Genome Biology

METHOD Open Access

sn-spMF: matrix factorization informs tissue-specific genetic regulation of gene expression

Yuan He[1], Surya B. Chhetri[2,3], Marios Arvanitis[1,4], Kaushik Srinivasan[5], François Aguet[6], Kristin G. Ardlie[6], Alvaro N. Barbeira[7], Rodrigo Bonazzola[7], Hae Kyung Im[7], GTEx Consortium, Christopher D. Brown[8*] and Alexis Battle[1,5*]

*Correspondence:
chrbro@pennmedicine.upenn.edu;
ajbattle@jhu.edu
Please find the full list of authors in GTEx in Additional file 3
[8]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA
[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA
Full list of author information is available at the end of the article

Abstract

Genetic regulation of gene expression, revealed by expression quantitative trait loci (eQTLs), exhibits complex patterns of tissue-specific effects. Characterization of these patterns may allow us to better understand mechanisms of gene regulation and disease etiology. We develop a constrained matrix factorization model, sn-spMF, to learn patterns of tissue-sharing and apply it to 49 human tissues from the Genotype-Tissue Expression (GTEx) project. The learned factors reflect tissues with known biological similarity and identify transcription factors that may mediate tissue-specific effects. sn-spMF, available at https://github.com/heyuan7676/ts_eQTLs, can be applied to learn biologically interpretable patterns of eQTL tissue-specificity and generate testable mechanistic hypotheses.

Keywords: Matrix factorization, Ubiquitous eQTLs, Tissue-specific eQTLs, Transcription factors

Background

Understanding the genetic effects on gene expression is essential to characterizing the gene regulatory landscape and provides insights into the molecular basis of phenotypes. Expression quantitative trait locus (eQTL) studies using genotype and gene expression data have demonstrated that the genetic regulation of gene expression is pervasive ([1–5], the GTEx Consortium 2020, in submission). Additionally, numerous studies have leveraged eQTLs to characterize the molecular basis of complex phenotypic variation [6–10].

Tissues in the human body carry out universal cellular processes in addition to performing highly specialized functions, driven in large part by patterns of gene expression in each cell type [11, 12]. Characterizing the tissue-sharing and tissue-specificity of genetic effects on gene expression is therefore critical to understanding how genetic variation leads to phenotypic changes. Recent work has identified eQTLs across a broad range of human tissues. The Genotype-Tissue Expression (GTEx) project has collected eQTL data across 49 human tissues (Additional file 1: Figure S1), which provide an unprecedented opportunity to uncover the ubiquitous and tissue-specific patterns of genetic regulation of gene expression [1].

Several methods have been developed to capture the underlying tissue-specific architecture in eQTLs across tissues. The simplest such method is based on the effect sizes or *P* values of eQTLs to identify eQTLs specific to individual tissues or cell types [13, 14]. Such heuristic methods are computationally efficient, but require manual selection of numerous subjective thresholds that affect the interpretation of results. Statistical frameworks have been developed to jointly analyze eQTLs from different datasets, such as eQTL-BMA and Meta-Tissue [15, 16]. These methods are more computationally demanding but potentially more accurate in their estimation of tissue-specificity. However, neither class of methods addresses the underlying similarity of multiple tissues or conditions in datasets such as GTEx, which may arise from shared mechanism.

Genetic effects on gene expression are often shared across some, but not all, tissues. When defining tissue-specific patterns of eQTL effects, three issues need to be considered. First, patterns of shared effects across tissues are often not obvious a priori. Manually identifying relevant groupings of tissues or contexts is not always obvious or feasible. Second, these groupings are not necessarily mutually exclusive. A single tissue may naturally belong to two or more groups based on shared biology with both. Third, an eQTL may have effects in more than one group of tissues. For example, in GTEx, different regions of the brain often have shared eQTL effects. However, effects in cerebellar tissues sometimes align with the other brain regions, but are sometimes quite distinct. Similarly, while many eQTL effects are shared across a set of digestive tissues (esophagus, stomach, and colon), many effects are specific to different subsets of these tissues, and it is not obvious how they would be grouped manually.

Matrix factorization is a general method for automatically decomposing data into overlapping, learned patterns, and has been successfully applied in biological domains, such as modeling gene expression for overlapping sets of co-functional genes. Matrix factorization applied to eQTL statistics offers a flexible and natural approach for identifying underlying patterns across eQTLs that may indeed better reflect biological mechanisms which likewise act across related, non-mutually exclusive subsets of tissues, conditions, or samples [17]. Recently, matrix factorization has been applied in a Bayesian setting to capture the structure of genetic regulation in human tissues; however, specific modeling choices for factorizing eQTL effects in various domains remain to be comprehensively evaluated [18]. It is further unexplored what insights into regulatory mechanism and functional consequences can be gained by evaluating these complex patterns of ubiquitous and tissue-specific eQTL effects.

In this study, we propose a constrained matrix factorization model called weighted semi-nonnegative sparse matrix factorization (sn-spMF) and apply it to analyze eQTLs across 49 human tissues from the GTEx consortium. We learn a lower-dimensional representation of eQTL effects across tissues, capturing both tissue-shared and tissue-specific patterns of eQTL activity. We leverage this atlas of ubiquitous and tissue-specific eQTLs to begin to characterize the regulatory mechanisms that underlie this specificity, and compare this approach to standard methods of identifying tissue-specific eQTLs.

We demonstrate that the ubiquitous and tissue-specific eQTLs exhibit distinct patterns of cis-regulatory element enrichment and identify specific TFs that appear to drive tissue-specific genetic effects.

Results

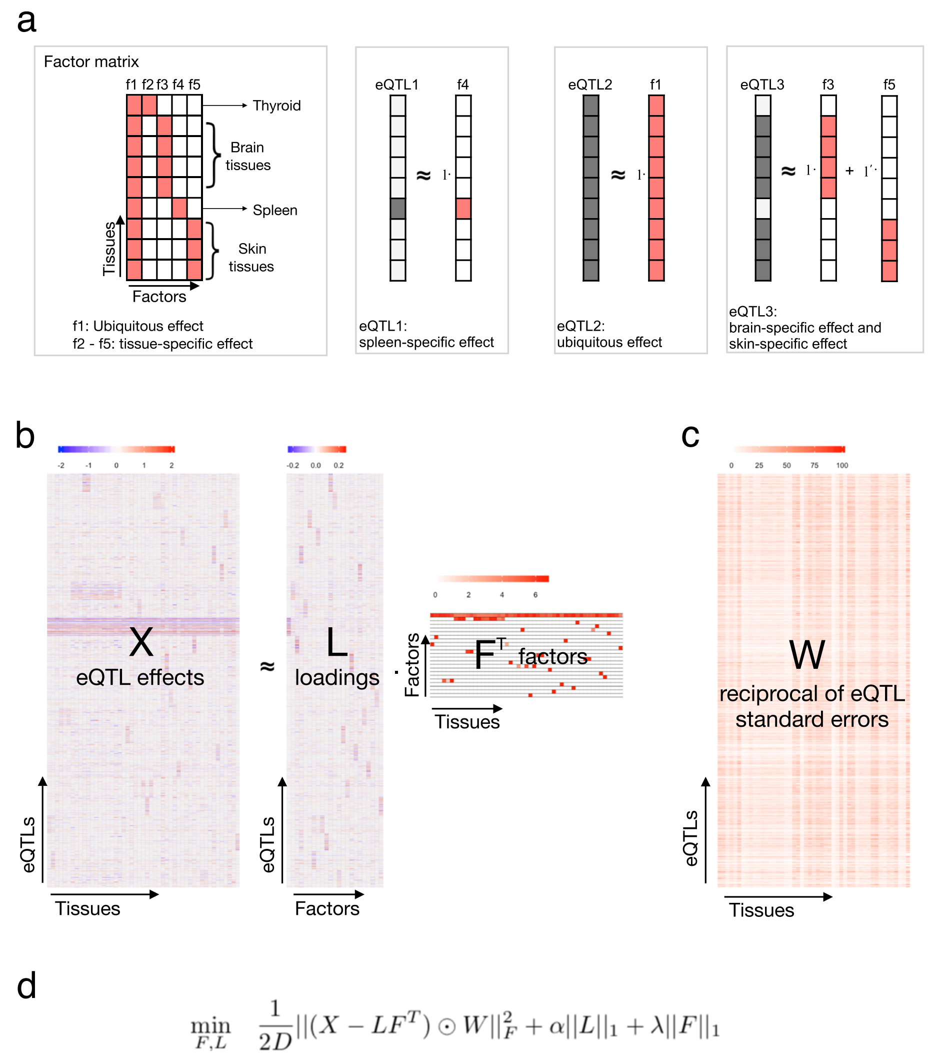
Matrix factorization of multi-tissue eQTL effects

The effect of eQTL variants on gene expression varies across tissues, as has been previously observed [1, 2, 19]. To better understand common patterns of genetic impact across tissues and to characterize the mechanisms that underlie tissue-specificity, we developed and applied a matrix factorization model called semi-nonnegative sparse matrix factorization (sn-spMF). The model overall seeks to decompose an input matrix of eQTL effect sizes in each tissue (regression parameters from a linear model for eQTL mapping) into underlying patterns of tissue-sharing and tissue-specificity. This model assumes that the effect size vector of one eQTL across tissues can be approximated as a linear combination (weighted sum) of learned "factors," where every factor is a vector representing one common pattern of eQTL effect sizes across tissues (Fig. 1a). When many entries in the factor are small or zero, as our model will enforce, a factor points to a subset of tissues that are commonly affected by the same eQTLs. Then, for a given eQTL, the loadings, or "weights," on each factor reflect how strongly that eQTL's effects are explained by that factor (and corresponding non-zero tissues). Given a multi-tissue dataset of eQTL association statistics as input, we identified a set of explanatory tissue factors by minimizing an objective function combining two components: (1) a weighted squared error term that captures how well the learned weights and factors reconstruct the observed eQTL effect sizes and (2) a regularization term that encourages sparsity, or many zero entries, in both factors and weights through an L1 penalty (Fig. 1b). Since it has previously been shown that inconsistent directions of effect for eQTLs will often arise from allelic heterogeneity rather than true sharing [20, 21], we constrained factors to be nonnegative.

By optimizing the objective function using alternating least squares applied to the GTEx v8 data across 49 tissues, we learned a factor matrix F with 23 factors (see the "Methods" section, Additional file 1: Figure S1, S2). These factors can be categorized into two major types: a ubiquitous factor, which captures eQTLs with largely consistent effects across all 49 tissues, and tissue-specific factors, which reflect effects only found among subsets of individual tissues. Tissue-specific factors include two subtypes: 8 factors representing combinations of tissues and 14 factors representing single tissues. Each of the 8 multi-tissue factors involves closely related tissues. For example, factor 2 represents effects of eQTLs in 13 brain regions; factor 15 represents effects in transverse colon and small intestine. For interpretability, each factor is named based on the tissues it represents (Additional file 1: Figure S2). In total, 41 out of 49 tissues are represented by non-zero values in at least one tissue-specific factor. The 8 tissues that do not appear in any tissue-specific factor have significantly smaller sample sizes compared to the 41 tissues captured by one or more factors (two-sided t test P value $= 0.024$, Additional file 2: Table S1), and thus, fewer eQTLs are detected that are unique to those tissues.

Identification of ubiquitous and tissue-specific eQTLs using sn-spMF

For each individual eQTL, we identified the relevant patterns of tissue-sharing and tissue-specificity by estimating the contribution from each of our learned factors to the eQTL's



$$\min_{F,L} \quad \frac{1}{2D}||(X - LF^T) \odot W||_F^2 + \alpha||L||_1 + \lambda||F||_1$$

Fig. 1 Matrix factorization model to dissect eQTL effects across tissues. **a** Simplified examples of the relationship between eQTL effect sizes and factors. eQTL1: the effect of an eQTL in the spleen can be represented by a spleen-specific factor. eQTL2: the effect of an eQTL in all nine tissues can be summarized as a ubiquitous effect across all tissues. eQTL3: the effect of an eQTL in four brain tissues and three skin tissues can be summarized as the summation of brain-specific effect and skin-specific effect. **b** Learning factors underlying eQTL effects from GTEx. *X* matrix represents the effect size of eQTLs across tissues (see the "Methods" section). Patterns of tissue-sharing and tissue-specificity are observed in *X*. Matrix factorization is implemented to learn the factor matrix *F*, where each factor captures a pattern of eQTL effect sizes across tissues. **c** Matrix *W* represents the weights for each eQTL across tissues. Each weight is the reciprocal of the standard error. **d** The objective function in sn-spMF, where α and λ are sparsity penalty parameters, and *D* is the number of eQTLs

effect sizes, using a second pass of weighted linear regression (see the "Methods" section). The observed patterns of tissue-sharing and tissue-specificity and how they are decomposed by matrix factorization are illustrated in the four following examples. First, an eQTL for GLT1D1 is highly specific to the liver and loads only on the corresponding liver factor (Fig. 2a). Second, an eQTL for AATF loads on the brain tissue factor and the tibial nerve factor to explain its combined effect size profile (Fig. 2b). Although this eQTL has small effects (or large variance) in some brain subregions, the model is able to identify a brain-wide effect as a likely explanatory factor for this eQTL. Third, an eQTL for

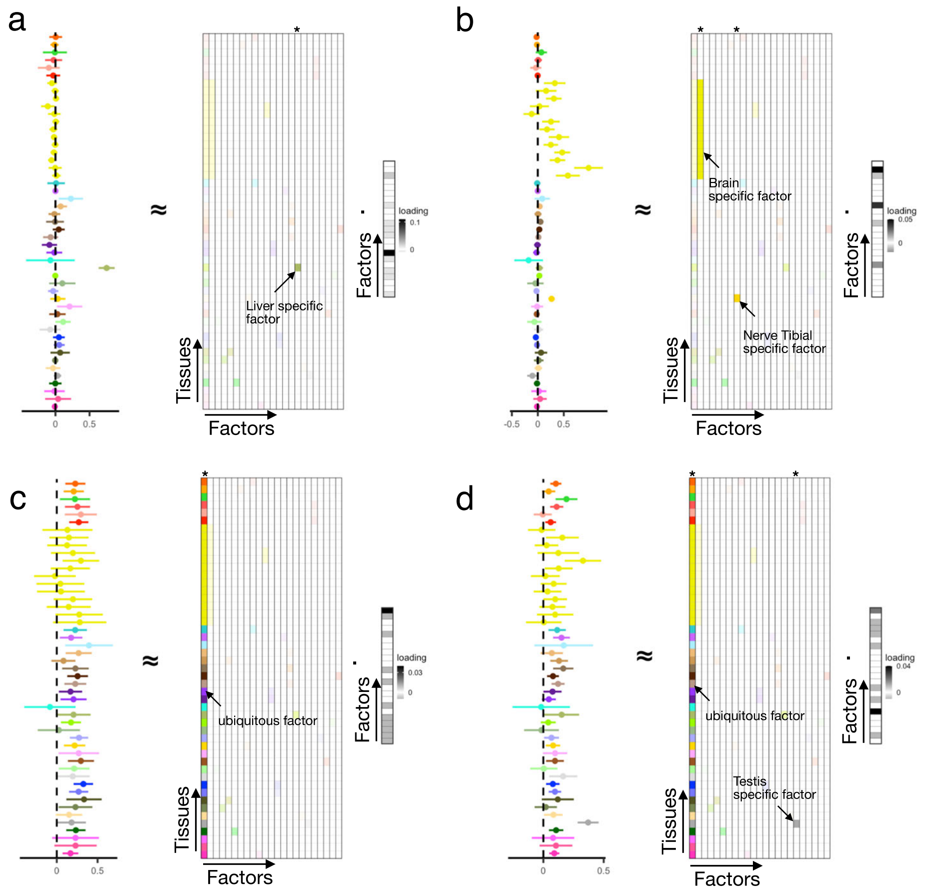


Fig. 2 Assignment of eQTLs to factors. Effect sizes and 95% confidence intervals of four eQTLs across 49 tissues are illustrated. The fitted linear combination of factors for the eQTL is displayed in gray scale at the right of each panel. Faded colors indicate factors with coefficients with FDR ≥ 0.05. Asterisk on the tissue indicates that this eQTL was significant with FDR < 0.05 in that tissue. **a** A liver-specific eQTL (GLT1D1-rs1012994). **b** An eQTL (AATF-rs76014915) with activity in brain tissues and tibial nerve. **c** A ubiquitous eQTL (U2AF1-rs234719). **d** An eQTL (CD14-rs2563249) with ubiquitous and testis-specific effects

U2AF1 with relatively consistent effects across tissues loads only on the ubiquitous factor (Fig. 2c). Finally, an eQTL for CD14 has consistent effects across all tissues in addition to a stronger effect specific to the testis (Fig. 2d).

In summary, 1,076,761 eQTLs (20% of tested eQTLs) load on the ubiquitous factor; we refer to these eQTLs as “ubiquitous eQTLs” (u-eQTLs). For each tissue-specific factor, 76,976 to 431,585 eQTLs (1.5 to 8.1% of tested eQTLs) have significant loadings; we call these eQTLs “tissue-specific eQTLs” (ts-eQTLs) (Fig. 3a, Additional file 2: Table S2). Identified ts-eQTLs do not appear to result from genes with low levels of tissue-specific gene expression (Figure S3). In total across factors, 2,821,650 eQTLs (53% of tested eQTLs) are found to use at least one tissue-specific factor (Fig. 3b). There are 638,784 eQTLs that load on both the ubiquitous factor and tissue-specific factors (59% of the u-eQTLs and 22% of the ts-eQTLs, Fig. 3c), indicating that in addition to a broad, shared effect across tissues, these eQTLs have a much stronger effect on expression in a particular subset of tissues. eQTLs tend to load on a small set of tissue-specific factors,

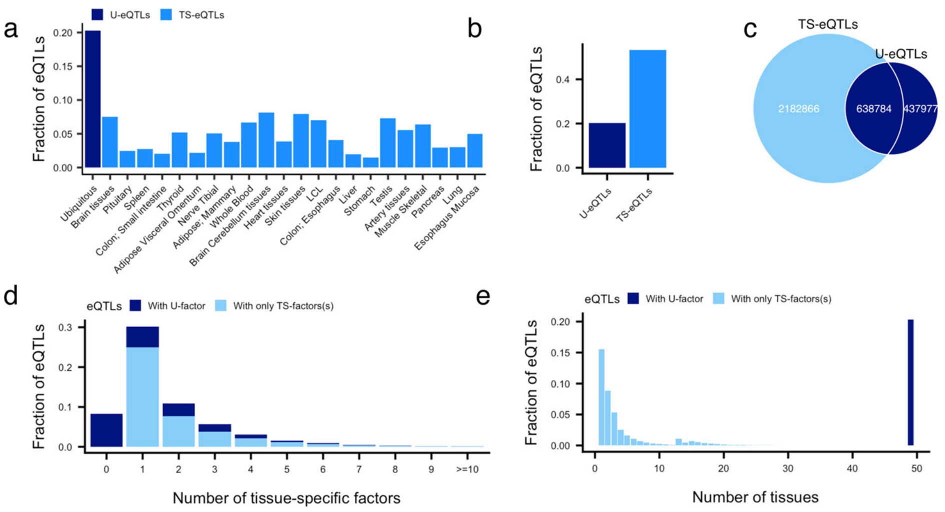


Fig. 3 Identification of tissue-specific and ubiquitous eQTLs. **a** Fraction of tested eQTLs that load on each factor. **b** Fraction of eQTLs that load on ubiquitous and tissue-specific factors. **c** The overlap of tested eQTLs that loaded on the ubiquitous factor (u-eQTLs) and any tissue-specific factor (ts-eQTLs). **d** Fraction of eQTLs that load on different numbers of tissue-specific factors. eQTLs that load with a specific number of ts-factors can fall into one of two categories: those with the ubiquitous factor and those with only ts-factors. The figure shows the fraction of tested eQTLs that load on each number of ts-factors with colors to show the contribution for each category. **e** Fraction of eQTLs with activity in different numbers of tissues. The numbers of unique tissues represented in the set of factors for each eQTL are summed

with 3,083,103 eQTLs (99% among the eQTLs loaded on at least one factor) using less than six tissue-specific factors (Fig. 3d).

The number of factors an eQTL loads on should provide a more biologically interpretable indication of the number of independent contexts in which an eQTL is active, rather than simply counting individual significant tissues. Datasets often contain multiple similar or even duplicate tissues, such as the thirteen brain regions in GTEx, or the two skin tissues that only differ by sun exposure. It may be misleading to count a neuron-specific eQTL as active in thirteen tissues, not at all comparable to a very general eQTL active in thirteen highly distinct tissues. Here, we demonstrate that eQTLs tend to be active in just a few factors, tailing off rapidly, but these factors sometimes correspond to numerous tissues (Fig. 3d, e), providing some interpretation for the familiar "U-shape" curve that has been reported previously ([22], the GTEx Consortium 2020, in submission). However, we note that 8 tissues are not significantly represented by any tissue-specific factor and, therefore, cannot be captured in this analysis (Additional file 2: Table S1).

Matrix factorization improves biological interpretation over heuristic methods of determining tissue relevance

The method most commonly used to identify ts-eQTLs is simply to apply heuristic thresholds based on effect sizes, *P* values, or meta-analysis results for individual tissues [13, 14, 16, 19]. If an eQTL statistic exceeds the chosen threshold for a given tissue, and remains below another threshold for other tissues, it is considered to be tissue-specific. None of these approaches consider common patterns of tissue-sharing and may obscure eQTL mechanisms shared across a subset of tissues (such as the brain or endothelium) unless they were manually predefined for investigation. Moreover, none of these approaches handle complex patterns of tissue-specificity, where an eQTL influences more than one tissue or predefined set, but is not universally shared.

Based on heuristic thresholds on individual tissue P values (heuristic$_1$, see the "Methods" section), we identified 312,502 u-eQTLs and between 1374 and 102,414 ts-eQTLs per tissue—far fewer eQTLs are confidently assigned to each category compared to results from sn-spMF (Additional file 1: Figure S4; Additional file 2: Table S2). This difference is partly because standard heuristic methods allow only one pattern (a single tissue or a ubiquitous effect) to be assigned to each eQTL, while matrix factorization allows multiple factors and tissues to be involved in explaining the effect size of an eQTL (Additional file 1: Figure S5). In addition, heuristic methods often miss small effects from similar tissues, while matrix factorization is able to aggregate effects for similar tissues (Fig. 2). We also tried manually grouping together tissues with clear shared biology and applying heuristic thresholds based on these (heuristic$_2$, see the "Methods" section, Additional file 2: Table S3), resulting in 175,637 u-eQTLs and between 1460 and 201,584 ts-eQTLs (Additional file 1: Figure S6, S7). In subsequent sections, we show that matrix factorization allows for the identification of more biologically coherent eQTLs than heuristic approaches by comparing sn-spMF to the standard approach defined by heuristic$_1$. We also show that manually defined tissue sets as in heuristic$_2$ offer only small gains over heuristic$_1$ and do not perform as well as matrix factorization either.

Tissue-specific eQTL gene function

To examine the functional relevance of ts-eQTL genes, we ran enrichment analysis using biological processes from the Gene Ontology (GO) project [23]. We first evaluated genes with ts-eQTLs and no u-eQTL. For sn-spMF, these eQTL genes are enriched for 546 unique GO terms at FDR < 0.05 (Additional file 1: Figure S8), and the top enriched GO terms are relevant to the corresponding tissues (Additional file 1: Figure S9, S10, S11). The ts-eQTL genes from heuristic methods, however, are less enriched in GO biological processes (at FDR < 0.05, 110 enriched for heuristic$_1$, 421 enriched for heuristic$_2$, Additional file 1: Figure S12).

After initial enrichment analysis, we used a more stringent definition of tissue-specificity to restrict the analysis to the genes most unique to each factor. For sn-spMF, we selected genes appearing in less than 6 tissue-specific factors (on average 252 genes per factor). A total of 64 unique GO terms are enriched at FDR < 0.1. The enriched GO terms are related to the matched tissue(s) of the eQTLs (Fig. 4). For example, five GO terms are enriched among liver-specific genes including four metabolic processes (for steroid, drug, uronic acid, and flavonoid) and response to xenobiotic stimulus, each relevant to liver function. For heuristic$_1$, we selected genes appearing in less than 7 tissues (on average 325 genes per tissue); for heuristic$_2$, we selected genes appearing in less than 6 subsets of tissues (on average 243 genes per subset), such that the gene sets are of comparable sizes. No GO term is enriched among these gene sets for heuristic$_1$, and one GO term is enriched for heuristic$_2$ (Additional file 1: Figure S12). These results indicate that sn-spMF is able to identify eQTL genes with biological functions relevant in the corresponding tissues more effectively than heuristic methods, even with comparably stringent definitions of tissue-specific eQTL genes providing similar numbers of genes for analysis.

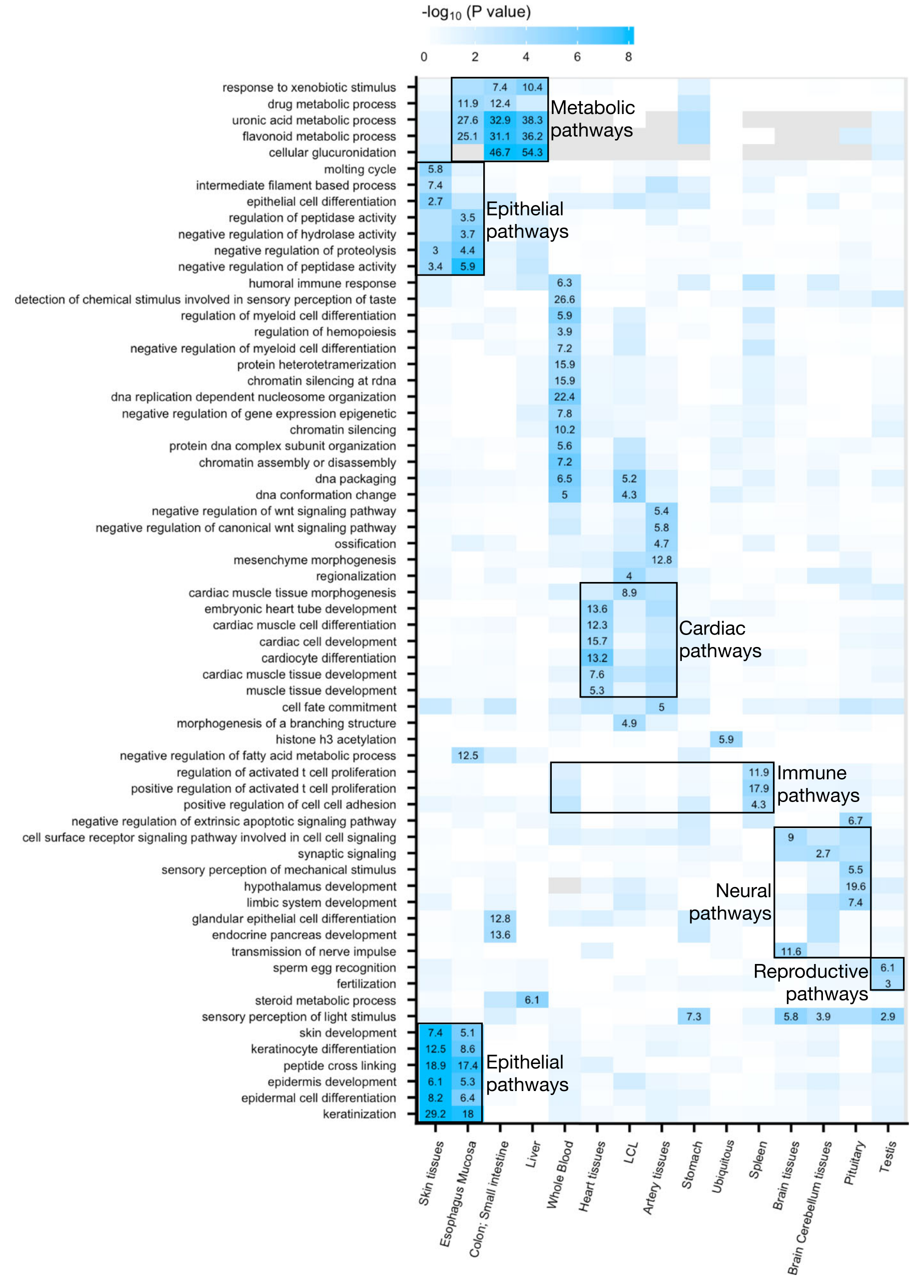


Fig. 4 Enriched GO terms for eQTL genes from sn-spMF at FDR < 0.1. Color represents the level of enrichment ($-\log_{10} P$ value). The significantly enriched GO terms are annotated by numbers representing the odds ratio. To compute the OR for each factor, background genes include all genes tested for the represented tissues in the factor. GO terms and factors are ordered by hierarchical clustering. Examples of relevant GO terms in related tissues are annotated

eQTL variant enrichment in cis-regulatory regions

eQTL variants are enriched in cis-regulatory elements, including cell type-specific promoters and enhancers [1, 24, 25]. Consistent with prior observations, u-eQTL variants identified by sn-spMF are more enriched in promoters (OR = 1.9, P value $< 2.2 \times 10^{-16}$) than ts-eQTL variants (OR = 1.5, P value $< 2.2 \times 10^{-16}$), while ts-eQTL variants are more strongly enriched in enhancers (OR = 1.3, P value = 8.5×10^{-12}) than u-eQTL variants (OR $=$ 1.0, P value $=$ 0.40, Additional file 1: Figure S13) [1, 26, 27]. Moreover, ts-eQTL variants are more likely than u-eQTLs to overlap enhancers whose activity is restricted to a small number of tissues (Additional file 1: Figure S14). Compared to sn-spMF, heuristically defined ts-eQTLs exhibit comparable enrichment magnitude in enhancers (for heuristic$_1$, OR = 1.3, P value = 7.8×10^{-8}; for heuristic$_2$, OR = 1.4, P value = 4.2×10^{-5}), but sn-spMF provides an order of magnitude more ts-eQTLs (Additional file 1: Figure S4, S6). While heuristic methods identify highly tissue-specific eQTLs by selecting those with effects clearly limited to a single tissue or a subset of tissues, sn-spMF identifies many more eQTLs relevant to each tissue-specific factor, each related to a shared set of cis-regulatory elements.

eQTL enrichment in transcription factor binding sites

To systematically assess whether eQTLs for each factor are enriched in binding sites for specific TFs, we performed enrichment analysis for each of the 579 TF motifs available in the JASPAR database [28]. As a proxy for TF binding sites (TFBS) in individual tissues, we identified TF motif instances overlapping predicted enhancers and promoters [29–32].

Enrichment analysis was performed separately for TFBS in promoters and TFBS in enhancers (see the "Methods" section). In promoters, u-eQTLs and ts-eQTLs are enriched for TFBS of 136 and 181 unique TFs (median $=$ 21 across factors), respectively (FDR $<$ 0.05, Fig. 5a, b). In enhancers, u-eQTLs and ts-eQTLs are enriched for TFBS of 39 and 264 unique TFs (median $=$ 41 across factors), respectively (FDR $<$ 0.05, Fig. 5a, b). Among these 264 TFs, 244 (92%) are enriched for fewer than six tissue-specific factors (Fig. 5c). Zero to 23% (among factors, median 4%) of TFs are enriched in both promoters and enhancers (Additional file 1: Figure S15). These results indicate that ts-eQTLs are more enriched in binding sites of particular TFs in enhancers than promoters, while u-eQTLs yield more enrichment in promoters than enhancers. The heuristic$_1$ approach for identifying ts-eQTLs yields only 5 TFs enriched in promoters and 47 TFs enriched in enhancers. Similarly, there are fewer TFs enriched for heuristic u-eQTLs (59 in promoters, and 8 in enhancers, Fig. 5a, Additional File 1: Figure S16). Heuristic$_2$ yields 9 TFs enriched in promoters and 51 TFs enriched in enhancers for ts-eQTLs, and 97 TFs enriched in promoters and 4 TFs enriched in enhancer for u-eQTLs. The relatively low enrichment of TFBS from heuristically identified eQTLs is presumably due to the much more limited number of eQTLs identified in each category.

Impact of matrix factorization methodological choices

In addition to our sn-spMF model, there are a variety of matrix factorization approaches available. Methodological choices include the selection of priors on loading and factor entries, which may encourage sparsity or other properties, nonnegativity constraints, and hyper-parameter selection.

We compared our method to several matrix factorization methods using simulated data (see the "Methods" section). We ran singular value decomposition (SVD) and nonnegative matrix factorization (NMF) as they are commonly used in matrix factorization. We also implemented matrix factorization with various constraints, including sparse SVD (SSVD), penalized matrix decomposition (PMD), softImpute, and nonparametric Bayesian sparse factor analysis (NBSPA) [33–36]. PMD penalizes the two decomposed matrices using either one penalty parameter scaled by the dimensions for

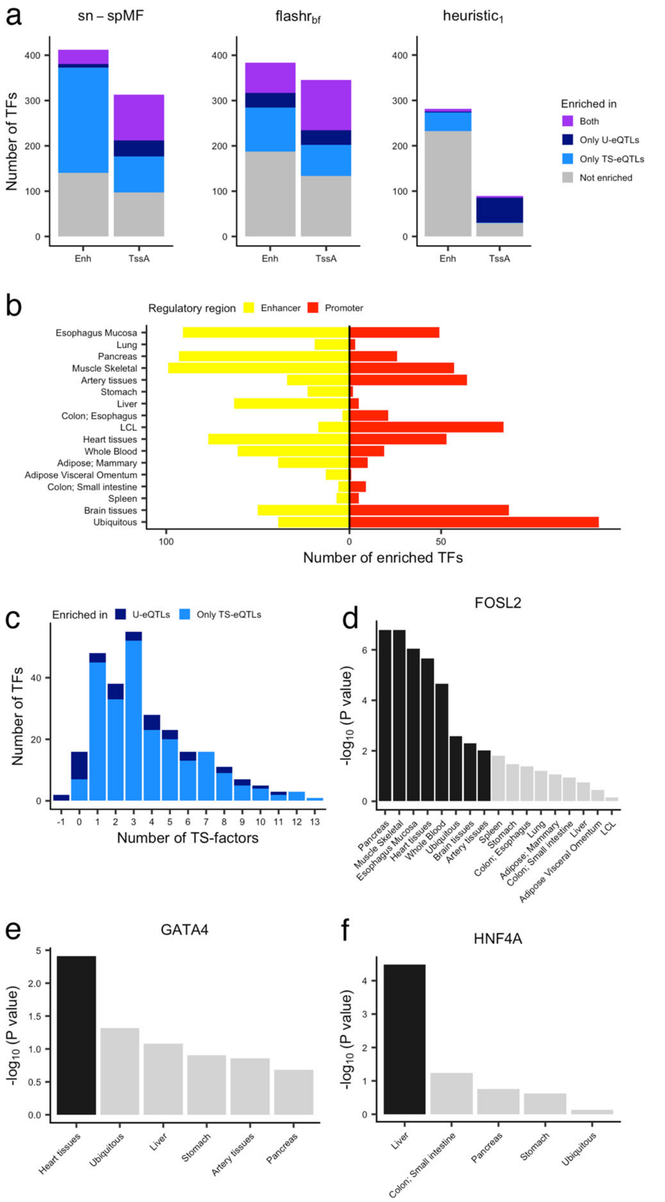


Fig. 5 Enrichment of TFBS for u-eQTLs and ts-eQTLs. **a** Number of TFs whose binding sites are enriched for eQTLs across factors at FDR < 0.05 for sn-spMF, flashr$_{bf}$, and heuristic$_1$ methods. Enh, enhancers; TssA, active transcription start sites. **b** Total number of TFs with binding sites enriched for either only u-eQTLs, or only ts-eQTLs, or both. **c** Distribution of the number of tissue-specific factors each TF is enriched in. **d**–**f** Enrichment for example TFs among eQTLs across each factor ($-\log_{10}(P \text{ value})$) where the TF was expressed in corresponding tissues for **d** FOSL2, **e** GATA4, and **f** HNF4A. Black bars represent that the BH-corrected *P* value is < 0.05

each decomposed matrix (PMD$_{CV1}$) or two separate penalty parameters (PMD$_{CV2}$). Finally, we applied flashr, a recent method which uses a Bayesian framework to automatically learn the sparse structure of effects across tissues [18]. Flashr was run with default setting (flashr$_{default}$), greedily adding factors followed by backfitting (flashr$_{bf}$) and with nonnegative priors (flashr$_{NN}$).To evaluate the performance of these methods on simulated data, we computed the correlation between the learned loadings and the true loadings, and the correlation between the learned factors and the true factors, as well as the precision and recall for true u-eQTLs and ts-eQTLs. We observed that sn-spMF and flashr$_{NN}$ achieve the most accurate loading matrix and factor matrix, and the highest precision and recall for correctly identifying u-eQLTs and ts-eQTLs (Additional file 1: Figure S17, S18), followed by other flashr approaches, NBSPA, and softImpute. Sparsity appears to confer some benefit in accuracy and interpretability of factors.

Based on strong performance in simulation, we also applied flashr methods to the GTEx data, each capturing both ubiquitous and sparse factors (Additional file 1: Figure S19). We first discuss flashr$_{bf}$, which displayed the strongest performance of the flashr methods on GTEx, in detail. Each flashr$_{bf}$ factor is somewhat more dense (more non-zero entries) than sn-spMF factors (Additional file 1: Figure S20, S21). We then identified flashr$_{bf}$ factors relevant to each eQTL using the same second pass linear regression pipeline as in sn-spMF. We thus identified 1,929,939 u-eQTLs and 69,594 to 929,009 ts-eQTLs.

Flashr$_{bf}$ ts-eQTL genes are comparably enriched for GO biological processes as sn-spMF factors, far exceeding heuristic ts-eQTL genes, with 593 enriched pathways (FDR < 0.05). However, flashr$_{bf}$ eQTL variants are not strongly enriched in enhancers (OR $= 1.1$, Additional file 1: Figure S22). This appears to be due to the denser flashr$_{bf}$ factors not isolating tissue-specific effects from ubiquitous effects as strongly. Assessing TF enrichment, however, because analysis is restricted to variants within enhancers identified in relevant tissues, is still able to identify enrichment for 197 TFBS across flashr$_{bf}$ factors (Fig. 5a). While regulatory element enrichment appears sensitive to matrix factorization methodological choices, both versions of matrix factorization show advantages over heuristic approaches for identifying tissue-relevant eQTL genes and for identifying particular transcription factors whose binding sites are impacted by ts-eQTL variants. Finally flashr$_{bf}$, does not include nonnegativity constraints on the factors, thus complicating interpretation of latent patterns and tissue-specificity. For example, we found that factors that contain tissues with different signs do not correspond well to patterns in the actual eQTL effect sizes—only 19–35% of eQTLs that mapped to such mixed sign factors actually display opposite sign eQTL effects in the corresponding tissues (Additional file 1: Figure S23).

For thorough comparison, we also applied other matrix factorization methods including flashr with default parameter setting (flashr$_{default}$), flashr with nonnegative prior (flashr$_{NN}$), softImpute, and PMD to the GTEx dataset (see the "Methods" section, Additional file 1: Figure S24 - S29). These methods did not offer performance gains over flashr$_{bf}$ or sn-spMF (Additional file 1: Figure S20, S21, S22, S30; Additional file 2: Table S4, S5, S6, S7). In particular, flashr$_{NN}$ provided sparse, interpretable tissue factors but suffered from multicollinearity making it difficult to distinguish ts-eQTLs from u-eQTLs (Additional file 1: Figure 25, [37, 38]). Overall, we conclude that the sparsity constraint on decomposed matrices is crucial to distinguish ts-eQTLs from u-eQTLs, and that depending on optimization approach, a nonnegativity constraint on factors can be helpful in interpreting the identified patterns of tissue-specificity.

Transcription factors enriched in u-eQTLs and ts-eQTLs

Given the limited systematic research on the consequences of genetic variation within tissue-specific TFBS, we examined the characteristics of TFBS enriched in ts-eQTLs for each factor and in u-eQTLs. We focused on the TFBS found within enhancers because of their generally increased tissue-specific functions (Additional file 1: Figure S13, S14). Binding sites for TFs with broad activity are enriched for u-eQTLs, such as CCAAT/enhancer-binding proteins (CEBPB, CEBPD, CEBPG), T-box 1 (TBX1), and AP-1 Transcription Factor Subunit FOSL2 [39–42] (Fig. 5d). The enrichment of these TFBS in u-eQTLs reflects their participation in a wide range of regulatory processes across tissues.

The enrichment of binding sites for 264 TFs in ts-eQTLs demonstrates their role in regulating gene expression in particular subsets of tissues corresponding to each factor. Among these, binding sites for 172 TFs display enrichment in ts-eQTLs for multiple factors with biologically plausible patterns across tissue groups. For example, hepatic nuclear factor HNF1A, known to be crucial for the development and function of the liver, pancreas, and gut epithelium, are enriched for the liver-specific eQTLs, pancreas-specific eQTLs, and ts-eQTLs for a factor reflecting the colon and small intestine [43, 44]. Furthermore, 92 TFBS are enriched in ts-eQTLs for one tissue-specific factor. Examples include binding sites for the well-characterized cardiac TF GATA4, which are enriched for heart-specific eQTLs [45, 46] (Fig. 5e); hepatocyte nuclear factor HNF4A, which are enriched for liver-specific eQTLs [47, 48] (Fig. 5f); and myogenic factor 4 MYOG, which are enriched for skeletal muscle-specific eQTLs [49] (Additional file 1: Figure S31). We continue to explore two TFs in more detail in the following sections. More examples of enriched TFs with previously characterized tissue-specific functions can be found in Additional file 1: Figure S31 and Additional file 2: Table S8.

Heart-specific eQTLs are enriched in GATA4 binding sites

Previous studies have demonstrated the essential roles of GATA4 in heart morphogenesis [50]. In mouse studies, GATA4 has been shown to recruit the histone acetyltransferase p300 in a tissue-specific manner in the heart [45]. This GATA4-p300 complex deposits H3K27ac at cardiac enhancers, thus stimulating transcription of genes necessary for heart development. In human, missense mutations in GATA4 are associated with multiple heart diseases such as cardiac septal defects and cardiomyopathy [51, 52]. However, common genetic variants affecting GATA4 TFBS have not previously been shown to be enriched for effects on expression in cardiac tissues. Binding sites of GATA4 in heart enhancers are enriched for heart-specific eQTLs (OR = 1.7, P value = 0.004, Fig. 5e), highlighting the importance of GATA4 in normal physiological conditions of the heart. Among the 48 genes loading on the heart-specific eQTL factor with variants located in TFBS of GATA4, we note that STAT3 has been reported to exhibit a crucial role in cardiomyocyte resistance to physiological stress stimuli [53].

Liver-specific eQTLs are enriched in HNF4A binding sites

Variants in liver-specific HNF4A binding sites are enriched for eQTLs loading on the liver-specific factor (OR = 2.9, P value = 3.3×10^{-5}, Fig. 5f). The enrichment of HN4FA binding sites has not been previously identified among liver eQTLs. HNF4A is an essential TF during liver organogenesis and development [47, 48] and harbors a missense mutation (rs1800961) strongly associated with liver relevant traits including high-density lipoprotein levels and total cholesterol [55–57] (Additional file 1: Figure S32).

With the availability of Chromatin Immunoprecipitation followed by high-throughput Sequencing (ChIP-seq) data for HNF4A in human liver tissues in ENCODE, we are able to directly map the genome-wide binding sites of HNF4A. Replicating the motif-based enrichment described above, liver-specific eQTLs are strongly enriched in HNF4A ChIP-seq peaks (OR = 3.6, *P* value $< 2.2 \times 10^{-16}$). The enrichment is not as strong in ts-eQTLs for other tissues (OR = 1.8 in the testis to 2.6 in the pancreas). Also, liver-specific eQTLs are significantly more enriched in HNF4A binding sites than are u-eQTLs (OR = 1.7, *P* value $< 2.2 \times 10^{-16}$).

We hypothesized that variants in HNF4A binding sites lead to liver-specific eQTLs via differential binding of HNF4A. We quantified allele-specific binding (ASB) of HNF4A and, as a tissue-shared control, CTCF (see the "Methods" section). Liver-specific eQTLs are indeed significantly enriched for ASB of HNF4A (OR = 1.4, *P* value = 0.003), but not CTCF (OR = 0.8, *P* value = 0.4). This finding supports the possibility that the enrichment of liver-specific eQTLs in HNF4A motifs reflects altered binding affinity of HNF4A at these eQTL variants, providing a testable hypothesis for experimental validation.

Example eQTL variant in HNF4A binding site relevant to liver phenotypes

Among the liver-specific eQTLs identified by sn-spMF, rs9987289 exhibits significant ASB for HNF4A (Fig. 6a,b, Additional file 1: Figure S33). The *A* allele is associated with increased HNF4A binding (ChIP-seq read ratio = 7.7, two-tailed binomial test *P* value $= 8.8 \times 10^{-5}$) and with significantly lower expression of the eGene TNKS (Fig. 6b, c). HNF4A may act as a repressor of TNKS, and these data suggest that the *A* allele of rs9987289 may act by increasing binding of HNF4A and therefore reducing expression levels of TNKS. Though HNF4A has been widely reported as a transcriptional activator, it has also been associated with transcriptional repression [58–62] (Fig. 6d). Rs9987289 is located in a flanking active promoter (TssAFlank) region surrounded by enhancers in liver, while it is found in quiescent or heterochromatin regions in all 13 non-liver tissues where HNF4A is expressed (Additional file 1: Figure S34, S35).

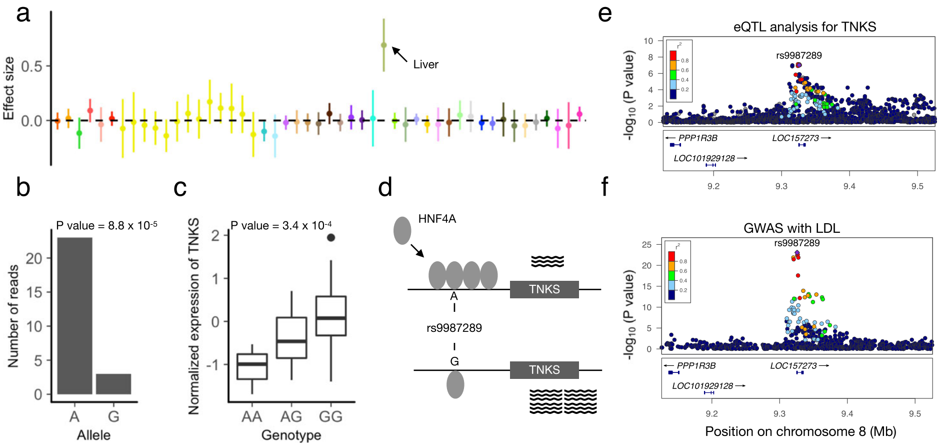


Fig. 6 Example liver-specific eQTL, TNKS-rs9987289, in a TFBS of HNF4A that co-localizes with liver-specific phenotypes. **a** Effect size and 95% confidence interval of TNKS-rs9987289 across 49 tissues in GTEx. **b** Allele-specific HNF4A ChIP-seq reads over rs9987289 in the liver (see the "Methods" section, two-sided binomial test *P* value $= 8.8 \times 10^{-5}$). **c** Normalized expression levels of TNKS in the liver among individuals with different genotypes at rs9987289. *P* value $= 3.4 \times 10^{-4}$ from GTEx eQTL analysis. **d** Schematic illustration of hypothesized mechanism: allele-specific binding of HNF4A at rs9987289 and altered levels of expression of TNKS. **e** Manhattan plot (LocusZoom v0.4.8) [54] of TNKS expression levels in the liver around rs9987289. **f** Manhattan plot for LDL GWAS around rs9987289

Furthermore, rs9987289 is significantly associated with several liver-related phenotypes, including low-density lipoprotein (LDL) cholesterol levels and high-density lipoprotein (HDL) cholesterol levels [REF GTEx GWAS companion] [55] (Additional file 1: Figure S36). The liver eQTL of TNKS and the association statistics for LDL are strongly co-localized (posterior probability of shared causal signal between LDL and the eQTL = 0.94, with rs9987289 having the highest posterior of being the shared causal variant) [63] (Fig. 6e, f). Though TNKS has been widely recognized for its role in controlling telomere length, there is emerging evidence of TNKS participating in liver metabolism [64, 65].

Together, these results support the hypothesis that the tissue-specific regulatory effect of ts-eQTL variant rs9987289 in the liver may have phenotypic consequences: an active cis-regulatory element unique to the liver, allele-specific binding of liver TF HNF4A in hepatocytes, and finally co-localization of the eQTL effect with lipid GWAS hit. Such examples can provide testable hypotheses regarding multiple steps of the mechanism through which genetic variation may affect a high-level phenotype.

Discussion and conclusions

In this study, we explored the genomic context and potential mechanisms underlying tissue-specific effects of genetic variation by applying a constrained matrix factorization model (sn-spMF) to multi-tissue eQTL data from the GTEx project. Using sn-spMF, we learned factors representing the common patterns of eQTL sharing across tissues, such as factors corresponding to ubiquitous effects across all tissues and effects shared among only brain tissues or among muscle tissues. This allowed us to explore eQTL effects shared across overlapping subsets of tissues that share cis-regulatory mechanisms due to shared cell types or developmental origin, without having to manually prespecify each such pattern. These learned factors enabled us to evaluate potential mechanisms relevant to genetic effects following these patterns of tissue-sharing.

sn-spMF identified much larger sets of tissue-specific eQTLs than did heuristic methods. The ts-eQTLs from sn-spMF were also equally or more enriched for GO biological processes, transcription factor binding sites, and tissue-specific cis-regulatory elements than the heuristic ts-eQTLs. These results suggest that sn-spMF identifies larger numbers of ts-eQTLs that remain biologically coherent, offering an opportunity for novel mechanistic insights. Other versions of matrix factorization, such as flashr, also provide meaningful views of tissue-specificity. In particular, we note the flashr has the advantage of learning the parameters with less computational burden, compared to sn-spMF where a grid search is needed for tuning parameters.

There can be other definitions of the manually selected subsets of tissues. However, it is not clear how to choose the relevant tissues and the thresholds before we have learned the latent patterns. For example, it is not clear whether whole blood and spleen should be grouped into one factor, or used as two separate factors. Also, heuristic methods can be hard to implement in situations where we have little knowledge about the feature (in contrast to our knowledge of tissue similarity). For example, in a time-series data, it is typically unknown, a priori, how patterns change during the time course.

The large set of ts-eQTLs provided by sn-spMF enabled a detailed evaluation of eQTLs in transcription factor binding sites that was not possible from heuristic approaches.

We evaluated 76,976 to 431,585 ts-eQTLs for enrichment in promoter and enhancer elements, and were able to identify 181 and 264 TFs enriched among these, respectively. This list of 264 TFs enriched in ts-eQTL enhancers provides experimentally testable hypotheses about specific genetic variants within TFBS that alter expression in a tissue-specific fashion.

Matrix factorization is inherently limited by the eQTL data used as input to the method—any tissue that is underpowered or not well represented in the original eQTL dataset is unlikely to be captured strongly by a ts-eQTL factor with sn-spMF. Further, sn-spMF does not explicitly model linkage disequilibrium (LD) or consider allelic heterogeneity, rather it relies on the user to pre-select candidate causal variants using fine-mapping tools or other approaches. Additionally, many matrix factorization approaches, priors, and constraints remain to be explored that may capture different properties of the eQTL data than represented here. Different applications, such as time series or perturbation-response eQTL data, may ultimately benefit from specialized matrix factorization formulations [17].

In conclusion, we have developed a constrained matrix factorization model to learn patterns of eQTL tissue-specificity across 49 human tissues using data from GTEx v8. We observed improved enrichment of biologically relevant genes and cis-regulatory elements compared to heuristic methods. Matrix factorization also revealed the potential impact of ubiquitous TFs on ubiquitous eQTLs and provided a list of candidate TFs relevant to each tissue-specific set of eQTLs.

Methods

GTEx data

GTEx Release v8 project has collected both genotype data from whole genome sequencing (WGS) and RNA sequence (RNA-seq) from 838 people. Here, we analyze GTEx data from 15,253 samples, consisting of 47 tissues and two cell lines (the GTEx Consortium 2020, in submission). GTEx v8 data release includes cis-eQTL analyses that test for association between gene expression and variants within 1 MB of the genes' transcription start sites (TSS). Effect sizes of the eQTLs are represented by coefficients estimated in the linear model association tests.

Preprocessing and input data

To restrict the analysis to potential casual variants, we used cis-eQTLs that are in the 95% credible set for at least one tissue [66]. Specifically, for each eQTL gene, the credible set consists of eQTL variants that include the causal variant with 95% probability. In total, 5,301,827 eQTLs with 17,480 unique protein coding eQTL genes are included in the analysis. For these 5,301,827 eQTLs, we collected the effect size and standard error from univariate cis-eQTL analysis across tissues, based on the linear model association test results from GTEx (the GTEx Consortium 2020, in submission). Missing entries, corresponding to tissues where an eQTL variant-gene pair was not tested, were assigned weights of 0 and thus do not contribute to the objective function of sn-spMF. This avoids biasing towards shared eQTLs caused by removing data points with any missing data. Finally, the lead variants, within credible sets, with the most extreme geometric mean P values across tissues for the 17,480 eQTL genes were used as input (rows in matrix X and W) to learn the factor matrix (matrix F). Ultimately, only 17,480 of the original 5,301,827 eQTLs are used to learn the factor matrix. However, the learned sn-spMF representation can then be used to analyze any tested eQTL variant.

sn-spMF is able to learn the underlying patterns from a subset of representative eQTL summary statistics. In our case, we restricted to credible set variants with the strongest signals across tissues, as described above. Other users may choose another representative subset of variants of interest based on their preferred methods for selecting likely causal variants or lead variants, but regardless, sn-spMF does not require summary statistics for every tested variant to learn relevant factors.

Lower-dimensional representation of eQTL effects

eQTL effects across tissues can be represented by a matrix $X_{D \times T}$ where D is the number of eQTLs and T is the number of tissues. Each entry is the regression parameter obtained from eQTL association testing of one variant/gene pair in one tissue, in the case of GTEx based on a standard linear model. Each row is then the effect of one eQTL across all tissues, and each column is the effect of all eQTLs for one tissue. The effect values are real-valued and can be positive or negative. A lower-dimensional representation of the effect matrix X can be written based on a factor matrix $F_{T \times K}$ and a loading matrix $L_{D \times K}$ such that $X \approx LF^T$ (Fig. 1).

Weighted semi-nonnegative sparse matrix factorization algorithm sn-spMF

In order to describe the eQTL effects, we designed a matrix factorization objective function with several features: (1) *A penalty on a weighted sum of residuals*: in order to account for uncertainty in effect size estimates, the residual for each data point was weighted by the reciprocal of its standard error. In this way, data points with more certain eQTL effect sizes have more influence over optimal parameter estimates. Missing values in the input data were assigned a weight of zero and thus do not influence the value of the objective. (2) *Sparsity*: to alleviate over-fitting, an l1 penalty was applied to the decomposed matrices. (3) *Semi-nonnegativity of the decomposed matrices*: the factors capture the pattern of effects across tissues, and thus, it was a natural constraint to make the factors nonnegative for ease of interpretation. At the same time, because the input matrix has mixed signs, there was no such constraint on the loading matrix. The objective function was formulated as below:

$$\min_{F,L} \quad \frac{1}{2D}||(X - LF^T) \odot W||_F^2 + \alpha||L||_1 + \lambda||F||_1$$

where F is nonnegative, W is the element-wise reciprocal of the standard error of the eQTLs, D is the number of data points (in this case the number of eQTLs), and α and λ are the penalty parameters.

This objective function is biconvex, that is, convex only in F or in only L given the other, but not convex in both jointly. We used alternating least squares (ALS) with gradient descent to optimize the objective (Algorithm 1, implemented in `R version 3.5.1`, [67, 68]). At each iteration, we fixed F and updated L, and then fixed L and updated F. The update was finished when the Frobenius norm of difference in F between two iterations was < 0.01. In each update step, the optimization problem was a linear regression with constraints. Since the solution to linear regression was guaranteed to minimize the sum of mean squared error and penalty, the cost function monotonically decreased.

Algorithm 1 Weighted semi-nonnegative sparse matrix factorization algorithm (sn-spMF)

```
Input: $X_{D\times T}$
Output: $L_{D\times K}, F_{T\times K}$
Randomly Initialize nonnegative $F$
while not converged do
    for $i = 1 \ldots D$ do
        $l_i \leftarrow \min_{l_i} || \left(x_i - l_i F^T\right) \odot w_i||_F^2 + \alpha||l_i||_1$
        which is equivalent to
        $l_i \leftarrow \min_{l_i} ||x_i \odot w_i - l_i \left(F^T \mathrm{diag}(w_i)\right) ||_F^2 + \alpha||l_i||_1$
    end for
    for $j = 1 \ldots T$ do
        $f_j \leftarrow \min_{f_j} || \left(x_j - f_j L^T\right) \odot w_j||_F^2 + \lambda||f_j||_1, ||f_j|| \geq 0$
        which is equivalent to
        $f_j \leftarrow \min_{f_j} || \left(x_j \odot w_j - f_j \left(L^T \mathrm{diag}(w_j)\right)\right) ||_F^2 + \lambda||f_j||_1, ||f_j|| \geq 0$
    end for
end while
```

Model selection

In the sn-spMF model, we need to set hyper-parameters including the rank of the decomposition (K) and the sparsity penalty (α, λ). We evaluated K within $[20, 25, 30, 35, 40]$, and α and λ within $[4.9, 24.5, 49, 245, 490]$. These ranges were chosen by considering the number of tissues in GTEx to define plausible values for K and by manual inspection of solutions for widely varying α and λ to avoid high-resolution search for ranges of these hyper-parameters that resulted in clearly implausible solutions, such as lack of sparsity or large numbers of empty, un-utilized factors.

Within these chosen search spaces, we evaluated sn-spMF models for all combinations of K, α, and λ using (1) a previously defined criterion of matrix factorization stability and (2) independence of the learned factors, which represents adequate sparsity. Considering the stochastic nature of matrix factorization, Brunet et al. proposed a method looking for the most stable factorization result, and this method has been applied in various studies [69, 70]. We obtained the consensus matrix C after 30 runs with random initialization for each model. The values in C are between 0 and 1, representing the proportion of runs in which a pair of tissues are assigned to the same factor. Using the C matrix, we computed the cophenetic correlation which is used to measure the degree of dispersion for the C matrix. Higher cophenetic correlation indicates a more stable factor matrix.

Evaluating the runs for all combinations of hyper-parameter settings, we first eliminated some settings of K. Here, for each observed mean number of learned, non-empty factors K' (which may be less than the input K), we aggregated across the different settings of λ and α and computed the median cophenetic correlation [69]. We eliminated from consideration any settings of K corresponding to a K' with a median cophenetic correlation < 0.9. Next, among the remaining individual settings, we eliminated any cophenetic correlation < 0.9. Last, among these apparently stable settings, we selected the final hyper-parameters based on the minimum Pearson correlation between pairs of factors, to encourage independent factors and a level of sparsity that matches independent signals in the data. Here, we computed the Pearson correlation for each pair of factors, took the Frobenius norm of the pairwise correlation matrix, and averaged this across the 30 randomly initialized runs for the same setting. Documented code and examples of the model-selection process are available on Github (https://github.com/heyuan7676/ts_eQTLs)

Assignment of eQTLs to factors

After we have learned the factors, we identify a set of relevant factors for each eQTL using weighted linear regression. Specifically, for each eQTL, a weighted linear regression of the form $x = FL$ is fit, where x is the vector of eQTL effect sizes across tissues, F is the factors learned from sn-spMF, and L are the regression coefficients. Weights w are incorporated, where w_t is the reciprocal of the standard error for the eQTL effect size x_t in tissue t. Weighted linear regression using standard error in this manner is a common approach allowing data points with high uncertainty to have less influence on the regression parameter estimates [71]. Statistical significance of each factor for the eQTL is determined according to P values based on the standard t test from this linear regression. To alleviate the multiple testing burden, we removed the eQTLs for which the variants were in perfect LD ($R^2 = 1$) with variants from another eQTL before running regression for the remaining 3,601,800 eQTLs [72]. We applied the Benjamini-Hochberg correction to get the q value for every factor for each eQTL [73]. We then mapped the q value back to all 5,301,827 eQTLs where the SNPs are in an LD block with the tested SNPs for the same gene. We observed that occasionally, there were factors assigned negative regression coefficients when the actual observed effect sizes in the corresponding tissues were positive, or vice versa. This discrepancy arose due to collinearity between the factors, and in such cases, the discrepant factors were not included for downstream analysis. We also removed those factors that caused one tissue to have an oppositely signed small effect (absolute Z-score < 3, or P value > 0.00135) when compared to the factor where this eQTL has the strongest effect; such discrepancies may often reflect allelic heterogeneity or LD contamination rather than true opposite effects from the same causal variant [20, 21]

Background SNP-gene pairs

For enrichment analyses, random SNP-gene pairs were sampled from all SNP-gene pairs to match for eQTLs by three criteria: (1) SNP MAF was matched to the eQTL variants' MAF, (2) distance from the SNP to transcription start sites (TSS) of the gene was matched to eQTL, and (3) a number of SNPs per gene were matched as in eQTLs.

Enrichment analysis of chromatin states

For each 5 bp window centered on each SNP, we identified overlapping (1) chromatin state predictions from the Roadmap Epigenomics project and (2) regions of open chromatin identified by DNAse-seq from ENCODE [29, 30, 74–76]. In Roadmap, chromatin states are predicted for each tissue or cell type that include enhancers, promoters, and transcribed regions. We used the standard 15-state Roadmap segmentations independently for each of the samples that were matched to GTEx tissues (Additional file 2: Table S9, S4). If a tissue had more than one dataset available, we merged the datasets using BEDTools [77]. For the datasets using genome assembly hg19, we used liftOver to map the peaks to GRCh38 [78]. We built the 2 × 2 contingency table for eQTLs from each factor and across the 15 chromatin states. In the table, the first row includes eQTL variants in the factor, and the second row includes randomly matched SNPs. The columns indicate the number of SNPs that are located in the tested chromatin state in the tested tissues. Both tissues matched for the factor and tissues not matched for the factor were tested. We then ran a one-sided Fisher's exact test for each contingency table and corrected the P values using BH-correction. To summarize the results across tissues and across factors, we used a random-effects model (`rma()` in R) to obtain the combined odds ratio and combined standard error [79].

Heuristic thresholding methods to derive u-eQTLs and ts-eQTLs

heuristic$_1$: We defined ts-eQTLs in one tissue as those with P value > 0.001 in at least 44 other tissues, and with P value $< 100\times$ the most extreme P value of the eGene in the tissue of interest, and within the credible set for that tissue. The thresholds were chosen such that we have a reasonable number of ts-eQTLs, and at the same time only eQTLs with a strong effect in the tissue of interest. u-eQTLs were restricted to those found in the credible sets for at least 5 tissues.

heuristic$_2$: Here, we defined ts-eQTLs in manually defined subsets of similar tissues (Additional file 2: Table S3). For each subset of N_k tissues, the ts-eQTLs were defined as those with P value >0.001 in at least $49 - N_k - 5$ other tissues, and with P value $<100\times$ the most extreme P value of the eGene in $\geq 50\%$ of the tissues in the subset. U-eQTLs were restricted to those found in at least 5 different subsets of tissues.

Simulation

We simulated data with $N = 100$ eQTLs, $T = 10$ tissues, and $K = 5$ factors with sparse loadings and nonnegative factors including a dense factor and four sparse factors. Nonzero values in the loadings were randomly drawn from a standard normal distribution. An error matrix E added noise to the input matrix such that $X = LF^T + E$. Values in E were randomly drawn from normal distribution with mean 0 and different levels of variance σ^2 ($\sigma^2 = 0.001, 0.01, 0.05, 0.1$). To evaluate the performance of multiple methods, we computed the correlation between the learned loadings/factors and the true simulated loadings/factors (factor orderings were permuted to reach the highest correlation for each method), and the relative root mean squared error: $RRMSE(\hat{X}, X) = \sqrt{\frac{\sum_{i,j}(\hat{X}_{i,j}-X_{i,j})^2}{\sum_{i,j} X_{i,j}^2}}$ [18].

Other matrix factorization methods

We ran singular value decomposition (SVD) using the R function `prcomp`, and nonnegative matrix factorization (NMF) using the R package `NMF` [80]. We ran sparse SVD (SSVD) using the R package `ssvd` [33, 81], penalized matrix decomposition (PMD) using the R package `PMA` [34, 82], and softImpute using the R package `softImpute` [35, 83]. We ran flashr using the R package `flashr` [18, 84].

SSVD is reported to be robust to tuning parameters, so we ran SSVD with the default settings [18, 33]. PMD penalizes the two decomposed matrices using either one penalty parameter scaled by the dimensions for each decomposed matrix (PMD$_{CV1}$) or two separate penalty parameters (PMD$_{CV2}$). We chose the tuning parameter by cross-validation, in both PMD$_{CV1}$ and PMD$_{CV2}$ [34]. softImpute has one parameter λ, and we chose it such that the factor matrix reaches the highest sparsity while preserving the rank [35]. To run default flashr, we ran *flashr*. To run flashr_{bf}, we initialized the rank 1 factor and loading using *flashr* ::: udv_{si} where the initial decomposition was done using softImpute (with penalty parameter $\lambda = 0$, [18, 83]). We then did a two-round fitting by first greedily adding factors (*flash_greedy_workhorse*) and then applying backfit (*flash_backfit_workhorse*). In flashr_{NN}, initialization was also done using *flashr* ::: udv_{si}, and nonnegative priors were imposed by setting $ebnm_{param} = list(l = list(mixcompdist = “normal'', optmethod = “mixSQP''), f = list(mixcompdist = “normal'', optmethod = “mixSQP''))$ [18].

Enrichment analysis of transcription factor binding sites

To examine the enrichment of TF binding sites in u-eQTLs and in ts-eQTLs, we constructed the 2×2 contingency tables across factors for each TF. For each TF, we first annotated its binding sites by overlapping tissue-specific enhancer predictions from Roadmap Epigenomics and its TFBS predictions on the genome from JASPAR [28–30]. We then restricted analysis to genes with at least one variant located in TFBS to avoid genes intrinsically lacking variants in TFBS. In the contingency table for each TF, the first row includes eQTLs, and the second row includes randomly matched SNP-gene pairs. For u-eQTLs, the columns indicate the number of genes with or without ubiquitous variants in the TFBS. For ts-eQTLs, first column indicates the number of genes with or without tissue-specific variants in the TFBS. One thing to note is that the TFBS were annotated using matched tissues for each factor. Fisher's exact test was performed for each of these contingency tables, and the *P* values were corrected using Benjamini-Hochberg [73].

For eQTLs from each factor, the analysis was done for TFs with median TPM > 1 in at least half of the corresponding tissues with available data. TFs with a total number of genes in TFBS < 10 were removed. The tissue-specificity of the enriched TFs is unlikely to result from filtering TFs based on expression level and the number of hits (Additional file 1: Figure S37, S38).

Identification of allele-specific binding sites using ChIP-seq data

FASTQ files from human liver samples of HNF4A and CTCF were downloaded from ENCODE web portal and aligned to the GRCh38 genome assembly using STAR [85] (Additional file 2: Table S11). Reads that mapped to variants in GTEx and passed WASP filters were extracted [86]. BAM files of the samples and controls from the same ENCODE repository were downloaded, and peak-calling was performed using MACS2 [87]. Only reads that mapped to peaks at *q* value < 0.1 were included, and ASB was computed for each variant with more than 10 reads by examining if the numbers of reads at each allele were significantly different, using a two-tailed binomial test. Variants with significant ASB events were called at FDR < 0.05 using Benjamini-Hochberg [73].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02129-6.

Additional file 1: Supplementary figure. Figure S1 - S38.

Additional file 2: Supplementary tables. Table S1 - S11.

Additional file 3: GTEx Consortium Information.

Additional file 4: Review history.

Peer review information

Yixin Yao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Acknowledgements

We thank Jessica Bonnie for editing the manuscript, and thank Princy Parsana for open source code.

Review history

The review history is available as Additional file 4.

Authors' contributions

AB, CDB, and YH designed and coordinated the project. YH performed the analysis. YH, AB, and CDB wrote the manuscript. SBC participated in the data analysis and critically revised the manuscript. MA participated in the analysis of GWAS hits. KS participated in running flashr. FA and KGA were responsible for the V8 data generation and cis-eQTL calling. ANB, RB, and HKI harmonized and imputed the GWAS summary statistics. AB and CDB conceived and supervised the study. All authors read and approved the final manuscript.

Authors' information

Twitter handles: @alexisjbattle (Alexis Battle), @casey6r0wn (Christopher D. Brown), @YuanHe7 (Yuan He), @hakyim (Hae Kyung Im), @francoisaguet (François Aguet), @kaushiksrins (Kaushik Srinivasan).

Funding

AB is supported by NIMH 1R01MH109905, NHGRI 1R01HG010480, and Searle Scholar's Program. CDB is supported by RF1AG05547701 and R01HL133218. MA is supported by T32 HL007227. FA and KGA are supported by GTEx program grants HHSN268201000029C, 5U41HG009494. HKI is supported by R01MH107666 and P30DK020595. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH). Additional funds from the National Cancer Institute; National Human Genome Research Institute (NHGRI); National Heart, Lung, and Blood Institute; National Institute on Drug Abuse; National Institute of Mental Health; and National Institute of Neurological Disorders and Stroke. Donors were enrolled at Biospecimen Source Sites funded by Leidos Biomedical, Inc. (Leidos) subcontracts to the National Disease Research Interchange (10XS170) and Roswell Park Cancer Institute (10XS171). The Laboratory, Data Analysis and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through a Leidos subcontract to Van Andel Institute (10ST1035). Additional data repository and project management provided by Leidos (HHSN261200800001E).

Availability of data and materials

The data and analyses presented in the current publication are based on the use of study data downloaded from the dbGaP website under phs000424.v8.p2 [88] and on the GTEx portal (http://gtexportal.org/). All the code used for the matrix factorization and mapping eQTLs is available, under the Creative Commons Attribution 4.0 International License, on Zenodo with the access code DOI: (https://doi.org/10.5281/zenodo.3969649) [89], and GitHub (https://github.com/heyuan7676/ts_eQTLs) [90]. Details of the license can be found here: http://creativecommons.org/licenses/by/4.0/. Data from ROADMAP and ENCODE project used in the analysis are listed in Additional file 2: Table S9, S10, S11.

Ethics approval and consent to participate

Not applicable. The raw data is published with the GTEx Consortium 2020, in submission.

Consent for publication

Not applicable

Competing interests

FA is an inventor on a patent application related to TensorQTL; HKI has received speaker honoraria from GSK and AbbVie.

Author details

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA. [2]HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806, USA. [3]Current Address: Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA. [4]Department of Medicine, Division of Cardiology, Johns Hopkins University, Baltimore, MD, 21287, USA. [5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, 21218, USA. [6]The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7]Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA. [8]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

Received: 30 September 2019 Accepted: 4 August 2020
Published online: 11 September 2020

References

1. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13. https://doi.org/10.1038/nature25160.
2. C Nica A, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K, K Hedman A, Bataille V, Bell J, Surdulescu G, S Dimas A, Ingle C, O Nestle F, Di Meglio P, Min J, Spector T. The architecture of gene regulatory variation across multiple human tissues: the muther study. PLoS Genet. 2011;7:1002003. https://doi.org/10.1371/journal.pgen.1002003.

3. Battle A, Mostafavi S, Zhu X, Potash J, Weissman M, Mccormick C, Haudenschild C, Beckman K, Shi J, Mei R, Urban A, B Montgomery S, F Levinson D, Koller D. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. Genome Res. 2013;24:. https://doi.org/10.1101/gr.155192.113.
4. Innocenti F, M Cooper G, Stanaway I, Gamazon E, D Smith J, Mirkov S, Ramirez J, Liu W, S Lin Y, Moloney C, Force Aldred S, D Trinklein N, Schuetz E, A Nickerson D, E Thummel K, J Rieder M, Rettie A, J Ratain M, J Cox N, Brown C. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. PLoS Genet. 2011;7:1002078. https://doi.org/10.1371/journal.pgen.1002078.
5. Gibbs J, P van der Brug M, Hernandez D, J Traynor B, A Nalls M, Lai S-L, Arepalli S, Dillman A, Rafferty I, Troncoso J, Johnson R, Ronald Zielke H, Ferrucci L, Longo D, Cookson M, B Singleton A. Abundant quantitative trait loci exist for dna methylation and gene expression in human brain. PLoS Genet. 2010;6:1000952. https://doi.org/10.1371/journal.pgen.1000952.
6. Tak YE, Farnham P. Making sense of gwas: using epigenomics and genome engineering to understand the functional relevance of snps in non-coding regions of the human genome. Epigenetics Chromatin. 2015;8. https://doi.org/10.1186/s13072-015-0050-4.
7. Pavlides J, Zhu Z, Gratten J, McRae A, Wray N, Yang J. Predicting gene targets from integrative analyses of summary data from gwas and eqtl studies for 28 human complex traits. Genome Med. 2016;8:. https://doi.org/10.1186/s13073-016-0338-4.
8. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson M, Powell J, W Montgomery G, E Goddard M, Wray N, M Visscher P, Yang J. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. Nat Genet. 2016;48:. https://doi.org/10.1038/ng.3538.
9. Ratnapriya R, A. Sosina O, Starostik M, Kwicklis M, J. Kapphahn R, Fritsche L, Walton A, Arvanitis M, Gieser L, Pietraszkiewicz A, Montezuma S, Chew E, Battle A, R. Abecasis G, Ferrington D, Chatterjee N, Swaroop A. Retinal transcriptome and eqtl analyses identify genes associated with age-related macular degeneration. Nat Genet. 2019;51:. https://doi.org/10.1038/s41588-019-0351-9.
10. Porcu E, Rüeger S, Lepik K, Santoni F, Reymond A, Kutalik Z. Mendelian randomization integrating gwas and eqtl data reveals genetic determinants of complex and clinical traits. Nat Commun. 2019;10:. https://doi.org/10.1038/s41467-019-10936-0.
11. Ongen H, Brown A, Delaneau O, Panousis N, Nica A, Little A, Dermitzakis E. Estimating the causal tissues for complex traits and diseases. Nat Genet. 2017;49:. https://doi.org/10.1038/ng.3981.
12. Finucane H, Reshef Y, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh P-R, Lareau C, Shoresh N, Genovese G, Saunders A, Macosko E, Pollack S, Perry J, Buenrostro J, Bernstein B, Raychaudhuri S, Mccarroll S, Price A. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018;50:621–9. https://doi.org/10.1038/s41588-018-0081-4.
13. Gutierrez-Arcelus M, Ongen H, Lappalainen T, B Montgomery S, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, Falconnet E, Bielser D, Gagnebin M, Giger T, Borel C, Letourneau A, Makrythanasis P, Guipponi M, Gehrig C, Dermitzakis E. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. PLoS Genet. 2015;11:1004958. https://doi.org/10.1371/journal.pgen.1004958.
14. Mckenzie M, K Henders A, Caracella A, Wray N, Powell J. Overlap of expression quantitative trait loci (eqtl) in human brain and blood. BMC Med Genet. 2014;7:31. https://doi.org/10.1186/1755-8794-7-31.
15. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eqtl analysis in multiple tissues. PLoS Genet. 2013;9:1003486. https://doi.org/10.1371/journal.pgen.1003486.
16. Sul JH, Han B, Ye C, Choi T, Eskin E. Effectively identifying eqtls from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet. 2013;9:1003491. https://doi.org/10.1371/journal.pgen.1003491.
17. Strober B, Elorbany R, Rhodes K, Krishnan N, Tayeb K, Battle A, Gilad Y. Dynamic genetic regulation of gene expression during cellular differentiation. Science. 2019;364:1287–90. https://doi.org/10.1126/science.aaw0040.
18. Wang W, Stephens M. Empirical bayes matrix factorization. arXiv:1802.06931. 2018.
19. S Dimas A, Deutsch S, Stranger B, B Montgomery S, Borel C, Attar H, Ingle C, Beazley C, Gutierrez-Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis E, Antonarakis S. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325:1246–50. https://doi.org/10.1126/science.1174148.
20. Wen X, Luca F, Pique-Regi R. Cross-population joint analysis of eqtls: fine mapping and functional annotation. PLoS Genet. 2015;11:1005176. https://doi.org/10.1371/journal.pgen.1005176.
21. Casale F, Horta D, Rakitsch B, Stegle O. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. PLoS Genet. 2017;13:1006693. https://doi.org/10.1371/journal.pgen.1006693.
22. G. Ardlie K, Deluca D, V. Segrè A, J. Sullivan T, Young T, T. Gelfand E, A. Trowbridge C, B. Maller J, Tukiainen T, Lek M, Ward L, Kheradpour P, Iriarte B, Meng Y, D. Palmer C, Esko T, Winckler W, N. Hirschhorn J, Kellis M, C. Lockhart N. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60. https://doi.org/10.1126/science.1262110.
23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
24. W Albert F, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16:. https://doi.org/10.1038/nrg3891.
25. Ongen H, Andersen C, B Bramsen J, Øster B, Rasmussen M, Ferreira P, Sandoval J, Vidal E, Whiffin N, Planchon A, Padioleau I, Bielser D, Romano L, Tomlinson I, S Houlston R, Esteller M, F Orntoft T, Dermitzakis E. Putative cis-regulatory drivers in colorectal cancer. Nature. 2014;512:. https://doi.org/10.1038/nature13602.
26. Heintzman N, Hon G, Hawkins D, Kheradpour P, Stark A, F Harp L, Ye Z, K Lee L, Stuart R, W Ching C, A Ching K, E Antosiewicz-Bourget J, Liu H, Zhang X, D Green R, Lobanenkov V, Stewart R, Thomson J, E Crawford G, Ren B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459:108–12. https://doi.org/10.1038/nature07829.

27. Ernst J, Kheradpour P, S Mikkelsen T, Shoresh N, Ward L, Epstein C, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, E Bernstein B. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9. https://doi.org/10.1038/nature09906.
28. Khan A, Fornes O, Stigliani A, Gheorghe M, Robin van der Lee JAC-M, Bessy A, Chèneby J, Kulkarni SR, Damir Baranasic GT, Arenillas DJ, Sandelin A, Vandepoele K, Lenhard B, Ballester B, Wasserman WW, Parcy F, Mathelier A. Jaspar 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46:260–6. https://doi.org/10.1093/nar/gkx1126.
29. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with chromhmm. Nat Protoc. 2017;12: 2017–124. https://doi.org/10.1038/nprot.2017.124.
30. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller M, Amin V, Whitaker J, Schultz M, Ward L, Sarkar A, Quon G, Sandstrom R, Eaton M, Wu Y-C, Lin Y. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30. https://doi.org/10.1038/nature14248.
31. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26. https://doi.org/10.1038/nrg3207.
32. W Whitfield T, Wang J, J Collins P, Partridge E, Force Aldred S, D Trinklein N, Myers R, Deng Z. Functional analysis of transcription factor binding sites in human promoters. Genome Biol. 2012;13:50. https://doi.org/10.1186/gb-2012-13-9-r50.
33. Yang D, Ma Z, Buja A. A sparse singular value decomposition method for high-dimensional data. J Comput Graph Stat. 2014;23. https://doi.org/10.1080/10618600.2013.858632.
34. Witten D, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostat (Oxford, England). 2009;10:515–34. https://doi.org/10.1093/biostatistics/kxp008.
35. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res. 2010;11:2287–322.
36. Knowles D, Ghahramani Z. Nonparametric bayesian sparse factor models with application to gene expression modeling. Ann Appl Stat. 2010;5:. https://doi.org/10.1214/10-AOAS435.
37. Yoo W, Mayberry R, Bae S, Singh K, He Q, Lillard J. A study of effects of multicollinearity in the multivariable analysis. Int J Appl Sci Technol. 2014;4:9–19.
38. Stine R. Graphical interpretation of variance inflation factors. Am Stat. 1995;49:53–56.
39. Lynch V, May G, Wagner G. Regulatory evolution through divergence of a phosphoswitch in the transcription factor cebpb (vol 480, pg 383, 2011). Nature. 2011;480:383–6. https://doi.org/10.1038/nature10595.
40. Ko C-Y, Chang W-C, Wang JM. Biological roles of ccaat/enhancer-binding protein delta during inflammation. J Biomed Sci. 2015;22:6. https://doi.org/10.1186/s12929-014-0110-2.
41. Papaioannou V. The t-box gene family: emerging roles in development, stem cells and cancer. Development. 2014;141:3819–33. https://doi.org/10.1242/dev.104471.
42. Hess J, Angel P, Schorpp-Kistner M. Ap-1 subunits: quarrel and harmony among siblings. J Cell Sci. 2005;117: 5965–73. https://doi.org/10.1242/jcs.01589.
43. Servitja J-M, Pignatelli M, Maestro M, Cardalda C, f boj S, Lozano J, Blanco E, Lafuente A, I McCarthy M, Sumoy L, Guigó R, Ferrer J. Hnf1 (mody3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver. Mol Cell Biol. 2009;29:2945–59. https://doi.org/10.1128/MCB.01389-08.
44. D'Angelo A, Bluteau O, Garcia-Gonzalez M, Gresh L, Doyen A, Garbay S, Robine S, Pontoglio M. Hepatocyte nuclear factor 1 and control terminal differentiation and cell fate commitment in the gut epithelium. Development. 2010;137:1573–82. https://doi.org/10.1242/dev.044420.
45. He A, Gu F, Hu Y, Ma Q, Ye LY, Akiyama JA, Visel A, Pennacchio LA, Pu WT. Dynamic gata4 enhancers shape the chromatin landscape central to heart development and disease. Nat Commun. 2014;5:4907. https://doi.org/10.1038/ncomms5907.
46. Ang Y-S, Rivas R, Ribeiro A, Srivas R, Rivera J, Stone N, Pratt K, Mohamed T, Fu J-D, Spencer C, Tippens N, Li M, Narasimha A, Radzinsky E, Moon-Grady A, Yu H, Pruitt B, Snyder M, Srivastava D. Disease model of gata4 mutation reveals transcription factor cooperativity in human cardiogenesis. Cell. 2016;167:. https://doi.org/10.1016/j.cell.2016.11.033.
47. P. Hayhurst G, Lee Y-H, Lambert G, M. Ward J, J. Gonzalez F. Hepatocyte nuclear factor 4 (nuclear receptor 2a1) is essential for maintenance of hepatic gene expression and lipid homeostasis. Mol Cell Biol. 2001;21:1393–403. https://doi.org/10.1128/MCB.21.4.1393-1403.2001.
48. Parviz F, Matullo C, D Garrison W, Savatski L, W Adamson J, Ning G, Kaestner K, Rossi J, S Zaret K, Duncan S. Hepatocyte nuclear factor 4 controls the development of a hepatic epithelium and liver morphogenesis. Nat Genet. 2003;34:292–6. https://doi.org/10.1038/ng1175.
49. Hasty P, Bradley A, Morris JH, Edmondson DG, Venuti JM, Olson EN, Klein WH. Muscle deficiency and neonatal death in mice with a targeted mutation in the myogenin gene. Nature. 1993;364:501–6. https://doi.org/10.1038/364501a0.
50. Kuo C, Morrisey E, Anandappa R, Sigrist K, Lu M, Parmacek MS, Soudais C, Leiden JM. Gata4 transcription factor is required for ventral morphogenesis and heart tube formation. Genes Dev. 1997;11:1048–60. https://doi.org/10.1101/gad.11.8.1048.
51. Chen J, Qi B, Zhao J, Liu W, Duan R, Zhang M. A novel mutation of gata4 (k300t) associated with familial atrial septal defect. Gene. 2015;575:. https://doi.org/10.1016/j.gene.2015.09.021.
52. Li J, Liu W-D, Yang Z-L, Yuan F, Xu L, Li R-G, Yang Y-Q. Gene. 2014;548:. https://doi.org/10.1016/j.gene.2014.07.022.
53. Haghikia A, Ricke-Hoch M, Stapel B, Gorst I, Hilfiker-Kleiner D. Stat3, a key regulator of cell-to-cell communication in the heart. Cardiovasc Res. 2014;102:. https://doi.org/10.1093/cvr/cvu034.
54. Pruim R, P Welch R, Sanna S, Teslovich T, S Chines P, P Gliedt T, Boehnke M, R Abecasis G, Willer C. Locuszoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:2336–7. https://doi.org/10.1093/bioinformatics/btq419.

55. Willer C, M Schmidt E, Sengupta S, M Peloso G, Gustafsson S, Kanoni S, Ganna A, Chen J, L Buchkovich M, Mora S, Beckmann J, Bragg-Gresham J, Chang H-Y, Demirkan A, den Hertog H, Do R, A Donnelly L, B Ehret G, Esko T, R Abecasis G. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45:. https://doi.org/10.1038/ng.2797.
56. Mahajan A, Taliun D, Thurner M, R. Robertson N, M. Torres J, William Rayner N, J. Payne A, Steinthorsdottir V, A. Scott R, Grarup N, Cook J, M. Schmidt E, Wuttke M, Sarnowski C, Mägi R, Nano J, Gieger C, Trompet S, Lecoeur C. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018;50:. https://doi.org/10.1038/s41588-018-0241-6.
57. Watanabe K, Stringer S, Frei O, Mirkov M, Leeuw C, Polderman T, Sluis S, Andreassen O, Neale B, Posthuma D. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet. 2019;51:1–10. https://doi.org/10.1038/s41588-019-0481-0.
58. Delaforest A, Di Furio F, Jing R, Ludwig-Kubinski A, Twaroski K, Urick A, Pulakanti K, Rao S, Duncan S. Hnf4a regulates the formation of hepatic progenitor cells from human ipsc-derived endoderm by facilitating efficient recruitment of rna pol ii. Genes. 2018;10:21. https://doi.org/10.3390/genes10010021.
59. Qu M, Duffy T, Hirota T, Kay S. Nuclear receptor hnf4a transrepresses clock:bmal1 and modulates tissue-specific circadian networks. Proc Natl Acad Sci. 2018;115:201816411. https://doi.org/10.1073/pnas.1816411115.
60. Tremblay A, Lamarche B, Lemelin V, Hoos L, Benjannet S, Seidah N, Davis H, Couture P. Atorvastatin increases intestinal expression of npc1l1 in hyperlipidemic men. J Lipid Res. 2011;52:558–65. https://doi.org/10.1194/jlr.M011080.
61. Ouyang H, Qin Y, Liu Y, Xie Y, Liu J. Prox1 directly interacts with lsd1 and recruits the lsd1/nurd complex to epigenetically co-repress cyp7a1 transcription. PLoS ONE. 2013;8:62192. https://doi.org/10.1371/journal.pone.0062192.
62. Guo D, Dong L-y, Wu Y, Yang L, An W. Down-regulation of hepatic nuclear factor 4-alpha on expression of human hepatic stimulator substance via its action on the proximal promoter in hepg2 cells. Biochem J. 2008;415:111–21. https://doi.org/10.1042/BJ20080221.
63. Giambartolomei C, Vukcevic D, Schadt E, Franke L, Hingorani A, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:1004383. https://doi.org/10.1371/journal.pgen.1004383.
64. Cook B, Dynek J, Chang W, Shostak G, Smith S. Role for the related poly(adp-ribose) polymerases tankyrase 1 and 2 at human telomeres. Mol Cell Biol. 2002;22:332–42. https://doi.org/10.1128/MCB.22.1.332-342.2002.
65. Li N, Wang Y, Neri S, Zhen Y, Fong LWR, Qiao Y, Li X, Chen Z, Stephan C, Deng W, Ye R, Jiang W, Zhang S, Yu Y, Hung M-C, Chen J, Lin SH. Tankyrase disrupts metabolic homeostasis and promotes tumorigenesis by inhibiting lkb1-ampk signalling. Nat Commun. 2019;10:4363. https://doi.org/10.1038/s41467-019-12377-1.
66. Wakefield J. Bayes factors for genome-wide association studies: comparison with p-values. Genet Epidemiol. 2009;33:79–86. https://doi.org/10.1002/gepi.20359.
67. Goeman JJ, Meijer RJ, Chaturvedi N. Penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. 2018. R package version 0.9-51.
68. Goeman JJ. L1 penalized estimation in the cox proportional hazards model. Biom J. 2010;52:14. https://doi.org/10.1002/bimj.200900028.
69. Brunet J-P, Tamayo P, R Golub T, P Mesirov J. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci. 2004;101:4164–9. https://doi.org/10.1073/pnas.0308531101.
70. Wu S, Joseph A, S. Hammonds A, E. Celniker S, Yu B, Frise E. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proc Natl Acad Sci. 2016;113:201521171. https://doi.org/10.1073/pnas.1521171113.
71. Rawlings J, Pantula S, Dickey D. Applied regression analysis - a research tool: second edition. New York: Springer; 1998.
72. Purcell S, Neale B, Todd-Brown K, Thomas L, A.R. Ferreira M, Bender D, Maller J, Sklar P, I W de Bakker P, Daly M, C Sham P. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75. https://doi.org/10.1086/519795.
73. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
74. Ernst J, Kellis M. Chromhmm: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6. https://doi.org/10.1038/nmeth.1906.
75. Dunham I, Birney E, Lajoie B, Sanyal A, Dong X, Greven M, Xinying L, Wang J, W. Whitfield T, Zhuang J, Dekker J, Deng Z, Jain G, ENCODE PC. An integrated encyclopedia of dna elements in the human genome. Nature. 2012;489. https://doi.org/10.1038/nature11247.
76. Davis C, Hitz B, A Sloan C, Chan E, M Davidson J, Gabdank I, Hilton J, Jain K, K Baymuradov U, K Narayanan A, Onate K, Graham K, R Miyasato S, Dreszer T, Seth Strattan J, Jolanki O, Y Tanaka F, Michael Cherry J. The encyclopedia of dna elements (encode): data portal update. Nucleic Acids Res. 2017;46:. https://doi.org/10.1093/nar/gkx1081.
77. Quinlan A, M Hall I. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
78. James Kent W, W Sugnet C, Furey T, M Roskin K, Pringle T, Zahler A, Haussler D. The human genome browser at ucsc. Genome Res. 2002;12:996–1006. https://doi.org/10.1101/gr.229102.
79. Viechtbauer W. Conducting meta-analyses in r with the metafor package. J Stat Softw. 2010;36:. https://doi.org/10.18637/jss.v036.i03.
80. Gaujoux R, Seoighe C. The package NMF: manual pages. CRAN. 2017. http://cran.r-project.org/package=NMF. R package version 0.23.6.
81. Yang D. ssvd: sparse SVD. 2013. https://CRAN.R-project.org/package=ssvd. R package version 1.0.
82. Witten D, Tibshirani R, Gross S, Narasimhan B. PMA: Penalized Multivariate Analysis. 2019. https://CRAN.R-project.org/package=PMA. R package version 1.1.
83. Hastie T, Mazumder R. softimpute: matrix completion via iterative soft-thresholded svd. 2015. R package version 1.4.

84. Stephens M, Wang W, Willwerscheid J. flashr: empirical Bayes matrix factorization. 2019. http://github.com/stephenslab/flashr. R package version 0.6-7.
85. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras T. Star: ultrafast universal rna-seq aligner. Bioinformatics. 2012;29. https://doi.org/10.1093/bioinformatics/bts635.
86. Geijn B, McVicker G, Gilad Y, Pritchard J. Wasp: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015;12:. https://doi.org/10.1038/nmeth.3582.
87. Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nusbaum C, Myers R, Brown M, Li W, Liu S. Model-based analysis of chip-seq (macs). Genome Biol. 2008;9:137. https://doi.org/10.1186/gb-2008-9-9-r137.
88. GTEx Consortium. Genotype-Tissue Expression Project (GTEx). dbGaP. 2020. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2.
89. He Y, Chhetri SB, Arvanitis M, Srinivasan K, Aguet F, Ardlie KG, Barbeira AN, Bonazzola R, Im HK, GTEx Consortium, Brown CD, Battle A. Extended data for sn-spMF: matrix factorization informs tissue-specific genetic regulation of gene expression. Zenodo. 2020. https://doi.org/10.5281/zenodo.3969649.
90. He Y, Chhetri SB, Arvanitis M, Srinivasan K, Aguet F, Ardlie KG, Barbeira AN, Bonazzola R, Im HK, GTEx Consortium, Brown CD, Battle A. Learn latent factors using sn-spMF. Github. 2020. https://github.com/heyuan7676/ts_eQTLs.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.