Fig. 6**).

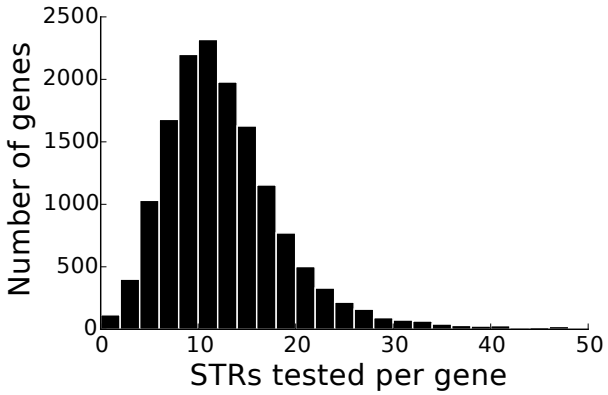
3 Supplementary Figures

3.1 Supplementary Figure 1: STR genotype errors reduce power to detect eSTR associations



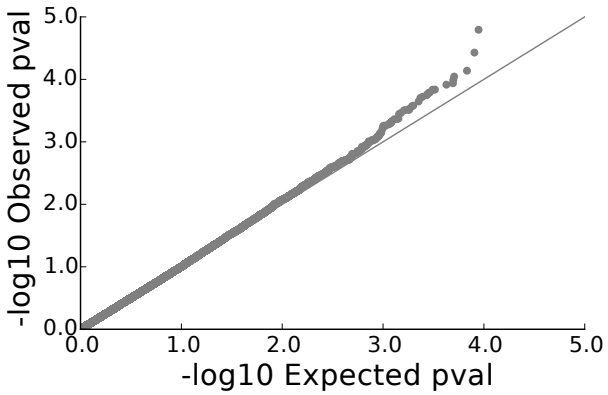
STR genotyping errors reduce power to detect eSTR associations. **a**. Power to detect associations and **b**. estimated variance explained for different simulated values of variance explained by the STR. (black: observed capillary electrophoresis genotypes, blue: lobSTR genotypes).

3.2 Supplementary Figure 2: Number of STRs tested per gene



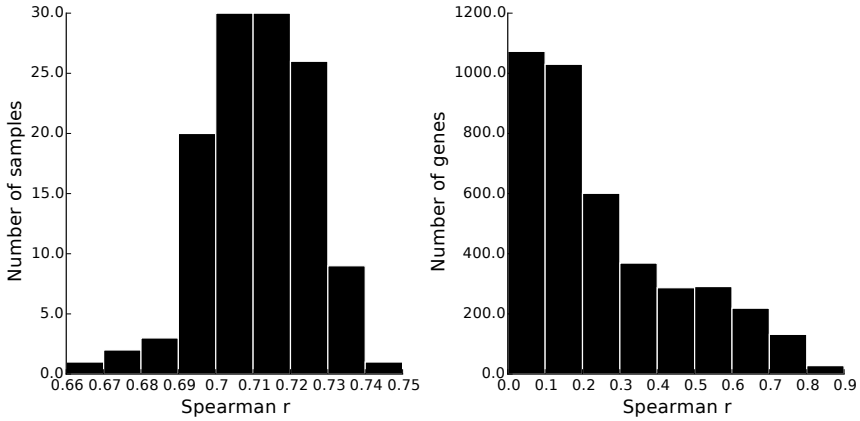
Number of STRs tested per gene. Histogram gives the number of STRs within 100kb of each gene that passed quality filters and were included in the eSTR analysis.

3.3 Supplementary Figure 3: Unlinked controls follow the null



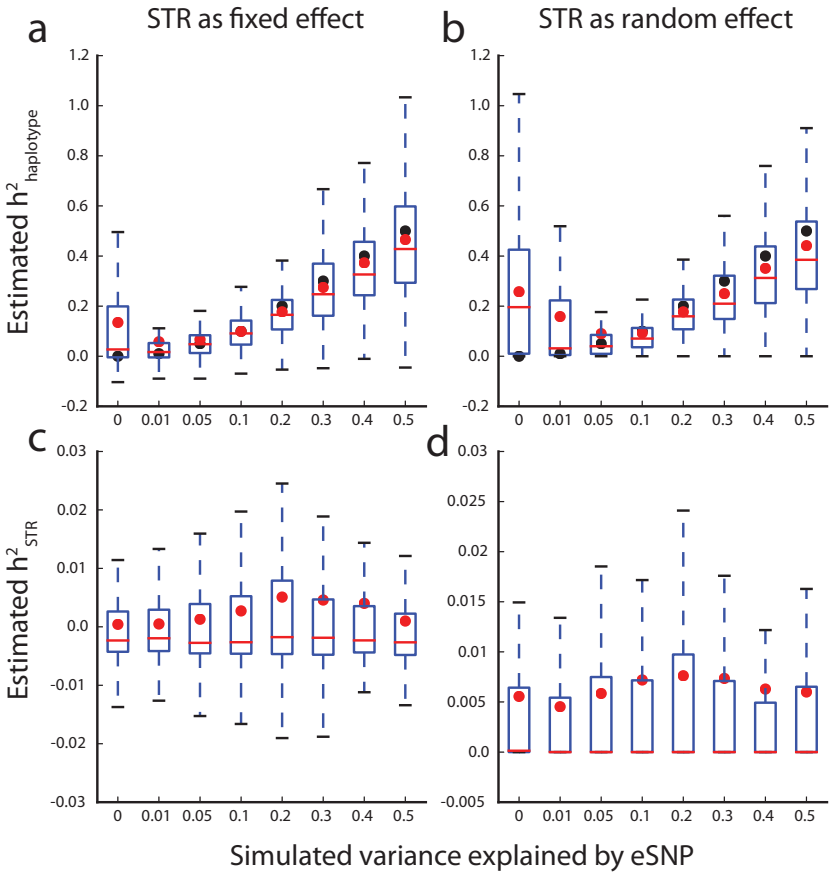
Unlinked controls follow the null. QQ plot of association tests between random unlinked STRs and genes.

3.4 Supplementary Figure 4: Expression values are moderately reproducible across platforms



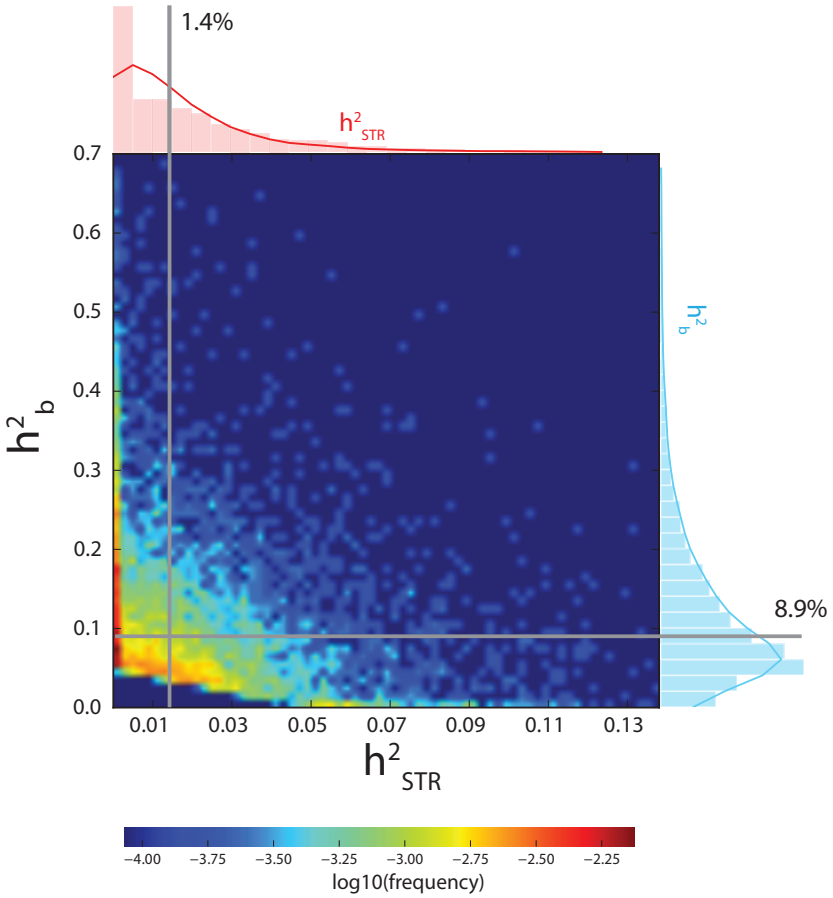
Expression values are moderately reproducible across platforms. **a**. Distribution of Spearman rank correlation coefficients between gene expression profiles of individuals measured on microarray vs. RNA-sequencing platforms. **b**. Distribution of Spearman rank correlation coefficients between the order of individuals ranked by expression levels across transcripts measured using microarray vs. RNA-sequencing platforms.

3.5 Supplementary Figure 5: Variance partitioning simulations



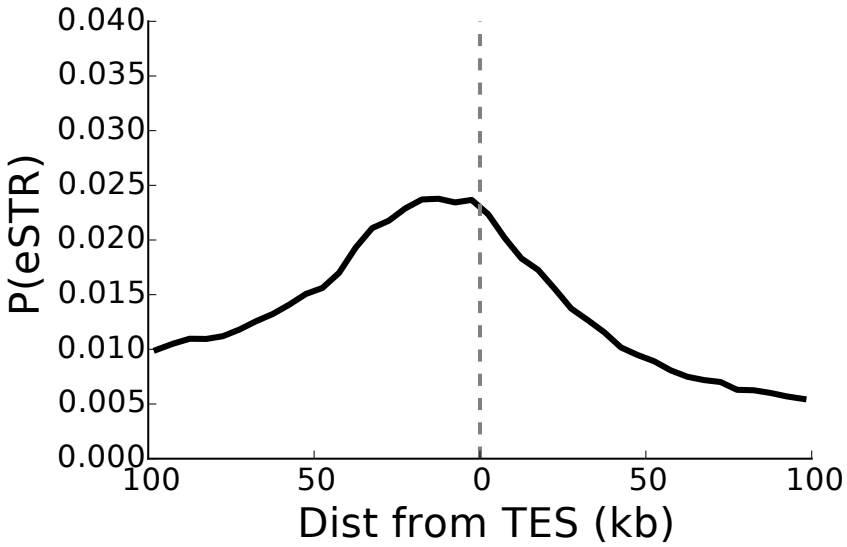
Variance partitioning simulations Plots show variance partitioning results from simulations in which each gene has a single causal eSNP. (**a**&**b**) The distributions of h_b^2 (**c**&**d**) The distributions of h_{STR}^2 (**a**&**c**) The LMM simulations with STRs as fixed effects (**b**&**d**) The LMM simulations with STRs as random effects (**a**-**d**) Black points denote the true value of the variance explained by the causal SNP. Red dots denote the average value of the estimator. Red bars denote the median value of the estimator. The figure shows that the median values of the best STRs are insensitive to the presence of a strong SNP eQTL.

3.6 Supplementary Figure 6: Partitioning variance when treating the STR as a random effect



Partitioning variance when treaing the STR as a random effect. The heatmap shows the distribution of h^2_{STR} and h^2_b for each gene. Dashed gray lines give the medians of each distribution.

3.7 Supplementary Figure 7: Enrichment of eSTRs at transcription end sites



eSTRs are enriched near the transcription end site (TES). For each distance bin around the TES, the plot shows the percentage of STRs in that bin that were called as significant eSTRs.

4 Supplementary Tables

4.1 Supplementary Table 1: Significant eSTRs

See file Gymrek_etal_signifcant_estrs.tab.

4.2 Supplementary Table 2: Distribution of motif lengths in eSTRs vs. all STRs

Period	Num. eSTRs	% eSTRs	Num. all STRs	% all STRs	Enrichment	Pval
2	951	50.2%	50,184	62.0%	0.81	1.0
3	223	11.8%	7,369	9.1%	1.29	4.8×10^{-5}
4	516	27.2%	17,938	22.2%	1.23	8.2×10^{-8}
5	166	8.8%	4,466	5.5%	1.59	3.9×10^{-9}
6	39	2.1%	1,023	1.3%	1.63	2.4×10^{-3}

Distribution of motif lengths in eSTRs vs. all STRs. Distribution of motif lengths in all unique eSTR loci vs. all unique STR loci included in the analysis after applying quality filters.

4.3 Supplementary Table 3: Distribution of motifs in eSTRs vs. all STRs

Motif	Num. eSTRs	% eSTRs	Num. all STRs	% all STRs	Enrichment	Pval
AAAAAC	17	0.9%	217	0.3%	3.35	1.7×10^{-5}
AATC	10	0.5%	152	0.2%	2.81	3.2×10^{-3}
AAAAC	94	5.0%	1,822	2.2%	2.20	1.8×10^{-12}
AAC	95	5.0%	2,056	2.5%	1.97	5.0×10^{-10}
AAAC	173	9.1%	3,995	4.9%	1.85	9.0×10^{-15}
AAAG	47	2.4%	1,179	1.5%	1.70	3.6×10^{-4}
AAG	10	0.5%	285	0.4%	1.50	0.13
AAAAG	15	0.8%	449	0.6%	1.43	0.11
ATCC	16	0.8%	488	0.6%	1.40	0.11
ATC	10	0.5%	392	0.5%	1.09	0.44
AG	128	6.8%	5,174	6.3%	1.06	0.27
AAAT	198	10.4%	8,073	10.0%	1.05	0.25
AAAAT	35	1.8%	1,451	1.8%	1.03	0.45
AATG	16	0.8%	676	0.8%	1.01	0.52
AAT	74	3.9%	3,678	4.5%	0.86	0.92
AT	161	8.5%	8,775	10.8%	0.78	0.99
AC	662	34.9%	36,206	44.7%	0.78	1.0
AGAT	16	0.8%	1,561	1.9%	0.44	1.0

Distribution of motifs in eSTRs vs. all STRs. Distribution of motifs in all unique eSTR loci vs. all unique STR loci included in the analysis after applying quality filters. Only motifs for which there were at least 10 eSTRs are shown. Motifs were converted to canonical format as described in [8].

4.4 Supplementary Table 4: Distribution of genomic locations of eSTRs vs. all STRs

Annotation	Num. eSTRs	% eSTRs	Num. all STRs	% all STRs	Enrichment	Pval
Coding	13	0.7%	157	0.2%	3.54	9.1×10^{-5}
5' UTR	51	2.7%	897	1.1%	2.43	1.0×10^{-8}
Exon	127	6.7%	2,452	3.0%	2.21	1.5×10^{-16}
3' UTR	77	4.1%	1,569	1.9%	2.10	1.7×10^{-9}
Neargene (5')	335	17.7%	7,357	9.1%	1.95	1.5×10^{-32}
Neargene (3')	326	17.2%	7,399	9.1%	1.88	4.5×10^{-29}
Intron	1,314	69.3%	52,326	64.6%	1.07	6.1×10^{-6}
Intergenic	395	20.8%	23,373	28.9%	0.72	1.00

Distribution of genomic locations of eSTRs vs. all STRs. Annotations were compiled using Ensembl version 71. "Exon" refers to both coding and non-coding exons and untranslated regions. "Neargene" refers to regions within of a gene. "Intergenic" refers to STRs not falling into any other annotation. Note some STRs may overlap multiple annotations.

4.5 Supplementary Table 5: Heritability of gene expression explained by STRs vs. SNPs in each LMM

	h^2_b	h^2_{STR}	h^2_{STR}/h^2_{cis}
eSTR genes - (STR fixed)	0.1203 (0.1139-0.1259)	0.0180 (0.0166-0.0199)	0.1230 (0.1106-0.1420)
eSTR genes - (STR random)	0.1229 (0.1159-0.1295)	0.0200 (0.0178-0.0216)	0.1288 (0.1179-0.1451)
Moderate *cis* h^2_{cis} (STR fixed)	0.0910 (0.0884-0.0938)	0.0145 (0.0137-0.0151)	0.1283 (0.1222-0.1346)
Moderate *cis* h^2_{cis} (STR random)	0.0892 (0.0865-0.0918)	0.0143 (0.0137-0.0149)	0.1245 (0.1184-0.1309)

Heritability of gene expression explained by STRs vs. SNPs in each LMM. Values show the median and 95% confidence interval of the median across all eSTR-containing genes and genes with moderate *cis* heritability ($\geq$5%). h^2_b denotes the variance explained by all common *cis* bi-allelic markers, h^2_{STR} denotes the variance explained by the best STR for each gene, and $h^2_{cis} = h^2_{STR} + h^2_b$.

5 References

References

[1] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, Sep 2007.

[2] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet.*, 2(12):e190, Dec 2006.

[3] Alan Dabney and John D. Storey. *qvalue: Q-value estimation for false discovery rate control*. R package version 1.38.0.

[4] N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, Dec 2002.

[5] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, Sep 2008.

[6] E. Choy, R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena, P. L. De Jager, S. Y. Shaw, C. S. Wolfish, J. M. Slavik, C. Cotsapas, M. Rivas, E. T. Dermitzakis, E. Cahir-McFarland, E. Kieff, D. Hafler, M. J. Daly, and D. Altshuler. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.*, 4(11):e1000287, Nov 2008.

[7] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88(1):76–82, Jan 2011.

[8] T. Willems, M. Gymrek, G. Highnam, D. Mittelman, and Y. Erlich. The landscape of human STR variation. *Genome Res.*, 24(11):1894–1904, Nov 2014.