

Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs

Christopher D. Brown^{1*}, Lara M. Mangravite², Barbara E. Engelhardt^{3,4,5*}

1 Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **2** Sage Bionetworks, Seattle, Washington, United States of America, **3** Biostatistics & Bioinformatics Department, Duke University, Durham, North Carolina, United States of America, **4** Department of Statistical Science, Duke University, Durham, North Carolina, United States of America, **5** Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina, United States of America

Abstract

Genetic variants in cis-regulatory elements or trans-acting regulators frequently influence the quantity and spatiotemporal distribution of gene transcription. Recent interest in expression quantitative trait locus (eQTL) mapping has paralleled the adoption of genome-wide association studies (GWAS) for the analysis of complex traits and disease in humans. Under the hypothesis that many GWAS associations tag non-coding SNPs with small effects, and that these SNPs exert phenotypic control by modifying gene expression, it has become common to interpret GWAS associations using eQTL data. To fully exploit the mechanistic interpretability of eQTL-GWAS comparisons, an improved understanding of the genetic architecture and causal mechanisms of cell type specificity of eQTLs is required. We address this need by performing an eQTL analysis in three parts: first we identified eQTLs from eleven studies on seven cell types; then we integrated eQTL data with cis-regulatory element (CRE) data from the ENCODE project; finally we built a set of classifiers to predict the cell type specificity of eQTLs. The cell type specificity of eQTLs is associated with eQTL SNP overlap with hundreds of cell type specific CRE classes, including enhancer, promoter, and repressive chromatin marks, regions of open chromatin, and many classes of DNA binding proteins. These associations provide insight into the molecular mechanisms generating the cell type specificity of eQTLs and the mode of regulation of corresponding eQTLs. Using a random forest classifier with cell specific CRE-SNP overlap as features, we demonstrate the feasibility of predicting the cell type specificity of eQTLs. We then demonstrate that CREs from a trait-associated cell type can be used to annotate GWAS associations in the absence of eQTL data for that cell type. We anticipate that such integrative, predictive modeling of cell specificity will improve our ability to understand the mechanistic basis of human complex phenotypic variation.

Citation: Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS Genet* 9(8): e1003649. doi:10.1371/journal.pgen.1003649

Editor: Greg Gibson, Georgia Institute of Technology, United States of America

Received: October 4, 2012; **Accepted:** June 4, 2013; **Published:** August 1, 2013

Copyright: © 2013 Brown et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: BEE was funded through NIH grant HG006265. Generation of the Cholesterol and Pharmacogenetics (CAP) data and LMM were funded by a grant from the National Institutes of Health, U01 HL69757. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chrbro@penn.edu (CDB); barbara.engelhardt@duke.edu (BEE)

Introduction

The precise spatial and temporal control of gene transcription is critical for biological processes, as evidenced by the causal role of gene expression perturbation in many human diseases [1–3]. Gene expression is controlled by regulatory proteins, RNAs, and the cell type specific cis-regulatory elements with which they interact. Genetic variation within cis-regulatory elements (CREs, e.g., transcription promoters, enhancers, or insulators) can affect gene expression in a cell type specific manner. An extensive body of work, performed in a variety of cell types in both humans and model organisms, has demonstrated that genetic variants that impact gene expression, or expression quantitative trait loci (eQTLs), are common and exist in both *cis* (local) and *trans* (over long genetic distances) [3–6]. Over 85% of genotype-phenotype associations found in genome-wide association studies (GWAS) are with non-coding single nucleotide polymorphisms (SNPs), making their mechanistic interpretation more challenging. It is possible that these associated SNPs tag causal coding SNPs; however,

numerous compelling lines of evidence [2,7–11] demonstrate that regulatory SNPs have causal roles in many complex human phenotypes. This is further supported by the finding that GWAS associations are enriched within DNase I hypersensitive (DHS) sites [12] and eQTL SNPs [13,14], and by several elegant GWAS follow up studies that have mechanistically tied disease associations with SNPs that cause the misregulation of gene expression [15,16].

Although eQTLs are increasingly used to provide mechanistic interpretations for human disease associations, the cell type specificity of eQTLs presents a problem. Because the cell type from which a given physiological phenotype arises may not be known, and because eQTL data exist for a limited number of cell types, it is critical to quantify and understand the mechanisms generating cell type specific eQTLs. For example, if a GWAS identifies a set of SNPs associated with risk of type II diabetes, the researcher must choose a target cell type to develop a mechanistic model of the molecular phenotype that causes the gross physiological change. One can imagine that the relevant cell type might be adipose tissue, liver, pancreas, or another hormone-regulating

Author Summary

When interpreting genome-wide association studies showing that specific genetic variants are associated with disease risk, scientists look for a link between the genetic variant and a biological mechanism behind that disease. One functional mechanism is that the genetic variant may influence gene transcription via a co-localized genomic regulatory element, such as a transcription factor binding site within an open chromatin region. Often this type of regulation occurs in some cell types but not others. In this study, we look across eleven gene expression studies with seven cell types and consider how genetic transcription regulators, or eQTLs, replicate within and between cell types. We identify pervasive allelic heterogeneity, or transcriptional control of a single gene by multiple, independent eQTLs. We integrate extensive data on cell type specific regulatory elements from ENCODE to identify general methods of transcription regulation through enrichment of eQTLs within regulatory elements. We also build a classifier to predict eQTL replication across cell types. The results in this paper present a path to an integrative, predictive approach to improve our ability to understand the mechanistic basis of human phenotypic variation.

tissue. Furthermore, if the GWAS SNP produces a molecular phenotype (i.e., is an eQTL) in lymphoblastoid cell lines (*LCLs*), it is not necessarily the case that the SNP will generate a similar molecular phenotype in the cell type of interest. Furthermore, there are many examples of cell types with particular relevance to common diseases, for example dopaminergic neurons and Parkinson's disease, that lack comprehensive eQTL data or catalogs of CREs. The utility of eQTLs for complex trait interpretation will therefore be improved by a more thorough annotation of their cell type specificity.

While several studies have quantified the reproducibility of eQTLs within or between cell types derived from the same or different individuals [17–28] the determinants of eQTL cell specificity are still largely unknown. We address this need in this study by analyzing cell specific eQTLs collected from eleven studies performed in seven different cell types and by integrating these data with cell specific CRE data to mechanistically interpret cell specific eQTLs. We used Bayesian regression models to identify all cis-linked SNPs that are independently associated with each gene expression trait in each study. In an effort to identify the functional determinants of eQTL cell specificity, we quantified the associations between eQTL SNPs and 526 CRE data sets, many of which were derived from the cell types used in eQTL discovery and are known to function in a cell type specific manner (e.g., transcription factor binding sites (*TFBSs*), DHS sites). We further considered the relationship between eQTL SNP-CRE overlap and the cell type specificity of eQTL replication. Lastly, we built a series of classifiers to predict the cell type specificity of eQTLs in the absence of additional gene expression data and to predict the function of GWAS SNPs with phenotypes relevant to cell types lacking eQTL data. We believe these predictive models will facilitate more substantial mechanistic analyses of individual SNPs by enabling the integration of disease genetics and regulatory SNPs with the thousands of genomic data sets that have been produced by projects like ENCODE [29,30].

Results

A uniform analysis of cis-eQTLs across seven cell types

In an effort to comprehensively characterize eQTL reproducibility within and between different cell types, we gathered publicly available data sets that included both gene expression and genotype data. This collection included eleven studies from seven unique cell types (Table 1) [17,26,31–33]. To ensure the data from each eQTL study were comparable, we uniformly processed all raw data by developing a standardized analysis pipeline that was designed to marginalize the effect of study design differences on the identified eQTLs (see Methods). Genotype data, regardless of array type, were subjected to standard quality control filters. Missing and unobserved genotypes were subsequently imputed to the SNPs in the HapMap phase 2 CEPH panel (3,907,239 SNPs) using BIMBAM [34,35]. Each gene expression array was uniformly re-annotated; probe sequences were aligned to the human reference genome (hg18) and to RefSeq gene models. Within each array platform, multiple probes mapping to a single gene were clustered as in previous work [26]. Only uniquely aligned probes that did not overlap known, common polymorphisms were included in our analysis.

We chose to control for the confounding effects of both known covariates and unknown factors by removing the effects of principal components (*PCs*; Figure S1, Table S1) [36,37]. Given that a diverse set of demographic (e.g., age, sex), environmental (e.g., BMI, drug use), and technical (e.g., post-mortem interval, array batch, ozone levels, identity of the technician who handled the arrays) variables are known to be associated with gene expression measurements and to confound eQTL ascertainment [26,36,37] we felt it was critical to control for these effects in the most consistent way possible prior to eQTL mapping. Across the diverse set of studies examined here, the covariate annotation ranges from non-existent to detailed. To address this non-uniformity, we analyzed each data set with the same approach, irrespective of covariate annotation. Multiple independent studies demonstrate the effectiveness of controlling for latent variables with respect to eQTL ascertainment; indeed, controlling for *PCs* substantially increases power to detect cis-eQTLs within these studies [26]. Importantly, it has also been demonstrated that each of these eQTL discoveries is also more likely to replicate across studies [26].

We projected residual expression variation to the quantiles of a standard normal distribution to control for outliers, and we used these projected values as the quantitative traits for association mapping, which was performed in each study set using the same HapMap phase 2 CEPH SNP panel. We evaluated evidence for gene expression-genotype associations in terms of Bayes factors (*BFs*) using BIMBAM [34,38], as *BFs* have been shown to be more robust to SNPs with small minor allele frequencies (*MAF*) than *p*-values [34,39].

We identified, for each gene expression trait, the most highly associated SNP within each local linkage disequilibrium (*LD*) block. We tested the independence of each SNP by multivariate regression (Figure S2) and took the union of the independently associated SNPs for each gene. We refer to, for example, the first and second most significant, independently associated SNPs as *primary* and *secondary* SNPs, respectively, and we refer to the set of primary SNPs as first *tier*, or tier 1, extending in the straightforward way through tier 4. We do not consider tiers beyond the fourth tier because of lack of statistical power. For each study, and within each tier, we independently estimated false discovery rates

Table 1. Study Information.

Study label	TLA	Tissue	N	N genes	PMID	Accession	Platform	Genotype
CAP_LCL	CPL	LCLs	480	18718	20339536	GSE36868	GPL6883-5509	ILMN 310K & ILMN QUAD
Stranger_LCL	STL	LCLs	210	15752	17289997	GSE6536	GPL2507	NA
Harvard_cerebellum	HCE	cerebellum	540	18263	NA	syn4505	GPL4372	GPL14932
Harvard_prefrontal_cortex	HPC	prefrontal cortex	678	18257	NA	syn4505	GPL4372	GPL14932
Harvard_visual_cortex	HVC	visual cortex	463	18263	NA	syn4505	GPL4372	GPL14932
GenCord_fibroblast	GCF	blood fibroblasts	83	16691	19644074	GSE17080	GPL6884	GPL6982
GenCord_LCL	GCL	LCLs	85	16691	19644074	GSE17080	GPL6884	GPL6982
GenCord_tcell	GCT	blood t cells	85	16691	19644074	GSE17080	GPL6884	GPL6982
UChicago_liver	CLI	liver	206	16236	21637794	GSE26106	GPL4133	GPL8887
Merck_liver	MLI	liver	266	18234	18462017	GSE9588	GPL4372	GPL3720&GPL3718&GPL6987
Myers_brain	MBR	brain	193	11707	17982457	GSE8919	GPL2700	GPL3720&GPL3718

Accession numbers are from the GEO database when prefixed with *GSE* and from the Synapse database when prefixed with *syn*. *Study label* is the name used to refer to the study throughout the paper. *TLA* is the three letter acronym used to reference the study in figures. *CAP* stands for the Cholesterol and Pharmacogenetics Trial [55,74].

doi:10.1371/journal.pgen.1003649.t001

(*FDRs*) by permutation. Although we computed a *BF* for every SNP-gene pair, we limit our subsequent analysis to cis-linked SNPs, or SNPs within 1 Mb of the transcription start site (*TSS*) or transcription end site (*TES*) of a gene. While we have standardized analysis and reporting across studies, we have not considered the scope of differences in eQTL discovery based on alternative data analysis pipelines.

Across these studies, we observe between 585 and 5433 genes with eQTLs ($FDR \leq 5\%$), corresponding to $\log_{10} BF$ thresholds between 2.70 and 3.86 (Figures 1A–C, Table 2, Table S2). As expected, studies with larger sample sizes and replicate gene expression measurements identified more eQTLs at a given *FDR* threshold (Figure 1D; $p < 3.95 \times 10^{-6}$ and $p < 1.58 \times 10^{-4}$, respectively, by multivariate ANOVA). Indeed, across the eleven studies analyzed, >95% of the variance in the proportion of genes with eQTLs can be explained by sample size and expression replication. The per study effect size distribution is also consistent with the expectation that larger studies identify eQTLs with smaller effect sizes (Figures S3, S4). We expect that future eQTL studies with larger sample sizes (even from previously examined cell types) will identify additional eQTLs with smaller effects. We find that, despite study heterogeneity, the relationship between *BF* and *FDR* is quite uniform across studies (Figure 1A). As demonstrated in previous studies [40,41], eQTL SNPs are highly enriched at the transcription start site (*TSS*) of the associated gene (Figure 1F).

Across all eleven studies, 29% (3225 of 11081) of eQTL associated genes are independently associated with at least two SNPs in at least one study ($FDR \leq 5\%$; Figures 1C and 1E, Figures S3, S4). Within each study, the fraction of eQTL-associated genes with two or more independently associated SNPs ranges from 3–22% ($FDR \leq 5\%$). Our search for *allelic heterogeneity*, or multiple SNPs in the same locus that are independently associated with a single trait, appears to be power-limited and our estimates of its frequency should be taken as a lower bound; larger sample sizes will identify additional heterogeneity (Figure 1E), as the relationship appears almost identical to the linear relationship between genes with eQTLs and sample size (Figure 1D). As with tier 1 eQTL discovery, sample size and replicate expression measurements are significantly associated with the fraction of genes with an associated eQTL SNP exhibiting allelic heterogeneity (AH);

$p < 4.34 \times 10^{-6}$ and $p < 8.63 \times 10^{-4}$, respectively, by multivariate ANOVA). Tier 2 eQTL SNPs reside significantly further from the associated gene *TSS* than tier 1 eQTL SNPs (Figure 1D). For example, in the CAP_LCL eQTL data set, the median absolute distances between the *TSS* and tier 1 and tier 2 eQTL SNPs are 64 and 165 kb, respectively (Wilcoxon signed rank test $p < 2.2 \times 10^{-16}$).

Cis-eQTL replication within and between cell types

We next investigated the cell type specificity of eQTLs, comparing eQTLs both within and between cell types. *Cell type specific eQTLs* are defined here as eQTL SNPs that replicate across studies of the same cell type but fail to replicate across studies of different cell types. Given the broad array of technical and biological factors associated with the reproducibility of eQTLs [21,22,26,36], our analysis of eQTL replication focused on three specific comparison trios:

1. CAP_LCL versus Stranger_LCL and Merck_liver
2. UChicago_liver versus Merck_liver and Stranger_LCL
3. Harvard_cerebellum versus Myers_brain and Stranger_LCL.

Each trio of comparisons enabled the simultaneous quantification of within and between cell type eQTL reproducibility. Each of the six studies above used different expression platforms and were composed entirely of independent samples. Although post hoc comparisons between heterogeneous studies will have limitations, we found there to be substantial scientific merit to using the full breadth of data available while being completely forthcoming about both our comparison methods and those limitations. We note that, despite this heterogeneity, the conclusions highlighted below are largely independent of the particular discovery cohort, replication cohort, or cell type (Figure 2, Figures S5, S6). These specific trios were chosen for comparative analysis based on the following criteria: (i) two or more studies in our analysis included only these three cell types; (ii) of the studies that included these three cell types, we chose those with the largest sample size, and (iii) LCLs and liver are valuable in this comparative context because of the substantial amount of ENCODE data available for GM12878 and HepG2 cells. We note that the Myers_brain study includes samples from several different brain cell types, a minority

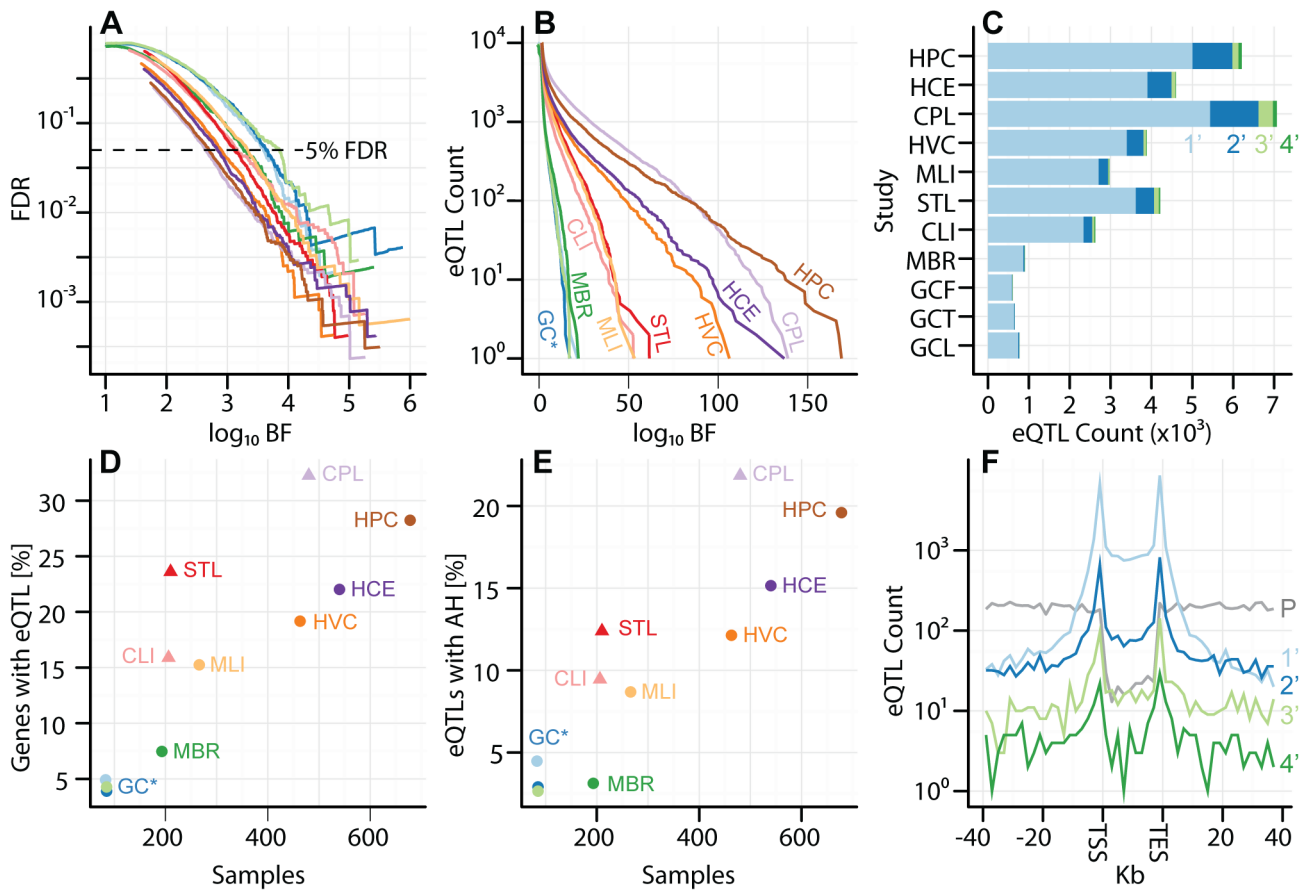


Figure 1. Uniform analysis of multi-cell type eQTL data sets. Studies are labeled by their acronym from Table 1. (A) Plot of $\log_{10} FDR$ (y-axis) as a function of $\log_{10} BF$ (x-axis), for each study as a separate line of a different color, as indicated in panels B, D, and E. Dashed line represents $FDR \leq 5\%$. (B) Plot of \log_{10} eQTL counts as function of $\log_{10} BF$, for all studies. (C) eQTL count (x-axis) by tier, for tiers 1–4 (light blue, dark blue, light green, and dark green, respectively), with separate bars for each study (y-axis). (D) Fraction of genes with a significant eQTL SNP (y-axis; thresholded at $FDR \leq 5\%$), as function of sample size (x-axis). Each study is plotted in a distinct color, as indicated with labels. Studies with replicate expression measurements are depicted as triangles, studies without as circles. (E) Fraction of genes with a significant eQTL that have more than one independently associated SNP (y-axis; thresholded at $FDR \leq 5\%$), as a function of sample size (x-axis). Each study is plotted in a distinct color. Studies with replicate expression measurements are depicted as triangles, studies without as circles. (F) Histogram of eQTL counts by tier (y-axis; colors as in panel C), summed across studies, as a function of their distance to the gene transcription start and end sites (x-axis; gene split into 10 bins). P (grey) line depicts the counts of first tier eQTL SNPs from a permutation, to illustrate the background distribution of tested SNPs.
doi:10.1371/journal.pgen.1003649.g001

Table 2. Study-specific cis-eQTLs and $\log_{10} BF$ cutoff values for 1%, 5%, 10%, and 20% FDRs.

Study	Tissue	1%		5%		10%		20%	
		$\log_{10} BF$	eQTLs	$\log_{10} BF$	eQTLs	$\log_{10} BF$	eQTLs	$\log_{10} BF$	eQTLs
GenCord	fibroblasts	3.16	566	3.58	772	2.35	916	1.99	1292
GenCord	t cells	3.40	450	2.85	596	2.47	749	2.07	1076
GenCord	LCLs	3.37	441	2.72	649	2.46	782	2.06	1111
Harvard	cerebellum	3.18	3367	2.59	4065	2.29	4595	1.95	5547
Harvard	prefrontal cortex	3.21	4331	2.51	5189	2.24	5775	1.88	6833
Harvard	visual cortex	3.24	2872	2.63	3469	2.29	4095	1.96	5040
Merck	liver	3.52	2333	2.90	2828	2.55	3272	2.21	4078
Myers	brain	3.17	688	2.61	888	2.30	1076	1.99	1408
UChicago	liver	3.29	1951	2.60	2543	2.25	3005	1.93	3687
Stranger	LCLs	3.32	3147	2.67	3759	2.37	4167	2.06	4695
CAP	LCLs	3.09	5094	2.42	5810	2.14	6335	1.82	7235

The cutoff values for each FDR were determined via permutation; see Methods for details.
doi:10.1371/journal.pgen.1003649.t002

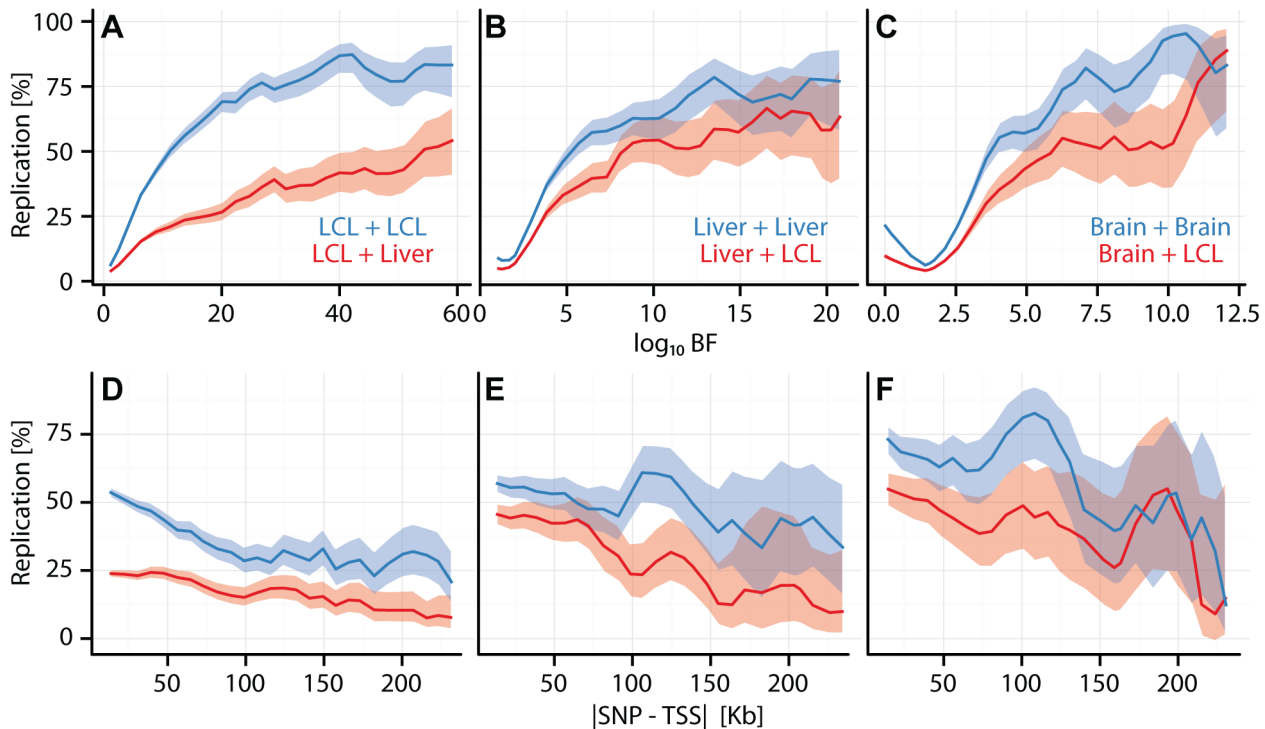


Figure 2. Cell type specific eQTL replication frequencies. (A, B, C) eQTL replication frequency (y-axis) as a function of discovery significance (x-axis: $\log_{10} BF$). SNPs were grouped into 30 equally spaced bins by BF. (D, E, F) eQTL replication frequency (y-axis; thresholded at 5% *FDR*) as a function of SNP position ($|SNP - TSS|$) (x-axis). Cis-eQTL SNPs within 250 kb of the TSS were grouped into 30 equally spaced bins. (A, D) Replication frequencies for CAP_LCL eQTLs in Stranger_LCLs (blue) and Merck_liver (red). (B, E) Replication frequencies for UChicago_liver eQTLs in Merck_liver (blue) and Stranger_LCL (red). (C, F) Replication frequencies for Myers_brain eQTLs in Harvard_cerebellum (blue) and Stranger_LCL (red). In all panels, bold lines depict percentage of SNP-gene pairs with $\log_{10} BF \geq 1$ per bin, and ribbons depict 95% confidence interval. doi:10.1371/journal.pgen.1003649.g002

of which were cerebellum, implying that the cell type matching in comparison 3 above is inexact.

Consistent with previous observations[17,18,24], cis-eQTLs are more likely to replicate across studies within the same cell type than they are to replicate between different cell types (e.g., in CAP_LCL: McNemar's test $p < 2.2 \times 10^{-16}$). Beyond the trios listed above (Figures S5, S6), replication frequencies vary broadly. Two variables have large effects on replication: sample size for the replication cohort (which is well correlated with statistical power), and genetic independence of the samples (i.e., whether the two cell types were derived from the same or different individuals).

Within a given comparison, eQTL replication frequency is associated with a number of factors. For example, within and between cell type replication of CAP_LCL eQTLs is positively associated with discovery significance (within: $p < 2.2 \times 10^{-16}$, between: $p < 1.78 \times 10^{-11}$, quantified by multivariate logistic regression, Equation (3)) and negatively associated with absolute distance to the TSS (Figure S7; within: $p < 2.2 \times 10^{-16}$, between: $p < 2.94 \times 10^{-6}$) and with eQTL tier (within: $p < 2.49 \times 10^{-11}$, between: $p < 4.06 \times 10^{-11}$), while differences in allele frequency across studies does not have a major effect (Figure S8). We found that as the level of discovery significance increases, the likelihood that the eQTL replicates in both matched and unmatched cell types also increases, implying that cell type specific eQTLs tend to have smaller effects (Figure S9). After controlling for discovery significance, effect size is not significantly associated with replication frequency. Similar to previous reports (see Figure S6 from [26]), alternative post hoc replication metrics (e.g., correlation of effect sizes) produce qualitatively similar results. To assess

the effects of model parameters and post hoc comparison thresholds, we applied a bivariate Bayesian regression model to a subset of our studies (Figure S10; see Methods). The results of these more formal bivariate analyses are qualitatively similar to those obtained from post hoc comparisons: the fraction of cell type specific cis-eQTLs decreases with increasing discovery significance and cell specific eQTL SNPs reside further from the TSS.

eQTL SNP tier is significantly associated with eQTL replication frequencies; tier 1 eQTL SNPs are more reproducible than additional independently associated SNPs (Figure S11; e.g., CAP_LCLs: Fisher's exact test $p < 2.2 \times 10^{-16}$). Additionally, first tier eQTL SNPs are significantly less likely to be cell type specific, relative to additional independently associated SNPs (e.g., CAP_LCLs: Fisher's exact test $p < 1.23 \times 10^{-5}$). Therefore, for any given gene, the first tier eQTL SNP is more likely to be TSS-proximal, of large effect, and observed in additional cell types, as compared to additional independent eQTL SNPs, which are more likely to be specific to the discovery cell type, have smaller effect sizes, and reside further from the TSS.

eQTL SNPs are associated with many classes of cis-regulatory elements

We next sought to investigate the biological characteristics associated with the reproducibility and cell specificity of eQTLs. To do this, we quantified the overlap between cis-eQTL SNPs and 526 genomic features associated with functional cis-regulatory elements (CREs), including DHS sites, chromatin marks, and binding sites for transcription factors and other DNA associated regulatory proteins (see Table S3 for full list of data sets). We

categorized regions of open or activating chromatin, and regions of transcription factor or DNA protein binding as *activating* CREs, and regions of repetitive, repressive, or heterochromatic chromatin domains as *repressive* CREs, to draw a contrast between genomic regions where transcription factor binding is frequent and regions where it is discouraged or unlikely. We focused analyses of LCL eQTL SNPs on 166 CRE data sets produced in LCLs (primarily GM12878) and analyses of liver eQTLs on 150 CRE data sets produced in HepG2 cells, a well-characterized, if imperfect, proxy for hepatocyte biology. We note that the quantification of eQTL SNP-CRE overlap enrichments is inherently conservative, given that the boundaries of most genomically defined CRE types are imprecise and that eQTL SNPs are typically tag SNPs, rather than the exact causal variants.

We quantified the enrichment or depletion of eQTL SNPs, relative to the full set of CEU HapMap phase 2 SNPs tested for eQTL associations, within each class of CRE by multivariate logistic regression, controlling for the SNP to TSS distance and the expression level of the associated gene (see Methods and Equation (2)). Consistent with the hypothesis that many eQTL SNPs exert their effect by modifying the biochemical function of CREs, cis-eQTLs are known to be enriched for overlaps with several classes of CREs, including DHS sites (Figure 3A and [42,43]). Moreover, cis-eQTLs have been shown to be depleted within regions in which a CTCF binding site lies between the eQTL SNP and the target gene TSS (Figure 4G and [42,43]).

We further extend these observations by demonstrating that LCL eQTL SNPs are associated ($p < 0.05$, quantified by Equation (2)) with 135/166 LCL derived CRE data sets, liver eQTL SNPs are associated with 79/150 HepG2 derived CRE data sets, and cerebellum eQTL SNPs are associated with 1/1 cerebellum derived CRE data set (Figures S12, S13). Almost universally, eQTL SNPs are enriched within regions of activating CREs (Figure 3; Tables S4, S5, S6) and depleted within repressive CREs (Figure 4, Figures S12, S13, S14, Tables S4, S5, S6). LCL eQTL SNPs are enriched within 134/141 activating CREs while being depleted within 18/23 repressive CREs (Fisher's exact test $p < 2.93 \times 10^{-14}$). Liver eQTL SNPs display a similar enrichment within activating CREs and depletion within repressive CREs (Fisher's exact test $p < 4.02 \times 10^{-10}$).

The pattern of eQTL SNP-CRE enrichment displays significant spatial structure and is typically consistent with the known biology of the CRE (Figures 3–4, Figure S14). For example, eQTL-CRE enrichment peaks immediately adjacent to the TSS for several classes of activating CREs, including H3K4me3 and H2A.Z. Alternatively, eQTL enrichment increases throughout the gene body within H3K36me3 domains, and is more uniformly distributed within H3K4me1 domains. In contrast, we find that eQTL SNP overlap with heterochromatin, repressive chromatin, or repetitive regions is typically most highly depleted through the gene body (e.g., Figures 4A and D). Similarly, we find that intervening CTCF sites are most depleted immediately upstream of the gene TSS, but the decay of this depletion is intriguingly asymmetrical about the TSS (Figure 4G).

Tiers 2–4 eQTL SNPs are themselves also associated with numerous CRE classes. For example, primary and secondary CAP_LCL eQTL SNPs are associated with 134/166 and 100/166 LCL CRE classes, respectively (Table S7; multivariate logistic regression (Equation (4)), controlling for distance to TSS and gene expression levels, $p < 0.05$). Interestingly, CTCF binding sites are significantly enriched *between* primary and secondary eQTL SNPs (Figure S15; multivariate logistic regression (Equation (5)), controlling for inter-SNP distance, SNP-TSS distance, and the presence of intervening TSSs and recombination hotspots,

$p < 1.95 \times 10^{-11}$). Independently associated primary and secondary eQTL SNPs separated by less than 20 kb are more than twice as likely to have a CTCF binding site as similarly spaced background cis-SNPs (55.7% versus 22.6%). CTCF binding sites are enriched between alternative promoters in human [44] and *Drosophila melanogaster* [45], supporting the hypothesis that CTCF binding sites frequently demarcate independent CREs for the same gene. These observations, combined with the replication rates of eQTL SNPs in tiers 2–4 (Figure S11), suggest that eQTL SNPs independently associated with the same gene frequently tag SNPs affecting the biochemical function of distinct CREs that, in turn, independently regulate transcription, rather than SNPs tagging the same causal regulatory variant.

eQTL-regulatory element overlap is frequently cell type specific

Previous investigations have suggested several plausible mechanisms underlying the cell type specificity of eQTLs [18,24,27]. For example, given the known cell type specificity of regulatory protein binding sites and local chromatin environment [29,46,47], it is plausible that a SNP that disrupts a TFBS would have different downstream effects if it were found within a region of open, activating chromatin as opposed to a region of repressive chromatin. Although the current resolution of CRE and tag eQTL SNP data sets make this hypothesis difficult to test directly for individual SNPs, we sought to quantify the frequency, in aggregate, with which eQTL SNPs overlap CREs that differ between cell types using our identified eQTLs and cell specific CRE data from the ENCODE project.

We assessed the cell type specificity of eQTL-CRE overlap by quantifying the fraction of eQTL SNPs overlapping a CRE derived from the same cell type, relative to the fraction overlapping a CRE derived from a second cell type. When the frequency of overlap between an eQTL SNP set and a matched and unmatched cell type CRE is significantly different (McNemar's test $p < 0.05$), we consider the overlap to be *cell specific*, and the eQTL SNPs to be differentially represented in that CRE type. LCL and liver eQTL SNPs were tested for overlap with 105 CRE data sets available from both LCLs and HepG2 cells, while cerebellum eQTLs were tested for overlap with a single pair of cerebellum and LCL CRE data sets. We find that eQTL SNP-CRE overlap is frequently cell specific (see Figures 3–5 for examples, Tables S4, S5, S6 for full results). The observed specificity of eQTL-CRE overlap is not dependent on analysis thresholds and is recapitulated in bivariate analyses of eQTL data sets (Figure S10).

LCL and liver eQTL SNPs are differentially represented (McNemar's test $p < 0.05$) in 97/105 and 86/105 CRE data types, respectively; moreover, cerebellum eQTLs are over-represented in cerebellum derived DHS sites relative to LCL DHS sites (Table S6). For example, 8.2% and 4.2% of CAP_LCL eQTL SNPs overlap LCL and HepG2 derived chromHMM strong enhancers, respectively (McNemar's test $p < 2.2 \times 10^{-16}$). Notably, the canonical biochemical function of the CRE class is predictive of the pattern of cell type specific eQTL-CRE overlap. eQTL SNPs are more likely to overlap activating CREs and less likely to overlap repressive CREs derived from the same cell type as the eQTL discovery data (repeated measures logistic regression $p < 4.63 \times 10^{-3}$; Tables S4, S5, S6). Consistent with previous observations [44,45,48], the proportion of eQTL SNP - TSS pairs with intervening insulators is remarkably consistent across cell types, suggesting that CTCF binding sites do not substantially affect cell-specific eQTL function (Figure 4I, Figure S16, Tables S4, S5, S6).

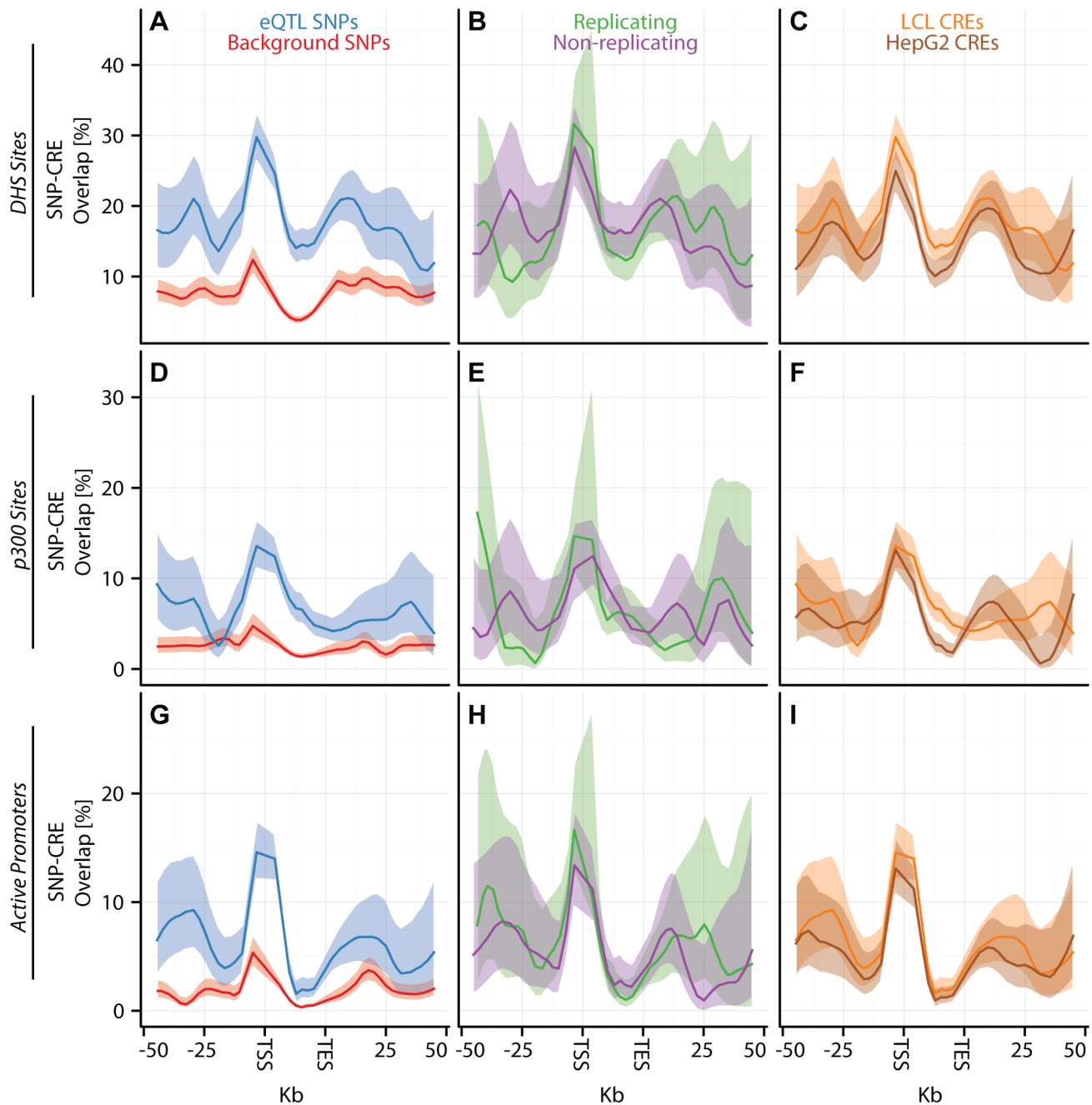


Figure 3. eQTL SNPs are enriched within activating cis-regulatory elements. (A–I) CAP_LCL eQTL SNP ($FDR \leq 5\%$) overlap with predicted cis-regulatory elements. Each row of panels depicts overlap with distinct CRE data sets: (A–C) DNase hypersensitive sites, (D–F) p300 binding sites, (G–I) chromHMM predicted active promoters. In each panel, SNPs are grouped into 25 equally spaced bins within the 50 kb upstream and downstream of the TSS and TES, and 10 bins between the TSS and TES. Each bin is plotted along the x-axis. Bold lines depict the percentage, per bin, of SNPs overlapping the CRE class, ribbons depict 95% confidence interval. Each column of panels depicts a distinct SNP set contrast. (A,D,G) Observed eQTL SNPs (blue) and randomly drawn cis-linked SNPs at expressed genes (red). (B,E,H) eQTL SNPs that replicate in Stranger_LCL ($\log_{10} BF \geq 1$) (green) and SNPs that fail to replicate (purple). (C,F,I) CAP_LCL eQTL SNP overlap with CREs derived from the LCL line GM12878 (orange) and HepG2 cells (brown). doi:10.1371/journal.pgen.1003649.g003

Genetic architecture of cell type specific eQTLs

We next examined the hypothesis that the CRE landscape is a major determinant of eQTL specificity, and found that eQTL SNPs that overlap cell type specific CREs are more likely to be cell type specific than are eQTL SNPs that overlap shared CREs (Fisher's exact test $p < 0.05$) for 44/105, 44/105, and 1/1 CRE

data sets in LCLs, liver, and cerebellum, respectively (Tables S4, S5, S6). For example, LCL eQTLs are significantly more likely to be cell type specific (i.e., replicate in an independent cohort of LCLs but fail to replicate in the liver) when they overlap an LCL-derived p300 binding site, but do not overlap a HepG2-derived p300 binding site (Fisher's exact test $p < 0.0027$). To illustrate a

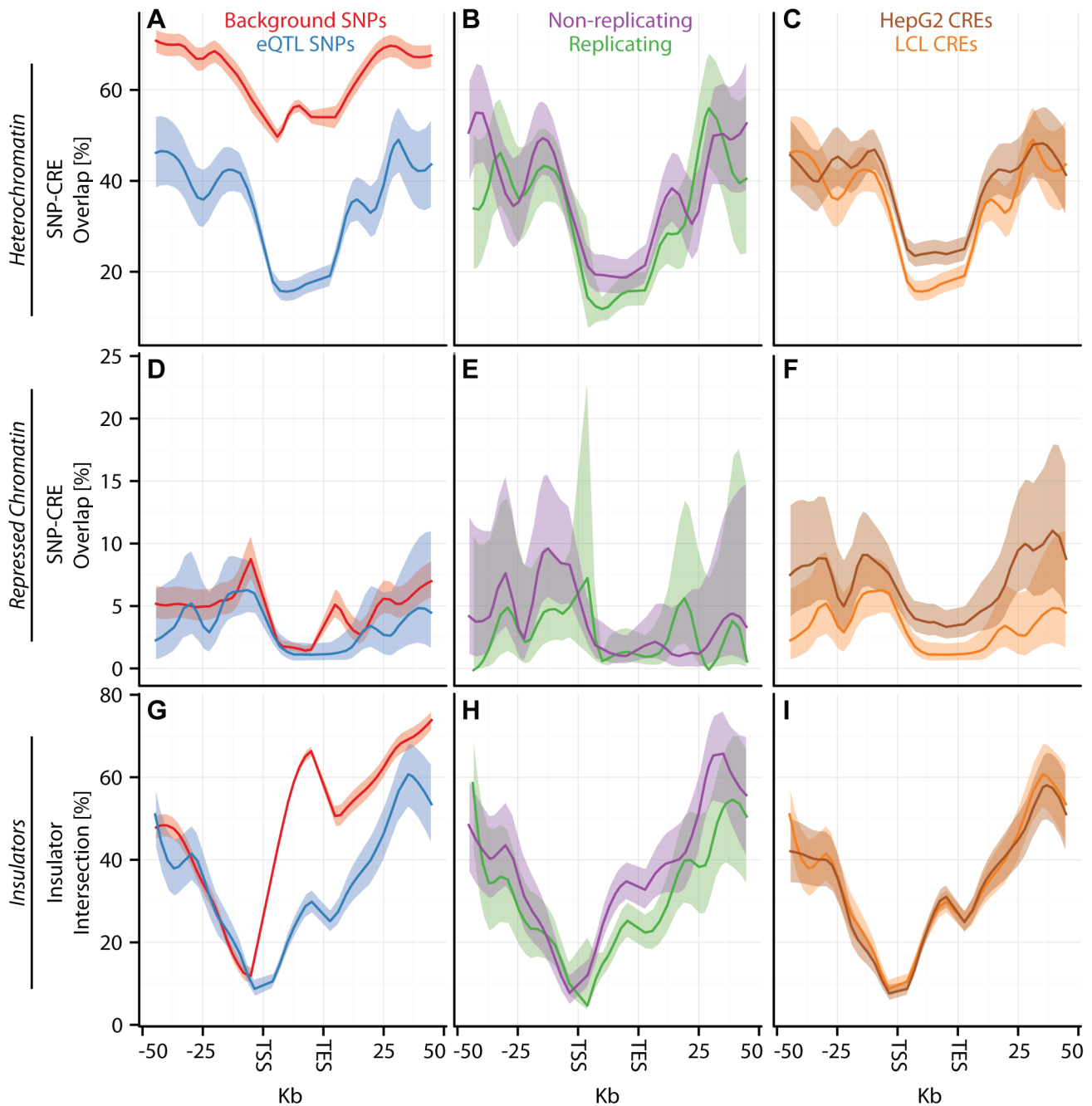


Figure 4. eQTL SNPs are depleted within repressive chromatin contexts. (A–I) CAP_LCL eQTL SNP ($FDR \leq 5\%$) overlap with predicted cis-regulatory elements. (A–C) eQTL SNP overlap with chromHMM predicted heterochromatin, (D–F) eQTL SNP overlap with chromHMM predicted repressive chromatin, (G–I) eQTL SNP-TSS pairs with an intervening CTCF binding site. In each panel, SNPs are grouped into 25 equally spaced bins within the 50 kb upstream and downstream of the TSS and TES, and 10 bins between the TSS and TES. Each bin is plotted along the x-axis. Bold lines depict bin percentage, ribbons depict 95% confidence interval. Each column of panels depicts a distinct SNP set contrast. (A,D,G) Observed eQTL SNPs (blue) and randomly drawn cis-linked SNPs at expressed genes (red). (B,E,H) eQTL SNPs that replicate in Stranger_LCL ($\log_{10} BF \geq 1$) (green) and SNPs that fail to replicate (purple). (C,F,I) CAP_LCL eQTL SNP overlap with CREs derived from the LCL line GM12878 (orange) and HepG2 cells (brown). doi:10.1371/journal.pgen.1003649.g004

specific example, we examined the pattern of cell specific eQTL SNP-CRE overlap at the SORT1 locus, a well characterized myocardial infarction risk locus (Figure 6). Consistent with previous observations [15], we find a liver specific eQTL association approximately 40 kb downstream of the SORT1 gene, which overlaps a cluster of predicted enhancers that are present in HepG2 cells but not LCLs.

Prediction of eQTL replication across cell types

Given the association between cell type specific eQTLs and cell type specific cis-regulatory elements, we sought to test our ability to use CRE data in conjunction with genomic location information to predict the cell type specificity of eQTLs. We trained a random forest classifier on a large set of SNP features, including SNP position, effect size, cell type specific CRE overlap, and non-

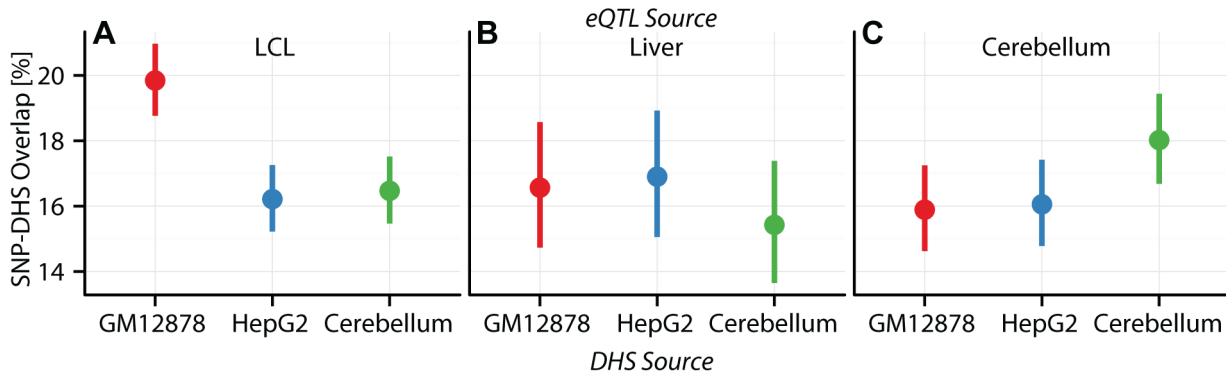


Figure 5. Cell specificity of eQTL SNP-CRE overlap illustrated with DNase hypersensitivity data. Percentage (dots) and 95% confidence interval (lines) of (A) CAP_LCL, (B) UChicago_liver, and (C) Harvard_cerebellum eQTL SNPs overlapping DHS sites (y-axis) derived from the LCL cell line GM12878 (red), the HepG2 cell line (blue), and the cerebellum (green). doi:10.1371/journal.pgen.1003649.g005

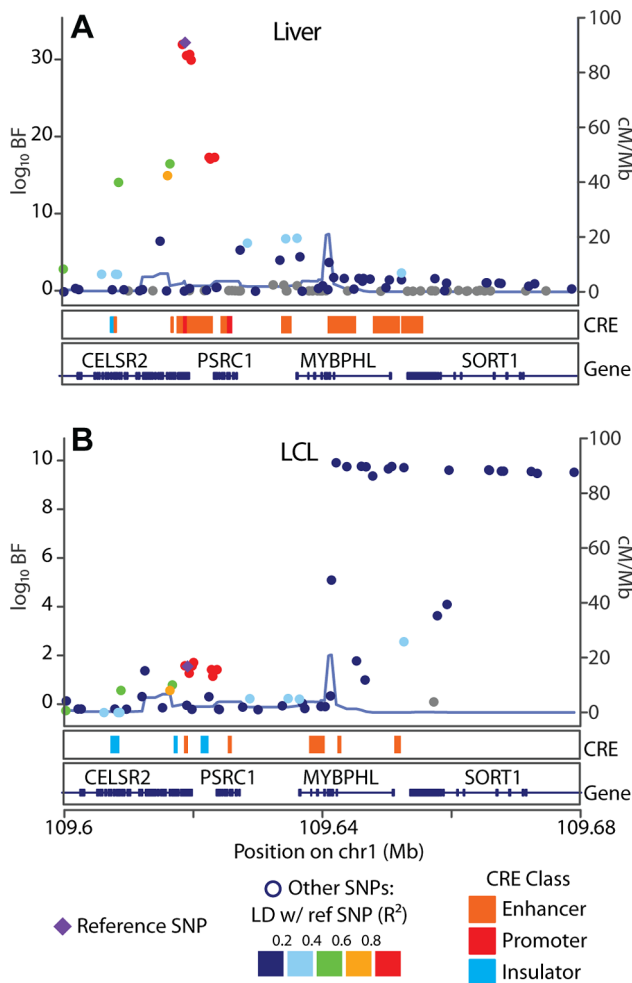


Figure 6. SORT1 eQTL illustrates mechanisms underlying cell specificity of eQTLs. Associations between (A) UChicago_liver and (B) CAP_LCL SORT1 expression and cis-linked SNPs (left y-axis; $\log_{10} BF$), plotted as points by SNP genomic coordinates (x-axis). Blue line overlaying the manhattan plot is the estimate of the local recombination rate (right y-axis; cM/Mb). Points are colored by level of LD (see legend below) with the reference SNP (purple diamond). Below each manhattan plot are boxes depicting the location of chromHMM predicted promoters (red), enhancers (orange), and insulators (blue). Below CRE predictions are RefSeq gene models. doi:10.1371/journal.pgen.1003649.g006

cell type specific genomic elements (CRE features listed in Table S3) to predict whether each eQTL SNP association would replicate in a second study or not (i.e., binary class). The classifier accurately predicts within cell type eQTL replication, between cell type eQTL replication, and cell type specific eQTL replication. We validated the classifier with 10-fold cross validation and demonstrated that its accuracy is dependent upon the inclusion of cell specific CRE data (Figure 7, Figure S17, Table 3). For example, the area under the ROC curves (*AUC*) for within LCL replication, between LCL and liver replication, and LCL specific replication were 0.79, 0.73, and 0.67, respectively. We believe that the predictive performance of the random forest on this problem reflects the fact that random forests are capable of capturing interactions among the features. Given a broad collection of chromatin state and regulatory factor binding site data sets, such as is available for a large number of cell types in the ENCODE project database, it is now possible to predict whether a given eQTL association exists in a different, specific cell type, in the absence of eQTL data from the second cell type.

We further quantified the contribution of each feature to prediction accuracy (see Methods). Across all training sets, eQTL discovery significance, SNP to TSS distance, and discovery cell type gene expression level contribute substantially to prediction accuracy (Table S8). Consistent with the relative cell type specificity of CTCF binding sites and chromatin marks discussed above, CREs vary considerably in the degree to which they are useful in predicting within or between cell type replication. Cell type specific differences in activating and repressive chromatin states overlapping eQTLs contribute substantially to the accuracy of predictions of between cell type eQTL replication and cell type specific replication. In contrast, CTCF binding site data are less informative for predicting cell specific eQTL replication, as might be expected considering its function appears less cell type specific.

Integration of CREs and eQTLs for GWAS interpretation

While the use of eQTL data to aid in the interpretation of GWAS results has proven extremely useful [7,8,11,16], these data are most useful when they are derived from the cell type that is most relevant to the disease of interest. Because eQTL data exist from a limited number of cell types, they often have limited utility for annotating the cell type of interest. We investigated whether cell specific CRE overlap would be useful to predict whether disease associated SNPs function as eQTLs in cell types lacking eQTL data. We extracted from the NHGRI GWAS catalog [49] SNPs associated with diseases arising from breast, kidney, or lung

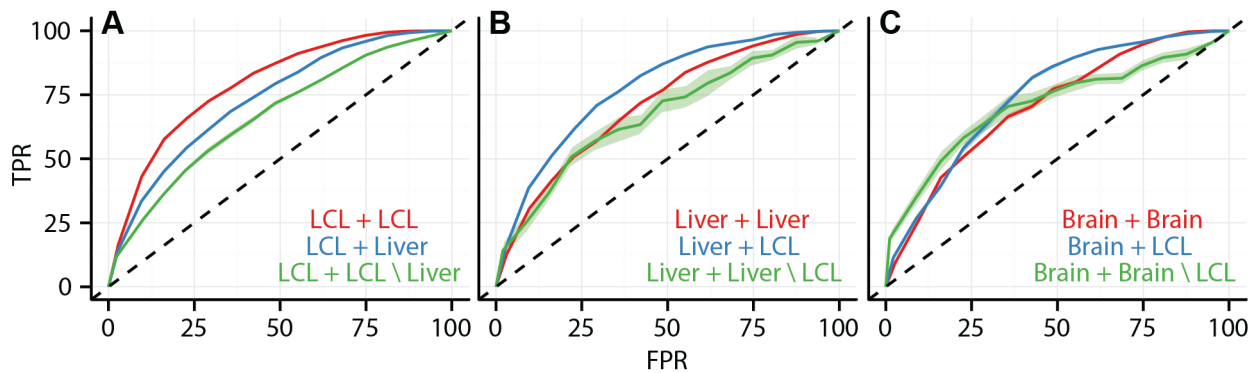


Figure 7. Data integration predicts cell type specificity of eQTLs. ROC curves depicting the performance of a random forest classifier to predict within cell type reproducibility (red), between cell type reproducibility (blue), and within cell type specific reproducibility (green). Predictions plotted separately for (A) LCL/LCL/Liver, (B) Liver/Liver/LCL, and (C) Brain/Brain/LCL. The classifier was trained on a diverse collection of CREs (see Methods and Supplement for complete data set description). True positive rates (y-axis) and false positive rates (x-axis) were quantified by ten fold cross validation.

doi:10.1371/journal.pgen.1003649.g007

tissue. We trained a random forest classifier to differentiate eQTL SNPs from non-eQTL SNPs in liver and LCLs using as features HepG2 and LCL CRE-SNP overlap (see Methods). We then applied this classifier to each GWAS SNP set, using CRE data from HMEC, NHEK, and NHLF cells to model breast, kidney, and lung function, respectively, to see if there was an enrichment for predicted eQTL SNPs when the CRE data represented the cell type of interest. These predictions were compared with predictions using LCL and HepG2 CREs as features (Figure S18). In each case, CREs from the disease-relevant cell type are more likely to predict that GWAS SNPs function as eQTLs (Wilcoxon signed rank tests: breast $p < 0.06$, kidney $p < 7.64 \times 10^{-6}$, lung $p < 6.33 \times 10^{-6}$). Given the hypothesis that a substantial number of GWAS SNPs function by modifying gene expression, these results support the hypothesis that integrated eQTL-CRE modeling can aid in the annotation of GWAS results arising from a wide variety of cell types.

Discussion

The integrative analyses presented here provide new insights into the patterns of cis-eQTL replication within and between cell types, while controlling for biological and technical variation.

Table 3. Accuracy of random forest classifier predictions.

Prediction	AUC	Accuracy
CAP_LCL \in Stranger_LCL	0.79	0.74
CAP_LCL \in Merck_liver	0.73	0.82
CAP_LCL \in (Stranger_LCL\Merck_liver)	0.67	0.71
UChicago_liver \in Merck_liver	0.74	0.70
UChicago_liver \in Stranger_LCL	0.77	0.73
UChicago_liver \in (Merck_liver\Stranger_LCL)	0.71	0.66
Harvard_cerebellum \in Myers_brain	0.71	0.83
Harvard_cerebellum \in Stranger_LCL	0.77	0.81
Harvard_cerebellum \in (Myers_brain\Stranger_LCL)	0.68	0.60

For the prediction task in the left-most column, we include the area under the curve (AUC) and prediction accuracy.
doi:10.1371/journal.pgen.1003649.t003

Several notable results emerge from our analyses. We demonstrate that eQTLs are more likely to overlap activating CREs and less likely to overlap repressive CREs when they are ascertained from the same cell type versus different cell types. Cis-eQTL SNPs overlapping most classes of activating cis-regulatory elements are significantly more likely to replicate in independent studies. Conversely, eQTL SNPs that overlap repetitive or repressive chromatin states and eQTL SNP-gene pairs that are intersected by insulators are significantly less reproducible. Cis-eQTL SNP-CRE overlap is also significantly more predictive of eQTL reproducibility when the CRE data are derived from the same cell type as the gene expression data. Furthermore, eQTL SNPs that overlap cell type specific CREs are significantly enriched for cell type specific eQTLs, suggesting specific regulatory mechanisms for those cell type specific eQTL associations. The observed relationship between eQTL SNP reproducibility and CRE overlap led us to test the hypothesis that SNP-CRE overlap could be used to predict the cell type specificity of eQTLs in the absence of additional gene expression or genotype data. While we see room for substantial improvement, we believe that the successful validation of this hypothesis with a random forest classifier will enable improved interpretation of genome-wide association study results.

After a GWAS is performed, it is now common practice to search eQTL databases to determine whether SNPs of interest are eQTLs for the cell type relevant to the study phenotype. When the SNPs are not known eQTLs in the cell type of interest, typically the line of reasoning is dropped; however, it is possible that the specific cell type was not tested, that the relevant SNP-gene pair was not interrogated, or that the sample size was too small for a cell type specific eQTL study to substantiate the SNP as an eQTL. Instead, if the researcher finds that the SNP is an eQTL in an alternative cell type, our classifier can be applied to determine the likelihood of the SNP being an eQTL in the cell type of interest when there are CRE data available for the relevant (or related) cell type. Furthermore, the genomic location can be scrutinized relative to known CREs to identify specific CRE types that may explain the mechanism by which gene transcription and downstream phenotype are regulated. Conversely, given those same GWAS hits, using these predictions one might be able to identify the physiologically relevant cell type based on overlap with (predicted or known) cell specific eQTLs.

Methods

Genotype preparation

Genotype data were downloaded from public databases or individual investigators as summarized in Table 1. Genotype and quality control filtering was performed with plink [50]. Individuals with a call rate of less than 90% were removed. SNPs with a call rate of less than 90% were classified as missing and later imputed. SNPs deviating from HWE were removed ($p < 1 \times 10^{-4}$).

Merck_liver study genotypes were previously imputed by MACH [51]; we extracted from the full set of SNPs only those that were >90% unimputed (and we removed all of the imputed genotypes from individuals in the non-imputed SNPs) to represent the original genotyping data. This set we filtered using identical criteria as above.

The HapMap phase 2 individuals fully sequenced genotypes were downloaded from the Impute2 website [52]; we matched the genotypes to individuals by comparing individual SNPs to genotypes from the indexed individuals. We filtered these genotypes as above. We removed ungenotyped individuals in the Harvard study, leaving us with 540 individuals with cerebellum tissue data, 678 individuals with prefrontal cortex tissue data, and 463 individuals with visual cortex tissue data.

Genotype imputation

Genotypes were imputed using BIMBAM [35]. We imputed genotypes up to the HapMap phase 2 CEPH 3.8×10^6 SNP set. BIMBAM removes SNPs with a minor allele frequency (MAF) less than 0.01 or missing SNPs by default. For the studies with Caucasian only participants, we used only the 60 unrelated CEPH individuals for imputation. For the UChicago_liver study, which has 27 African American individuals (of 206 subjects total), we used both the CEPH and the 60 unrelated YRI individuals as a reference set. For the Stranger_LCL study on a subset of HapMap phase 2 individuals, we used the CEPH, YRI, and the 90 unrelated JPT and CHB individuals' genotypes and did not impute. Mean values for the imputed genotypes were used for association and other downstream analyses [38].

Gene expression preparation, normalization, and processing

For each expression array platform, probe sequences were aligned to the human reference genome (hg18) and the RefSeq transcript set. Probes with only one genomic alignment with $\geq 90\%$ identity to the reference genome, over the full length of the probe, were considered to be uniquely aligned. Probes that failed to align to the genome but did have at least one alignment with at least 90% identity to a RefSeq transcript were further considered to be adequately aligned. All other probes were removed from further analyses. Based on genomic alignment coordinates and RefSeq gene annotations, each aligned probe was assigned to a RefSeq gene. We further searched the genomic locations of each probe alignment for the presence of common polymorphisms, as defined by dbSNP131 and the One Thousand Genomes Project (8/4/2010 release [53]).

Where appropriate, we defined a lower expression level boundary, above which we considered a gene to be expressed. Genes falling below the expression threshold were removed from further analyses. Low expression thresholds were defined either on the basis of negative control probes, exogenous RNA spike in probes, or the observed relationship between probe mean expression and variance.

Gene expression data from each study were prepared independently as follows. Poorly extracted, non-uniform, outlier, or other

flagged features were treated as missing data. Where appropriate, background signal intensities were subtracted. Negative adjusted intensities were set to one half the minimum positive value on the array. Background corrected intensities were \log_2 transformed. Missing data were imputed using the k -nearest neighbors algorithm ($k=10$), as implemented in the R package impute [54]. Each array was quantile transformed to the overall average empirical distribution across all arrays. Across all arrays within a study, each probe's expression values were transformed to the quantiles of the standard normal distribution (or *quantile normalized*). Transformation to standard normal avoids potential problems due to outliers or other deviations from normality in later association tests [55]. We controlled for known and unknown sources of non-genetic variation by correcting these data using their principle components (PCs), identified using the R function `pca` from the R package `pcaMethods` [36,37]. For each matrix of gene expression data, we computed the percentage of variance explained (PVE) for each PC, which is a monotone decreasing function, and controlled for PCs until the difference in the PVE by the subsequent PC was $< 2.5 \times 10^{-5}$ (Table S1). We jointly controlled for the PCs in the gene expression data by taking the residuals from a linear model with these PCs as covariates. For one data set with high quality covariate annotation (UChicago_liver), we have quantified the correlation between each PC and available covariates (Figure S1), demonstrating that the PCs extracted are capable of capturing heterogeneity arising from such confounding variables. Finally, we quantile normalized these residuals within each probe and used these normalized data in our subsequent analyses.

For genes with multiple probes on a single array, we used the R package `mclust` [56] to cluster the p probes $\times n$ samples matrix of expression levels. We allowed up to $\min(4, p)$ clusters per gene. Within each probe cluster, we used the per individual, PC corrected mean of the different probes as a proxy for the gene expression level for that collection of probes. Each probe cluster was modeled downstream independently, under the assumption that uncorrelated probe sets represent either independent transcript isoforms or poorly performing probes.

Two studies included non-European individuals: Stranger_LCL included individuals with European, Yoruban, and East Asian ancestry, and UChicago_liver included 200 individuals with European ancestry and 25 African Americans. To minimize the effect that differences in allele frequencies between populations may have on false positive discoveries and false negative replication, we analyzed these data with a modified protocol. Within a multi-population study, the distribution of gene expression measurements from each array was normalized to the average distribution of expression measurements across all arrays. For each population separately, each probe was then quantile normalized and measurements across populations were pooled [57]. When assessing replication across studies, we require that the SNP has a MAF $> 1\%$ in both studies. Therefore alleles that are polymorphic in one population but fixed in another will not result in false negative replication. We note that replication frequencies are similar between mixed population and single population studies (Figures S5, S6). Furthermore, cross study replication is not substantially affected by differences in allele frequency across studies (Figure S8).

eQTL mapping

We used Bayesian regression, as implemented in BIMBAM [34,38] to quantify the association between each SNP and residual gene expression data for each gene across each sample from each study. We used default parameters, which average over different plausible effect sizes for additive and dominant models. We used

the mean imputed genotype for all studies except Stranger_LCLs, in which we used assayed genotypes for each individual because the individuals did not require imputation.

Multiple cell type eQTL mapping

We used multi-trait Bayesian linear regression models to jointly test for eQTLs with specific models of differential association in paired samples [58,59]. We computed the Bayes factors for five models as compared with the null model (no eQTL in either cell type):

1. *stable* model, where the eQTL is associated with gene expression in both cell types,
2. *A not B* model, where the eQTL is associated with gene expression in one of the two cell types,
3. *B not A* model, where the eQTL is associated with gene expression in one of the two cell types,
4. *A and B* model, where the eQTL is associated with gene expression differently in the two cell types, and
5. *opposite B* model, where the eQTL is associated with gene expression in both cell types, but the direction of the effect is opposite.

To combine these individual Bayes factors, we calculated an integrated Bayes factor (iBF) summarizing evidence for a differential eQTL as:

$$iBF = \frac{2 * \max(BF_{AnotB}, BF_{BnotA}, BF_{AandB}, BF_{AoppositeB})}{BF_{stable} + 1} \quad (1)$$

eQTLs with $iBF > 1$ (i.e., there was more evidence supporting a differential eQTL than a stable eQTL) were further categorized by the specific model with the largest BF, although we note that the power to identify each of the models is different in this construction. Further analysis was conducted only on eQTLs with $FDR \leq 5\%$.

Summarizing eQTLs

We identified the SNP with the largest $\log_{10} BF$ for each gene, and also identified the cis-SNP with the largest $\log_{10} BF$ in each LD block around a gene. We defined LD blocks using the HapMap recombination rates [60] in which each SNP interval with ≥ 10 cM/Mb defines the boundaries of an LD block. For each of these associations, we note the chromosome and location of the gene and SNP, major and minor allele of the SNP, the LD block index, the MAF of the SNP in the sample population, the r^2 value of the fit of the linear model between the imputed values for the SNP and the rounded values of the SNP, the magnitude and direction of the association (β) and the r^2 value of the fit of the linear model, the number of exons and the average length of the gene exons, number of probes corresponding to the probe cluster, mean gene expression value for those probes, and the $\log_{10} BF$ for this association. We determined the maximum MAF for all SNPs overlapping expression array probe alignment coordinates using One Thousand Genomes Project data and dbSNP131. We note that non-replication within cell types is not driven by differences of MAF between the studies (Figure S8 is a cross-study eQTL MAF comparison). For all downstream comparisons between studies, we considered only expressed gene-SNP pairs in common between the two studies.

Evaluating FDR by permutation

To evaluate the FDR for each study, we permuted the sample indices on the gene expression data identically across genes within

each study, and ran association mapping on these permuted data. Then, for each cutoff $\log_{10} BF$, we conservatively computed FDR by the number of associations identified in the original data at that cutoff divided by the number of associations identified in the permuted data at that cutoff (Table S9). We performed a single permutation for each study because of the resources required to run a single complete permuted association test. Unless otherwise noted, eQTL results in the text refer to associations significant at $FDR \leq 5\%$.

Multivariate analysis for allelic heterogeneity

Given the computational requirements of performing $\sim 20,000$ conditional QTL scans in eleven studies, we implemented a two step approach to identify independently associated SNPs at each gene. For each gene probe cluster, we identified the most highly associated SNP in each LD block within 1Mb of the gene's transcription start site (TSS) or transcription end site (TES). We subsequently recomputed the BF of each SNP association by Bayesian multivariate regression to quantify the conditional independence of the effects of the associated SNPs [61]. Finally we took the union of the identified SNPs from all probe clusters for a single gene; when this set had more than one SNP below the appropriate $FDR \leq 5\%$ cutoff, the gene is said to exhibit allelic heterogeneity.

We compared this approach to forward stepwise regression for a subset of CAP_LCL genes with significant allelic heterogeneity ($n=696$ genes; $FDR \leq 5\%$) [62,63]. We implemented, in R, a custom forward stepwise regression with all SNPs within 1 Mb of a gene's TSS and TES to identify the set of independently associated SNPs. In particular, starting from the model with a gene probe cluster as the response variable and no covariates, we included a SNP in the model when it most improved the Bayesian Information Criterion (BIC) of the fitted model. The BIC is a criterion for model selection that takes into account both the likelihood of the model and the number of parameters (here, eQTL SNPs), so as to avoid excessive parameters that may result in overfitting. We continued to identify a single SNP that most improves the BIC and included that SNP in the model until a SNP resulted in a non-improvement of the BIC, when we stopped. As in the LD-based method, we took the union of the identified SNPs from all probe clusters for a single gene. Although forward selection is not exhaustive, it is tractable, and these results represent a more complete (but still conservative) estimate of the allelic heterogeneity for a particular gene relative to our method of identification (for full results comparison, see Table S10). In particular, we note that this method is an approximation in the presence of interactions [64], but we consider this case to be beyond the scope of this analysis. Our LD-based allelic heterogeneity ascertainment generally appears to underestimate the number of independently associated SNPs per gene, however this may be a result of the thresholds for SNP inclusion between the two methods being different (Figure S19, Table S10).

Analysis of Gene Ontology annotation enrichment

We performed Gene Ontology (GO) [65] enrichment analyses using DAVID software [20]. We considered GO Biological Process ontology, Molecular Function ontology, and KEGG pathway [66] enrichment terms with $FDR \leq 20\%$ (Table S11). We considered CAP_LCL, UChicago_liver, and Harvard_cerebellum as eQTL discovery data sets, and looked for enrichment among genes with multiple independent eQTLs in these data, and also looked for enrichment among genes that did and did not replicate within cell type and between cell type studies as above.

Replication quantification

A SNP-gene association was considered ‘replicated’ if the $\log_{10} BF \geq 1.0$ in the target study. Only SNP-gene pairs that were tested in a given replication cohort were considered when calculating the replication frequency in that study. In other words, if a SNP failed quality control or if a gene was not represented on a particular gene expression microarray platform, the SNP-gene pair was not considered when calculating replication. Similar to previous observations (see Figure S10H in [26]) we do not observe a substantial effect of cross-study allele frequency differences on eQTL replication (Figure S8).

Comparison between eQTLs and functional genomic data sets

When available, CRE data from the CEU HapMap LCL line GM12878, the hepatocellular carcinoma HepG2 cell line, and primary cerebellum tissue were used to represent LCLs, liver, and cerebellum respectively. These data sets (fully listed in Table S3) were all downloaded directly from the ENCODE data coordination center at UCSC. Cell type independent data sets were also used, including CpG islands [67], GERP evolutionary constrained elements [68], clustered DNase hypersensitive sites, and clustered transcription factor binding sites. A second chromatin structure segmentation (Segway [69]) data set, derived from the integration of K562 data sets, was also included in the regulatory element feature set. When available, precomputed ‘peak calls’ were used. If more than one replicate (i.e., peak calls generated in the same lab using the same antibody), was available, peak calls were merged by taking the union of all elements.

Given that a putative eQTL SNP does not necessarily represent the causal genetic variant, but rather is a ‘tag’ for the causal variant in substantial linkage disequilibrium, we classified a SNP as overlapping a given genomic element if it was either contained within the element or found within 500 bp of the element boundary. CTCF sites were classified as eQTL interrupting if the midpoint of the CTCF binding site was between the eQTL SNP and the TSS of the associated gene.

To test if eQTL SNPs are enriched within each class of putative cis-regulatory element, we first modeled the background distribution of cis-linked SNPs as follows. The background set of SNPs was selected to match the set of SNPs that were tested for eQTL association. All CEU HapMap phase 2 SNPs that lie within 1 Mb of a RefSeq gene model TSS or TES, or lie within a RefSeq gene model were included. Each such SNP that was cis-linked to more than one RefSeq gene model was randomly assigned to one such gene. As with the eQTL SNP set, the background SNP set was tested for overlap with any element in each genomic feature class.

For each eQTL or background SNP (i), we modeled the probability of cis-regulatory element overlap (p_i) by logistic regression:

$$p_i = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 D_i + \beta_2 G_i + \beta_3 I_i)\}} \quad (2)$$

In this equation, we controlled for the effects of SNP position ($D_i = \log_{10}(|\text{SNP}_i - \text{TSS}_i| + 1)$) and the expression level of the associated gene (G_i), in order to quantify the difference in overlap frequency between observed eQTL SNPs and those drawn from the background SNP distribution (denoted by the indicator variable I_i). This and other logistic regression models were run using the R glm function. To assess the significance of the effect of each covariate, we computed a Z-score and, from that, a two-tailed p-value by comparison to the normal distribution.

Similarly, for each eQTL SNP (i), we model the effect of cis-regulatory element overlap (C_i) on the probability of within cell type replication (p_i) by logistic regression:

$$p_i = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 M_i + \beta_2 T_i + \beta_3 D_i + \beta_4 C_i)\}} \quad (3)$$

In this equation, we controlled for the effects of SNP position and the significance of the eQTL association ($M_i = \log_{10} BF$, as assessed by Bayesian multivariate regression) and the tier of the associated SNP (T_i), in order to quantify the enrichment of within cell type replicating eQTLs in cis-regulatory elements.

We tested for an enrichment of second tier SNPs within CRE data sets, relative to the expectation from the background distribution of cis-linked SNPs, by modeling, for each SNP (i), the probability of cis-regulatory element overlap (p_i) by logistic regression:

$$p_i = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 D_i + \beta_2 G_i + \beta_3 T_i)\}} \quad (4)$$

where we quantify the difference in overlap frequency between background SNPs, first tier SNPs, and second tier SNPs with the indicator variable T_i . Enrichment of second tier SNPs, relative to background, was quantified by testing for the significance of the difference between the β_3 estimates.

To test for an enrichment of CTCF sites between SNPs independently associated with the same gene expression trait (i.e., SNPs tagging the allelic heterogeneity at a given locus), we used the background distribution of cis-linked SNPs as defined above; however, for each RefSeq gene, two cis-linked SNPs were selected at random. This collection of background SNP pairs was contrasted with the first and second tier eQTLs at all genes for which the secondary eQTL was significant at a $FDR \leq 5\%$ threshold. For each SNP pair (j), we modeled the probability of the presence of an intervening insulator (p_j) by logistic regression:

$$p_j = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 DS_j + \beta_2 DT_j + \beta_3 HS_j + \beta_4 TS_j + \beta_5 I_j)\}} \quad (5)$$

In this equation, we controlled for the inter-SNP distance (DS_j), SNP pair position relative to the gene TSS ($DT_j = \min_k (|\text{SNP}_{j,k} - \text{TSS}_{j,k}|)$, in which k indexes each SNP in the pair), the presence of an intervening recombination hotspot (HS_j), and the presence of an intervening TSS (TS_j), in order to quantify the difference in insulator frequency between observed eQTL SNP pairs and those drawn from the background SNP distribution (denoted by the indicator variable I_j). We note that because the majority of SNP pairs that are independently associated with a gene expression trait span a recombination hotspot, we have attempted to control for any bias this may introduce by including the presence of such a hotspot in the model.

Predicting eQTLs with random forests

We built a classifier to predict within cell type, between cell type, and cell type specific replication using random forests [70]. A random forest is an ensemble classifier that uses the mode of predictions from a large number of decision trees built by bootstrapping the training set. For each pair of comparisons, we can train a random forest classifier to predict the binary replication outcome given a set of features about the genomic location of the eQTL and the CREs for the corresponding cell types (see Table

S3 for complete set). We used the random forest classifier and computed variable importance using the R randomForest package [71]. We performed 10-fold cross validation to compute generalization error. We used the R package ROCR [72] to build the ROC curves and compute the AUC. We also computed $accuracy = \frac{TP + TN}{P + N}$.

To predict the function of GWAS SNPs from cell types lacking eQTL data, we built a random forest classifier trained on cell specific CRE data. We first trained the classifier to predict whether a SNP is an eQTL or not an eQTL in LCLs and liver by selecting the most highly associated SNP for each gene ($FDR < 5\%$) in these two cell types and selecting a background set of 100,000 SNPs that were not associated with gene expression in any study ($\log_{10}BF < 1$). We selected a reduced set of CRE data (chromHMM segments, H2A.Z, EZH2, DHS, CTCF) that were available for LCLs, HepG2, NHEK, NHLF, and HMEC cells. We initially trained the classifier using as features the CRE data from LCLs and HepG2 to predict whether a SNP is an eQTL in LCLs and liver. After training, the model was applied to sets of GWAS SNPs that were associated with phenotypes with relevance to breast, kidney, and lung (hand annotated from the NHGRI GWAS catalog [49] downloaded on January 10, 2013). In particular, the classifier was used to predict the probability that each GWAS SNP was an eQTL or not, using as features CRE data from the matched cell type (HMEC for breast, NHEK for kidney, NHLF for lung) or from an unmatched cell type (LCL or HepG2). Distributions of eQTL probabilities between matched and unmatched cell types were compared by Wilcoxon's signed rank test.

Additional statistical analyses

2×2 tests of unpaired categorical data were quantified by Fisher's exact test. 2×2 tests of paired categorical data were quantified by McNemar's test. Tests of paired interval data were quantified by Wilcoxon's signed rank test.

Supporting Information

Figure S1 Sample principal components (PCs) capture many known study covariates in UChicago_liver study. Heat map indicating the absolute Pearson's correlation between sample covariates (x-axis) and the first twenty PCs (y-axis). Age, for example, is captured well by the 17th PC and smoking status is captured best by the 5th PC. Sample ID appears correlated with the 3rd PC because the IDs were ordered by when the sample was processed, which is well correlated with batch and other known gene expression confounders. (TIF)

Figure S2 Identification of allelic heterogeneity. Univariate $\log_{10}BF$ (x-axis) versus multivariate $\log_{10}BF$ (y-axis) as a function of linkage disequilibrium (r^2 ; color as indicated in scale bar) between the primary and secondary SNP. All tier two SNPs are plotted for each study independently, in each panel, as labeled at top. Note SNPs in higher LD with the primary SNP ('bluer' points) tend to produce greater drops from the univariate $\log_{10}BF$. (TIF)

Figure S3 Distribution of eQTL effect sizes by study and SNP tier. Violin plots of the distribution of effect sizes for significant ($FDR \leq 5\%$) eQTLs. Data are plotted separately for each study, as indicated by plot color and sample labels at left. SNPs from each tier are plotted separately in each facet, as labeled at top. Note studies are ordered on the y-axis by sample size. (TIF)

Figure S4 eQTL effect size and significance by tier. eQTL significance ($\log_{10}BF$; y-axis) as a function of SNP effect size (x-axis) and SNP tier. Data from each study and tier are plotted in independent facets, as labeled at top and right, respectively. Each eQTL SNP set was independently thresholded at $FDR \leq 5\%$ by permutation. (TIF)

Figure S5 eQTL replication frequencies, by observation significance, across all cohorts. eQTL replication (y-axis), as a function of discovery significance (x-axis: $\log_{10}BF$). Each column of facets depicts the set of eQTL SNPs discovered in each of 11 studies, as labeled at top. Within each column, replication frequencies are plotted separately for each replication study set, in rows and in a different color, as labeled at the right. SNPs are binned along the x-axis into 30 equally spaced intervals. Per bin replication frequencies are plotted as bold lines, 95% confidence intervals are plotted as ribbons. (TIF)

Figure S6 eQTL replication frequencies, by SNP position, across all cohorts. eQTL replication (y-axis), as a function of SNP position (x-axis: $|\text{SNP} - \text{TSS}|$). Each column of facets depicts the set of eQTL SNPs discovered in each of 11 studies, as labeled at top. Within each column, replication frequencies are plotted separately for each replication study set, in rows and in a different color, as labeled at right. SNPs are binned along the x-axis into 30 equally spaced intervals. Per bin replication frequencies are plotted as bold lines, 95% confidence intervals are plotted as ribbons. (TIF)

Figure S7 eQTL replication by SNP to TSS distance, conditioned on significance. Within cell type eQTL replication (y-axis), as a function of absolute SNP to TSS distance (x-axis). Replication frequencies are displayed separately for eQTLs with $\log_{10}BF < 3$ (red), $3 < \log_{10}BF < 6$ (blue), and $\log_{10}BF > 6$ (green). SNPs have been binned along the x-axis into 15 equally spaced intervals. eQTL SNP-CRE overlaps per bin are plotted as bold lines, 95% confidence intervals are plotted as ribbons. Each facet depicts a different study set comparison (labeled at top). (TIF)

Figure S8 Minor allele frequency differences and replication. Minor allele frequencies in the two different populations of replicating (left column) and non-replicating (right column) eQTLs. The size of the point corresponds to the $\log_{10}BF$ of the eQTL in CAP_LCL; the color represents the $\log_{10}BF$ in the replication study, where blue is higher and grey (in the right-hand column) is < 1.0 . Top row: MAFs compared between eQTLs discovered in CAP_LCL (x-axis) that replicate (left) or fail to replicate (right) in HM2_LCL (y-axis). Bottom row: MAFs compared between eQTLs discovered in CAP_LCL (x-axis) that replicate in HM2_LCL and that replicate (left) or fail to replicate (right) in Merck_liver. (TIF)

Figure S9 Extended eQTL replication plots. eQTL replication (y-axis), as a function of discovery significance (x-axis; left facet set) or absolute SNP to TSS distance (x-axis; right facet set; thresholded at $FDR \leq 5\%$), as indicated at the bottom. Replication frequencies are displayed separately for within cell type (red), between cell type (blue), within but not between cell type (i.e., *cell specific*; green), and within and between replication (purple). SNPs have been binned along the x-axis in 30 equally spaced intervals. eQTL SNP-CRE overlaps per bin are plotted as bold lines, 95% confidence intervals are plotted as ribbons. Each column of facets depicts a different study set comparison, as labeled at the top, for

LCLs (left column), liver (middle column), and brain (right column).

(TIF)

Figure S10 Bivariate Bayesian regression recapitulates results of post hoc comparisons. eQTL SNPs were classified by their best fitting model (denoted at left and by color). Results from bivariate cis-eQTL mapping are on top, results from post hoc models are on bottom. (A) Overlap of SNPs in each class with cerebellum DHS sites. Points denote overlap percentage, lines denote 95% CI. (B) Distribution of $\log_{10}BF$ (x-axis) for eQTL SNPs in each model class. Box plots denote median, inter-quartile range, and 95% CI. (C) Distribution of absolute distances (x-axis; \log_{10} scale) between each SNP and its associated gene's TSS.

(TIF)

Figure S11 eQTL replication conditional on SNP tier. eQTL replication (y-axis), as a function of discovery significance (x-axis; $\log_{10}BF$; left set of panels) or absolute SNP to TSS distance (x-axis, right set of panels). Replication frequencies are displayed separately for tier 1 (red) and tiers 2–4 (blue). SNPs have been binned along the x-axis into 30 equally spaced intervals. eQTL SNP-CRE overlaps per bin are plotted as bold lines, 95% confidence intervals are plotted as ribbons. Each panel depicts a different study set comparison, as labeled at the top of each panel. Within cell type and between cell type replication frequencies are plotted along the left and right columns, respectively, for LCLs (top row), liver (middle row), and brain (right row).

(TIF)

Figure S12 Overlap of full data set of ENCODE LCL CREs with CAP_LCL eQTLs. CAP_LCL eQTL SNP-CRE overlap (y-axis; tier means as points, 95% confidence interval as lines), with SNPs from each tier plotted in separate panels, as indicated at right. Each CRE class is plotted in a separate panel, as labeled at top. As applicable, within each panel, overlaps are plotted separately for LCL CREs (red circles) and HepG2 CREs (blue triangles). In cases where multiple CRE data sets exist for the same CRE class (e.g., H3K27me3 marks), overlaps were calculated from each data set independently and over plotted, with a jitter.

(TIF)

Figure S13 Overlap of full data set of ENCODE HepG2 CREs with UChicago_liver eQTLs. UChicago_liver eQTL SNP-CRE overlap (y-axis; tier means as points, 95% confidence interval as lines), with SNPs from each tier plotted in separate panels, as indicated at right. Each CRE class is plotted in a separate panel, as labeled at top. As applicable, within each panel, overlaps are plotted separately for HepGe CREs (red circles) and LCL CREs (blue triangles). In cases where multiple CRE data sets exist for the same CRE class (e.g., H3K27me3 marks), overlaps were calculated from each data set independently and over plotted, with a jitter.

(TIF)

Figure S14 CAP_LCL eQTL associations with 142 LCL derived CRE data sets. CAP_LCL eQTL SNP ($FDR \leq 5\%$) overlap with predicted cis-regulatory elements. Each panel depicts overlap with distinct CRE data sets, as labeled at top. In each panel, SNPs are grouped into 25 equally spaced bins within the 50 kb upstream and downstream of the TSS and TES, and 10 bins between the TSS and TES. Each bin is plotted along the x-axis. Bold lines depict the percentage, per bin, of SNPs overlapping the CRE class, ribbons depict 95% confidence interval. Observed eQTL SNPs are plotted in blue and randomly drawn cis-linked SNPs at expressed genes in red.

(TIF)

Figure S15 CTCF binding sites are enriched between SNPs independently associated with the same gene expression trait. Percentage of primary and secondary CAP_LCL (red) and Stranger_LCL (blue) LCL eQTL SNP pairs that have an intervening CRE (y-axis; SNP pairs were binned by the distance between them, bold lines depict bin frequency, ribbons depict 95% confidence interval) as a function of the absolute distance between the SNPs (x-axis). Randomly drawn cis-linked SNPs are displayed in green. Each panel depicts the analysis of a different CRE data set, as labeled at top, including CTCF, SMC3, and Rad21, which have each been shown to mark enhancer blocking insulators [73], DHS sites, which promiscuously mark insulators and other CRE classes, chromHMM defined heterochromatin regions, and p300 sites.

(TIF)

Figure S16 Overlap of ENCODE data sets across cell types. Overlap between CRE data sets (y-axis), as a function of absolute CRE to TSS distance (x-axis). Data are plotted separately for comparisons between two different LCL lines (red; e.g., between GM12878 and GM060990) and between LCLs (GM12878) and HepG2 cells (blue). CREs are binned into 30 equally spaced intervals along the x-axis. Per bin CRE overlaps are plotted as bold lines, 95% confidence intervals are plotted as ribbons. Data are plotted separately for each of six different CRE types, as labeled at the top of each panel. Note the striking difference between the cell specificity of CTCF binding sites and each other CRE class.

(TIF)

Figure S17 CRE data improve accuracy of cell specific eQTL prediction. ROC curves depicting the performance of random forest classifiers to predict within cell type specific reproducibility, either trained with CRE data (red) or without CRE data (blue). Facets depict predictions for LCL, liver, and brain eQTL SNPs (labeled at top). True positive rates (y-axis) and false positive rates (x-axis) were quantified by tenfold cross validation. AUCs from models with and without CRE training, respectively, were 0.67 and 0.61 (LCL), 0.71 and 0.57 (liver), and 0.68 and 0.57 (brain).

(TIF)

Figure S18 Regulatory element overlap predicts GWAS SNP function. Boxplot of the distribution of random forest classifier predictions. A classifier was trained to discriminate LCL and liver eQTL SNPs from non-eQTL SNPs on the basis of SNP overlap with CREs from LCLs and HepG2 cells. The trained classifier was then applied to SNPs associated with phenotypes of relevance to breast, kidney, and lung function (facets labeled at top). The probability that each SNP is an eQTL (y-axis, as box plot) was calculated using CREs from matched and unmatched cell types (listed on the x-axis and color coded). Asterisks denote significant differences in probability distributions (Wilcoxon signed ranks test, $p < 1 \times 10^{-5}$).

(TIF)

Figure S19 Quantification of allelic heterogeneity by forward stepwise regression and an LD-based method. Forward stepwise regression was applied to 696 genes with allelic heterogeneity as identified by the LD-based method described in the text. For each gene, the resulting pair of models are contrasted as follows. At left, multivariate $\log_{10}BF$ of the FSR model (y-axis) is plotted as function of the multivariate $\log_{10}BF$ of the LD based model (x-axis). At right, the Bayesian information criterion (BIC) of the FSR model (y-axis) is plotted as function of the BIC of the LD based model (x-axis). Circle size and color depicts the number of SNPs identified by the FSR and LD-based models, respectively (blue = 2

SNPs, green = 3 SNPs, red = 4 SNPs). We note that, in one gene, *NAN7L*, FSR found fewer independent eQTLs than our LD-based method; interestingly, the tertiary eQTL SNP for this gene had a lower univariate $\log_{10} BF$ than the conditional $\log_{10} BF$ (1.91 versus 2.32), implying a possible SNP-SNP interaction. For 12 genes (2%) the two methods found the same numbers of independent eQTL SNPs, and for 683 genes (98%) FSR identified additional eQTL SNPs (Table S9).

(TIF)

Table S1 Numbers of PCs removed from each study gene expression data. Columns include the study name, the cell type, the number of PCs removed (PCs), and the percentage variance, which refers to the percentage variance explained by the last PC removed.

(TXT)

Table S2 eQTL SNP replication table. Summary of eQTL findings and replication across all included studies. eQTL SNPs were included in the table if, from a given study, they were the most highly associated cis-linked SNP within an LD block and if their $\log_{10} BF \geq 1$. eQTL SNPs are listed in separate rows and associated data are in columns, as labeled in the first row and denote the following:

1. REF STUDY. eQTL discovery study.
2. GENE. RefSeq gene identifier.
3. CHR. Chromosome of the eQTL SNP and associated gene expression trait.
4. STRAND. Annotated strand of the associated transcript.
5. TSS. Genomic coordinate of the transcription start site.
6. TES. Genomic coordinate of the transcription end site.
7. RSID. Identity of the eQTL SNP.
8. HS INDEX. Identity of the LD block containing the SNP.
9. RSCOORD. Genomic coordinate of the SNP.
10. TIER. Tier of the SNP with respect the associated gene.
11. UBF. Univariate $\log_{10} BF$ of the eQTL SNP-gene expression trait association.
12. MBF. Multivariate $\log_{10} BF$ of the eQTL SNP-gene expression trait association, controlling for all SNPs in lower tiers.
13. BETA. Estimate of the effect size per minor allele of the SNP.
14. GEX. Mean gene expression level across all samples in the discovery study.
15. PROBE MAF. Maximum minor allele frequency of SNPs overlapping the coordinates of the gene expression probe.
16. NALN. Number of high quality alignments between the gene expression probe and hg18.
17. STL, MBR, CLI, CPL, HVC, GCT, GCF, HCE, HPC, MLI, GCL. Each subsequent column denotes the measure of univariate replication between the SNP and gene expression trait in the study set indicated by the TLA, where:

$$U = \begin{cases} \log_{10} BF & \text{if } \log_{10} BF \geq 1 \\ 0 & \text{if } \log_{10} BF < 1 \end{cases}$$

(BZ2)

Table S3 CRE data set details. File name, element summary type, cell type, experiment type, experiment class, experiment name, control type, file format, and two indicators denoting whether the data set was tested for overlap with eQTL SNPs or for intersection between the eQTL SNP and gene TSS (i.e., is an insulator), or both.

(TXT)

Table S4 CAP_LCL eQTL SNP enrichment/depletion within CREs: (a) relative to background, (b) replicating versus non-replicating, (c) same versus different cell types. For Tables S4, S5, and S6, eQTL SNPs ($FDR \leq 5\%$) were tested for overlap with 526 CRE data sets, each listed in a separate row. Columns are labeled in the first row and denote the following:

1. CRE type. Name of the tested CRE, including the cell type of origin, as applicable.
2. Cell A Overlap. Fraction of eQTL SNPs that overlap a CRE of this class, where the CRE was derived from cell type A (the cell type indicated in column 1).
3. Cell B Overlap. Fraction of eQTL SNPs that overlap a CRE of this class, where the CRE was derived from cell type B (LCLs for liver and cerebellum eQTLs, HepG2 cells for LCL eQTLs).
4. AB Overlap Test McNemar PV. McNemar's test p-value run on the 2×2 table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where a denotes the number of SNPs overlapping a CRE from cell type A, b denotes the number of SNPs *not* overlapping a CRE from cell type A, c denotes the number of SNPs overlapping a CRE from cell type B, d denotes the number of SNPs *not* overlapping a CRE from cell type B.
5. A Overlap Test Beta. β_3 from Eqn. (2), for CREs derived from cell type A.
6. A Overlap Test SE. Standard error of the estimate of β_3 from Eqn. (2).
7. A Overlap Test Z. Z-score of the significance of β_3 from Eqn. (2).
8. A Overlap Test PV. P-value of the significance of β_3 from Eqn. (2).
9. A Overlap Test AIC. Akaike information criterion of Eqn. (2).
10. B Overlap Test Beta. β_3 from Eqn. (2), for CREs derived from cell type B.
11. B Overlap Test SE. Standard error of the estimate of β_3 from Eqn. (2).
12. B Overlap Test Z. Z-score of the significance of β_3 from Eqn. (2).
13. B Overlap Test PV. P-value of the significance of β_3 from Eqn. (2).
14. B Overlap Test AIC. Akaike information criterion of Eqn. (2).
15. A Overlap Rep OR. Odds ratio derived from a Fisher's exact test performed on the 2×2 table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where a denotes the number of within cell type replicating SNPs overlapping a CRE from cell type A, b denotes the number of within cell type replicating SNPs *not* overlapping a CRE from cell type A, c denotes the number of non-replicating SNPs overlapping a CRE from cell type A, d denotes the number of non-replicating SNPs *not* overlapping a CRE from cell type A.
16. A Overlap Rep PV. P-value of the Fisher's exact test above.
17. A Overlap Rep Beta. β_4 from Eqn. (3), for CREs derived from cell type A.

18. A Overlap Rep SE. Standard error of the estimate of β_4 from Eqn. (3).
 19. A Overlap Rep Z. Z-score of the significance of β_4 from Eqn. (3).
 20. A Overlap Rep PV. P-value of the significance of β_4 from Eqn. (3).
 21. A Overlap Rep AIC. Akaike information criterion of Eqn.(3).
 22. B Overlap Rep OR. Odds ratio derived from a Fisher's exact test performed on the 2×2 table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where a denotes the number of within cell type replicating SNPs overlapping a CRE from cell type B, b denotes the number of within cell type replicating SNPs *not* overlapping a CRE from cell type B, c denotes the number of non-replicating SNPs overlapping a CRE from cell type B, d denotes the number of non-replicating SNPs *not* overlapping a CRE from cell type B.
 23. B Overlap Rep PV. P-value of the Fisher's exact test above.
 24. B Overlap Rep Beta. β_4 from Eqn. (3) for CREs derived from cell type B.
 25. B Overlap Rep SE. Standard error of the estimate of β_4 from Eqn. (3).
 26. B Overlap Rep Z. Z-score of the significance of β_4 from Eqn. (3).
 27. B Overlap Rep PV. P-value of the significance of β_4 from Eqn. (3).
 28. B Overlap Rep AIC. Akaike information criterion of Eqn. (3).
 29. AB Overlap Rep pos OR. Odds ratio derived from a Fisher's exact test performed on the 2×2 table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where a denotes the number of SNPs that replicate in cell type A but not B that overlap a CRE found in cell type A but not B, b denotes the number of SNPs that replicate in cell type A but not B that overlap a CRE found in cell type A and B, c denotes the number of SNPs that replicate in cell type A and B that overlap a CRE found in cell type A but not B, d denotes the number of SNPs that replicate in cell type A and B that overlap a CRE found in cell type A and B.
 30. AB Overlap Rep pos PV. P-value of the Fisher's exact test above.
 31. AB Overlap Rep neg OR. Odds ratio derived from a Fisher's exact test performed on the 2×2 table $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where a denotes the number of SNPs that replicate in cell type A but not B that overlap a CRE found in cell type B but not A, b denotes the number of SNPs that replicate in cell type A but not B that overlap a CRE found in cell type A and B, c denotes the number of SNPs that replicate in cell type A and B that overlap a CRE found in cell type B but not A, d denotes the number of SNPs that replicate in cell type A and B that overlap a CRE found in cell type A and B.
 32. AB Overlap Rep neg PV. P-value of the Fisher's exact test above.
- (TXT)

Table S5 UChicago_liver eQTL SNP enrichment/depletion within CREs: (a) relative to background, (b) replicating versus non-

replicating, (c) same versus different cell types. Details are the same as Table S4.

(TXT)

Table S6 Harvard_cerebellum eQTL SNP enrichment/depletion within CREs: (a) relative to background, (b) replicating versus non-replicating, (c) same versus different cell types. Details are the same as Table S4.

(TXT)

Table S7 AH model comparison table. Details of tests for independent enrichment of primary and secondary tier eQTL SNPs within CREs. CRE data sets are listed in separate rows, columns are labeled with a header in the first row as:

1. 1 Overlap Test Beta. Primary tier β_3 , relative to background SNPs, from Eqn. (4).
2. 1 Overlap Test SE. Standard error of the estimate of the primary tier β_3 from Eqn. (4).
3. 1 Overlap Test Z. Z-score of the significance of the primary tier β_3 from Eqn. (4).
4. 1 Overlap Test PV. P-value of the significance of the primary tier β_3 from Eqn. (4).
5. 2 Overlap Test Beta. Second tier β_3 , relative to background SNPs, from Eqn. (4).
6. 2 Overlap Test SE. Standard error of the estimate of the second tier β_3 from Eqn. (4).
7. 2 Overlap Test Z. Z-score of the significance of the second tier β_3 from Eqn. (4).
8. 2 Overlap Test PV. P-value of the significance of the second tier β_3 from Eqn. (4).

(TXT)

Table S8 Random forest classifier feature importance. Measures of feature importance for each variable for each classification task. Columns are:

1. classification task,
2. variable name,
3. variable importance. For each tree, the prediction accuracy on the out- of-bag portion of the data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all trees, and normalized by the standard error..

(TXT)

Table S9 The \log_{10} *BF* cutoffs and total numbers of eQTLs at $FDR \leq 5\%$ for each tier of AH in each study. FDR cutoff was assessed by permutation.

(TXT)

Table S10 Comparison of LD-based method and forward stepwise regression on subset of CAP_LCL genes with AH. Columns are: Gene names (note that these represent only single gene probe clusters, and not combined results across probes); the number of eQTLs for the LD-based method, the BIC score for the LD-based method, the multivariate \log_{10} *BF* for the LD-based method, the set of RSIDs for the LD-based method, the number of eQTLs for FSR, the BIC score for FSR, the multivariate \log_{10} *BF* for FSR, the set of RSIDs for FSR.

(TXT)

Table S11 Gene Ontology enrichment results. First two columns denote the eQTL discovery study and cell type; second two columns the eQTL comparison study and cell type. The *Genes* column denotes the type of enrichment analysis performed, where

multiple eQTLs indicates the enrichment of terms in genes with allelic heterogeneity with a background of genes with one or more eQTLs, *eQTLs* indicates the enrichment of terms in genes with eQTLs, with a background of all of the tested genes in the study, *NR* indicates the enrichment of terms in genes with eQTLs that do not reproduce across cell types, with the background of all eQTLs identified in the discovery data. The *Type* column indicates what term is enriched, where *KEGG* indicates a KEGG pathway, *MF* denotes a molecular function in GO, and *BP* indicates biological process. Any of the comparisons not mentioned (for the three types or the different study comparisons discussed) did not come up with any statistically significant enrichment ($FDR \leq 20\%$). (TXT)

Acknowledgments

The authors would like to thank Manolis Dermitzakis for access to the GenCode study data, and Greg Crawford for pre-release access to DNase I

References

- Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* 10: 184–94.
- Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. *Nature* 452: 423–8.
- Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in genetics* 24: 408–415.
- Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296: 752–5.
- Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422: 297–302.
- Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430: 743–7.
- De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, et al. (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312: 1215–7.
- Small KS, Hedman AK, Grundberg E, Nica AC, Thorleifsson G, et al. (2011) Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genetics* 43: 561–4.
- Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics* 39: 1208–16.
- Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448: 470–3.
- Emison ES, Garcia-Barcelo M, Grice EA, Lantieri F, Amiel J, et al. (2010) Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *American journal of human genetics* 87: 60–74.
- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 337: 1190–1195.
- Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* 6: e1000888.
- Fraser HB, Xie X (2009) Common polymorphic transcript variation in human disease. *Genome Research* 19: 567–75.
- Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466: 714–9.
- Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, et al. (2011) 9p21 DNA variants associated with coronary artery disease impair interferon- γ signalling response. *Nature* 470: 264–8.
- Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325: 1246–50.
- Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature Genetics* 44: 502–10.
- Powell JE, Henders AK, McRae AF, Wright MJ, Martin NG, et al. (2012) Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *PLoS One* 7: e45666.
- Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, et al. (2009) Extracting biological meaning from large gene lists with DAVID. *Current protocols in bioinformatics* Chapter 13: Unit 13.11.
- van Nas A, Ingram-Drake L, Sinshheimer JS, Wang SS, Schadt EE, et al. (2010) Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics* 185: 1059–68.
- Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics* 7: e1002003.
- Ding J, Gudjonsson JE, Liang L, Stuart PE, Li Y, et al. (2010) Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *American journal of human genetics* 87: 779–89.
- Fu J, Wolfs MGM, Deelen P, Westra HJ, Fehrmann RSN, et al. (2012) Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genetics* 8: e1002431.
- Gerrits A, Li Y, Tesson BM, Bystrikykh LV, Weersing E, et al. (2009) Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genetics* 5: e1000692.
- Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, et al. (2011) Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genetics* 7: e1002078.
- Heap GA, Trynka G, Jansen RC, Bruinenberg M, Swertz MA, et al. (2009) Complex nature of SNP genotype effects on gene expression in primary human leukocytes. *BMC medical genomics* 2: 1.
- Flutre T, Wen X, Pritchard J, Stephens M (2013) A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genetics* 9: e1003486.
- Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489: 75–82.
- Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315: 848–853.
- Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. *Nature Genetics* 39: 1494–9.
- Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* 6: e107.
- Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* 3: e114.
- Scheet P, Stephens M (2006) A fast and exible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics* 78: 629–44.
- Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 3: 1724–35.
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464: 768–772.
- Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. *PLoS Genetics* 4: e1000279.
- Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* 10: 681–90.
- Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics* 4: e1000214.
- Veyrieras JB, Gaffney DJ, Pickrell JK, Gilad Y, Stephens M, et al. (2012) Exon-specific QTLs skew the inferred distribution of expression QTLs detected using gene expression array data. *PLoS One* 7: e30629.
- Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* 482: 390–4.
- Gaffney DJ, Veyrieras JB, Degner JF, Pique-Regi R, Pai AA, et al. (2012) Dissecting the regulatory architecture of gene expression QTLs. *Genome biology* 13: R7.

44. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, et al. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128: 1231–45.
45. Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, et al. (2010) A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genetics* 6: e1000814.
46. Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43–9.
47. Cooper SJ, Trinklein ND, Nguyen L, Myers RM (2007) Serum response factor binding sites differ in three human cell types. *Genome Research* 17: 136–44.
48. Chen H, Tian Y, Shu W, Bo X, Wang S (2012) Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One* 7: e41374.
49. Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106: 9362–7.
50. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* 81: 559–75.
51. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology* 34: 816–34.
52. Howie B, Marchini J, Stephens M (2011) Genotype Imputation with Thousands of Genomes. *G3: Genes, Genomes, Genetics* 1: 457–470.
53. 1 kg (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–73.
54. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17: 520–525.
55. Barber MJ, Mangravite LM, Hyde CL, Chasman DI, Smith JD, et al. (2010) Genome-wide association of lipid-lowering response to statins in combined study populations. *PLoS One* 5: e9763.
56. Fraley C, Raftery A, Wehrens R (2005) Incremental Model-Based Clustering for Large Datasets With Small Clusters. *Journal of Computational and Graphical Statistics* 14: 529–546.
57. Kudaravalli S, Veyrieras JB, Stranger BE, Dermizakis ET, Pritchard JK (2009) Gene expression levels are a target of recent natural selection in the human genome. *Molecular biology and evolution* 26: 649–58.
58. Maranville JC, Luca F, Richards AL, Wen X, Witonsky DB, et al. (2011) Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. *PLoS Genetics* 7: e1002162.
59. Petretto E, Bottolo L, Langley SR, Heinig M, McDermott-Roe C, et al. (2010) New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS computational biology* 6: e1000737.
60. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–61.
61. Wen X, Stephens M (2011) Bayesian Methods for Genetic Association Analysis with Heterogeneous Subgroups: from Meta-Analyses to Gene-Environment Interactions. arXiv:1111.1210v2.
62. Wang S, Yehya N, Schadt EE, Wang H, Drake Ta, et al. (2006) Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics* 2: e15.
63. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, et al. (2012) Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLoS Genetics* 8: e1002639.
64. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genome-wide expression in yeast. *PLoS Biology* 3: e267.
65. Ashburner M, Ball C, Blake J (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25–29.
66. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* 40: D109–14.
67. Gardiner-Garden M, Frommer M (1987) CpG Islands in vertebrate genomes. *Journal of Molecular Biology* 196: 261–282.
68. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglu S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* 15: 901–13.
69. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, et al. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods* 9: 473–476.
70. Breiman L (2001) Random Forests. *Machine Learning* 45: 5–32.
71. Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R news* 2: 18–22.
72. Sing T, Sander O, Beerwinkler N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21: 3940–1.
73. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, et al. (2008) Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* 451: 796–801.
74. Simon JA, Lin F, Hulley SB, Blanche PJ, Waters D, et al. (2006) Phenotypic predictors of response to simvastatin therapy among African-Americans and Caucasians: the Cholesterol and Pharmacogenetics (CAP) Study. *The American journal of cardiology* 97: 843–50.