nature genetics

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

Melissa Gymrek[1–4], Thomas Willems[1,4,5], Audrey Guilmatre[6,7], Haoyang Zeng[8], Barak Markus[1], Stoyan Georgiev[9], Mark J Daly[3,10], Alkes L Price[3,11,12], Jonathan K Pritchard[9,13], Andrew J Sharp[6] & Yaniv Erlich[1,4,14,15]

The contribution of repetitive elements to quantitative human traits is largely unknown. Here we report a genome-wide survey of the contribution of short tandem repeats (STRs), which constitute 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 that of linked SNPs and indels and found that eSTRs contribute 10–15% of the *cis* heritability mediated by all common variants. Further functional genomic analyses showed that eSTRs are enriched in conserved regions, colocalize with regulatory elements and may modulate certain histone modifications. By analyzing known genome-wide association study (GWAS) signals and searching for new associations in 1,685 whole genomes from deeply phenotyped individuals, we found that eSTRs are enriched in various clinically relevant conditions. These results highlight the contribution of STRs to the genetic architecture of quantitative human traits.

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[1–3]. These observations suggest that understanding the genetic architecture of transcriptomes may provide insights into the cellular-level mediators underlying complex traits[4–6]. So far, eQTL mapping studies have mainly focused on SNPs and, to a lesser extent, biallelic indels and copy number variations (CNVs) as determinants of gene expression[7–11]. 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% of the heritability unexplained[8,12]. It has been speculated that such heritability gaps could indicate the involvement of repetitive elements that are not well tagged by common SNPs[13,14].

To augment the repertoire of eQTL classes, we focused on STRs, which constitute one of the most polymorphic and abundant types of repetitive elements in the human genome[15,16]. These loci consist of periodic DNA motifs of 2–6 bp spanning a median length of around 25 bp. 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, resulting in mutation rates that are orders of magnitude higher than those for most other variant types[15,17]. Over 40 Mendelian disorders, such as Huntington disease, are attributed to STR mutations, with most caused by large expansions of trinucleotide coding repeats[18].

Several properties of STRs suggest that they may have a regulatory role. *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]. *In vivo* experiments have reported specific examples of STR variations that control gene expression across a wide range of taxa, including *Haemophilus influenzae*[26], *Saccharomyces cerevisiae*[27], *Arabidopsis thaliana*[28] and *Microtus ochrogaster*[29]. Recent studies reported that dinucleotide repeats are a hallmark of enhancers in *Drosophila melanogaster* and are enriched in putative enhancers in humans[30]. Human promoters also disproportionately harbor STRs[31], and the presence of STRs in promoters or transcribed regions greatly increases the divergence of gene expression profiles across great apes[32], suggesting that STRs have a key role in the evolution of expression. Indeed, several candidate gene studies in humans reported that STR variations modulate gene expression[19,33–37] and alternative splicing[23,38,39]. A recent study found that a GWAS signal for Ewing sarcoma is a sequence variant in an AAGG repeat that increases binding of the EWSR1-FLI1 oncoprotein, resulting in *EGR2*

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA. [2]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, 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, New York, USA. [5]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [7]Department of Pediatric Hematology, Robert Debré Hospital, Paris, France. [8]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [9]Department of Genetics and Biology, Stanford University, Stanford, California, USA. [10]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [12]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [13]Howard Hughes Medical Institute, Chevy Chase, Maryland, USA. [14]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, New York, USA. [15]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, USA. Correspondence should be addressed to Y.E. (yaniv@cs.columbia.edu).

Received 17 August; accepted 12 November; published online 7 December 2015; doi:10.1038/ng.3461

overexpression[40]. Despite the accumulating evidence, there has been no systematic evaluation of the contribution of STRs to gene expression in humans.

To this end, we conducted a genome-wide analysis of STRs that affect expression of nearby genes, termed eSTRs, in lymphoblastoid cell lines (LCLs), an *ex vivo* model commonly used for eQTL studies. Next, we used several statistical genetic and functional genomic analyses to show that hundreds of these eSTRs are predicted to be functional. Finally, we tested the involvement of eSTRs in clinically relevant phenotypes.

RESULTS

Initial genome-wide discovery of eSTRs

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[9] project and whose whole genomes were sequenced by the 1000 Genomes Project[41]. STR genotypes were obtained from our previous study[42] 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[43]. 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 variation in STR locus lengths in the 1000 Genomes Project data sets, as determined by large-scale validation using 5,000 STR genotype calls obtained by capillary electrophoresis, the gold standard for STR genotyping[42]. 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 in comparison to the capillary calls. This process showed that STR genotype errors reduce the power to detect eSTRs by 30–50% but, notably, do not create spurious associations (**Supplementary Fig. 1** and **Supplementary Note**).

To detect eSTR associations, we regressed gene expression on STR dosage, defined as the sum of the lengths for the two STR alleles in each individual. We decided to use this measure because of previous findings that reported a linear trend between STR length and gene expression[19,34,36] or disease phenotypes[44,45]. As covariates, we included sex, population structure and other technical parameters (**Fig. 1a** and **Supplementary Note**). 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 (Online Methods) and were within 100 kb of the transcription start and end sites of the gene transcripts as annotated by Ensembl[46]. 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 false discovery rate (FDR) ≤ 5%) (**Fig. 1b** and **Supplementary Data Set 1**). The majority of these were di- and tetranucleotide STRs (**Supplementary Tables 1** and **2**). Only 13 eSTRs fell in coding exons, but eSTRs were nonetheless strongly enriched in 5′ UTRs ($P = 1.0 \times 10^{-8}$), 3′ UTRs ($P = 1.7 \times 10^{-9}$) and regions near genes ($P < 1 \times 10^{-28}$) in comparison to all STRs analyzed (**Supplementary Table 3**). Overall, there was no bias in direction of effect (**Supplementary Table 4**). We also 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** and **Supplementary Fig. 3**). Both negative controls produced the uniform *P*-value distributions expected under the null hypothesis of no association. These findings provide support for the absence of spurious associations due to inflation of the test statistic or the presence of uncorrected population structure. To assess the effect of low sequencing coverage on our results, we generated high-coverage targeted sequencing data for 2,472 promoter STRs and repeated the eSTR analysis (Online Methods). We found that the association results were largely reproducible across data sets, with 80% of the eSTRs tested showing the same direction of effect as in the original analysis ($P = 9.9 \times 10^{-12}$; $n = 126$) (**Supplementary Fig. 4** and **Supplementary Note**). Three previous reports described candidate gene studies of eSTRs and involved STRs that were tested in our framework[19,36,47]. Our genome-wide

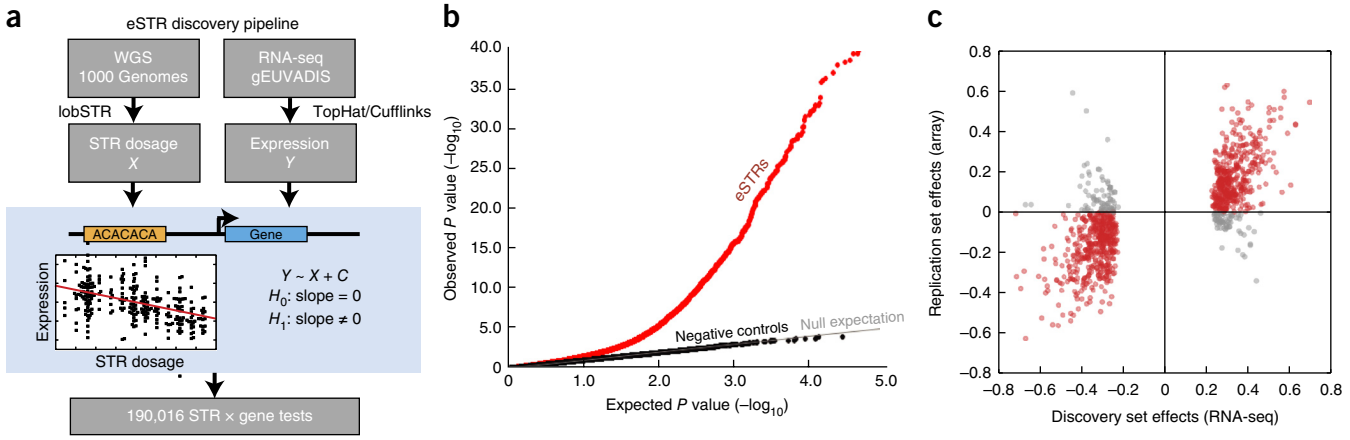

Figure 1 eSTR discovery and replication. (**a**) eSTR discovery pipeline. An association test using linear regression was performed between STR dosage and expression level for every STR within 100 kb of a gene. WGS, whole-genome sequencing; RNA-seq, RNA sequencing; *C*, covariates; H_0, null hypothesis; H_1, alternative hypothesis. (**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, and red dots give the results of the observed association tests. (**c**) Comparison of eSTR effect sizes as Pearson correlations in the discovery data set versus the replication data set. Red points denote eSTRs whose directions of effect were concordant in the two data sets, and gray points denote eSTRs with discordant directions of effect.

approach was able to replicate the association between *TP5313* and the pentanucleotide STR in the 5′ UTR of the gene and showed the same direction of effect. However, the associations for the other two candidate genes did not meet the multiple hypothesis–corrected *P*-value threshold (**Supplementary Table 5**).

The initial discovery set of eSTRs was largely reproducible in an independent set of individuals using 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[48]. These individuals belong to cohorts with African, Asian, European and Mexican ancestry, enabling testing of associations in a largely distinct set of populations. The Illumina expression array allowed us to test 882 eSTRs of the 2,060 identified above. The association signals of 734 of the 882 (83%) eSTRs tested showed the same direction of effect in both data sets (sign-test $P = 2.7 \times 10^{-94}$), and the effect sizes were strongly correlated ($R = 0.73$, $P = 1.4 \times 10^{-149}$) (**Fig. 1c**), despite only moderate reproducibility of expression profiles across platforms (**Supplementary Fig. 5** and **Supplementary Note**). For comparison, only 54% of non-eSTRs showed the same direction of effect in both data sets, close to the expected value of 50% for null associations. Overall, these results show that eSTR association signals are robust and reproducible across populations and expression assay technologies.

Partitioning the contribution of eSTRs 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[42,49,50], but there are known examples of STRs tagging GWAS SNPs[51].

To address this question, we partitioned the relative contributions of eSTRs versus those of all common (minor allele frequency (MAF) ≥1%) biallelic SNPs, indels and structural variants in the *cis* region of each gene using a linear mixed model (LMM) (**Fig. 2a**). Multiple studies have used this approach to measure the total contribution of common variants to the heritability of quantitative traits and to partition the contributions of different classes of variants[52,53]. Taking a similar approach, we included two types of effects for each gene: a random effect (h^2_b), which captures all common biallelic loci detected within 100 kb of the gene, and a fixed effect (h^2_{STR}), which captures the lead STR. To test whether other causal variants in the local region could inflate the estimate of the STR contribution, we simulated gene expression with one or two causal SNP eQTLs per gene while preserving the local haplotype structure. In this negative-control scenario, the LMM correctly reported median $h^2_{STR}/h^2_{cis} \approx 0$ across all conditions (**Supplementary Figs. 6** and **7**, and **Supplementary Note**), where $h^2_{cis} = h^2_b + h^2_{STR}$. This suggests that other causal variants in LD do not inflate the estimates of the relative contribution of STRs. However, simulations based on capillary electrophoresis data suggest that the variance explained by STRs is downwardly biased in the presence of genotyping errors (**Supplementary Fig. 8** and **Supplementary Note**), suggesting that 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 h^2_{STR}/h^2_{cis} ratio of 12.3% (95% confidence interval (CI) = 11.1–14.2%; n = 1,928) (**Supplementary Table 6**). We repeated the same analysis for genes with at least moderate (≥5%) *cis* heritability (Online 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[54] and to obtain a transcriptome-wide perspective on the role of STRs in gene expression (**Fig. 2c**). In this set of genes, eSTRs contributed about 13% (95% CI = 12.2–13.5%; n = 6,272) of the genetic variance attributed to common polymorphisms in *cis*. The median h^2_{STR} was 1.45% of the total variance in expression, whereas the median h^2_b was 9.10% (**Supplementary Table 6**). Repeating the analysis performed while considering STRs as a random effect gave highly similar results (**Supplementary Fig. 9**, **Supplementary Table 7** and **Supplementary Note**). Considering these results together, this analysis shows that STR variations explain a sizeable component of variation in gene expression after controlling for all variants that are well tagged by common biallelic markers in the *cis* region.

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 using whole-genome sequencing data for 311 individuals who were part of the discovery set to identify common eSNPs (MAF ≥5%) within 100 kb of each gene. This process identified 4,290 genes with an eSNP (gene-level FDR ≤ 5%). We then reanalyzed the eSTR association signals while conditioning on the genotype of the most significant eSNP for each gene (**Fig. 3a**). For each eSTR, we ascertained the subset of individuals who were homozygous for the major allele of the lead eSNP in the region. If the eSTR simply tags this eSNP, its effect when conditioned on that of the eSNP should be randomly distributed in comparison to its unconditioned effect. Alternatively, if the eSTR is causal, the direction of the conditioned effect should match that of the unconditioned

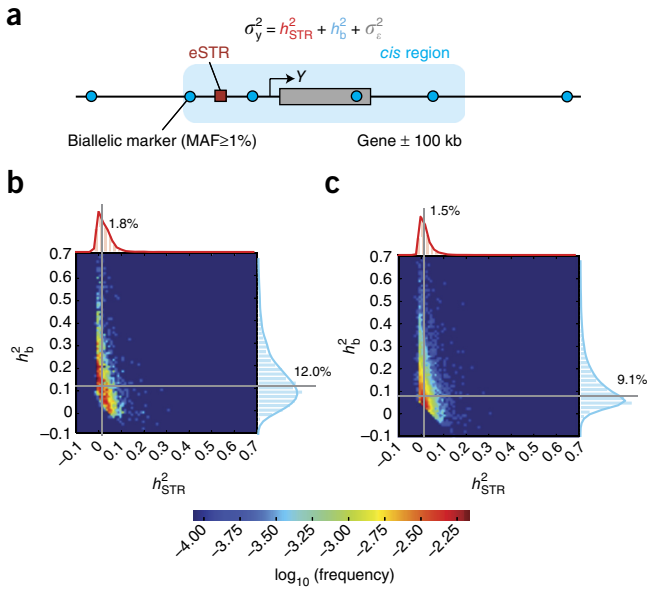

Figure 2 Variance partitioning using linear mixed models. (**a**) The normalized variance in the expression of gene *Y* was modeled as the contribution to variance of the best eSTR and all common biallelic markers in the *cis* region (±100 kb from the gene boundaries). (**b**,**c**) Heat maps show the joint distributions of the variance explained by eSTRs (*x* axis) and by the *cis* region (*y* axis). Gray lines denote the median variance explained. (**b**) Variance partitioning across genes with a significant eSTR in the discovery set. (**c**) Variance partitioning across genes with moderate *cis* heritability.

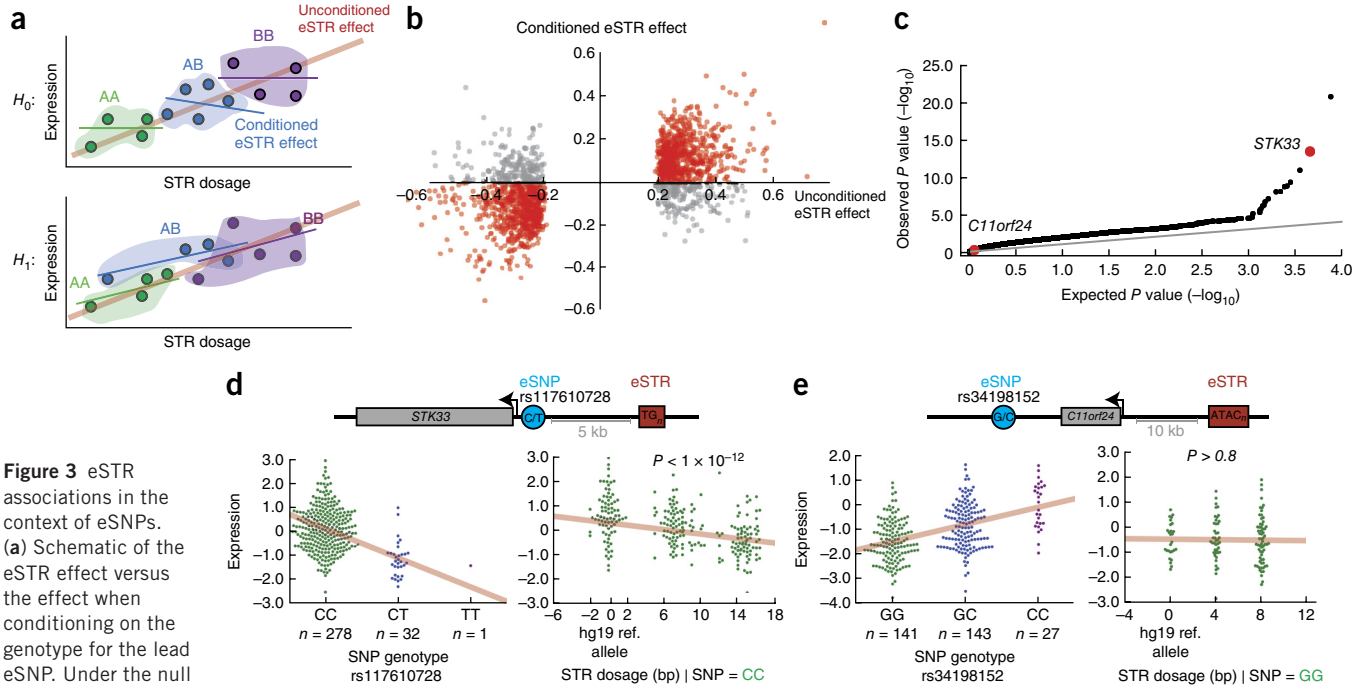

Figure 3 eSTR associations in the context of eSNPs. (**a**) Schematic of the eSTR effect versus the effect when conditioning on the genotype for the lead eSNP. Under the null expectation (H_0), the original association (red line) comes from mere tagging of the eSNP. Thus, the eSTR effect disappears when restricting to a group of individuals (dots) with the same eSNP genotype (colored backgrounds). Under the alternative hypothesis (H_1), the effect is concordant between the original and conditioned associations. (**b**) Original eSTR effect versus conditioned eSTR effect. Red points denote eSTRs whose directions of effect were concordant in both data sets, and gray points denote eSTRs with discordant directions of effect. (**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 strong explanatory value beyond that of the lead eSNP (blue circle), according to ANOVA. When conditioning on individuals who are homozygous for the C allele of the eSNP (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 who are homozygous for the G allele of the eSNP (bottom left; green dots), the STR effect is lost (bottom right).

effect. We conducted this analysis for eSTR loci where at least 25 individuals were homozygous for the lead eSNP and for which these individuals had at least two unique STR genotypes (1,856 loci). After conditioning on the lead eSNP, the direction of effect for 1,395 loci (75%) was identical to that in the original analysis (sign-test $P < 4.2 \times 10^{-109}$) and the effect sizes from the two analyses were significantly correlated ($R = 0.52$, $P = 3.2 \times 10^{-130}$) (**Fig. 3b**). This result further supports the additional role of eSTRs beyond that of traditional *cis*-eQTLs.

We also found that hundreds of eSTRs in the discovery set provide additional explanatory value for gene expression beyond that provided by the lead eSNP. Model comparison using ANOVA showed that, for 23% of the cases, a model with an eSTR significantly improved the variance in gene expression explained over the model considering only the lead eSNP (FDR < 5%) (**Fig. 3c–e** and Online Methods). In combination 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 fully attributed to tagging of the lead eSNP. Given the reduced quality of STR genotypes in comparison to SNP genotypes, this analysis is likely to underestimate the true contribution of STRs. Nonetheless, our results provide concrete examples for hundreds of associations in which an eSTR increases the variance explained by a lead eSNP.

Integrative genomic evidence for a functional role of eSTRs

To provide further evidence of the regulatory role of eSTRs, we analyzed them in the context of functional genomics data. First, we assessed the potential functionality of STR regions by measuring signatures of purifying selection, as previous studies reported that putatively causal eSNPs are slightly enriched in conserved regions[55]. We inspected sequence conservation[56] across 46 vertebrates in the sequences upstream and downstream of the eSTRs in our discovery data set (**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 the 1-kb window centered on each eSTR was slightly but significantly higher ($P < 0.03$) than for a comparable region for the control STR. Tightening the window size to shorter stretches of ±50 bp resulted in 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 the eSTRs discovered by our association pipeline reside in regions exposed to relatively higher purifying selection, further suggesting a functional role.

eSTRs substantially colocalize with functional elements. They show the strongest enrichment closest to TSSs (**Fig. 4b**) and, to a lesser extent, in or near predicted enhancers (**Supplementary Fig. 10**). We also investigated the colocalization of eSTRs with histone modifications as annotated by the Encyclopedia of DNA Elements (ENCODE) Consortium[7] in LCLs. eSTRs were strongly enriched in peaks for histone modifications associated with regulatory regions (trimethylation of histone H3 at lysine 4 (H3K4me3) and acetylation of histone H3 at lysine 27 (H3K27ac) and lysine 9 (H3K9ac)) and transcribed regions (trimethylation of histone H3 at lysine 36 (H3K36me3)) and were

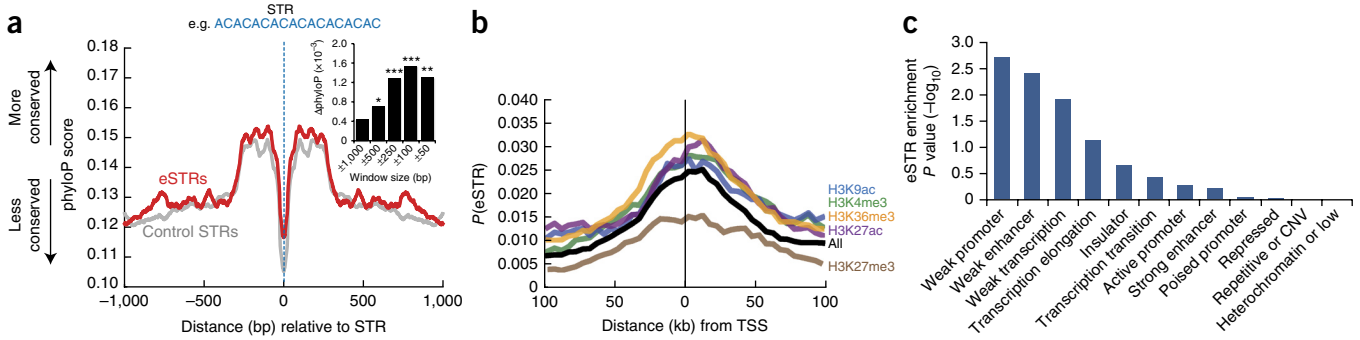

Figure 4 Conservation and epigenetic analysis of eSTR loci. (**a**) Median phyloP conservation score as a function of distance from the STR. Red, eSTR loci; gray, matched control STRs. The inset shows the difference in the phyloP conservation score between eSTRs and matched control STRs as a function of window size around the STR. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$. (**b**) Probability that an STR scores as an eSTR in the discovery set as a function of distance from the TSS. eSTRs show clustering around the TSS (black line). Conditioning on the presence of a histone mark (colored lines) significantly modulated the probability that an STR is an eSTR. (**c**) Enrichment of eSTRs in different chromatin states.

depleted in repressed regions (trimethylation of histone H3 at lysine 27 (H3K27me3)) (**Fig. 4b**). To test the significance of these signals, we constructed a null distribution for each histone modification by measuring the colocalization of eSTRs with randomly shifted histone peaks, similar to the procedure used by Trynka *et al.*[57]. This null distribution controls for the co-occurrence of eSTRs and histone peaks due to their proximity to other causal variants. We found that eSTR–histone peak colocalizations were significant (weakest *P* value < 0.01) after the peak-shifting procedure, suggesting that these results stem from the eSTRs themselves (**Supplementary Table 8**). We also performed a peak-shifting analysis using ChromHMM annotations[58] (**Fig. 4c**), which indicated that eSTRs are most strongly enriched in 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.

We also found that variations in eSTR length seem to modulate the presence of certain histone marks (Online Methods and **Supplementary Fig. 11**). We introduced different eSTR alleles to GERV[59], a machine learning approach that examines the effect of DNA sequence on histone marks. This process found that eSTRs have significantly greater effects than control STRs on predicted regulatory regions (H3K4me3, $P = 0.00109$; DNase I hypersensitivity, $P = 0.00045$; H3K9ac, $P = 0.00462$) and transcribed regions (H3K36me3, $P = 0.01336$). These results are consistent with the analysis of chromatin modifications above. Notably, because 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 for the functionality of eSTRs and suggest histone mark modulation as a potential mechanism.

The potential role of eSTRs in human conditions

Encouraged by the evidence for the regulatory role of eSTRs, we wondered about their potential involvement in clinically relevant conditions. First, we tested whether genes implicated by previous GWAS listed in the National Human Genome Research Institute (NHGRI) GWAS catalog[60] are enriched for eSTR-associated genes. We focused on seven complex disorders: rheumatoid arthritis, Crohn's disease, type 1 diabetes, type 2 diabetes, blood pressure, bipolar disorder and coronary artery disease. The first three conditions have a strong autoimmune component, rendering them more relevant to the LCL data used for eSTR discovery. To create a proper null, we compared the overlap of genes with an eSTR to randomly chosen sets of genes matched to the tested GWAS genes on the basis of both gene expression level in LCLs and *cis* heritability.

We found that GWAS genes for Crohn's disease were significantly ($P < 0.001$) enriched for eSTR hits (**Fig. 5a** and **Supplementary Fig. 12**). Moderate enrichment for eSTRs ($P = 0.074$) was found in GWAS genes for rheumatoid arthritis, consistent with the known role of immune function in these traits. Enrichments were 2–3 times higher for autoimmune diseases than for the other conditions (average overlap of 6%). Interestingly, for seven overlapping genes, the eSTRs explained more variance in gene expression than the lead eSNP for the gene. Furthermore, for close to 30 genes, a joint model of the lead eSTR and eSNP explained significantly more variance in gene expression than the eSNP alone, raising the possibility of an etiological role for the eSTR.

Next, we performed an association study using eSTRs to further test the hypothesis that eSTRs underlie clinically relevant phenotypes. For this analysis, we turned to ~1,700 unrelated individuals who were sequenced to medium coverage (6×) with 100-bp paired-end reads using Illumina sequencing as part of the TwinsUK cohort of the UK10K project[61] and were phenotyped for a wide array of quantitative traits, primarily blood metabolites and anthropometric traits. Although most of these traits are not directly related to the immune system, we hypothesized that, similar to other eQTLs[3], some of the discovered eSTRs are shared across tissues and could have a role in additional tissues. After genotyping STRs with lobSTR, we tested for association between eSTRs and each of the 38 reported phenotypes while controlling for sex, age and population structure. To enrich for STR loci that are likely to be causal for variation in gene expression, we restricted analysis to eSTRs that significantly improved the explained variance in gene expression over a model with the lead eSNP alone. In total, we obtained 499 eSTRs after applying this condition and excluding eSTRs that were genotyped in <1,000 individuals.

We identified 12 significant associations (FDR per phenotype < 10%) between eSTRs and the clinical phenotypes in the TwinsUK data (**Fig. 5b** and **Supplementary Table 9**). Only one association overlapped a known GWAS hit—an AAAC repeat at 4p16 was associated with decreased expression of *SLC2A9* and increased uric acid concentration in serum samples from TwinsUK, matching results from previous studies with SNPs[62–65]. The other 11 associations involved changes in blood metabolites, such as albumin and C-reactive protein levels, and physical traits, such as diastolic blood pressure and FEV_1 lung function, and have yet to be described in GWAS catalogs,

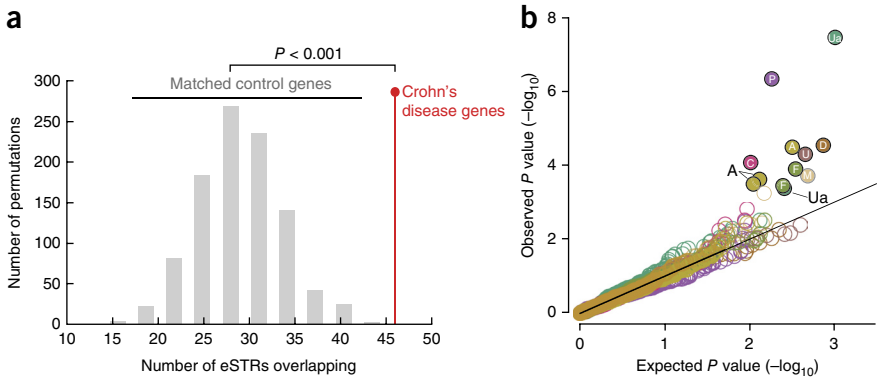

Figure 5 Association of eSTRs with clinical phenotypes. (**a**) Overlap between eSTRs and Crohn's disease GWAS genes (red) versus random subsets of genes (gray) matched on the basis of expression and heritability profiles in LCLs to the disease-associated genes. (**b**) Quantile-quantile plots of eSTR associations in the TwinsUK data. Only traits with significant (FDR < 0.1) associations are plotted. Closed circles, significant; open circles, not significant. A, albumin; C, C-reactive protein; D, diastolic blood pressure, F, FVC; M, mean corpuscular volume; P, phosphate; U, urea; Ua, uric acid.

suggesting new loci. We caution that full validation of each of these associations will require replication in additional cohorts. Nonetheless, as we were mainly interested in the overall trend for eSTRs, we repeated the association of the 38 phenotypes in the TwinsUK cohort with a similar number of random STR loci matched on the basis of distance to the TSS, repeat motif and number of genotyped samples. One hundred rounds of bootstrapping showed that eSTRs produced significantly more associations than the matched STR controls (mean for controls, 6.8 associations at FDR <10%; z-test $P < 1.8 \times 10^{-16}$). Repeating this test with a more stringent FDR cutoff of 5% gave a similar picture: the eSTRs produced six associations passing this threshold (**Supplementary Table 9**), significantly more that the matched STR controls (mean for controls of 3.2 associations at FDR <5%; $P < 1.1 \times 10^{-5}$). Taken together, our results show that eSTR signals are enriched in clinical phenotypes, both in known and potentially novel GWAS hits. These results could inform future efforts for disease mapping studies.

DISCUSSION

Repetitive elements have often been considered as neutral, with no phenotypic consequences[16]. Together with the technical difficulties in analyzing these regions, this notion has led large-scale genetic studies to largely overlook the potential contribution of repeats to human phenotypes. Our study focused on STRs, one of the most polymorphic classes of loci that comprise 1% of the human genome. Despite STRs being less abundant than SNPs, previous studies have shown that they are enriched in promoters and enhancers, where they frequently induce multiple–base pair variations, increasing the prior expectation of their ability to explain variation in gene expression. Following on these observations, we conducted a genome-wide scan for the contribution of STRs to gene expression. Our scan identified over 2,000 potential eSTRs and found that eSTRs contribute on average about 10–15% of the *cis* heritability of gene expression attributed to common (MAF ≥1%) polymorphisms. Functional genomics analyses provided further support for the predicted causal role of eSTRs. Finally, we found that eSTRs are enriched in clinically relevant phenotypes.

We hypothesize that there are more eSTRs to find in the genome, as our analysis had several technical limitations. First, the higher genotyping error rates for STRs in comparison to SNPs limited our power to detect eSTRs and likely downwardly biased their estimated contribution in the LMM and ANOVA analyses. 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[42]. Second, on the basis of previous findings in humans[19,34,36], our association tests focused on a linear relationship between STR length and gene expression. However, experimental work in yeast reported that certain loci exhibit nonlinear relationships between STR length and expression[27], which are unlikely to be captured in our current analysis. Finally, our association pipeline takes into account only the length polymorphisms of STRs and does not distinguish the effect of sequence variations inside STR alleles with identical lengths (dubbed homoplastic alleles[66]). Addressing these technical complexities would likely require phased STR haplotypes and longer sequence reads, which are currently unavailable, for large sample sizes. We envision that recent advancements in sequencing technologies[67] will further expand the catalog of eSTRs.

Despite these technical limitations, our findings show that repetitive elements in the human genome extensively contribute to variation in expression and are enriched in clinically relevant phenotypes. Our results are consistent with a recent study that reported that haplotypes of common SNPs, which capture genetic variants poorly tagged by current genotype panels, can explain substantially more heritability than common SNPs alone[68]. We anticipate that integrating the analysis of repetitive elements, specifically STR variations, will explain additional heritability and will lead to the discovery of new genetic variants relevant to human conditions.

URLs. lobSTR, http://lobstr.teamerlich.org; gEUVADIS, http://www.geuvadis.org/; 1000 Genomes Project Phase 1 genotypes, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521/; HapMap Consortium draft release 3 genotypes, ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3_r3/plink_format/; liftOver, http://genome.ucsc.edu/cgi-bin/hgLiftOver; European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), http://www.ebi.ac.uk/; phyloP track for hg19, ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/phyloP46way-Placental.txt.gz; UCSC Genome Browser, http://genome.ucsc.edu/; ENCODE chromatin state track for GM12878, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/wgEncodeBroadHmmGm12878HMM.bed.gz; ENCODE histone modification peaks for GM12878, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/; National Human Genome Research Institute (NHGRI) GWAS catalog, https://www.genome.gov/26525384; qvalue R package, https://www.bioconductor.org/packages/release/bioc/html/qvalue.html; statsmodels Python package, http://statsmodels.sourceforge.net/. Code and data used for this manuscript are available on Github at https://github.com/mgymrek/estrs under the GPLv3 license.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank T. Lappalainen, A. Goren, T. Hashimoto and D. Zielinksi for useful comments and discussions. M.G. was supported by the National Defense Science and 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.), National Institute of Justice (NIJ) grant 2014-DN-BX-K089 (Y.E. and T.W.), and US National Institutes of Health (NIH) grants 1U01HG007037 (H.Z.), R01MH084703 (J.K.P.), R01HG006399 (A.L.P.), HG006696 (A.J.S.), DA033660 (A.J.S.) and MH097018 (A.J.S.) and by research grant 6-FY13-92 from the March of Dimes Foundation (A.J.S.).

AUTHOR CONTRIBUTIONS

M.G. and Y.E. conceived the study. M.G., T.W., H.Z., B.M. and Y.E. performed analyses. A.G. performed experimental work to generate high-coverage sequencing data for promoter STRs. S.G., M.J.D., A.L.P. and J.K.P. provided statistical input. A.J.S. contributed data and analyses. M.G., T.W. and Y.E. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
2. Moffatt, M.F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
3. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
4. Nica, A.C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6**, e1000895 (2010).
5. 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).
6. Ward, L.D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
7. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
8. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
9. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
10. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
11. Montgomery, S.B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23**, 749–761 (2013).
12. Wright, F.A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
13. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
14. Press, M.O., Carlson, K.D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* **30**, 504–512 (2014).
15. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
16. Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44**, 445–477 (2010).
17. Weber, J.L. & Wong, C. Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128 (1993).
18. Mirkin, S.M. Expandable DNA repeats and human disease. *Nature* **447**, 932–940 (2007).
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**, 315–320 (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. USA* **102**, 3800–3804 (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**, 2803–2809 (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**, 4069–4079 (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. USA* **101**, 3504–3509 (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**, 1988–1998 (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. USA* **98**, 8985–8990 (2001).
26. Weiser, J.N., Love, J.M. & Moxon, E.R. The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* **59**, 657–665 (1989).
27. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. & Verstrepen, K.J. Unstable tandem repeats in promoters confer transcriptional evolvability. *Science* **324**, 1213–1216 (2009).
28. Sureshkumar, S. *et al.* A genetic defect caused by a triplet repeat expansion in *Arabidopsis thaliana*. *Science* **323**, 1060–1063 (2009).
29. Hammock, E.A. & Young, L.J. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**, 1630–1634 (2005).
30. Yáñez-Cuna, J.O. *et al.* Dissection of thousands of cell type–specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.* **24**, 1147–1156 (2014).
31. Sawaya, S. *et al.* Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS ONE* **8**, e54710 (2013).
32. Bilgin Sonay, T. *et al.* Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res.* **25**, 1591–1599 (2015).
33. Borel, C. *et al.* Tandem repeat sequence variation as causative *cis*-eQTLs for protein-coding gene expression variation: the case of *CSTB*. *Hum. Mutat.* **33**, 1302–1309 (2012).
34. 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**, 13176–13180 (1999).
35. Rockman, M.V. & Wray, G.A. Abundant raw material for *cis*-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).
36. Shimajiri, S. *et al.* Shortened microsatellite $d(CA)_{21}$ sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett.* **455**, 70–74 (1999).
37. Warpeha, K.M. *et al.* Genotyping and functional analysis of a polymorphic $(CCTTT)_n$ repeat of *NOS2A* in diabetic retinopathy. *FASEB J.* **13**, 1825–1832 (1999).
38. Hui, J., Stangl, K., Lane, W.S. & Bindereif, A. HnRNP L stimulates splicing of the e*NOS* gene by binding to variable-length CA repeats. *Nat. Struct. Biol.* **10**, 33–37 (2003).
39. Sathasivam, K. *et al.* Aberrant splicing of *HTT* generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. USA* **110**, 2366–2370 (2013).
40. Grünewald, T.G. *et al.* Chimeric EWSR1-FLI1 regulates the Ewing sarcoma susceptibility gene *EGR2* via a GGAA microsatellite. *Nat. Genet.* **47**, 1073–1078 (2015).
41. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
42. Willems, T. *et al.* The landscape of human STR variation. *Genome Res.* **24**, 1894–1904 (2014).
43. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
44. Duyao, M. *et al.* Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* **4**, 387–392 (1993).
45. 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**, 301–304 (1992).
46. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
47. 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**, 13176–13180 (1999).
48. Stranger, B.E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
49. Payseur, B.A., Place, M. & Weber, J.L. Linkage disequilibrium between STRPs and SNPs across the human genome. *Am. J. Hum. Genet.* **82**, 1039–1050 (2008).
50. Sawaya, S., Jones, M. & Keller, M. Linkage disequilibrium between single nucleotide polymorphisms and hypermutable loci. *bioRxiv* doi:10.1101/020909 (2015).
51. 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).
52. Gusev, A. *et al.* Regulatory variants explain much more heritability than coding variants across 11 common diseases. *bioRxiv* doi:10.1101/004309 (2014).
53. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
54. Ioannidis, J.P. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).

55. Gaffney, D.J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
56. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
57. Trynka, G. *et al.* Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).
58. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
59. Zeng, H., Hashimoto, T., Kang, D.D. & Gifford, D.K. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* doi:10.1093/bioinformatics/btv565 (17 October 2015).
60. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
61. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
62. Döring, A. *et al.* SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. *Nat. Genet.* **40**, 430–436 (2008).
63. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* **40**, 437–442 (2008).
64. Wallace, C. *et al.* Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia. *Am. J. Hum. Genet.* **82**, 139–149 (2008).
65. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
66. Weber, J.L. & Broman, K.W. 7 Genotyping for human whole-genome scans: past, present, and future. *Adv. Genet.* **42**, 77–96 (2001).
67. Chaisson, M.J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
68. Bhatia, G. *et al.* Haplotypes of common SNPs can explain missing heritability of complex diseases. *bioRxiv* doi:10.1101/022418 (2015).

ONLINE METHODS

Genotype data sets. lobSTR genotypes were generated for the phase 1 individuals from the 1000 Genomes Project as described in ref. 42. 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. SNPs were converted to hg19 coordinates using the liftOver tool and filtered using PLINK[69] to include only individuals for whom both expression array data and STR calls were available. Throughout this manuscript, all coordinates and genomic data are referenced according to hg19.

Targeted sequencing of promoter region STRs. We applied a previously published method using capture and high-throughput sequencing[70] to sequence 2,472 STRs located in gene promoters (TSS ± 1 kb) in 120 HapMap individuals of European (58 CEU individuals) and African (62 YRI individuals) ancestry. Briefly, the method uses a custom NimbleGen EZ Capture system to enrich the genomic sequence flanking and sometimes including the target STRs to be genotyped before sequencing using an Illumina HiSeq 2000 instrument. We multiplexed 24 individuals per sequencing lane and used 100-bp single-end reads. We used lobSTR version 3.0.3 to genotype STRs in these samples.

Expression data sets. RNA sequencing data sets from 311 HapMap LCLs for which STR and SNP genotypes were also available were obtained from the gEUVADIS Consortium (EMBL-EBI ArrayExpress experiment E-GEUV-1). Raw FASTQ files containing paired-end 100-bp Illumina reads were downloaded from the European Bioinformatics Institute (EBI). The hg19 Ensembl transcriptome annotation was downloaded as a GTF file from the UCSC Genome Browser[71,72] ensGene table. The RNA sequencing reads were mapped to the Ensembl transcriptome using TopHat v2.0.7 (ref. 73) with default parameters. Gene expression levels were quantified using Cufflinks v2.0.2 (ref. 74) with default parameters and supplied with the GTF file for Ensembl reference version 71. Genes with median fragments per kilobase of transcript per million mapped reads (FPKM) values of 0 were removed, leaving 23,803 genes. We restricted analysis to protein-coding genes, including 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 LCLs from the EBI website (EMBL-EBI ArrayExpress experiment E-MTAB-264). These data sets contain two replicates each for 730 unrelated individuals from eight HapMap populations (YRI, CEU, CHB, JPT, GIH, MEX, MKK and LWK) and were generated as described by Stranger *et al.*[75]. Background-corrected and summarized probe set intensities (from Illumina software) contained values for 7,655 probes. Additionally, probes containing common SNPs were removed[76]. 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 population membership, gene expression heterogeneity and population structure (**Supplementary Note**). Adjusted expression values were used as input to the eSTR analysis. To restrict to STR loci with high-quality calls, we filtered the call set 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 fewer 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 a heterozygosity of at least 0.3. This left 80,980 STRs within 100 kb of a gene expressed in our LCL data set.

A linear model was used to test for association between normalized STR dosage and expression for each STR within 100 kb of a gene. Dosage was defined as the sum of the deviations in the STR allele lengths from the hg19 reference sequence. For example, if the hg19 reference for an STR is 20 bp and the two alleles called are 22 bp and 16 bp, the dosage is equal to (22 – 20) + (16 – 20) = –2 bp. STR genotypes were *z* score normalized to have a mean of 0 and a variance of 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:

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

where $\vec{y}_g = \left(y_{g,1}, y_{g,2}, \ldots, y_{g,n}\right)^T$, with $y_{g,i}$ being the normalized covariate-corrected expression of gene g in individual i and n being the number of individuals, α_g is the mean expression level of individuals homozygous for the reference allele, $\beta_{j,g}$ is the effect of the allelic dosage of STR locus j on gene g, $\vec{x}_j = \left(x_{j,1}, \ldots, x_{j,n}\right)^T$, with $x_{j,i}$ being the normalized allelic dosage of STR locus j in the ith individual, and $\vec{\varepsilon}_{j,g}$ is a random vector of length n whose entries are drawn from $N\left(0, \sigma^2_{\varepsilon,j,g}\right)$, with $\sigma^2_{\varepsilon,j,g}$ being 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: $\beta_{j,g} = 0$ versus H_1: $\beta_{j,g} \neq 0$ using a standard t test. We controlled for a gene-level FDR of 5%.

Controlling for gene-level false discovery rate. We controlled for a gene-level 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 Bonferroni 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 Bonferroni correction on a subset of genes. We then used this list of adjusted P values as input to the qvalue R package to determine all genes with an FDR of at most 5%.

Partitioning heritability using linear mixed models. For each gene, we used an LMM to partition heritability between the lead explanatory STR and other *cis* variants. We used a model of the form:

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

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

We used the GCTA program[77] to determine the restricted maximum-likelihood (REML) of estimates $\beta_{j,g}$ and $\sigma^2_{u_g}$. To obtain 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:

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

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

This equation shows the relationship:

$$\sigma^2_{\mathrm{p}} = h^2_{\mathrm{STR}} + h^2_b + \sigma^2_{\varepsilon}$$

where σ^2_{p} is the phenotypic variance, which is equal to 1; h^2_{STR} is the variance explained by the STR, which is equal to $\beta^2_{j,g}\mathrm{var}\left(\vec{x}_j\right) = \beta^2_{j,g}$ because

doi:10.1038/ng.3461

the STR genotypes were scaled to have a mean of 0 and a variance of 1; and h^2_b is the variance explained by biallelic variants in the *cis* region. h^2_b is approximately equal to $\sigma^2_{u_g}$ because 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 $\hat{h}^2_{STR} = E\left[\beta^2_{j,g}\right] = \hat{\beta}^2_{j,g} - SE^2$, 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 $\hat{\beta}_{j,g} \sim N\left(\beta_{j,g}, SE\right)$. We estimated the percent of phenotypic variance explained by biallelic markers as $\hat{h}^2_b$. Note that, 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 for 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 lead eSNP. We identified eSNPs 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 eSNPs after replacing the vector $\vec{x}_j$ with a vector of SNP genotypes (0, 1 or 2 reference alleles) that was *z* score normalized to have a mean of 0 and a variance of 1. To determine whether our eSTR signal was independent of the lead eSNP at each gene, we repeated association tests between STR dosages and expression levels while holding the genotype of the SNP with the most significant association to that gene constant. For this analysis, we identified all samples at each gene that were either homozygous reference or homozygous non-reference for the lead 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 fewer 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 directions of effect before and after conditioning on the lead SNP were more concordant than expected by chance.

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

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

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

where α_g is the mean expression value for the reference haplotype, $\vec{y}_g$ is a vector of expression values for gene *g*, $\beta_{\text{eSNP},g}$ is the effect of the eSNP on gene *g*, $\beta_{\text{eSTR},g}$ is the effect of the eSTR on gene *g*, $\vec{x}_{\text{eSNP},g}$ is a vector of genotypes for the lead eSNP for gene *g*, $\vec{x}_{\text{eSTR},g}$ is a vector of genotypes for the best eSTR for gene *g* and $\vec{\varepsilon}_{j,g}$ represents the residual term. A major caveat is that the eSNP data set has significantly more power to detect associations than the eSTR data set owing 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 performed significantly better than the SNP-only model. We obtained the ANOVA *P* value for each gene and used the qvalue package in R 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 1,000 random sets of STRs with matched distributions of distance to the nearest TSS. For each STR, we determined the mean phyloP score for a given window size centered on the STR. The *P* value given represents the percentage of random sets whose mean phyloP score was greater than the mean score for the observed eSTR set.

Enrichment of STRs and eSTRs in predicted enhancers. H3K27ac peak data produced by the ENCODE Project[7] were used to determine predicted enhancers in GM12878. Peaks were downloaded from the UCSC Genome Browser and converted to hg19 coordinates using the liftOver tool. Any peak overlapping with sequence within 3 kb of a TSS was removed to exclude promoter regions from the analysis.

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 than randomly chosen variants, naive enrichment tests of eSTRs versus randomly chosen control regions may return strong enrichments simply because of the proximity of eSTRs 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 lead 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 percentage of STRs overlapping each annotation. We used these null distributions to calculate empirical *P* values for the enrichment of eSTRs in each annotation.

Effects of eSTRs on modulating regulatory elements. One potential mechanism by which eSTRs may act is by modulating epigenetic properties. The GERV (Generative Evaluation of Regulatory Variants)[59] model predicts the results of chromatin immunoprecipitation and sequencing (ChIP-seq) experiments directly from genomic sequences and optional covariates such as DNase-seq data. We used the non-covariate version of this technique to assess the effect of STR variations on the occupancy of chromatin marks.

GERV builds on a *k*-mer–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 data set, we generated sequences representing each observed allele. We filtered out STRs with interruptions in the repeat motif because 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 GERV 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 randomly chosen negative-control loci. Control loci were chosen to match the distribution of repeat lengths and absolute signal for each mark in the reference genome. We used a Mann-Whitney rank test to compare the magnitudes of the slopes between the eSTR and control sets for each mark.

Overlap of eSTR and GWAS genes. Aggregate results for seven common diseases (rheumatoid arthritis, Crohn's disease, type 1 diabetes, type 2 diabetes, blood pressure, bipolar disorder and coronary artery disease) were downloaded from the NHGRI GWAS catalog, accessed on 12 June 2015. Relevant genes were selected from the columns "Reported Gene(s)" and "Mapped_Gene". To generate a null distribution, we chose 1,000 sets of randomly selected genes matched to eSTR-associated genes on the basis of expression in LCLs (difference in FPKM <10) and *cis* heritability (difference in variance explained by *cis* biallelic variants <5%). We compared the overlap of GWAS genes with eSTR genes versus the 1,000 control sets to determine an empirical *P* value.

eSTR associations with human traits. To generate STR genotypes for each of the individuals in the UK10K TwinsUK data set, we ran lobSTR v2.0.3 on each BAM file using the options fft-windowsize = 16, fft-window-step = 4 and bwaq = 15. The resulting BAM files were analyzed using v2.0.3 of the lobSTR allelotyper with default options, resulting in STR genotypes for 1,685 individuals.

We then performed an association test between each STR and each phenotype. To control for population structure, we adjusted STR dosages and phenotypes for the top ten ancestry principal components based on common SNPs (MAF 5%) after LD pruning. Principal components were computed using EIGENSTRAT[78] v5.0.1. Phenotypes were further adjusted for the age at which the phenotype was measured. Association tests were performed between the adjusted dosages and the quantile-normalized adjusted phenotypes.

We were able to analyze the TwinsUK cohort for the following 38 phenotypes (in parentheses, the PMID reference given by TwinsUK to describe the phenotype measurement procedure): albumin (19209234), alkaline phosphatase (19209234), apolipoprotein A-I (15379757), apolipoprotein B (15379757), bicarbonate, bilirubin (19209234), body mass index, creatinine (11017953), diastolic blood pressure (16249458), heart rate (19587794), FEV_1 (17989158), FEV_1/FVC ratio (17989158), FVC (17989158), γ-glutamyl transpeptidase (19209234), glucose (19209234), high-density lipoprotein (19016618), standing height (17559308), hemoglobin (19862010), hip circumference (17228025), homocysteine (18280483), C-reactive protein (21300955), insulin (16402267), mean corpuscular volume (19862010), packed cell volume (10607722), phosphate (12193151), platelet count (19221038), red blood cell count (19820697), sodium (18179892), systolic blood pressure (16249458), total cholesterol (19820914), triglycerides (15379757), urea (18179892), uric acid (19209234), waist circumference (17228025), white blood cell count (19820697), weight (17016694) and waist-hip ratio.

We then examined the association in the 666 eSTR loci for which the eSTR significantly improved the variance in gene expression when combined with the lead eSNP (nominal ANOVA $P < 0.05$). Of these eSTRs, 499 were genotyped in >1,000 participants. For each phenotype, q values were calculated by adjusting the P values using the Benjamini-Hochberg procedure. Only hits with a q value <0.1 are reported.

69. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
70. Guilmatre, A., Highnam, G., Borel, C., Mittelman, D. & Sharp, A.J. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum. Mutat.* **34**, 1304–1311 (2013).
71. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014).
72. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
73. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
74. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
75. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
76. 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).
77. 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).
78. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

doi:10.1038/ng.3461