

Research Article



Telescoping bimodal latent Dirichlet allocation to identify expression QTLs across tissues

Ariel DH Gewirtz¹ , F William Townes², Barbara E Engelhardt^{2,3}

Expression quantitative trait loci (eQTLs), or single-nucleotide polymorphisms that affect average gene expression levels, provide important insights into context-specific gene regulation. Classic eQTL analyses use one-to-one association tests, which test gene-variant pairs individually and ignore correlations induced by gene regulatory networks and linkage disequilibrium. Probabilistic topic models, such as latent Dirichlet allocation, estimate latent topics for a collection of count observations. Prior multimodal frameworks that bridge genotype and expression data assume matched sample numbers between modalities. However, many data sets have a nested structure where one individual has several associated gene expression samples and a single germline genotype vector. Here, we build a telescoping bimodal latent Dirichlet allocation (TBLDA) framework to learn shared topics across gene expression and genotype data that allows multiple RNA sequencing samples to correspond to a single individual's genotype. By using raw count data, our model avoids possible adulteration via normalization procedures. Ancestral structure is captured in a genotype-specific latent space, effectively removing it from shared components. Using GTEx v8 expression data across 10 tissues and genotype data, we show that the estimated topics capture meaningful and robust biological signal in both modalities and identify associations within and across tissue types. We identify 4,645 cis-eQTLs and 995 trans-eQTLs by conducting eQTL mapping between the most informative features in each topic. Our TBLDA model is able to identify associations using raw sequencing count data when the samples in two separate data modalities are matched one-to-many, as is often the case in biological data. Our code is freely available at <https://github.com/gewirtz/TBLDA>.

DOI [10.26508/lsa.202101297](https://doi.org/10.26508/lsa.202101297) | Received 11 November 2021 | Revised 15 July 2022 | Accepted 18 July 2022 | Published online 17 August 2022

Introduction

Genomic differences, such as single-nucleotide polymorphisms (SNPs), among individuals are important drivers of gene expression

variability. Much previous work has focused on discovering expression quantitative trait loci (eQTLs), which capture associations between the number of copies of a minor allele present at a given genomic locus and the expression level of a single gene (GTEx Consortium, 2017; GTEx Consortium, 2020). However, a one-to-one mapping of genes to SNPs is too simplistic given the reality of biological interactions and the availability of many observations per individual. Pleiotropy, gene regulatory networks with biological redundancy and feedback loops, and linkage disequilibrium (LD) blocks of highly correlated SNPs all contribute to a complex and dynamic biological regulatory system.

From a statistical perspective, performing genome-wide one-to-one association tests yields an astronomical multiple-testing burden for trans-eQTLs, where the agnostic approach examines every interchromosomal gene and SNP combination. Statistical power is further reduced because trans-eQTLs, or eQTLs where the regulatory SNP is on a different chromosome than the gene that it regulates, often have smaller effect sizes than cis-eQTLs or eQTLs where the regulatory SNP is local to the gene that it regulates (Petretto et al, 2006). One method to reduce the effective number of tests is to cluster correlated SNPs and genes and compare the averaged cluster signals versus testing for every possible marginal association.

Probabilistic topic models, such as latent Dirichlet allocation (LDA), are unsupervised machine learning methods that were initially introduced in natural language processing (Blei et al, 2003) and in statistical genetics as models of ancestry (Pritchard et al, 2000). LDA finds latent topics via soft clustering of feature counts over many samples while simultaneously estimating each sample's topic membership proportions. More recently, these types of models have been applied to gene expression data with gene counts as features. The topics estimated by these models represent interpretable underlying biology such as cell type or developmental stage and have been used in QTL mapping as the quantitative traits themselves (Hore et al, 2016; Dey et al, 2017).

We hypothesized that multimodal topic modeling could identify clusters of covarying genes and SNPs. Existing methods have used Dirichlet process mixture models to integrate two data modalities (Savage et al, 2010), but nonparametric Bayesian models tend to be

¹Lewis-Sigler Institute of Integrative Genomics, Princeton University, Princeton, NJ, USA ²Department of Computer Science, Princeton University, Princeton, NJ, USA
³Gladstone Institutes, San Francisco, CA, USA

Correspondence: barbara.engelhardt@gladstone.ucsf.edu

too computationally intense for larger data sets such as modern genotype arrays, which capture millions of SNPs. Argelaguet et al (2018) designed a factor model framework (MOFA) to jointly model multiple data modalities, allowing various data likelihoods via link functions. However, relevant methods assume that the modalities are measured on the same samples such that there is a single observation from each individual in each modality (Savage et al, 2010; Virtanen et al, 2012; Zhao et al, 2016; Argelaguet et al, 2018; Li & Gaynanova, 2018; Ash et al, 2021). Many earlier methods also require gene expression data to be normalized, potentially adulterating true signals or spuriously adding false ones (Robinson et al, 2010; Love et al, 2014; Hore et al, 2016; Hicks et al, 2018).

In this work, we create a probabilistic model to find shared structure between gene expression and genotype data. Our model uses raw sequencing read counts and is designed for a nested data structure where although samples are paired, modalities may have different numbers of samples from each subject. This is often the case when we have many samples of gene expression from a particular donor—as in the GTEx data with multiple tissue samples per donor and also for single-cell RNA sequencing samples with multiple cells per subject—but a single germline genotype vector. We apply our unsupervised model to GTEx v8 data and use known sample tissue labels and cell type enrichment scores post hoc to interpret the biological context of the estimated components (GTEx Consortium, 2020). To demonstrate the model's ability to find shared variation between data modalities, we conduct eQTL mapping using the most informative features in each topic to find both known and novel—and tissue-specific and general—cis- and trans-eQTLs.

Results

We applied our telescoping bimodal LDA (TBLDA) model to gene expression data for the 10 tissues with the highest number of samples from the v8 GTEx data release and to the genotypes from all individuals who contributed to at least one of the samples (Table 1). We took advantage of the known GTEx covariates to interpret biological variation captured in the model factors and ensure relevant signal was found.

First, we checked that ancestral structure, using reported ancestry as a proxy, was not associated with the estimated shared factors. As expected, ancestral structure was largely controlled for because it is captured in the genotype-specific portion of the model (median absolute value factor-ancestry point biserial correlation coefficient < 0.01). This contrasts with the genotype-specific factors, each of which had a point biserial correlation coefficient of at least 0.41 (ranging to 0.99) with at least one reported race.

Next, we looked for signal from one of the top sources of known variation in the data set, tissue of origin, by identifying factors active in specific tissues. We found 15,439 tissue-factor associations via the inner product of each factor and a tissue indicator vector, considering inner products greater than 40 to be tissue-associated (Fig S1 and Table 1). Tissue sample size was strongly correlated with the number of tissue-associated factors (Kendall's rank correlation $\tau = 0.64$, $P < 0.01$), which suggests that certain tissues may have underpowered downstream analyses. Overall runs, whole blood, and skeletal muscle had the most associated factors (2,599 and 2,649, respectively), whereas the tibial nerve had the fewest (905). The skin (sun-exposed), skin (not sun-exposed), lung, subcutaneous adipose, and thyroid had the weakest associations (median inner products between 60.7 and 69.7); whole blood, tibial nerve, and esophagus mucosa had the strongest (median inner products between 88.6 and 109.6). Accordingly, whole blood and skeletal muscle samples allocate most of their topic membership into tissue-specific factors (Fig 1).

To explore the robustness of these tissue-associated factors, we compiled sets of the top-ranked features that frequently appeared across factors associated with a common tissue (see the Extended Methods section). Taken together, the two skin tissues had the largest group with 6,515 genes, whereas whole blood had the largest number of unique genes considering the other sets (1,983). All tissue-associated robust gene sets were enriched for functionally relevant biological process Gene Ontology (GO) sets (Benjamini Hochberg [BH] FDR < 0.1, Table S1). Furthermore, all relevant robust tissue gene sets (except for tibial artery) contained most of the tissue-specific transcription factors (TFs) present in the overall analysis (whole blood 10/12, thyroid 6/8, esophagus 7/8, skins 12/14, lung 3/4, nerve 3/3, skeletal muscle 12/12, tibial artery 0/2,

Table 1. Overall summary statistics and available data for each of the 10 tissues included in the analysis.

Tissue	Sample size	Num. tissue-associated factors	Cell type enrichment scores
Subcutaneous adipose	581	1,536	Adipocytes
Tibial artery	584	1,504	
Esophagus mucosa	497	1,254	Keratinocytes, epithelial cells
Lung	515	1,095	Epithelial cells
Skeletal muscle	706	2,649	Myocytes
Tibial nerve	532	905	
Skin (not sun-exposed)	517	1,225	Keratinocytes, epithelial cells
Skin (sun-exposed)	605	1,390	Keratinocytes, epithelial cells
Thyroid	574	1,282	Epithelial cells
Whole blood	670	2,599	Neutrophils

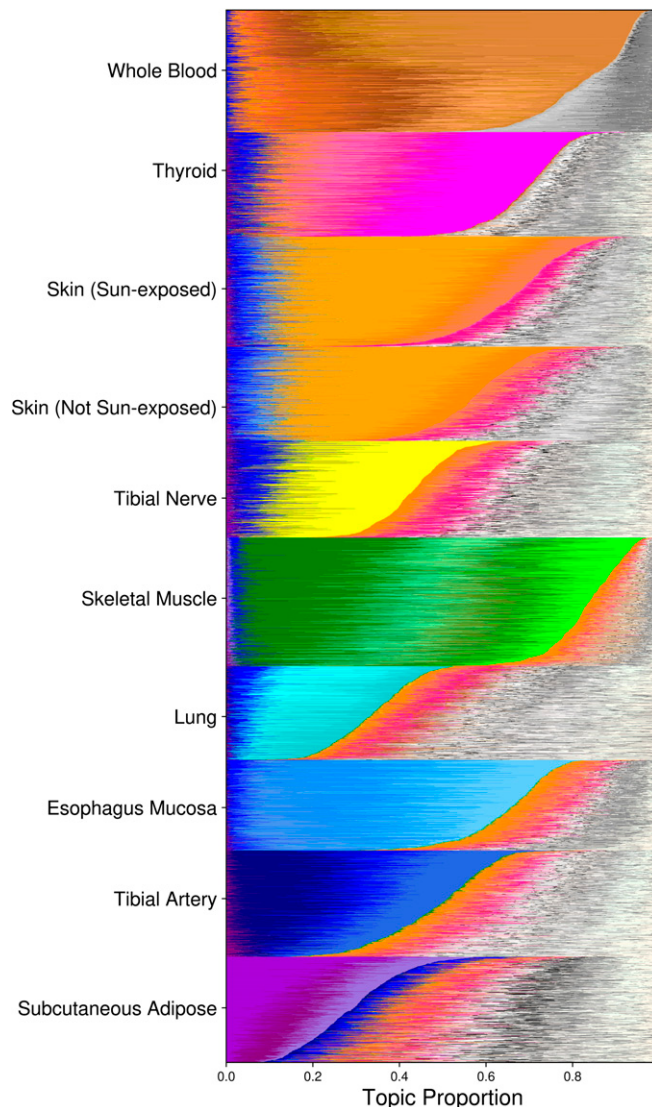


Figure 1. The estimated TBLDA topics capture tissue-specific signal. Each row depicts the expected sample-topic proportion for one sample for the model fit using genes on chromosome 19 and single-nucleotide polymorphisms on chromosome 22; samples are sorted by tissue. Tissue-associated topics are colored in a tissue-wise family color scheme. The remaining topics are drawn using a random gray scale.

subcutaneous adipose 2/2; see the Extended Methods section), whereas only one gene set included another tissue's TF (subcutaneous adipose contained one TF from whole blood) (Sonawane et al, 2017). We note that the tissue-specific TFs were also determined using GTEx data, and our analysis leads to results consistent with these prior results on these same GTEx data. This demonstrates that our model consistently found topics that captured important tissue-specific biological variation including functional pathways and tissue-specific regulatory activity.

Next, we used a compilation of 63 SNP classes describing general annotations shared across tissues from the LDSC (Bulik-Sullivan et al, 2015) data repository to explore functional regulatory enrichments among tissue-associated SNPs. The union of all robust

tissue-associated SNPs was enriched for 20 SNP classes (Fisher's exact test, BH FDR < 0.1) including TFBS ENCODE (BH FDR < 0.016), SuperEnhancer Hnisz (BH FDR $\leq 3.4 \times 10^{-3}$) and active enhancer-associated H3K27ac Hnisz (BH FDR $\leq 9.8 \times 10^{-5}$), and H3K4me1 Trynka (BH FDR $\leq 2.6 \times 10^{-5}$). In particular, several single tissue-associated SNP sets are associated with DGF ENCODE, DHS_Trynka.extend.500, H3K4me1 Trynka, H3K27ac Hnisz, and Enhancer Hoffman (eight, six, five, five, and three tissues, respectively, of the 10 total; BH FDR < 0.1; Fig 2 and Table S2). These SNP set enrichments from our model show that TBLDA identifies functional connections between genotype and gene transcription; these enrichments are intriguing because trans-eQTLs are known to be associated with enhancer activity (GTEx Consortium, 2017).

We then investigated whether the SNP sets are clustered together in particular genomic regions. The union of all tissue-associated SNPs was not enriched in any chromosomal regions using a bin size of 250,000 bp and a sliding window of 100,000 bp, but there were 143 tissue-specific genomic bin enrichments (Fisher's exact test, FDR < 0.1; Fig 2 and Table S3). Notably, 48 regions on chromosome four were enriched for the robust SNP set associated with subcutaneous adipose (BH FDR < 0.05). This highlights the ability of TBLDA to identify jointly functional genomic regions even when the SNP data have been LD-pruned.

Although the GTEx data provide the ground truth of each sample's origin tissue, this is not the case across all data sets. Thus, we next evaluated whether our model could recover robust components across relevant runs in an unsupervised manner. To do this, we ran our model 484 times, once for each pair of chromosomes in the GTEx v8 data, and identified shared components across these runs (see the Extended Methods section). Across all runs, we recovered 197 clusters of robust genotype factors and 1,799 groups of robust gene expression factors. Loadings that were well-correlated with each other across runs tended to cluster by tissue; 81 of the robust genotype clusters and 468 of the robust gene expression clusters included factors that were associated with the same tissue (Fig 3). Only 14 and 75 of the robust genotype and expression clusters, respectively, did not include tissue-associated factors. The presence of these tissue-associated robust genotype components demonstrates that TBLDA identifies interactions between the data modalities versus separate structure within each modality.

Because of the nature of bulk RNA-seq expression data, the GTEx samples average expression over heterogeneous tissue samples containing various cell types. We computed the Kendall correlation between cell type enrichment scores and factor values to determine whether factors capture sample cell type composition (Fig S2). We use estimated enrichment scores for bulk cell deconvolution across five cell types (adipocytes, keratinocytes, epithelial cells, myocytes, and neutrophils) in 8 tissues for a total of 11 tissue and cell type pairs (Table 1) (GTEx Consortium, 2020). Enrichment scores for 8/11 pairs of tissue and cell types were well captured by at least one factor (maximum abs (Kendall τ) > 0.5). This suggests that the TBLDA components often represent cell type-specific processes within tissue samples.

To test whether traditional eQTLs ascertained using univariate tests are captured by TBLDA, we ran a linear model for association between the top 10% most informative SNPs and genes on common

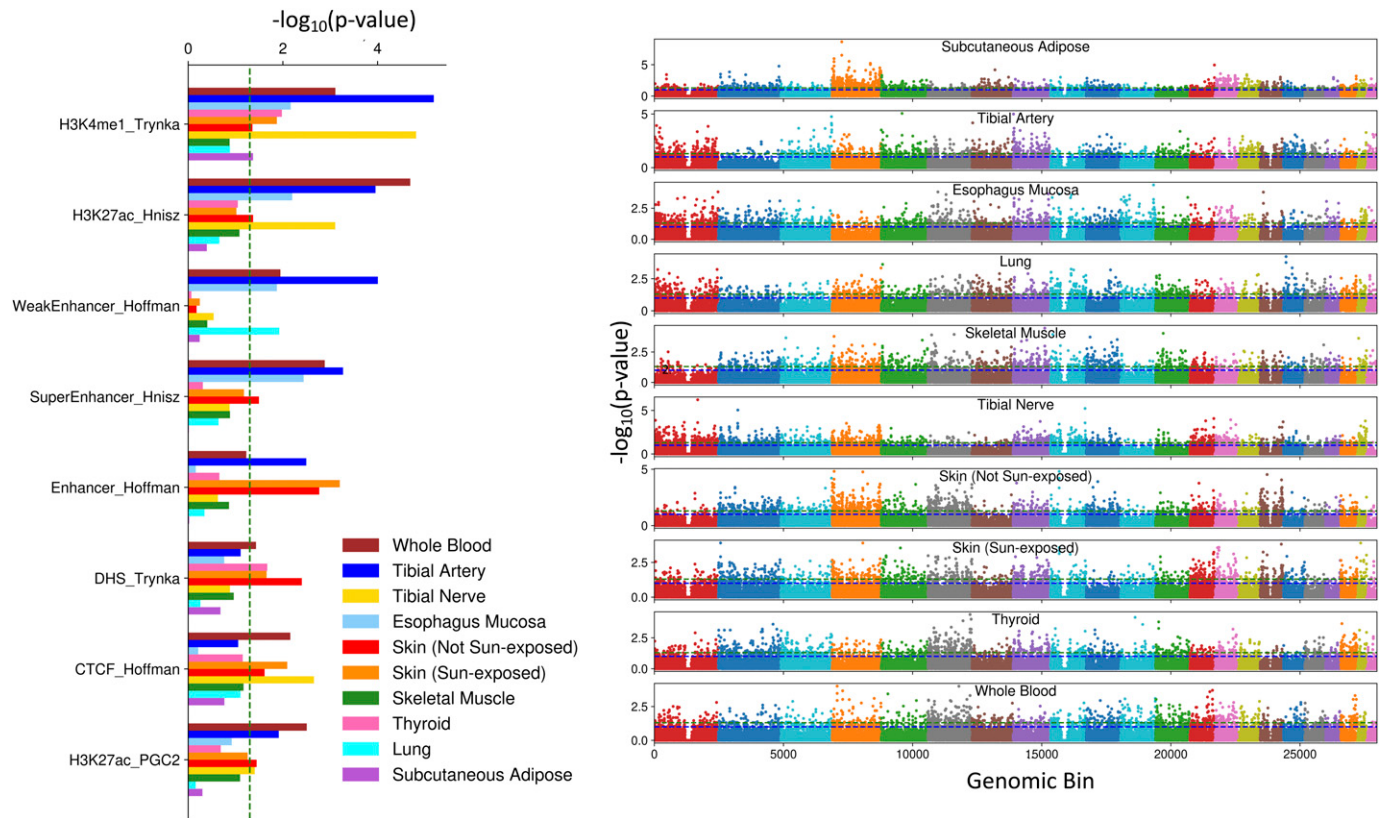


Figure 2. Robust tissue-associated single-nucleotide polymorphism (SNP) sets are enriched for DNA markers and localized throughout the genome. Left: enrichment via Fisher's exact test of eight of the 63 LDSC SNP classes across all robust tissue-associated SNP sets. Right: each tissue's associated SNP set was tested for genomic localization via Fisher's exact test. The blue and green dotted lines are drawn at P -value thresholds of 0.1 and 0.05, respectively. The colors mark the division between ordered chromosomes, with chromosome one on the far left.

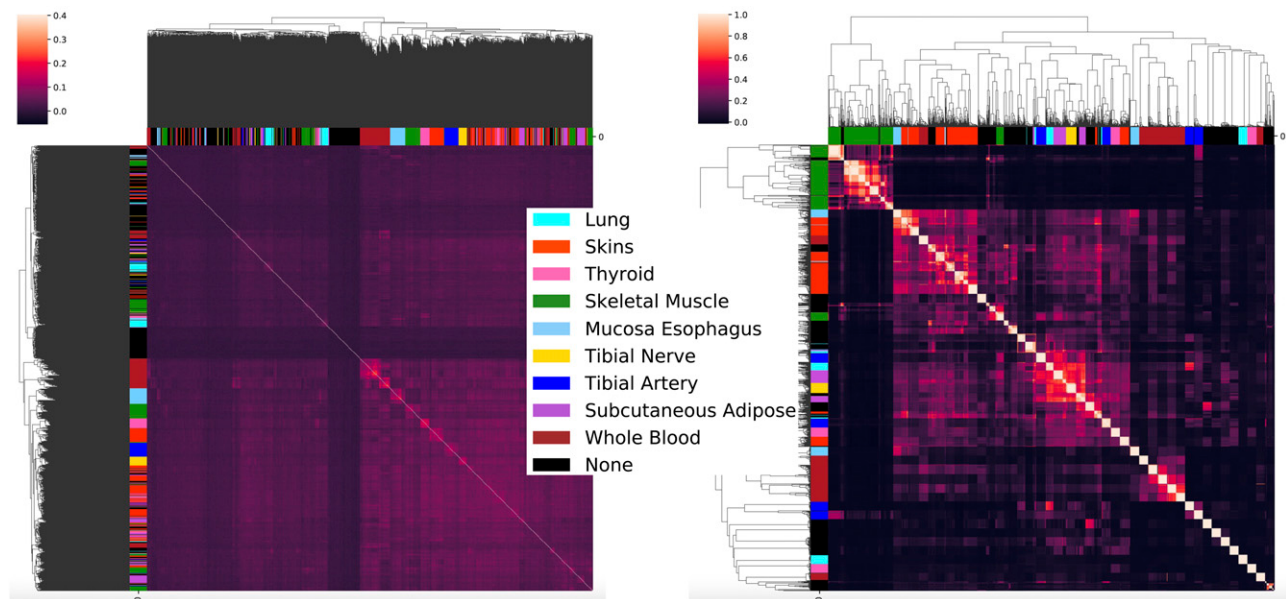


Figure 3. The TBLDA model estimates robust factors across independent runs. Cluster maps of the pairwise Pearson correlations between loadings from all runs that used features from chromosome two. The color bars associated with the axes label the topic's strongest tissue association, if any. Left: correlations calculated using the residuals after regressing coded MAF out from the expectation of the genotype loadings. Right: correlations between the expected value of gene expression loadings.



Figure 4. Characterization of cis- and trans-expression quantitative trait loci (eQTLs) between top-ranked features in each factor.

Top: for each of the 484 model runs, the ordered true MatrixEQTL association $-\log_{10}(P\text{-values})$ (y-axis) are plotted against ordered $-\log_{10}(P\text{-values})$ from tests using the same features but permuted expression and covariate data (x-axis). Clear cis-eQTL enrichment is present across all intrachromosomal runs. Points at which the ordered

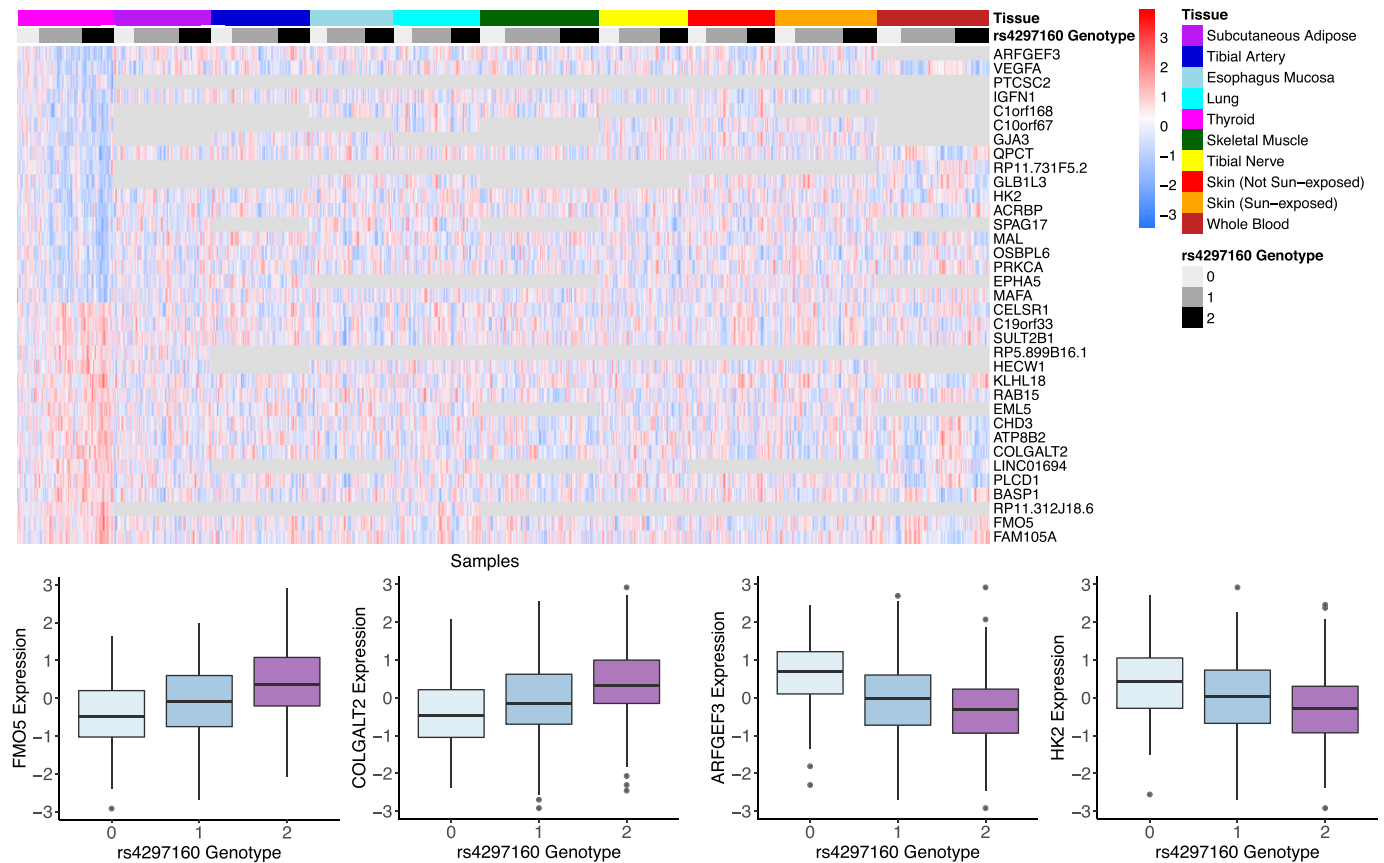


Figure 5. Exploring the 34 trans-eGenes associated with a single locus in thyroid.

Top: heatmap of quantile-normalized expression values for the trans eGenes associated with rs4297160 along with *PTCSC2*. Bottom: visualization of four of the trans-expression quantitative trait loci.

factors for each tissue separately using MatrixEQTL (Shabalín, 2012) on held-out data (further referred to as the multivariate testing approach; Fig 4). Of 10,855,277 total tests, we found 4,645 cis-eQTLs at BH FDR < 0.1 across all 10 tissues including 2,149 unique eVariants and 1,868 unique eGenes (Table S4). Most of these cis-eQTLs (3,923) affect protein-coding genes, with a minority (722) acting on long intergenic noncoding RNA (lincRNA) genes (Fig 4). Thyroid had the most cis-eQTLs, 692, followed by whole blood with 657.

We also discovered 995 trans-eQTLs at BH FDR < 0.1, which include 901 unique trans-eGenes and 979 unique trans-eVariants (Table S4). In contrast to both the data, which consist of 86% protein-coding genes, and the cis-eQTLs, trans-eQTLs have an approximately equal number of lincRNA- and protein-coding eGenes (517 and 478, respectively; Fig 4). Although these numbers may seem surprising, lincRNAs localize to the nucleus and are chromatin-associated, often acting in trans through chromatin modifiers (Holdt et al, 2013; Hacisuleyman et al, 2014; Cao et al, 2021). However, trans-eQTLs found using bulk RNA-seq can appear because of sample cell type proportions (Võsa et al, 2021). Because

lincRNAs show more cell type-specific expression than protein-coding transcripts, this could contribute to the imbalanced numbers of noncoding and coding trans eGenes (Liu et al, 2016; Grassi et al, 2021). The skin (not sun-exposed) had the highest number of trans-eQTLs (124), followed by the lung with 105. Not surprisingly (GTEx Consortium, 2017), the cis-eQTL enrichment is much stronger and more consistent than the trans-eQTL enrichment (Fig 4). These eQTL mapping results highlight the associations between SNPs and genes loaded onto a common factor and suggest that traditional eQTL candidates may be identified using the TBLDA factors.

Next, we restricted our analysis within each tissue to the respective tissue's associated factors. We found 746 cis-eQTLs (618 unique eGenes and 662 unique eVariants) and 939 trans-eQTLs (853 unique trans-eGenes and 933 unique trans-eVariants) at BH FDR < 0.1 (Table S5). Whole blood had the most cis-eQTLs and the most trans-eQTLs. Similarly, to the full analysis above, the trans-eQTLs have an approximately equal split of lincRNA and protein-coding eGenes, whereas the cis-eGenes are mostly protein-coding (Fig S3). Of the discoveries, 877 (93.4%) of the trans-eQTLs and 324 (43.4%) of

true $-\log_{10}(P\text{-value})$ is greater than the maximum permuted $-\log_{10}(P\text{-value})$ are colored in red to highlight deviation. Bottom: histograms depicting the numbers of trans- (top) and cis- (bottom) eQTLs mapped per-tissue and split by gene type.

the cis-eQTLs were novel, meaning not below the significance threshold in the unrestricted multivariate test using all common factors. The fact that 933 (93.8%) of the trans-eQTLs found by the full multivariate test were not found in the tissue-associated factors indicates that most of the trans associations found by the model are not in specific tissue-factor pairs.

Thus, to increase power to find trans-eQTLs shared across tissues, we next limited association tests to features in general factors that were not linked to any tissue. This approach yielded 964 trans-eQTLs (863 unique trans-eGenes and 945 unique trans-eVariants) and 2,901 cis-eQTLs (1,210 unique eGenes and 1,355 unique eVariants; Fig S3 and Table S6). Here, the tibial nerve had the most trans-eQTLs (113), whereas the thyroid produced the most cis-eQTLs (415). The reduction in test numbers allowed 387 new trans-eQTLs and 327 cis-eQTLs to move below the significance threshold relative to the unrestricted multivariate test.

Inferred covariates such as PEER factors are known to capture and thus inadvertently control for broad regulatory effects that may have a true genetic basis, potentially removing broad trans-eQTL signals (Rakitsch & Stegle, 2016; GTEx Consortium, 2017). To test whether factors in our model find these kinds of regulatory hotspots, we ran the same eQTL mapping as before except excluding all PEER factors from the covariate matrix. This resulted in fewer total cis- and trans-eQTLs (2,456 and 882, respectively, at BH FDR < 0.1; Table S7). However, the proportion of unique eVariants to trans-eQTLs versus including PEER factors was lower (0.91 versus 0.98), suggesting that, to some extent, PEER factors do remove trans-acting pleiotropic signals that are captured by our model. In line with their supposed mechanisms of action, 69.1% (1,698) of these cis-eQTLs overlapped with our prior analysis controlling for PEER factors, whereas only 2.3% (20) of these trans-eQTLs were also found when controlling for PEER factors in the association analysis.

Next, we explored the overlap of our eQTLs and the GTEx consortium cis- and trans-eQTL list, produced by the consortium through an exhaustive tissue-specific testing approach (GTEx Consortium, 2020). A majority (4,645, 98.3%) of the multivariate TBLDA cis-eQTLs were in the GTEx cis-eQTL list. Of the multivariate trans-eQTLs, just one overlapped with the 2,629 genome-wide GTEx trans-eQTLs in the top 10 tissues. However, 26/438 (5.9%) of the GTEx trans-eQTLs in the relevant tissue, skin (sun-exposed), included that common eGene (*ALDH3B2*). Although we fail to capture this extended signal because we use an LD-pruned SNP set, our model still groups the gene together with its genomic hotspot. Furthermore, although the eVariant is not shared, 37/439 (8.4%) of the GTEx skeletal muscle trans-eQTLs include *RP11-65J3.3*, which we identify as a trans-eGene in that tissue. Taken together, these results suggest that our approach finds overlapping cis-eQTL signals but expands our ability to identify broad-acting trans-eQTLs in these bulk data.

One interesting example from the model fit using all GTEx samples is the trans eVariant rs4297160, which is associated with both *MAPRE3* (P -value $P \leq 6.6 \times 10^{-11}$) and *ARFGEF3* (P -value $P \leq 2.2 \times 10^{-16}$) in the thyroid and sits in the 9q22 locus (Fig 5). Specifically, rs4297160 is located within the lincRNA gene *PTCSC2*, which has been linked to a predisposition for papillary thyroid cancer (He et al, 2015). The 9q22 locus houses the thyroid-specific TF *FOXO1*, which shares a bidirectional promoter with *PTCSC2* (Wang et al, 2017).

Furthermore, the 9q22 locus was previously found associated in trans with *ARFGEF3* in the thyroid (GTEx Consortium, 2017). Notably, PEER factors were shown to capture and therefore control for broad regulatory signals from that locus (GTEx Consortium, 2017); in line with this, in association tests from the model trained on all GTEx data and run without PEER factors as covariates, rs4297160 was a trans-eVariant for 34 different genes in thyroid, including *HECW1* (regulates the degradation of thyroid transcription factor 1 [Liu et al, 2019]), *COLGALT2* (down-regulated in patients with thyroid orbitopathy [Khong et al, 2015]), and *FMO5* (expressed in endocrine cells that produce hormones that regulate metabolism [Xu, 2017]; Fig 5). These 34 trans-eGenes are enriched in the HIF-1 signaling KEGG pathway and two SP1 TF motifs, lending support for transcriptional co-regulation (g:Profiler [Raudvere et al, 2019] adjusted P -value < 1.1×10^{-2} for all).

We evaluated the increase in statistical power compared with the univariate approach because of our reduced multiple testing burden. The cis-eQTL P -values with BH FDR < 0.1 from our method have a different distribution from the GTEx cis-eQTLs found via exhaustive search (Kolmogorov-Smirnov test statistic 0.11, P -value $P \leq 1.0 \times 10^{-16}$). Our cis-eQTLs found via TBLDA are a subset of all true associations, and they tend to have more moderate associations than the set of GTEx cis-eQTLs (Fig S4). We expect this because TBLDA factors identify associated groups eQTLs, each of which may only have a small univariate effect size. Moreover, because we are computing P -values on a small held-out sample, we cannot achieve the statistical significance for the same association test as that test applied to a larger sample.

Discussion

In this paper, we present a probabilistic telescoping bimodal latent Dirichlet allocation (TBLDA) model that uncovers shared latent factors between bulk RNA-seq expression and genotype data when there is no one-to-one mapping among the samples for each data modality. The model takes raw counts as input, which avoids any potential data skewing because of normalization. We fit the model in an unsupervised manner using gene expression data from 10 tissues in the GTEx v8 release and matched donor genotypes. We intentionally exclude hard-coding tissue labels into the TBLDA because different tissues have a range of overlapping cell types, meaning that samples from certain tissues will share varying proportions of underlying processes and eQTLs. In addition, in large data sets, some samples may be mislabeled. Using known GTEx covariates, we established that the recovered topics reflected meaningful biology such as sample cell type proportion (Fig S2). Robust gene sets in tissue-associated factors were enriched for functionally relevant pathways (Table S1). Causal eVariants identified by our method are known to be enriched in a variety of genomic regulatory regions (Albert & Kruglyak, 2015); top-ranked robust tissue-associated SNP sets in our model were likewise enriched, demonstrating motifs of known eVariants (Table S2).

Running linear association tests on top-ranked features from each factor using MatrixEQTL (Shabalin, 2012) on a held-out test set,

we found 4,645 cis-eQTLs and 995 trans-eQTLs at BH FDR 0.1. By restricting association tests to the top features per factor in our model, we decrease the multiple testing burden and increase power for mapping trans-eQTLs on a small held-out test set. This is demonstrated by the fact that 994 of our trans-eQTLs were not identified in the exhaustive genome-wide GTEx analysis. A critical caveat of our approach is that, with a finite number of topics, we do not expect the model to capture all true eQTLs; however, we show that it does reproducibly identify novel and functionally relevant eQTLs. Taken together, these results demonstrate that our method successfully learns biologically meaningful shared topics across gene expression and genotype data. TBLDA is a natural framework to investigate cell type-specific eQTLs using single-cell RNA sequencing data, and we are currently exploring this promising future direction.

There are several potential points of contention in our model. First, although the model's probabilistic nature provides important measures of uncertainty for noisy genomic data, because of our inference procedure, the posterior should be interpreted with caution because variational inference is known to underestimate the posterior variance (Giordano & Broderick, 2015 Preprint). Second, because we do not include a private subspace for gene expression, true latent components that reflect expression-specific variation such as batch effects will be forced to contribute to the modality-shared factors. We believe this is important to retain signal for broad regulatory effects that especially affect trans-eQTL discovery. Nevertheless, if the model is used in a context such as single-cell RNA sequencing, where there are known and strong expression-specific covariates such as batch effects, this design choice should be reconsidered. Furthermore, a natural question that arises for all parametric latent factor models is how to determine the number of topics. We stress that there is no "correct" topic number and the user will want to make a reasonable trade-off between computational speed for inference and the granularity of signal captured. In practice, we recommend anywhere from 20 to 150 factors depending on the size of the data set. Given these qualities, natural extensions to the model include adding latent or semi-supervised expression-specific topics and extending it to a nonparametric framework.

Materials and Methods

Given a genotype matrix and an RNA sequencing expression matrix, our goal is to find latent factors that capture groups of SNPs and genes that covary across samples. We have two input matrices: a RNA-seq count matrix $X \in \mathbb{R}^{G \times L}$ for G genes across L samples and a genotype matrix $Y \in \mathbb{R}^{S \times N}$ in minor allele dosage format (0,1,2) for S SNPs across N individuals. We henceforth refer to genes and SNPs as features. Each individual $i \in 1, \dots, N$ contributes at least one sample, and every sample $\ell \in 1, \dots, L$ comes from exactly one known individual; this is the telescoping property of the data.

We define K latent topics where (i) each sample ℓ has topic membership proportion $\phi_\ell \in S^K$ and (ii) each individual i has topic membership proportion $\theta_i \in S^K$, where S^K denotes the K -dimensional simplex where all values are positive and sum to 1 (Figs 6 and

S5). Topics are modeled as distributions over features, where, similar to LDA, gene expression topics $\psi_k \in S^G$ with $k = 1, \dots, K$ are located on the simplex (Blei et al, 2003). The expression portion of the model that describes gene probabilities π_ℓ is:

$$\psi_k \sim \text{Dirichlet}(\xi) \quad (1)$$

$$\phi_\ell \sim \text{Dirichlet}(\sigma) \quad (2)$$

$$\pi_\ell = \Psi \phi_\ell \quad (3)$$

$$\mathbf{x}_\ell \sim \text{Multinomial}(C_\ell, \pi_\ell) \quad (4)$$

where C_ℓ is the observed total read count in sample ℓ and the matrix Ψ comprises the concatenated column vectors ψ_k with $k = 1, \dots, K$.

The sample to individual mapping is encoded in $\omega_i \in \{0,1\}^L$, an indicator vector for individual i , where $\omega_{i\ell} = 1$ if sample ℓ originates from individual i . The "telescoping" portion of the model projects the shared factors in Φ (the matrix formed by concatenating all ϕ_ℓ , $\ell = 1, \dots, L$) between the sample and individual spaces:

$$\theta_i = \frac{1}{\sum_{\ell=1}^L \omega_{i\ell}} \Phi \omega_i \quad (5)$$

In contrast to the expression topics ψ_k , genotype topics are modeled independently over SNPs and consist of independent λ_{jk} as in the structure model (Pritchard et al, 2000):

$$\lambda_{jk} \sim \text{Beta}(\zeta_j, \gamma_j) \quad (6)$$

We include a modality-specific (private) subspace for genotype to control for ancestral structure in mixed-population samples, which consists of nonnegative factor and loadings matrices β and τ (see the Extended Methods section) given Q ancestry factors. The weight of the private versus shared genotype subspaces is determined by $0 \leq \alpha \leq 1$ and learned during inference. The model does not include a gene expression-specific latent space to avoid losing any broad regulatory signal that is genotype-dependent (Rakitsch & Stegle, 2016), as is often the case with transeQTLs (GTEx Consortium, 2017). This final portion of the model, which covers minor allele probabilities ρ_{ij} , is as follows:

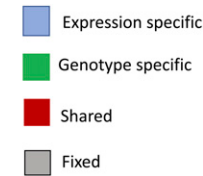
$$\alpha \sim \text{Uniform}(\delta, \mu) \quad (7)$$

$$\rho_{ij} = \alpha \sum_{q=1}^Q \beta_{jq} \tau_{qi} + (1 - \alpha) \left(\sum_{k=1}^K \lambda_{jk} \theta_{ki} \right) \quad (8)$$

$$y_{ij} \sim \text{Binomial}(2, \rho_{ij}) \quad (9)$$

We use stochastic variational inference to compute posterior estimates for Φ , Λ , Ψ , and α (see the Extended Methods section for details).

In this multimodal version of LDA, the same latent factors are shared across data modalities, allowing features of each modality to be directly linked together. Critically, because our framework directly models count data, we avoid spurious or distorted signals



Explicit representation of the model described in Materials and Methods section Equations (1)–(9) with dimensions drawn out. Each portion of the model is color-coded according to modality. Gray represents the known mapping between samples and individuals. The ancestry portion is striped because it is learned before fitting the shared model portion.

Features that have higher weights within a topic ($\psi_{r, \lambda_{jk}}$) have a larger relative contribution. However, the proportion of total counts for each gene varies widely. Genes with higher counts may dominate certain topics merely because of their high expression levels, overshadowing lower expressed genes that are actually more informative for that topic compared with others. Consequently, instead of using the raw expected loadings, we determine the importance of each feature across topics by ranking the average 2-Wasserstein distance between the posterior variational distributions. This allows us to control for both average feature counts and varying uncertainty in model estimation by using the full information provided in the posterior estimates. In particular, the 2-Wasserstein value, also known as the “Earth-movers distance,” is specifically designed as a distance metric between two empirical densities, unlike KL divergence, which is not a symmetric measure and often performs poorly when the two densities are defined on unequal basis measures. SNP minor allele frequency (MAF) is much less variable than gene total counts. To control for allele counts, we rank SNPs after regressing out the coded MAF in each loading (see the Extended Methods section for details).

Feature selection

After [Jo et al \(2016\)](#) Preprint, we used plink 1.9 ([Purcell et al, 2007](#)) to trim the GTEx v8 whole-genome sequencing SNP sets such that no two SNPs within a 200-Kb window have a Pearson correlation ≥ 0.2 . SNPs with imputed genotypes were removed, yielding 202,111 remaining SNPs across 831 individuals. All gencode v26 autosomal lincRNA and protein-coding genes from the 5,781 samples with genotypes were considered. We retained the 19,534 autosomal genes with a median RNAseq v1.1.9 ([Graubert et al. 2021](#)) read

- (i) σ : This is the hyperparameter for the sample-topic proportion ϕ_r . This vector should be kept symmetrical, and we recommend running the model with an uninformative prior where $\sigma = 1$. If the user wants to fit a model where samples comprise many topics, they should running the model with a more concentrated prior, setting $\sigma > 1$. Conversely, if the user wants each sample to be drawn from only a few topics, they can set a sparser prior where $\sigma < 1$.
- (ii) ζ, γ : These are the hyperparameters for the SNP loadings λ_{jr} . Although we set them equal to one for an uninformative prior, users could also set them to be less than one for a sparsity-inducing prior. In that case, a common choice would be setting the hyperparameters to $1/K$, where K is the number of topics. We recommend trying an uninformative prior first to let the

data lead, and if topic collapse is observed, next try the sparsity-inducing prior.

- (iii) ξ : This is the hyperparameter for the gene loadings ϕ_k . We follow the same guidance as for ζ and γ , except that ξ is the parameter vector for the Dirichlet, which generalizes the beta.
- (iv) μ, δ : These are the hyperparameters for α , which controls the mixture proportion for the genotype-specific versus shared space. The model will naturally put more weight on the separate genotype portion, and we want to restrict this so that TBLDA learns more shared structure. Thus, we want to restrain α from getting too close to zero or one. In practice, we recommend setting $\mu \geq 0.1$ and $\delta \leq 0.8$.

Running TBLDA on samples with additional technical covariates As a general practice for downstream model analysis, we recommend that users identify factors associated with all covariates (e.g., following the described methodology for tissue and cell type enrichment score associations). For example, when given technical covariates such as batch, users may remove all batch-associated factors from downstream analysis.

Feature ranking

After regressing out allele frequency, we take the top 10% of SNPs from each loading with the highest absolute value residuals. The 10% of genes from each loading with the highest 2-Wasserstein distances are considered the top gene features. Because the feature numbers vary by chromosome, runs have differing numbers of top features associated with their factors.

Functional enrichment data

The tissue-specific TF list originated from Table S3 in [Sonawane et al \(2017\)](#). To conduct GSEA, we used all biological process terms from GO v6.2 that had at least three genes in common with our analysis feature set. We used LDSC's baselineLD v2.1 ([Bulik-Sullivan et al, 2015](#)) genome annotations to compute SNP set enrichments. We did not consider MAF bin classes.

Tissue-associated genes

For each tissue, the set of robust tissue-associated genes consists of the genes that are top-ranked in at least one tissue-associated factor across all TBLDA runs.

Tissue-associated SNPs

For each tissue, we calculated the 75th percentile of the distribution of total tissue-associated factors across all runs that each top-ranked SNP is associated with. SNPs that are top-ranked in at least the 75th percentile of each tissue's associated factors across all runs comprise the set of robust tissue-associated SNPs.

eQTL pipeline

We held out two randomly selected samples from all individuals who contributed four or more samples to use for the eQTL pipeline and used all remaining samples to fit TBLDA. We used MatrixEQTL v2.3 ([Shabalov, 2012](#)) with modelLINEAR to run the eQTL testing. Expression for all genes that passed a 0.8 mappability filter was quantile-normalized as input. Sex, PCR, platform, the top five genotype principal components, and the top 60 PEER factors per

tissue were included as covariates. FDR was computed using the Benjamini–Hochberg procedure over each run for protein-coding and lincRNA genes separately. We note that the 0.1 FDR we use is more lenient than the 0.05 FDR threshold used in the v6p GTEx paper ([GTEx Consortium, 2020](#)), although the v8 GTEx trans-eQTLs we compare to were also identified using a 0.1 FDR threshold.

Robust components

We computed the correlation of each factor loading with all other loadings from runs on the same chromosome. Any factor with more than two loading Kendall $\tau > 0.15$ for SNPs and three Pearson $r^2 > 0.95$ for genes was flagged—along with the highly correlated factors—as a robust component. For each robust component, we averaged the constituent loadings to produce a representative factor loading. All components whose representative loadings exceeded $r^2 > 0.95$ were further collapsed into a single robust component.

Data access

All raw sequencing and genotype data from GTEx v8 used in this study can be found in dbGaP under accession number [phs000424.v8.p2](#).

Supplementary Information

Supplementary information is available at <https://doi.org/10.26508/lsa.202101297>.

Acknowledgements

ADH Gewirtz, FW Townes, and BE Engelhardt were funded by Helmsley Trust grant AWD1006624, NIH NCI 5U2CCA233195, NIH NHLBI R01 HL133218, and NSF CAREER AWD1005627.

Author Contributions

ADH Gewirtz: conceptualization, data curation, software, formal analysis, validation, investigation, visualization, methodology, and writing—original draft, review, and editing.

FW Townes: conceptualization, supervision, and methodology.

BE Engelhardt: conceptualization, resources, supervision, funding acquisition, methodology, project administration, and writing—original draft, review, and editing.

Conflict of Interest Statement

BE Engelhardt is on the SAB of Creyon Bio, ArrePath, and Freenome.

References

Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16: 197–212. doi:[10.1038/nrg3891](https://doi.org/10.1038/nrg3891)

- Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni J, Buettner F, Huber W, Stegle O (2018) Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 14: e8124. doi:[10.15252/msb.20178124](https://doi.org/10.15252/msb.20178124)
- Ash JT, Darnell G, Munro D, Engelhardt BE (2021) Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nat Commun* 12: 1609. doi:[10.1038/s41467-021-21727-x](https://doi.org/10.1038/s41467-021-21727-x)
- Bingham E, Chen J, Jankowiak M, Obermeyer F, Pradhan N, Karaletsos T, Singh R, Szerlip P, Horsfall P, Goodman N (2019) Pyro: Deep universal probabilistic programming. *J Mach Learn Res* 20: 28–1286. <https://jmlr.csail.mit.edu/papers/volume20/18-403/18-403.pdf>
- Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. *J Machine Learn Res* 3: 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson N, Daly MJ, Price AL, Neale BM, (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47: 291–295. doi:[10.1038/ng.3211](https://doi.org/10.1038/ng.3211)
- Cao H, Xu D, Cai Y, Han X, Tang L, Gao F, Qi Y, Cai D, Wang H, Ri M, et al (2021) Very long intergenic non-coding (vlinc) rnas directly regulate multiple genes in cis and trans. *BMC Biol* 19: 108. doi:[10.1186/s12915-021-01044-x](https://doi.org/10.1186/s12915-021-01044-x)
- Dey KK, Hsiao CJ, Stephens M (2017) Visualizing the structure of rna-seq expression data using grade of membership models. *PLoS Genet* 13: e1006599. doi:[10.1371/journal.pgen.1006599](https://doi.org/10.1371/journal.pgen.1006599)
- Giordano R, Broderick T (2015) Covariance matrices and influence scores for mean field variational bayes. *arXiv*. doi:[10.48550/arXiv.1502.07685](https://doi.org/10.48550/arXiv.1502.07685) (Preprint posted February 26, 2015)
- Gopalan P, Hao W, Blei DM, Storey JD (2016) Scaling probabilistic models of genetic variation to millions of humans. *Nat Genet* 48: 1587–1590. doi:[10.1038/ng.3710](https://doi.org/10.1038/ng.3710)
- Grassi L, Izuogu OG, Jorge NA, Seyres D, Bustamante M, Burden F, Farrow S, Farahi N, Martin FJ, Frankish A, et al (2021) Cell type-specific novel long non-coding RNA and circular RNA in the BLUEPRINT hematopoietic transcriptomes atlas. *Haematologica* 106: 2613–2623. doi:[10.3324/haematol.2019.238147](https://doi.org/10.3324/haematol.2019.238147)
- Graubert A, Aguet F, Ravi A, Ardlie KG, Getz G (2021) Rna-seq 2: Efficient rna-seq quality control and quantification for large cohorts. *Bioinformatics* 37: 3048–3050. doi:[10.1093/bioinformatics/btab135](https://doi.org/10.1093/bioinformatics/btab135)
- GTEX Consortium (2017) Genetic effects on gene expression across human tissues. *Nature* 550: 204–213. doi:[10.1038/nature24277](https://doi.org/10.1038/nature24277)
- GTEX Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369: 1318–1330. doi:[10.1126/science.aaz1776](https://doi.org/10.1126/science.aaz1776)
- Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, McClanahan P, Hendrickson DG, Sauvageau M, Kelley DR, et al (2014) Topological organization of multichromosomal regions by the long intergenic noncoding rna firre. *Nat Struct Mol Biol* 21: 198–206. doi:[10.1038/nsmb.2764](https://doi.org/10.1038/nsmb.2764)
- He H, Li W, Liyanarachchi S, Jendrzewski J, Srinivas M, Davuluri RV, Nagy R, de la Chapelle A (2015) Genetic predisposition to papillary thyroid carcinoma: Involvement of foxe1, tshr, and a novel lincrna gene, ptcsc2. *J Clin Endocrinol Metab* 100: 164–172. doi:[10.1210/jc.2014-2147](https://doi.org/10.1210/jc.2014-2147)
- Hicks SC, Townes FW, Teng M, Irizarry RA (2018) Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics* 19: 562–578. doi:[10.1093/biostatistics/kxx053](https://doi.org/10.1093/biostatistics/kxx053)
- Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, Finstermeier K, Stahringer A, Wilfert W, Beutner F, et al (2013) Alu elements in anril non-coding rna at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet* 9: 1003588. doi:[10.1371/journal.pgen.1003588](https://doi.org/10.1371/journal.pgen.1003588)
- Hore V, Viñuela A, Buil A, Knight J, McCarthy MI, Small K, Marchini J (2016) Tensor decomposition for multi-tissue gene expression experiments. *Nat Genet* 48: 1094–1100. doi:[10.1038/ng.3624](https://doi.org/10.1038/ng.3624)
- Jo B, He Y, Strober BJ, Parsana P, Aguet F, Brown AA, Castel SE, Gamazon ER, Gewirtz A, Gliner G, et al (2016) Distant regulatory effects of genetic variation in multiple human tissues. *BioRxiv*. doi:[10.1101/074419](https://doi.org/10.1101/074419) (Preprint posted September 9, 2016).
- Khong J, Wang L, Smyth G, McNab A, Hardy T, Selva D, Llamas B, Jung C, Sharma S, Burdon K, et al (2015) Differential gene expression profiling of orbital adipose tissue in thyroid orbitopathy. *Invest Ophthalmol Vis Sci* 56: 6438–6447. doi:[10.1167/iov.15-17185](https://doi.org/10.1167/iov.15-17185)
- Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788–791. doi:[10.1038/44565](https://doi.org/10.1038/44565)
- Li G, Gaynanova I (2018) A general framework for association analysis of heterogeneous data. *Ann Appl Stat* 12: 1700–1726. doi:[10.1214/17-aos1127](https://doi.org/10.1214/17-aos1127)
- Liu J, Dong S, Li L, Wang H, Zhao J, Zhao Y (2019) The e3 ubiquitin ligase hecw1 targets thyroid transcription factor 1 (ttf1/nkx2.1) for its degradation in the ubiquitin-proteasome system. *Cell Signal* 58: 91–98. doi:[10.1186/s13059-016-0932-1](https://doi.org/10.1186/s13059-016-0932-1)
- Liu SJ, Nowakowski TJ, Pollen AA, Lui JH, Horlbeck MA, Attenello FJ, He D, Weissman JS, Kriegstein AR, Diaz AA, et al (2016) Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* 17: 1–17. doi:[10.1186/s13059-016-0932-1](https://doi.org/10.1186/s13059-016-0932-1)
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol* 15: 550. doi:[10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8)
- Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, et al (2006) Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet* 2: 172. doi:[10.1371/journal.pgen.0020172](https://doi.org/10.1371/journal.pgen.0020172)
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959. doi:[10.1093/genetics/155.2.945](https://doi.org/10.1093/genetics/155.2.945)
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al (2007) Plink: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575. doi:[10.1086/519795](https://doi.org/10.1086/519795)
- Rakitsch B, Stegle O (2016) Modelling local gene networks increases power to detect trans-acting genetic effects on gene expression. *Genome Biol* 17: 33. doi:[10.1186/s13059-016-0895-2](https://doi.org/10.1186/s13059-016-0895-2)
- Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res* 47: 191–198. doi:[10.1093/nar/gkz369](https://doi.org/10.1093/nar/gkz369)
- Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140. doi:[10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616)
- Savage RS, Ghahramani Z, Griffin JE, de la Cruz BJ, Wild DL (2010) Discovering transcriptional modules by bayesian data integration. *Bioinformatics* 26: 158–167. doi:[10.1093/bioinformatics/btq210](https://doi.org/10.1093/bioinformatics/btq210)
- Shabalin AA (2012) Matrix eqtl: Ultra fast eqtl analysis via large matrix operations. *Bioinformatics* 28: 1353–1358. doi:[10.1093/bioinformatics/bts163](https://doi.org/10.1093/bioinformatics/bts163)
- Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, DeMeo DL, Quackenbush J, Glass K, Kuijjer ML (2017) Understanding tissue-specific gene regulation. *Cell Rep* 21: 1077–1088. doi:[10.1016/j.celrep.2017.10.001](https://doi.org/10.1016/j.celrep.2017.10.001)
- Townes FW, Engelhardt BE (2021) Nonnegative spatial factorization. *arXiv*. doi:[10.48550/arXiv.2110.06122](https://doi.org/10.48550/arXiv.2110.06122) (Preprint posted October 12, 2021)
- Vösa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, et al (2021) Large-scale cis-and trans-

eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 53: 1300–1310. doi:[10.1038/s41588-021-00913-z](https://doi.org/10.1038/s41588-021-00913-z)

Virtanen S, Jia Y, Klami A, Darrell T (2012) Factorized multi-modal topic model. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. UAI'12*. AUAI Press: 843–851.

Wang Y, He H, Li W, Phay J, Shen R, Yu L, Hancioglu B, de la Chapelle A (2017) Myh9 binds to lncrna gene ptcsc2 and regulates foxe1 in the 9q22 thyroid cancer risk locus. *Proc Natl Acad Sci U S A* 114: 474–479. doi:[10.1073/pnas.1619917114](https://doi.org/10.1073/pnas.1619917114)

Xu Z (2017) Flavin-containing monooxygenase-3 and 5: Tissue distribution, age-related expression and regulation by endoplasmic reticulum

stress. *Master's Thesis, Department of Biomedical and Pharmaceutical Sciences, University of Rhode Island*.

Zhao S, Gao C, Mukherjee S, Engelhardt BE (2016) Bayesian group factor analysis with structured sparsity. *J Machine Learn Res* 17: 6868–6914. <https://alkesgroup.broadinstitute.org/LDSCORE/baselineLD.v2.1.annots/> <https://jmlr.org/papers/volume17/14-472/14-472.pdf>



License: This article is available under a Creative Commons License (Attribution 4.0 International, as described at <https://creativecommons.org/licenses/by/4.0/>).