Bayesian Integration of Genetics and Epigenetics Detects Causal Regulatory SNPs Underlying Expression Variability

Avinash Das
Michael Morley
University of Pennsylvania
Christine S. Moravec
W.H.W. Tang
Hakon Hakonarson
University of Pennsylvania

See next page for additional authors

Follow this and additional works at: https://repository.upenn.edu/statistics_papers

Part of the Business Commons, Genetics and Genomics Commons, Genetic Structures Commons, and the Statistics and Probability Commons

Recommended Citation
Das, A., Morley, M., Moravec, C. S., Tang, W., Hakonarson, H., Consortium, M., Margulies, K. B., Cappola, T. P., Jensen, S. T., & Hannenhalli, S. (2015). Bayesian Integration of Genetics and Epigenetics Detects Causal Regulatory SNPs Underlying Expression Variability. Nature Communications, 6 http://dx.doi.org/10.1038/ncomms9555

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/statistics_papers/617
For more information, please contact repository@pobox.upenn.edu.
Bayesian Integration of Genetics and Epigenetics Detects Causal Regulatory SNPs Underlying Expression Variability

Abstract
The standard expression quantitative trait loci (eQTL) detects polymorphisms associated with gene expression without revealing causality. We introduce a coupled Bayesian regression approach—eQTeL, which leverages epigenetic data to estimate regulatory and gene interaction potential, and identifies combination of regulatory single-nucleotide polymorphisms (SNPs) that explain the gene expression variance. On human heart data, eQTeL not only explains a significantly greater proportion of expression variance but also predicts gene expression more accurately than other methods. Based on realistic simulated data, we demonstrate that eQTeL accurately detects causal regulatory SNPs, including those with small effect sizes. Using various functional data, we show that SNPs detected by eQTeL are enriched for allele-specific protein binding and histone modifications, which potentially disrupt binding of core cardiac transcription factors and are spatially proximal to their target. eQTeL SNPs capture a substantial proportion of genetic determinants of expression variance and we estimate that 58% of these SNPs are putatively causal.

Disciplines
Business | Genetics and Genomics | Genetic Structures | Statistics and Probability

Author(s)
Avinash Das, Michael Morley, Christine S. Moravec, W.H.W. Tang, Hakon Hakonarson, MAGNet Consortium, Kenneth B. Margulies, Thomas P. Cappola, Shane T. Jensen, and Sridhar Hannenhalli

This technical report is available at ScholarlyCommons: https://repository.upenn.edu/statistics_papers/617
Bayesian integration of genetics and epigenetics detects causal regulatory SNPs underlying expression variability

Avinash Das1, Michael Morley2, Christine S. Moravec3, W.H.W. Tang3, Hakon Hakonarson4, MAGNet Consortium†, Kenneth B. Margulies2, Thomas P. Cappola2, Shane Jensen5 & Sridhar Hannenhalli1

The standard expression quantitative trait loci (eQTL) detects polymorphisms associated with gene expression without revealing causality. We introduce a coupled Bayesian regression approach—eQTeL, which leverages epigenetic data to estimate regulatory and gene interaction potential, and identifies combination of regulatory single-nucleotide polymorphisms (SNPs) that explain the gene expression variance. On human heart data, eQTeL not only explains a significantly greater proportion of expression variance but also predicts gene expression more accurately than other methods. Based on realistic simulated data, we demonstrate that eQTeL accurately detects causal regulatory SNPs, including those with small effect sizes. Using various functional data, we show that SNPs detected by eQTeL are enriched for allele-specific protein binding and histone modifications, which potentially disrupt binding of core cardiac transcription factors and are spatially proximal to their target. eQTeL SNPs capture a substantial proportion of genetic determinants of expression variance and we estimate that 58% of these SNPs are putatively causal.
Numerous expression quantitative trait loci (eQTL) studies have been performed to determine the cell-type-specific regulatory architecture of the human genome. However, since single-nucleotide polymorphisms (SNP) within a linkage disequilibrium (LD) region are statistically indistinguishable from each other, these studies essentially reveal LD blocks that are associated with a gene expression but do not reveal the potential causative regulatory SNPs, which limits the utility of these studies. The recent explosion of epigenetic data has made it possible to detect cell-type-specific regulatory regions, which can be used to distinguish regulatory SNPs from non-regulatory SNPs in LD blocks.

Recently, a few approaches have incorporated regulation specific epigenetic data into association studies. However, to prioritize eQTL SNPs, these methods have utilized the regulatory information either retrospectively or to estimate an empirical prior probability for the SNPs. Such approaches are prone to missing regulatory SNPs with small effects due to the severe multiple testing correction (or sparsity constraints). Furthermore, these approaches ignore interaction between the region harbouring the SNP and the target gene, which is useful in identifying regulators specific to a gene. Multiple SNPs are known to regulate single genes, yet many current methods limit the number of causal SNPs per gene to a single SNP. In this paper, we introduce a new method, expression quantitative trait enhancer loci (eQTeL), which addresses these limitations. It identifies combination of regulatory SNPs—including SNPs with small effect sizes—that jointly determine expression variance.

eQTeL is a fully Bayesian approach (Fig. 1), which infers cis regulatory polymorphisms underlying gene expression variability by integrating: (i) genotype and gene expression variance across individuals; (ii) epigenetic data in appropriate cell types; (iii) DNase I hypersensitivity (DHS) variance of SNPs and promoters across cell types; (iv) expression variance of genes across multiple cell types; (v) LD blocks; and (vi) imputed haplotypes inferred from the 1,000 Genomes Project. Our approach addresses a number of key methodological challenges. First, it systematically integrates three characteristics of a causal regulatory eQTL, that is, correlation with the target genes expression across individuals, the regulatory properties of the harbouring region, and interaction with the target gene. Second, it can account for heterogeneity of regulatory regions in terms of different combinations of epigenetic marks. Third, to learn the regulatory model, eQTeL leverages regulatory polymorphisms that are not associated with gene expression in addition to expression-regulators. Fourth, it interrogates the LD structure to find the optimal combination of explanatory SNPs. Fifth, it implements a hierarchical scheme to select a sparse set of SNPs, while simultaneously explaining a maximal fraction of gene expression variance. Finally, eQTeL is scalable to large datasets.

We statistically validated our method using human heart data as well as realistic simulated data and demonstrated that it can predict an individual’s expression from the genotype more accurately compared to other methods. SNPs identified by our method include regulatory SNPs with small effect sizes. Further assessment of functional relevance of identified SNPs suggest that they tend to (i) overlap a high resolution DNase footprint, (ii) have an allele-specific DNase footprint, (iii) preferentially disrupt putative binding of core cardiac regulators and (iv) be spatially proximal to their putative target gene. We also estimate

Figure 1 | Overview of eQTeL model. (a) Input and output of eQTeL. eQTeL takes genotype and gene expression across samples, epigenetic and interaction features for each SNP and LD block as input. It outputs regulatory SNPs and their target genes, their effect sizes and regulatory-interaction potentials, as well as estimated feature importance of each epigenetic and interaction feature. (b) eQTeL is composed of two coupled regression models (i) a Bayesian variable selection with informative priors models expression as a linear combination of SNPs. Given the regulatory and interaction priors, this hierarchical model first identifies LD blocks and then combinations of SNPs that explains expression variance and that also have high regulatory and interaction potentials. (ii) A Bayesian logistic regression specifies the regulatory and interaction potential as linear model of epigenetic and interaction features in semi-supervised manner. The logistic regression passes the regulatory and interaction potentials to the variable selection model, while the variable selection model passes expression-regulators to the logistic regression model.

© 2015 Macmillan Publishers Limited. All rights reserved.
that 58% of SNPs identified by eQTeL (which we call eeSNPs, Supplementary Data 1) are likely to be causal. Collectively, these results strongly suggest that eeSNPs have functional role.

Results
Quantitative Trait enhancer Loci (eQTeL) model. We first provide a broad overview of the eQTeL model and further details can be found in Methods. As illustrated in Fig. 1, eQTeL is composed of two Bayesian regression models, an expression model and a regulatory model, which are coupled through message passing. The expression model is a Bayesian variable selection model16,17, which explains the gene expression variance among samples as a linear function of SNP alleles. A distinct feature of the expression model is that it uses informative prior for each SNP, which depends on the SNPs regulatory and interaction potential. The regulatory model, which is common for all genes, uses a Bayesian logistic regression18 to estimate that informative prior as a probabilistic function of epigenetic and interaction features. Known expression regulators can be used to train the regulatory model, while an accurate model of regulatory and interaction potential can help to identify expression regulators. The expression model then passes current estimates of expression regulators to the regulatory model, which in passes current estimates of regulatory and interaction priors for each SNP back to the expression model. eQTeL starts with estimating expression regulators assuming equal priors for each SNP and then, using current estimates of expression-regulators, trains the regulatory-model. In turn, current estimates of regulatory and interaction potential are used as informative priors to re-estimate expression regulators. This iterative process continues until convergence. Thus, our eQTeL model gradually improves estimation accuracy by joint learning.

In our approach (see equations below and Methods for details), expression $Y$ relates to candidate SNPs $X$ via a standard normal linear model16,19,20 with noise $\sigma^2$. However, for each SNP $\beta$, its effect size is non-zero only if its regulatory-interaction indicator $\gamma$ is 1, which depends on a function $\phi(\theta)$ of regulatory-interaction potential $\theta$ (Methods). The potential $\theta$ of a SNP is modelled as a combination of (i) features for regulatory potential and (ii) features for SNP-gene interaction $P$, via a logistic function. Vector $\alpha$ represents feature weights that are shared across all genes, thus we learn a single genome-wide model of regulators. This choice of modelling $\alpha$ obviates the need to explicitly scale genetic and epigenetic factors.

$$
\begin{align*}
Y & \sim \mathcal{N}(X \cdot \beta, \sigma^2 I), \\
\gamma & \sim \text{Bern}(\phi(\theta)) \quad \text{VSNPs} \\
\theta & \sim \text{Bern(logistic}(E, P \cdot \alpha)) \quad \text{vSNPs}
\end{align*}
$$

We use Markov chain Monte Carlo21 to infer all model parameters jointly (Supplementary Note 1). At each iteration of the sampler, the decision whether a region is a regulator (that is, $\theta = 1$) depends not only on correlation between corresponding SNP and gene but also on the regulatory and interaction features, as well as the current estimates of feature weights. This leads to a semi-supervised clustering of SNPs into regulators and non-regulators (Supplementary Note 1). Our Markov chain Monte Carlo implementation explicitly uses LD block information to judiciously choose combination of regulatory SNPs by sampling over the model space hierarchically22,23 at the top level it explores combinations of LD blocks and at the lower level it explores the sparse set of SNPs within each LD block that optimally explain the expression-variance (Fig. 1, Methods, Supplementary Note 1, Supplementary Fig. 1). This approach results in a superior exploration of the model space relative to approaches that disregard the LD structure. eQTeL uses a Rao–Blackwell estimate of $\theta$ that improves the mixing rate (Supplementary Fig. 1) of the sampler and leads to robust competition between SNPs within a LD block (Fig. 1). Further, the overall sparsity constraint (equivalent to a multiple testing correction in non-Bayesian approaches) of eQTeL is controlled by two factors: (i) the fraction of SNPs that are interacting-regulators and (ii) the fraction of interacting-regulators that are expression-regulators. This allows for a less conservative sparsity constraint and makes it possible to identify SNPs with small effect sizes which are typically missed by alternative approaches because of severe multiple testing correction. eQTeL assumes Normal priors on $\alpha$. Finally, eQTeL implementation allows an option to select a subset of epigenetic factors important for estimating regulatory potential through Bayesian variable selection model.

eQTeL detects expression regulatory SNP in MAGNet. We applied eQTeL to genotype and gene expression data for 313 human hearts (obtained by MAGNet consortium (www.med.upenn.edu/magnet/)) and compared with the performance of other eQTL methods (Supplementary Notes 2 and 3). To determine regulatory and interaction potentials, we used 95 epigenetic and interaction features (Supplementary Fig. 2) for primary tissues and cell lines of heart from ENCODE and Roadmap Epigenome project10,12. For expediency we selected 1,880 genes with expression deemed to have a significant genetic component according to the univariate eQTL11,25.

Consistent with its ability to explain a greater expression variance, eQTeL also predicts expression of genes much more accurately compared with other methods (Fig. 2b). The mean (cross-validated) Pearson correlation coefficient between predicted and actual expression is $0.176 \pm 0.065$ (in contrast with 0.025 for eqtnminer8 and 0.088 for LASSO26). The bimodality of distribution of correlation coefficient implies that for a subset of genes, the expressions are highly predictable by eQTeL.

Because of its ability to discriminate among multiple SNPs based on regulatory and interaction potentials, eQTeL is expected to be much more advantageous on imputed data, which has a substantially greater number of linked SNPs. To confirm this, we imputed22 ~6.5 million SNPs using the 1,000 Genome Project data15. Note that each imputed SNP is derived from the reference SNPs using the linkage information, and cannot be any more associated (in a statistical sense) with the gene expression than the reference SNPs, and therefore are not expected to increase the explained variance (as evident from Fig. 2c). However, eQTeL with imputation is expected to improve detection of causal functional SNPs compared with the genotyped SNPs10,11. Therefore, restricting our search to potentially functional SNPs, imputed SNPs should explain the expression better. Restricting our analysis on only to SNPs mapped to a DNAse footprint (as a proxy for putative functional SNPs), the relative advantage of imputation with eQTeL becomes evident (Fig. 2c).

To validate eeSNPs in an independent cohort, we analysed expression and genotype of 85 left ventricle samples from GTEx1 (Supplementary Note 2). We note that compared to an exhaustive eQTL, eQTeL cannot identify novel associated loci, but instead is designed to identify putatively causal SNPs within an associated locus. We found that 18.9% of eGenes detected in MAGNet by alternative approaches because of severe multiple testing correction. Therefore it becomes imperative to use an integrative approach, such as eQTeL, in the presence of a large number of linked SNPs (Fig. 2c).
identified by eQTeL in MAGNet, we estimated the explained variance in GTEx. We repeated this for other methods while controlling for the number of eeSNPs as well as other regularization procedures. While, as expected due to the differences in the datasets, the cross-cohort explained variance is lower than that within MAGNet (Fig. 2b versus 2d), relative to other methods, eQTeL exhibits substantially and significantly greater (in both cases Wilcoxon test P value between eQTeL and other methods is $<1.0 \times 10^{-16}$) cross-data set generalizability (Fig. 2d, Supplementary Fig. 3).

**eQTeL detects causal SNPs in semi-synthetic data.** To demonstrate that eQTeL can accurately identify putatively causal SNPs, we use a synthetic data evaluation (Fig. 3a) (for additional details refer to Methods). We used 174,800 SNP probes along with their genotypes from 313 MAGNet samples that were within 1 MB from transcription start of 200 genes (Methods). Since regulatory region may have no effect on genes included in our analyses and yet can contribute to learning the regulatory-model, eQTeL makes a distinction between a regulator and a gene-specific expression-regulator. This distinction was made explicitly in our simulation by designating 1% of all SNPs as regulators (as an approximation of previous estimation in humans$^{28}$). We then used a frequency distribution of expression regulators per gene inferred from MAGNet data to randomly choose gene-specific expression-regulators for 200 genes. Using allele status of 313 samples for expression-regulators, we generated gene expression and added random noise such that expected explained variance from simulated data matched MAGNets explained variance (Fig. 2a). We generated the epigenetic features for each SNP using ENCODE epigenetic data and validated heart-enhancers from VISTA$^{6}$. Thus our simulated data closely parallels the experimental data.

Next we applied eQTeL to the simulated data. The precision-recall plot (Fig. 3b) shows that eQTeL significantly outperforms other methods. In fact, the performance of full-eQTeL is close to the theoretically best eQTeL model that uses the original feature weights (see Methods). The previous integrative method eqtnminer$^{8,9}$, the only other current method that uses epigenetic data in eQTL, shows only a modest increase in precision compared to methods that do not use epigenetic data.

The immediate effect of increase in precision of detecting expression regulators, especially for SNPs with high regulatory potential, is that eQTeL explains a significantly greater proportion of expression variability (Supplementary Fig. 4). There is also significant improvement in correlation between predicted expression and actual gene expression; mean correlation for eQTeL was 0.298 $\pm$ 0.02 (compared with 0.18 for eqtnminer and 0.23 for LASSO regression, Supplementary Fig. 5). Note that for this
analysis we controlled for the number of SNPs that were selected for each method, using the most explanatory respective SNPs for each method. Overall, eQTeL can accurately identify around 75% of putative causal SNPs (at 40% recall) reinforcing the fact that our method can identify substantial fraction of likely causal genetic determinants of transcriptomic variance.

**eQTeL detects SNPs with small effect sizes.** The statistical power to detect SNPs associated with expression variance (that is, the probability of correctly rejecting the null hypothesis that the SNP is not associated with gene expression) depends on various factors such as sample size, noise to signal ratio, number of hypothesis tested (number of SNPs) and effect size of SNP. The effect size, in turn, depends on the allele frequency of SNP, thus low allele frequency limits statistical power to detect regulatory SNPs.

Another advantage of eQTeL model is that it can detect SNPs with small effect sizes by distributing sparsity between: (a) sparsity in the number of regulators and, (b) sparsity in expression regulators among all regulators. eQTeL employs relaxed sparsity constraints for SNPs that have high regulatory potential and therefore the model has higher statistical power to retrieve a greater fraction of SNPs with low minor allele frequency (small effect sizes) compared to eqtnminer. Furthermore, eQTeLs statistical power to identify low minor allele frequency SNPs is greater among SNPs with high regulatory-interacting potential (labelled as eQTeL-high in Fig. 4). This trend of differential statistical power is also observed in simulated data, where we know the exact effect size of regulatory SNPs (Supplementary Fig. 6).

eQTeL leverages LD information to judiciously choose combinations of SNPs (per gene) which explains a much greater proportion of expression variance (details in Supplementary Note 2). The power to detect SNPs with low allele frequency is the primary reason that eQTeL captures substantial proportion of causal genetic determinants underlying transcriptomic variance. However, it should be noted that SNPs with small effect sizes are only detected by eQTeL if they have a high regulatory potential.

eQTeLs performance gain is potentially due to two factors: (i) integration of epigenetic data, (ii) allowing multiple causal variants per gene. We assessed relative contribution due to multiple SNPs per gene (Supplementary Fig. 7, Supplementary Note 2), supporting a contribution due to multiple explanatory SNPs. However, in the absence of epigenomic data, that is, when using standard LASSO, we do not see a performance gain, and in general, the performance is substantially worse than the performance of eQTeL. This suggests that allowing multiple SNPs per gene is useful specifically when functional information is used.

**eeSNPs lie within protein-bound genomic regions.** Putative causal regulatory SNPs are expected to be bound by regulatory proteins. Earlier studies have shown enrichment of regulatory
elements near causal SNPs\textsuperscript{7–9,11}. Since eQTeL and eqtnminer use epigenetic data, which is known to be correlated\textsuperscript{10} with protein binding, we expect to find enrichment of DNAse footprints near the identified regulatory SNPs. Using genome-wide high-resolution DNAse footprint data for 41 cell types\textsuperscript{31}, we obtained the fraction of eeSNPs (control SNPs) overlapping with a footprint; Note that DNAse footprints were not used in eQTeL so they could be used for validation. 76.3% of eeSNP have a footprint overlapping the eeSNP (Fig. 5), in contrast to 6.3% of in SNPs detected by eqtnminer that uses same epigenetic data as eQTeL. The performance of eqtnminer did not improve even if the best SNP per gene were chosen for this analysis. For SNPs chosen by LASSO, which does not use epigenetic data, only 5.95% of SNPs have overlapping DNAse footprints. Only 2% of SNPs identified by Lirnet (for 200 genes) overlap with the DNAse footprints (Supplementary Fig. 8). Notably, the DNAse footprint enrichment is high in the four heart-related cell types.

### eeSNPs exhibit binding and regulatory allele specificity.

To ascertain the functional role of eeSNPs, we checked whether the change of a SNPs allele would affect their regulatory properties (such as protein binding, histone modifications and so on). For each cell line, we selected heterozygous SNPs by inspecting genotyped data or pooled reads from different histone modifications, DNAse-seq and CTCF. We first assessed allelic differences in footprint reads for human cardiac myocyte (HCM) (see Methods). As shown in Fig. 6, the eeSNPs that overlap a footprint show significantly greater (with odd-ratio of $M = 3.005$ and $P \text{ value } < 3.83 \times 10^{-17}$) allele-specificity relative to SNPs identified by eqtnminer, consistent with eeSNP having a regulatory impact (allele-specificity comparison with LASSO is shown in Supplementary Fig. 10). For eeSNPs, we obtained 6.57-fold more reads mapping to the allele with more DNA-seq reads compared with the other allele (for eqtnminer, the average read difference was 1.8). We also found higher allele specificity for eeSNPs in other heart cell lines (Supplementary Fig. 11, HCF, SKMC) for DNASe-Seq reads. The trend of higher allelic specificity is also true in heart cell lines for histone modification H3K4me3, which is associated with active enhancers (Supplementary Fig. 12). Allele-specificity of eeSNPs suggests that they may underlie population variance in gene expression.

### eeSNPs are spatially proximal to their target gene.

The spatial proximity of eeSNP with its target promoter is a pre-requisite for cis-regulation. Spatial proximity has been experimentally determined using chromatin interaction analysis with paired-end tags (ChIA-PET) assays\textsuperscript{32}. Identified SNPs that were closer than 100 bps from their target promoters were excluded. We quantified spatial proximity of each eeSNP and its target by the number of...
Figure 6 | DNAse hypersensitivity at eeSNPs shows greater allele specificity in HCM. X axis: rank of DHS read counts, Y axis: absolute log-ratio of read counts mapping to the two alleles at a SNP. SNPs from different methods are selected similarly to Fig. 5. The analysis was performed on a subset of SNPs that were heterozygous in the sample. The median "white" lines represent LOESS (local regression) for each method. Confidences intervals for each median line is estimated using bootstrapping and are represented either by thin lines representing the LOESS of each bootstrap or by colored shades representing confidence intervals in terms of standard deviation of bootstraps. Note the allele-specificity at SNPs detected by eQTeL and eqtnminer remains the same even if we control for number of SNPs per gene.

Figure 7 | eeSNP-gene pairs are spatially proximal. X axis: the rank of eeSNP-gene distance (log 10), Y axis: ChIA-pet support. SNPs from eQTeL and eqtnminer are selected as in Fig. 8. The random SNP-gene pairs were selected so as to have the same distance distribution as for eeSNPs. SNP-gene pair closer to 100 bps were excluded. The median ‘white’ lines represent LOESS (local regression) for each method. Confidence was estimated for each method just as in Fig. 6.

Discussion

Here we have introduced a novel Bayesian approach, eQTeL, that integrates genetic and epigenetic data in a statistically consistent manner to identify putatively causal genetic variants underlying the expression variance. We have shown that (i) eQTeL identifies combinations of SNPs (eeSNPs) that, compared with other methods, explain substantially greater portion of expression variability, (ii) eQTeL is especially effective in identifying SNPs with small effect sizes, (iii) 58% of the identified eeSNPs are likely to be causal, (iv) eeSNPs can predict sample specific expression much more accurately, (v) eeSNPs are much more likely to be bound by a regulatory factor in an allele-specific manner, (vi) eeSNPs preferentially disrupt core cardiac transcription factor binding and (vii) eeSNPs tend to be spatially proximal to their target genes. Taken together, our results strongly suggest that eQTeL captures a substantial proportion of putative causal regulatory genetic determinants underlying transcriptomic variance.

It is important to note limitations of eQTeL. First, eQTeL can only detect cis-eQTL and not trans-eQTL. Second, like other model-based association methods, eQTeL’s computational speed is a bottleneck; however, using parallel cores and certain reasonable compromises in parameter estimation procedure, the computational burden can be substantially reduced. Third, eQTeL assumes normality of expression data, therefore the expression data needs to be pre-processed accordingly, which can be particularly problematic for certain kinds of high throughput data. Fourth, eQTeL can only detect SNPs with small effect size if they have high regulatory potential. Finally, eQTeL statistically
infers the potentially causal SNPs and further experimental validations are required to establish causality.

eQTeL can effectively resolve LD and discriminate putative regulatory SNPs from myriad associated SNPs. This lays a foundation for future experimental studies to characterize genetic variants underlying disease risk. Finally, eQTeL can be extended by integrating additional layer of molecular data—easily achieved in Bayesian framework—to directly infer SNP that causes disease.

Methods

Modelling regulatory-interaction potential. There are $R_1$ epigenetic features $E_i$ that were used to predict if a SNP $i$ lies in a regulatory region. In addition, we also have $R_2$ interaction features $P_i$ that are predictive of the interaction between SNP $i$ and gene $j$. We refer to a SNP that has high regulatory potential and high interaction potential as interacting-regulator, regardless of whether it associates with gene expression. Further, if the SNP is associated with gene expression, we refer to that SNP as expression-regulator. In our eQTeL approach, we model the regulatory-interaction potential $\theta_{ij}$ between SNP $i$ and gene $j$ as a combined function of epigenetic features $E_i$ and interaction features $P_i$. Specifically, we use a Bayesian logistic regression model:

$$0_i \sim \text{Bern}(\text{logistic}(F_0 \cdot \alpha)),$$

where $F_0$ is a concatenated set of features consisting of both $E_i$ and $P_i$, and Bern is the Bernoulli distribution. The coefficients $\alpha$ are shared across all genes.

Modelling gene expression. In our model, the expression of gene $j$ depends not only on the allele status of candidate SNPs, but also on the estimated regulatory-interaction potential of the SNP $i$ and gene $j$ pair. Specifically, given gene expression in $n$ samples $Y_i = (Y_i^1, \ldots, Y_i^n)$, we model the vector of expression $Y_i$ for gene $j$ as a linear function of the allele status for all candidate SNPs, $X = [X_1, \ldots, X_p]$ where $X_i$ is allele status of SNP $i$ over the $n$ samples:

$$Y_i | \beta, X, \gamma_i \sim \mathcal{N}(X_i \cdot \beta, \sigma^2 I),$$

where the effect $\beta_0$ of SNP $i$ on the expression of gene $j$ is non-zero only when indicator variable $\gamma_i = 1$. In other words, $\gamma_i = 1$ signifies whether SNP i is associated with the expression of gene $j$. $\beta_j$ is a subset of SNPs for which $\gamma_i = 1$.

If a SNP lies within a genomic region that is deemed to be (i) a regulator, and (ii) interacting with the target gene, then the SNP is likely to affect the gene’s expression. Thus, the regulatory-interaction potential for each pair of SNP $i$ and gene $j$ enters our expression model through the prior distribution on the indicator variables $\gamma_i$.

$$\gamma_i | \phi(\theta_i) \sim \text{Bern}(\phi(\theta_i)) \quad \forall \text{SNPs } i$$

where the function $\phi(\theta)$ is defined so that $\phi(\theta) = \pi^\theta (1-\pi)^{1-\theta}$ with $\pi$ being our prior probability for each SNP to be expression-regulator and let $\tau = \pi/1-\pi$ be the prior probability when the SNP does not reside in such a region, where $\mu$ is an amplification factor. An uniform prior for $\pi \in (m/\mu, M/\mu)$ is defined, where $m$ and $M$ are respectively the minimum and the maximum number of expected expression-regulators. However, no substantial difference in results was observed when we just fixed $\pi = m/\mu$ where $m$ is expected number of expression regulators. A value of $\mu = 100$ was used because performance of model was insensitive to choice of $\mu \in (100, 1000)$.

Because of severe multiple testing corrections, association studies miss many potential causal regulators that have relatively small effect on expression. In our eQTeL model, overall sparsity is controlled by two factors: (a) the fraction of SNPs which are interacting-regulators, that is, $E(0)$ and (b) the fraction of interacting-regulators which are expression-regulators, that is, $\pi$. This is because the overall sparsity is a product of the two factors, that is, $\log E(0) \approx E(0) \log \pi$ assuming $\hat{\tau} >> 1$. Thus, the effective sparsity constraints are less conservative on SNPs that lie within an interacting-regulator in our eQTeL model, which allows us to capture potential causal expression-regulator SNPs with small (but non-zero) effects on expression variance (Fig. 4 and Supplementary Fig. 6; refer to Supplementary Note 5).

We also employ a standard prior distribution, Zeller’s g-prior\(^2\), for our linear model parameters,

$$\beta_i | \gamma, c \sim \mathcal{N}(0, c \gamma (X_i'X_i)^{-1}), \quad p(\sigma^2) \propto 1/\sigma^2$$

and we also define the following prior distributions for the rest of the parameters as

$$c \sim IG(\frac{1}{2}, \frac{n}{2})$$

$$\alpha \sim \mathcal{N}(\mathbf{0}, 100 \cdot \mathbf{I})$$
The first element of \( \mathbf{x} \), \( x_{0} \), is the bias term, and \( b \) is the prior for \( x \). The prior is set to 0, except for \( b_{0} \) (the prior for \( n_{0} \)), which can be used to control the sparsity on the number of interacting-regulators. We expect 1% of all SNPs to be regulators. To achieve this level of sparsity in number of regulators, \( s_{0} \) was set to \( \log(e^{p-c}) \), where \( p \) is expected number of interacting-regulators, and was set to \( p/100.0 \). That is, \( b_{0} = \log(1999) \).

Refer to Supplementary Note 1 for the eQTL's inference algorithm, initialization and convergence criteria.

Cardiac expression data (MAGNet). Samples of cardiac tissue (\( n = 313 \)) were acquired from patients from the Myocardial Applied Genomics Network (MAGNet; www.med.upenn.edu/magnet/). Left ventricular free-wall tissue was harvested at the time of cardiac surgery from subjects with heart failure undergoing transplantation and from unused donor hearts. Genomic DNA was extracted using the Gentra PureGene Tissue Kit (Qiagen, CA) according to manufacturer’s instructions. Total RNA was extracted using the miKNeasy Kit (Qiagen) including DNase treatment. RNA concentration and quality was determined using the NanoVue Plus spectrophotometer (GE Healthcare) and the Agilent 2,100 RNA Nano Chip (Agilent). To assess gene expression, RNA was hybridized with Affymetrix GeneChip ST11 arrays using manufacturer instructions. CEL files were normalized with the robust multi-array analysis using the oligo package in Bioconductor.

To reduce potential batch effects, expression values were further adjusted using ComBat, an empirical Bayes method that estimates parameters for location and scale adjustment of each batch for each gene independently. Probe sets were removed when they displayed robust estimates for expression values \( < 0.8 \) on all arrays. This filtering yielded sets of genes present well above background levels in the human heart. Probeset showing no annotated cross hybridization potential were kept, leaving 15,395 probes for final analysis.

Selection of genes. The genes were selected such that they had at least one significantly associated SNP based on univariate-eQTL (Matrix eQTL). 1,880 epigenetic data and interaction features within 200 Kb was used to estimate it. To achieve this level of sparsity in number of interacting-regulators, we expect 1% of all SNPs to be regulators. To determine number of factors (K) to be removed using PEER, we used \( O_{10} \) because number of eGenes \( = O_{10} \). We chose \( K = 10 \) because number of eGenes plateaued at \( K = 10 \). Factors from PEER were removed from the expression and residual analysis or used for performing association studies. In certain cases, in such as trans-eQTL, a genetic factor can affect multiple SNPs and PEER might remove biologically relevant signal. However, since the aim of the gene is to identify cis-eQTL, that is local effects, we can safely use PEER.

To determine number of factors (K) to be removed using PEER, we used an approach similar to Lappalainen et al. We ran PEER for 16,271 Affymetrix gene probes from MAGNet using parameter \( K = 0, 3, 5, 10, 15 \) and 20; then we compared number of genes (eGenes) that have at least one SNPs significantly associated with expression (P value \( < 0.10 \)). We chose \( K = 10 \) because number of eGenes plateaued at \( K = 10 \). Factors from PEER were removed from the expression and residual analysis or used for further analyses.

Linear regression assumes normality of the expression data. We converted the residual data from PEER to standard normal distribution before performing the association analysis.

Genotypes and imputation for cardiac samples. DNA samples were genotyped using Affymetrix Genome Wide SNP Array 6.0 and analysed per manufacturer instructions. We applied quality control (QC) filters to exclude unreliable samples, samples with cryptic relatedness and samples that were not genetically inferred Caucasian. After QC filtering, 313 individuals remained. All analyses were conducted using software package PLINK. For the analysis reported here, we eliminated SNPs with genotype call rate \(< 90\%\), with minor allele frequency (MAF) \(< 0.15\%\), or if there was significant departure from Hardy-Weinberg equilibrium (\( P < 10^{-5} \)). A total of 360,046 SNPs passed QC and were available for analysis. To improve cross study comparisons, genotype imputation was performed using the Minimac (\( \nu 2012.11.16 \)) (ref. 27). Imputation results were filtered at an imputation quality threshold of 0.5 and a MAF threshold of 0.15. PLINK was used to infer LD block for the genotypes. Default setting of SNPs within 200 Kb was used to estimate it.

Epigenetic data and interaction features. Epigenetic data were obtained from ENCODE, Roadmap epigenome project and GEO database for following heart tissues: AooAF, HCM, HCP, fetal-hearts, adult-hearts, left ventricle, right ventricle, aorta, and right atrium. Because DNase I footprints were used to validate eSNPs, they were excluded from the feature importance (\( \phi \)) estimation of eQTeL. Supplementary Fig. 2 lists the epigenetic and interaction features, that were critical for identification of interacting-regulators. We assessed the importance of epigenetic factors directly overlapping each SNP within 500 bp flanking region (suffix .500 in Supplementary Fig. 2). We also assessed the importance of epigenetic factors in broader context of each SNP within 500 bp flanking region (suffix .500 in Supplementary Fig. 2). Interaction features between a gene-promoter and a region containing SNP were calculated using RNAseq and DHS data from 15 cell types (A549, B, H1hesc, Hepg2, Hsmm, K562, Nhek, Ag04450, Gm12878, Hela3, Hmec, Hucve, MeC7, NLhf and Slkensra). These features include correlation and absolute correlation between DHS of the region and DHS of the promoter (b) correlation and absolute correlation between DHS of the region and RNAseq FPKM of the gene.

Both epigenetic and interaction features were normalized to mean of 0 and standard deviation of 1. This implies that distribution of each of these features for a set of random SNPs were expected to have 0 mean and 1 s.d. Therefore, y axis in Supplementary Fig. 2 shows absolute enrichment over random-SNPs with units in s.d.

Estimating fraction of putatively causal eSNP. Using an approach similar to Lappalainen et al, we estimated proportion of eSNPs that are putatively causal. Clearly, an independent estimation of proportion of causal SNPs cannot rely on features used to identify eSNPs, or any other potentially correlated feature, such as footprints. Thus, for an independent estimate of the proportion of causal SNPs, we used potential TF binding disruption by a SNP allele. Following Lappalainen et al., using MatrixeQTL25, we first identified causal SNPs as follows. For each gene we identified best and second best associated SNPs, and the best SNP was deemed causal if (i) the best SNP association was significant (FDR \( < 10^{-5} \)) and (ii) the difference in association score (\(-\log_{10} \) P value) between the best and the second best SNPs was greater than a threshold (conservatively, 2.5, a Lappalainen et al.). For each TF motif, we obtained the disruption at each SNP (decrease in motif match scores due minor allele relative to major allele) thus obtaining two distributions, one for causal SNPs and another for the presumed non-functional background. By using distribution of motif disruption score for causal SNP, we identified TF motifs that are preferentially disrupted by causal SNPs. For each of such motif \( y \), we calculated an enrichment score \( c_{y,\text{causal}} \), which is the ratio of means of TF motif disruption score between the causal and a set of presumed non-causal SNPs. For motif \( y \), we similarly calculated the enrichment score for eSNPs \( c_{y,\text{eSNPs}} \). Following Lappalainen et al., we then estimated the fraction of causal SNPs to be causal as \( \frac{c_{y,\text{causal}}}{c_{y,\text{eSNPs}}} \). Supplementary Fig. 14 shows these proportion of eSNPs that is likely to be causal for all selected motifs, suggesting that overall 58% of eSNPs are putatively causal.

Functional explained variance and expression predictability was defined as explained variance by subset of expression-regulators that mapped to a DNAse I footprint.

Simulation study. Simulation was done on 200 genes. We used 174,800 SNPs (874 SNPs per each gene) for 313 samples from MAGNet genotype data. 1% of total SNPs were declared as enhancers. We estimated, number of causal regulatory SNPs and distribution of explained expression variance by genotype by running eQTL in MAGNet data. Using estimated number of causal regulators from MAGNet, expression-regulators were selected among enhancer per gene. Effect-size of each expression regulator was generated from \( N(0,1) \), that is finally being used to generate expression for each gene using a linear model. Finally a random noise was added such that explained variance by expression-regulators will be same as estimated from MAGNet data. For each regulator SNP, seven epigenetic features (DNase, H3K4me1, H3K4me3, F300, H3K27me3, H3K36me3 and H3K9me3) for heart were generated from distribution derived from validated heart enhancers. For all other SNP epigenetic features were generated from random SNPs background.

Motif binding score differential. For each of the 981 vertebrate TF motif from TRANSFAC database, we scanned the 50,000 flanking eSNPs (and for 10,000 control SNPs) randomly sampled from 300,000 SNPs for the presence of motif using pwmscan tool, separately for the major and the minor allele. Only the cases where at one of the two alleles had a motif hits (P value \( < 0.0002 \)) were further considered. For each such case, the difference in the binding score for the two alleles was computed, as the difference in log(P value). For each motif, the binding differential score for eSNPs and the control SNPs were compared using Wilcoxon test and the motifs which had at least 1.5-fold greater differential among eSNPs and a P value \( < 0.05 \) were identified.

DNase footprint enrichment. From31 we obtained a list of genomic locations, for 41 different cell-types, where significant evidence of in vitro protein binding event were detected using DNase-footprint. For each tissue, we calculated fraction of number of SNP that have a footprint in the 500 bp flanking it.
**Allletic imbalance and ChiA-PET analysis.** DNAase hypersensitivity (DHS-seq) reads for heart cells (HCm2 sample) were obtained and mapped to eSNPs (and control SNPs). Heterozygosity at each SNP locus was ascertained by the presence of multiple alleles among the reads mapping to the SNP location. For each such locus, the allletic imbalance was calculated as the difference in the number of reads mapped to each allelic.

The allletic imbalance was plotted against the overall signal intensity rank. ChiA-pet assay identified spatially proximal genomic regions where at least one of the region is bound by PolII. Because ChiA-pet data is unavailable for heart-related cell types, we pooled multiple ChiA-pet data from K562, Hela, NB4 and MC27. For each 50 bps flanking an eSNP (or control SNP) and the target promoter pair, number of ChiA-pet reads supporting the spatial proximity of the two loci were recorded. The ChiA-pet support for each SNP-gene pair was then compared for different methods after controlling for the genomic distance between the SNP and its target gene.

In Figs 6 and 7, median ‘white’ lines represent LOESS (local regression) for each method. Confidence interval for each median line is estimated using bootstrap and are shown in the s using either of following two ways: by thin lines representing LOESS of each bootstrap, or by coloured regions representing confidence intervals in terms of standard deviation of bootstraps.

**Software availability.** The implementation of eQTel with its source code is freely available at (www.cbcb.umd.edu/software/goal) as a R-package under MIT license. For details of other eQTL methods (Supplementary Note 3); expression explained variance and predictability (Supplementary Note 6); and scalability of eQTel (Supplementary Note 7) refer to Supplementary Notes.

**References**

1. Lonsdale, J. et al. The genome-phenotype expression (gtex) project. *Nat. Genet.* 45, 580–585 (2013).
2. Beyer, K. & Goldstein, J. When is nearest neighbour meaningful? Database TheoryICDT’99 (1999). URL: http://link.springer.com/chapter/10.1007/3-540-49257-7_15.
3. Kraft, P. & Hunter, D. Genetic risk prediction: are we there yet? *N. Engl. J. Med.* 360, 1791–1794 (2009).
4. Hirschhorn, J. N. Genome-wide association studies-illuminating biologic pathways. *N. Engl. J. Med.* 360, 1699–1701 (2009).
5. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* 30, 1095–1106 (2012).
6. Sahu, A. D. et al. in Pacific Symposium on BioComputing. Pacific Symposium on BioComputing 92–102 (World Scientific, 2012).
7. Karczewicz, K. J. et al. Systematic functional regulatory assessment of disease-associated variants. *Proc. Natl Acad. Sci. USA* 110, 9667–9672 (2013).
8. Gaffney, D. J. et al. Dissecting the regulatory architecture of gene expression qtls. *Genome Biol.* 13, R7 (2012).
9. Veyrieras, J.-B. et al. High-throughput mapping of expression-qTLs yields insight into human gene regulation. *PLoS Genet.* 4, e1000214 (2008).
10. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
11. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506–511 (2013).
12. Bernstein, B. et al. The Hoei NHM Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048 (2010).
13. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* 489, 75–82 (2012).
14. Purcell, S. et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
15. Durbin, R. M. et al. A map of human genome variation from population-scale sequencing. *Nature* 473, 544–554 (2011).
16. George, E. & McCulloch, R. Approaches for Bayesian variable selection. *Stat. Sin.* 7, 339–373 (1997).
17. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* 5, 1780–1815 (2011).
18. Polson, N., Scott, J. & Windle, J. Bayesian inference for logistic models using eQTeL (Supplementary Note 7) refer to Supplementary Notes.
19. Zhang, Y. et al. GATA and Nkx factors synergistically regulate tissue-specific gene expression and development in vivo. *Development* 134, 189–198 (2007).
20. Pritt, M. E. et al. Evidence for coregulation of myocardial gene expression by MEF2 and NFAT in human heart failure. *Circ. Cardiovasc. Genet.* 2, 212–219 (2009).
21. Irizarry, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Bioinformatics* 4, 249–264 (2003).
22. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 118–127 (2007).
23. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 4, 500–507 (2012).
24. Teshendorff, A. E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 27, 1496–1505 (2011).
25. Matys, V. et al. Transfac and its module transCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110 (2006).
26. Hannenhalli, S. & Levy, S. Promoter prediction in the human genome. *Bioinformatics* 17, 590–596 (2001).
27. Tishkarian, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. J. Roy. Stat. Soc. B.* 58, 267–288 (1996).
28. Quinlan, A. R. & Hall, M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).

**Acknowledgements**

We thank Olga Ponomarova, Justin Malin, Nishanth Nair and Shruti Sarla for their valuable feedback in the Manuscript. The work was supported by R01GM100335 to S.H and R01HG105993 to T.C and a subcontract thereof to S.H.

**Author contributions**

S.H and A.D conceived the project. A.D developed the Bayesian method under supervision of S.H. A.D devised the inference algorithm with help from S.I. S.H and A.D analysed the data and performed the analyses. M.M, C.M, W.T, H.H, K.M, T.C and other members of MAGNet Consortium generated the MAGNet data. S.H and A.D wrote the manuscript, with help from others.

**Additional information**

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**How to cite this article:** Das, A. et al. Bayesian integration of genetics and epigenetics detects causal regulatory SNPs underlying expression variability. *Nat. Commun.*: 68555 doi: 10.1038/ncomms9555 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/
MAGNet Consortium

Euan A. Ashley\textsuperscript{6}, Jeffrey Brandimarto\textsuperscript{7}, Ray Hu\textsuperscript{7}, Mingyao Li\textsuperscript{8}, Hongzhe Li\textsuperscript{8}, Yichuan Liu\textsuperscript{8}, Liming Qu\textsuperscript{7}, & Pablo Sanchez\textsuperscript{6}

\textsuperscript{6}Stanford Center for Inherited Cardiovascular Disease, Stanford University School of Medicine, Stanford, CA 94305, USA. \textsuperscript{7}Penn Cardiacvascular Institute and Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. \textsuperscript{8}Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA.