PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering

Chuan Gao[1], **Ian C. McDowell**[2], **Shiwen Zhao**[2], **Christopher D. Brown**[3], **Barbara E. Engelhardt**[4]*

1 Department of Statistical Science, Duke University, Durham, North Carolina, United States of America, **2** Program in Computational Biology and Bioinformatics, Duke University, Durham, North Carolina, United States of America, **3** Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **4** Department of Computer Science, Center for Statistics and Machine Learning, Princeton University, Princeton, New Jersey, United States of America

* bee@princeton.edu

Abstract

Identifying latent structure in high-dimensional genomic data is essential for exploring biological processes. Here, we consider recovering gene co-expression networks from gene expression data, where each network encodes relationships between genes that are co-regulated by shared biological mechanisms. To do this, we develop a Bayesian statistical model for *biclustering* to infer subsets of co-regulated genes that covary in all of the samples or in only a subset of the samples. Our biclustering method, *BicMix*, allows overcomplete representations of the data, computational tractability, and joint modeling of unknown confounders and biological signals. Compared with related biclustering methods, BicMix recovers latent structure with higher precision across diverse simulation scenarios as compared to state-of-the-art biclustering methods. Further, we develop a principled method to recover context specific gene co-expression networks from the estimated sparse biclustering matrices. We apply BicMix to breast cancer gene expression data and to gene expression data from a cardiovascular study cohort, and we recover gene co-expression networks that are differential across ER+ and ER- samples and across male and female samples. We apply BicMix to the Genotype-Tissue Expression (GTEx) pilot data, and we find tissue specific gene networks. We validate these findings by using our tissue specific networks to identify trans-eQTLs specific to one of four primary tissues.

Author Summary

Recovering gene co-expression networks from high-throughput experiments to measure gene expression levels is essential for understanding the genetic regulation of complex traits. It is often assumed for simplicity that gene co-expression networks are static across different contexts—e.g., drug exposure, genotype, tissue, age, and sex. The biological reality is that, along with differences in gene expression levels, there are differences in gene

OPEN ACCESS

Citation: Gao C, McDowell IC, Zhao S, Brown CD, Engelhardt BE (2016) Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering. PLoS Comput Biol 12(7): e1004791. doi:10.1371/journal.pcbi.1004791

Editor: Xianghong Jasmine Zhou, University of Southern California, UNITED STATES

Received: May 14, 2015

Accepted: February 3, 2016

Published: July 28, 2016

Data Availability Statement: The gene expression data from the breast cancer study are available from Bioconductor as BreastCancerNKI. The Cardiovascular and Pharmacogenetics (CAP) study gene expression data were acquired through GEO GSE36868. The Genotype-Tissue Expression (GTEx) pilot data are available through dbGaP, accession phs000424.v4. All code is publicly available at the Engelhardt Group Website: http://beehive.cs.princeton.edu/software/

Funding: BEE, CG and SZ were funded by NIH R00 HG006265 and NIH R01 MH101822. CDB was funded by NIH R01 MH101822. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

interactions across contexts. In this work, we describe a model for Bayesian biclustering, or recovering non-disjoint clusters of co-expressed genes in subsets of samples using gene expression level data. Using results from our biclustering model, we build gene co-expression networks jointly across all genes by computing the full regularized covariance matrix between all pairs of genes instead of testing each possible edge separately. Because biclustering recovers structure in subsets of the samples, we are able to recover gene co-expression networks that occur across all samples, that are differential across contexts (e.g., up-regulated in males and down-regulated in females), and that are unique to a context (e.g., only co-expressed in lung tissue). We illustrate the robustness of our approach and biologically validate the networks recovered from three different gene expression data sets.

Introduction

Cellular mechanisms tightly regulate gene transcription. Gene transcription is not independently regulated across genes: many of the mechanisms regulating transcription affect multiple genes simultaneously. Functional *gene modules* consist of subsets of genes that share similar expression patterns and perform coordinated cellular functions [1, 2]. This cluster-based description of gene expression fails to capture the informative co-expression patterns among genes within a gene module.

If we consider each gene as a vertex in a network, then pairs of genes within a gene module for which the correlation in expression levels cannot be explained by other genes may be connected by an undirected edge. Across all genes, these pairwise relationships constitute gene co-expression networks. Constructing these undirected gene networks, as compared to clustering genes into gene modules [3–6], provides rich detail about pairwise gene relationships. An even richer structure capturing these pairwise relationships would be a directed network of genes, but currently directed networks are computationally intractable to construct relative to undirected gene networks [7–10]. Our work describes a rigorous approach to recover undirected gene co-expression networks from gene expression data that uses a probabilistic latent factor model to quantify the relationships between all pairs of genes.

Several methods have been proposed to construct gene co-expression networks by partitioning a set of genes (and, in some cases, samples) into gene modules from which an undirected graph is elicited [11–14]. In most cases, gene partitioning creates disjoint sets of genes, implying that genes only participate in a single gene module. Biologically this assumption does not hold; the impact is that the gene networks built from methods that assume disjoint clusters are modular. These approaches are not probabilistic, and thus uncertainty in the network edges is not well characterized.

Alternatively, statistical latent factor models are often used to identify groups of co-regulated genes in gene expression data [15–18]. In particular, latent factor models decompose a matrix $\mathbf{Y} \in \Re^{p \times n}$ of p genes and n samples into the product of two matrices, $\mathbf{\Lambda} \in \Re^{p \times K}$, the factor loadings, and $\mathbf{X} \in \Re^{K \times n}$, the latent factor matrix with K latent factors, assuming independent Gaussian noise. Because it is costly to obtain and assay genome-wide gene expression levels in a single sample, most expression studies include observations of many more genes p than samples n [19, 20]. This so-called $p \gg n$ scenario limits our ability to find latent structure in this expansive, underconstrained space. High dimensional data suggests the use of strong regularization on the latent space to provide sufficient structure for the optimization to reach a robust solution. For example, we may regularize a latent space to discourage all but a few genes from contributing to a latent factor through a sparsity-inducing prior or penalty on the loading vectors [15, 21, 22]. Non-disjoint clusters of genes can be extracted from the resulting fitted sparse loading matrix by recovering all genes with non-zero loadings on the same factor [16]. Sparse latent factor models are more interpretable than their non-sparse counterparts because of this clustering effect.

Besides encouraging sparsity in the factor loading matrix, which results in non-disjoint clusters of genes that co-vary across all samples, one can also induce sparsity in the factor matrix, which results in non-disjoint subsets of samples within which subsets of genes uniquely exhibit co-variation. Statistically, this corresponds to regularizing both factor and loading matrices using priors that encourage zero-valued elements. Biologically, such a model recovers components that identify small numbers of correlated genes, where the correlation among the genes is exclusive to, for example, female samples. This statistical model encodes a general framework known as *biclustering* [23–35]. A biclustering model decomposes a matrix into clusters that each correspond to a subset of samples and a subset of features that exhibit latent structure unique to those subsets.

Gene expression levels have been shown to be sensitive to a number of environmental, biological, and technical covariates including experimental batch, sex, ethnicity, smoking status, or sample tissue heterogeneity [36, 37]. Methods to adjust the observation matrix to control the effects of these covariates without eliminating signals of interest have been proposed. Most attempts have been limited to correcting for confounding effects in a two-stage approach [20, 38, 39] or controlling for confounding effects jointly with association testing [40, 41]. The two-stage approach applied to estimates of co-expression networks has not been successful: often variation in expression levels of groups of co-expressed genes are captured in the estimates of confounding effects and controlled in the first stage, leading to false negatives [42].

In this paper, we develop a Bayesian statistical model for biclustering called *BicMix*. Our motivation behind developing this method was to identify large numbers of subsets of co-regulated genes capturing as many sources of gene transcription variation as possible within arbitrary subsets of the samples. Our biclustering model also includes non-sparse components to represent sources of transcription variation that affect all genes or all samples, which includes many types of confounding effects. We developed a simple but principled statistical method to reconstruct gene co-expression networks based on the regularized covariance matrices estimated using our biclustering model. This method recovers different types of gene co-expression networks, categorized by quantifying the contribution of each sample to the latent components: i) ubiquitous co-expression networks, ii) co-expression networks specific to a sample context, and iii) networks with differentially co-expressed genes across a sample context.

In this paper, we motivate and describe our Bayesian model for biclustering, BicMix. We validate our biclustering model on extensive simulations and compare biclustering with a number of state-of-the-art methods. We then apply our model to gene expression data without correcting for known or unknown confounders. In particular, we apply our biclustering model to gene expression levels measured in heterogeneous breast cancer tissue samples to recover a co-expression network that is differentially expressed across estrogen receptor positive and negative (ER+ and ER-) samples [43, 44]. Next, we apply our biclustering model to gene expression levels measured in lymphoblastoid cell lines (LCLs) from a cohort of patients in a cardiovascular disease study to identify co-expression networks with differential co-expression across males and females, and across smoking status [45]. Finally, we apply BicMix to the Genotype-Tissue Expression (GTEx) pilot data to elucidate tissue specific gene co-expression networks [46]. We validate the recovered networks by identifying tissue specific trans-eQTLs using the recovered tissue specific co-expression networks.

Results

Bayesian biclustering using BicMix

Biclustering was first introduced to detect clusters of states and years that showed similar voting patterns among Republicans in national elections [47] and was later referred to as *biclustering* in the context of identifying co-expressed genes in subsets of samples [23]. Biclustering has also been referred to as two mode clustering [48], subspace clustering [49, 50], or co-clustering [51] in various applied contexts. Biclustering was used successfully to explore latent sparse structure in different applied domains [52], including gene expression data [23, 24, 53–56], neuroscience [57], time series data [54], and recommendation systems [58]. Refer to comprehensive biclustering reviews for details [27, 59].

Biclustering approaches fall into four general categories. The first category assumes that each observed gene expression level for one sample is a linear combination of a mean effect, a row (gene) effect, and a column (sample) effect, some of which may be zero [23]. One approach in this category, *Plaid*, captures gene expression levels as a sum of many sparse submatrix components, where each submatrix includes non-zero values only for a subset of genes and subset of samples [30, 31]. The second category of biclustering methods uses hierarchical clustering to group together similar samples and features [3]. For example, samples may be clustered by considering some measure of feature similarity [24, 25, 28, 32, 35]. The third category of biclustering methods builds up biclusters by iteratively grouping features in a greedy way—e.g., identifying all genes that have correlated expression levels with a selected gene—and then removing samples that do not support that grouping [26]. The last category of biclustering methods uses Bayesian sparse factor analysis models [33]. These models decompose a gene expression matrix into two sparse matrices. Sparsity-inducing priors, such as the Laplace prior, are imposed on elements of both the loading and the factor matrices to induce zero-valued elements. K biclusters are recovered as the non-zero feature and sample components for each of the K latent components. Our approach falls into this last category of a sparse latent factor model for biclustering.

BicMix: Bayesian biclustering model. We developed a Bayesian biclustering model, *BicMix*, built on factor analysis with sparsity-inducing priors on both of the low dimensional matrices. In particular, we defined the following latent factor model for matrix $\mathbf{Y} \in \Re^{p\times n}$, which is the set of observations of p gene expression levels across n samples:

$$\mathbf{Y} = \mathbf{\Lambda X} + \epsilon \tag{1}$$

where $\mathbf{\Lambda} \in \Re^{p\times K}$ is the *loading matrix*, $\mathbf{X} \in \Re^{K\times n}$ is the *factor matrix*, $\epsilon \in \Re^{p\times n}$ is the residual error matrix, and K is fixed a priori. We assume that the residual error is independent across genes and samples and has a zero-mean multivariate Gaussian distribution with gene specific variance: $\epsilon_{\cdot,i} \sim \mathcal{N}_p(0, \Psi)$ for $i = 1, \ldots, n$, where $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$. While a value must be specified for K, the number of latent factors, this model removes factors that are unsupported in the data through a global sparsity-inducing prior, so K should be set as an overestimate of the number of latent factors (see Methods) [60].

To induce sparsity in both the factors (samples) and the loadings (genes), we used the three parameter beta ($\mathcal{TPB}$) distribution [61], which has been shown to be computationally efficient and to induce flexible modeling behavior. In previous work [60, 62], we included three layers of regularization via the $\mathcal{TPB}$ distribution to induce sparsity in the loading matrix. Here, we extended this model to include this same sparsity-inducing prior on the factor matrix (see Methods). With a flexible sparsity-inducing prior on both the factor and loading matrices, the model becomes a biclustering model, estimating subsets of genes for which co-variation is observed in a subset of samples. This structured prior produces favorable behavior in this

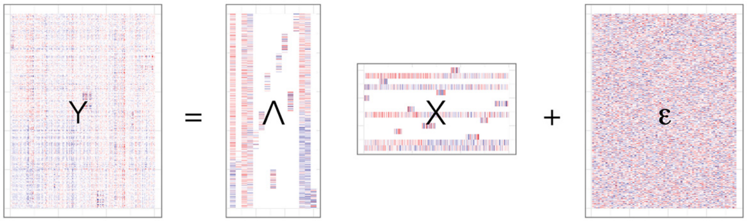

Fig 1. Schematic representation of the BicMix biclustering model. Ordered from left to right are, respectively, the $p \times n$ gene expression matrix **Y**, the $p \times K$ loading matrix Λ including both sparse and dense columns, the $K \times n$ factor matrix **X** including both sparse and dense rows, and the $p \times n$ residual matrix ϵ. Blue, red, and white entries in each matrix correspond to negative, positive, and zero values, respectively.

doi:10.1371/journal.pcbi.1004791.g001

biclustering model: i) the number of factors and loadings are effectively estimated from the data, because the sparsity-inducing prior removes unused factors; ii) each factor and corresponding loading has a different level of shrinkage applied to it, enabling a non-uniform level of sparsity and corresponding percentage of variance explained (PVE) for each factor and loading pair [60]; iii) neither the clusters of genes nor the clusters of samples are disjoint, so all genes and all samples may be in any number of clusters, or none; and iv) strong regularization allows overcomplete estimates of the response matrix, with possibly more factors K than samples n or genes p.

In gene expression data, observed covariates or unobserved confounding effects may systematically influence variation in the observation [38, 40, 41]. As in prior work, we tailored our sparsity-inducing prior to jointly model these often dense confounding effects [60]. In particular, we adapted our model so that the loadings and factors are drawn from a two-component mixture distribution, where each vector is either *sparse*—with zero elements—or *dense*—with no zero elements or all zero elements (Fig 1). We extracted information about whether a vector is sparse or dense directly from the fitted model parameters using the posterior mean of the mixture component assignment variables for each component $k = 1, \ldots, K$, where $z_k \in \{0, 1\}$ indicates a dense or a sparse loading and $o_k \in \{0, 1\}$ indicates a dense or a sparse factor. This two-component mixture in the prior distribution for the factors and loadings adds two favorable behaviors to the biclustering model. First, it jointly models covariates that regulate variation in most genes and also in few genes; we have found that confounding effects are often captured in the dense components as all samples and most genes are affected (e.g., batch effects, population structure) [37, 60]. Second, the mixture component has the effect of relaxing a computationally intractable space, enabling scalable parameter estimation in a Bayesian framework. Specifically, considering all possible subsets of genes and samples to identify biclusters is an intractably difficult problem; however, it is computationally tractable to first search over the space for which cluster membership is relaxed, represented as a continuous value between 0 and 1. Then we identify clusters by iteratively shrinking the small magnitude membership values to zero, which maps the continuous values to the binary representation of cluster membership. We estimated parameters in this model using both Markov chain Monte Carlo (MCMC) and a variational expectation-maximization (VEM) approach (see Methods).

Gene co-expression networks from biclusters. To construct an undirected gene network, we built a Gaussian Markov random field, or a Gaussian graphical model (GGM) [63], using the components estimated with our biclustering model (see Methods). In particular, regularized estimates of the gene-by-gene covariance matrix Ω may be computed from our parameter estimates as $\Omega = \mathbf{\Lambda}\, \Sigma\, \mathbf{\Lambda}^{\mathbf{T}} + \Psi$, where Σ is the covariance matrix for **X**. Factor analysis is often viewed as a method for low-rank covariance estimation by marginalizing over the factors, **X**. Furthermore, for any subset of components with sparse loading vectors, $A \subseteq \{1, \ldots, K\}$,

$\mathbf{\Omega_A} = \mathbf{\Lambda_A \Sigma_{A,A} \Lambda_A^T} + \Psi$, where $\Sigma_{A,A}$ is the covariance matrix for $\mathbf{X}_A$, estimates a regularized covariance matrix for the genes loaded on $\mathbf{\Lambda}_A$. Note that $\mathbf{\Omega}_A$ is generally both sparse and full rank; biclustering is a highly structured approach to estimating regularized covariance matrices [64]. The inverted covariance matrix is a symmetric precision matrix $\mathbf{\Delta_A} = \mathbf{\Omega_A^{-1}}$, where each element $\delta_{j,j'}$ can be transformed into the *partial correlation* between genes j and j',

$$cor(x_{j,\cdot}, x_{j',\cdot} \mid x_{\neg(j,j'),\cdot}) = -\frac{\delta_{j,j'}}{\sqrt{\delta_{j,j}\delta_{j',j'}}},$$

where $x_{\neg(j,\, j'),\cdot}$ indicates all genes in X that are neither j nor j'.

In a GGM, edges are defined as pairs of nodes for which the partial correlation is non-zero. Since each loading Λ_k, $k \in A$, specifies a cluster of genes, we do not invert the full covariance matrix, but instead invert the submatrix that corresponds to genes with non-zero loadings in those components. This approach avoids inducing non-zero precision estimates—and, correspondingly, edges in the GGM—between genes that never occur in the same bicluster. We used GeneNet [63] to test the precision matrix for significant edges. GeneNet assumes that the edges are drawn from a mixture of the null (no edge) and alternative (edge) distributions, $f(E) = \eta_0 f_0(E) + \eta_1 f_1(E)$, to calculate the probability of each edge being present or not. Practically, we selected edges with a probability > 0.8.

To recover co-variance networks specific to a subset of the samples, where the subset is characterized by context status, we chose the subset of components that contributes to this covariance matrix carefully. In particular, when we select subset A to include only components that have non-zero factor values for samples in a specific context (e.g., only female samples), we identify context specific covariance. When we select A such that all samples have a non-zero contribution to a component, we recover ubiquitous components. When we select A such that the mean rank of the factor values for one sample context is different than the mean rank of factor values for a different sample context—evaluated using a Wilcoxon signed-rank test—we identify components that are differential across the two contexts.

To combine results across runs of biclustering, we extracted co-expression network edges using this procedure separately for each run. Then we used ideas from an ensemble method, bootstrap aggregation (*bagging*) [65], and counted the number of times each edge for a specific network type was recovered across the runs. The final co-expression network contained edges that were identified in $\geq r$ runs.

Simulation results across biclustering methods

We validated our biclustering model using simulated data sets and compared results from five state-of-the-art biclustering methods. We simulated data from an alternative generative model for observation matrix $\mathbf{Y} = \mathbf{\Lambda}\, \mathbf{X} + \epsilon$, where $\mathbf{Y}$ has dimension $p = 500$ by $n = 300$ and $\epsilon_{i,j} \sim \mathcal{N}(0, \nu^{-1})$. Within this model, we simulated sparsity as follows: for each loading and factor, a number $m \in [5, 20]$ of elements were randomly selected and assigned values drawn from $\mathcal{N}(0, 2)$; the remaining elements were set to zero. We allowed components to share as many as five elements. Simulation 1 (Sim1) had ten sparse components. Simulation 2 (Sim2) had ten sparse components and five dense components, for which the loadings and factors were drawn from a $\mathcal{N}(0, 2)$ distribution. The components were shuffled so that a sparse loading may correspond to a dense factor, and vice versa. For both simulations we considered low and high noise scenarios: the residual variance parameter in the low noise (LN) setting was $\nu^{-1} = 1$ and, in the high noise (HN) setting, was $\nu^{-1} = 2$. Ten matrices Y were generated from each simulation scenario.

We ran BicMix and five other biclustering methods–Fabia [33], Plaid [30], CC [23], Bimax [27], and Spectral biclustering [34]. For all simulations, we ran BicMix by setting $a = b = 0.5$, $c = 1$, $d = 0.5$, $e = f = 1$ and $\nu = \xi = 1$ to promote substantial sparsity at the local level (i.e., the horseshoe prior [66]), weaker sparsity at the factor specific level (i.e., the Strawderman-Berger prior [67, 68]), and a uniform prior at the global level of the hierarchy [60]. The algorithm was initialized with warm start values by running MCMC for 500 iterations and using the final sample as the initial state for variational EM. For BicMix results, components that were classified as sparse have each element thresholded at 10^{-10}, because our parameter estimation methods converged to values near, but not exactly, zero. All other methods were run using their recommended settings (see Methods). For Sim2, we corrected the simulated data for the dense components by controlling for five principal components (PCs) before all other methods were run; without this initial correction for dense components, results from all five other biclustering methods were uninterpretable. For all runs, BicMix was initialized with $K = 50$ latent factors; all other methods were initialized with the correct number of sparse factors $K = 10$. For Fabia, we ran the software in two different ways. The results from running Fabia with the recommended settings are denoted as *Fabia*. We also set the sparsity threshold in Fabia to the number (from 100 quantiles of the uniform distribution over [0.1, 5]) that produced the closest match in the recovered matrices to the number of non-zero elements in the simulated data; we label these results *Fabia-truth*.

We used the *recovery and relevance score* (R&R score) [27] to measure the false discovery rate (FDR) and sensitivity of each method in recovering true biclusterings. Let the true set of sparse matrices be $\mathbf{M}_1$ and the estimated set of sparse matrices be $\mathbf{M}_2$; then the R&R score is calculated as:

$$\text{Rec} = \frac{1}{|M_1|} \sum_{b_1 \in M_1} \max_{b_2 \in M_2} \frac{b_1 \cap b_2}{b_1 \cup b_2}, \tag{2}$$

$$\text{Rel} = \frac{1}{|M_2|} \sum_{b_2 \in M_2} \max_{b_1 \in M_1} \frac{b_1 \cap b_2}{b_1 \cup b_2}. \tag{3}$$

Recovery quantifies the proportion of true clusters that are recovered (i.e., recall); *relevance* refers to the proportion of true clusters identified in the recovered clusters (i.e., precision). For BicMix, we applied this R&R score to the components for which both the loading and the factor vectors were sparse, which indicates a bicluster. For the doubly-sparse latent factor models, Fabia and BicMix, we also calculated a sparse stability index (SSI) [60] to compare the recovered and true matrices; SSI is invariant to label switching and scale, and falls between [0, 1] with 1 indicating perfect recovery.

For Sim1, we found that BicMix recovered the sparse loadings, sparse factors, and the biclusters well in the low noise scenario based on both R&R (Fig 2a) and SSI (Fig 2b). Fabia-truth had the second best performance based on R&R. For comparison, Fabia-truth achieved better R&R scores than Fabia (Fig 2a); the clustering results from BicMix dominated those from Fabia-truth, although there was only a small gain in relevance in the low noise Sim1 results for BicMix. Plaid showed high relevance for the recovered biclusters regardless of the noise level for Sim1, but at the expense of poor recovery scores. The remaining methods did not perform well in these simulations with respect to the R&R score for both low and high noise simulation scenarios.

For Sim2, BicMix correctly identified the sparse and dense components (Fig 2a), where a threshold of $\langle z_k \rangle > 0.9$ was used to determine when a loading k was dense. The performance of Fabia on Sim2 deteriorated substantially relative to its performance on Sim1, although the confounders were removed using principal components (PCs) and the correct number of factors

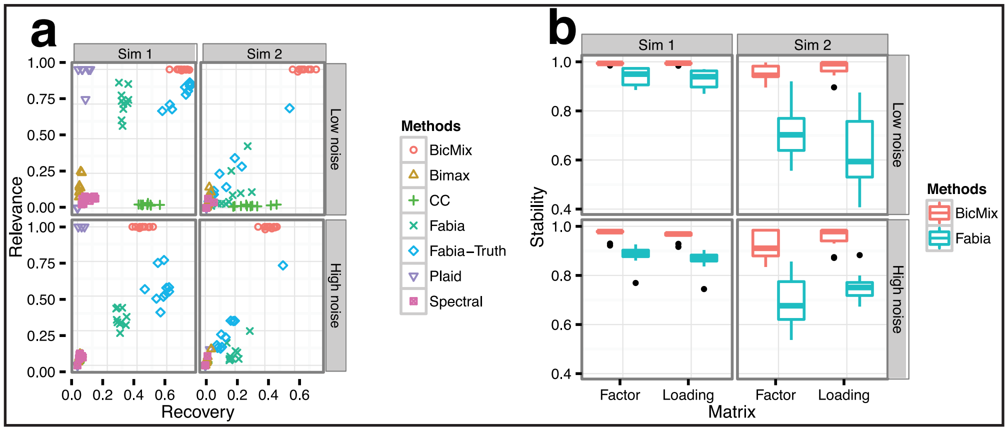

Fig 2. Comparison of BicMix with related methods. Top row: Simulation with low noise. Bottom row: Simulation with high noise. Left column: Sim1 with only sparse components. Right column: Sim2 with sparse and dense components. Panel a: Recovery score on the x-axis, relevance score on the y-axis for all methods in the legend. Panel b: Stability statistic (y-axis) for the sparse components recovered by BicMix and Fabia.

doi:10.1371/journal.pcbi.1004791.g002

was given. For both BicMix and Fabia, additional noise in the simulation made bicluster recovery more difficult, as shown in deterioration of the recovery score for both methods; however, unlike Fabia, the relevance score of the biclustering from BicMix was unaffected by additional noise in the Sim2 high noise scenario.

The other methods show inferior performance relative to BicMix and Fabia on Sim2. CC assumes that genes in each bicluster have constant expression values, which limits its ability to cluster genes with heterogeneous expression levels. Bimax assumes binary gene expression values (i.e., over- or under-expressed), which limits its utility for heterogeneous expression levels. Spectral biclustering imposes orthogonal constraints on the biclusters; this orthogonality assumption is violated in these simulations and also in gene expression data, where correlated sources of variation may impact similar subsets of genes.

Application of BicMix to three gene expression study data sets

We now turn to the application of BicMix to gene expression data from three studies. For each application, we first describe the biclustering results from BicMix. Then we show the interpretable networks that were recovered from the BicMix results and discuss network validation using cis- and trans-eQTLs.

Breast cancer network. We investigated a breast cancer data set that contains 24,158 genes assayed in 337 breast tumor samples [44, 69, 70] after removing genes that are >10% missing and imputing missing values of included genes [71] (see Methods). All patients in this data set had stage I or II breast cancer and were younger than 62 years old. Among the 337 patients, 193 had lymph-node negative disease and 144 had lymph-node positive disease; prognostic signatures such as *BRCA1* mutations, estrogen receptor status (ER), distant metastasis free survival (DMFS) were collected for all patients. We focused on building differential gene co-expression networks across ER positive (ER+) and ER negative (ER-) patients because of ER's prognostic value in profiling breast cancer patients [72]: cancer patients that are ER+ are more likely to respond to endocrine therapies than patients that are ER-. In these data, there are 249 ER+ and 88 ER- patients.

We ran *BicMix* on these data, setting $a = b = 0.5$, $c = 1$, $d = 0.5$, $e = f = 0.5$ and $\nu = \xi = 1$ as in the simulations; the initial number of components was set to $K = 300$. Starting from 1,000

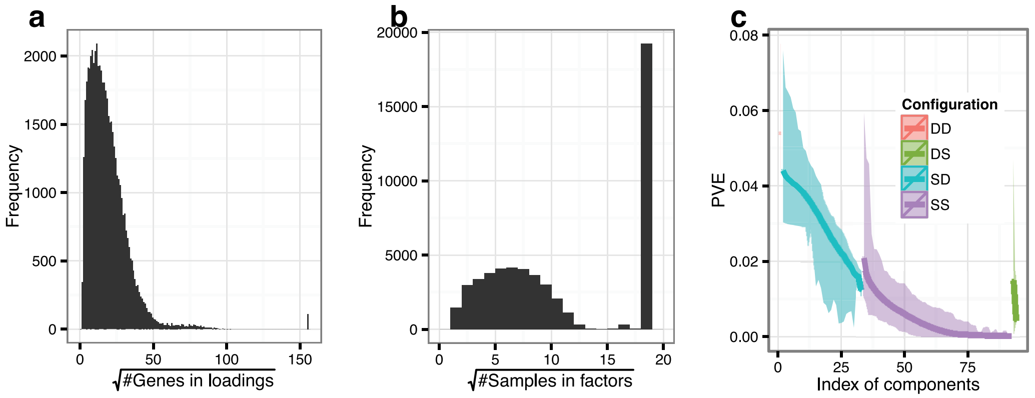

Fig 3. Distribution of the number of genes, the number of samples, and PVE in the breast cancer data. Panel a: Distribution of the number of genes with non-zero values in each of the 53,814 loadings. Panel b: Distribution of the number of samples with non-zero values in each of the 53,814 factors. Panel c: average PVE for the components sorted by PVE within each run. The middle lines show the median PVE, the ribbons show the range of the minimum and maximum PVE across 900 runs. For panels a and b, the peaks on the far right correspond to the number of the genes and samples for the dense loadings and dense factors.

doi:10.1371/journal.pcbi.1004791.g003

random values, BicMix was run until the total number of genes with non-zero loadings across components changed $\leq \frac{Kp}{100}$ over 100 iterations. Removing runs that did not converge, we recovered 53,814 components across 900 runs, of which 110 loadings and 19,239 factors were dense (Fig 3a and 3b).

The distribution of the number of genes in each sparse component was skewed to small numbers (Fig 3a). We categorized each component as one of four configurations: sparse gene loadings with sparse sample factors (SS), sparse gene loadings with dense sample factors (SD), dense gene loadings with sparse sample factors (DS), and dense gene loadings with dense sample factors (DD). SS components captured subsets of genes that are uniquely co-expressed in a subset of samples, which may be due to, e.g., a genotype or environmental context that impacts expression levels of a small number of genes. SD components capture subsets of genes that are differentially co-expressed among all samples, which may be due to, e.g., sex-differential expression or batch effects. DS components capture a subset of samples in which all genes have additional co-variation, which may be due to, e.g., sample contamination. DD components capture variation that occurs across all samples and affects co-variation across all genes, which may be due to, e.g., latent population structure.

For each run, we calculated the percentage of variance explained (PVE) per component k as $\frac{Tr\left(\Lambda_k \langle X_k X_k^T \rangle \Lambda_k^T\right)}{Tr\left(\mathbf{\Lambda} \langle \mathbf{X}\mathbf{X}^{\mathbf{T}} \rangle \mathbf{\Lambda}^{\mathbf{T}}\right)}$, where Tr denotes the trace operator. We ordered the components by PVE within each SS, SD, DS, and DD component category. We calculated the mean, maximum, and minimum values of the PVE-ordered, categorized components across the runs. Note that, because there are no orthogonality constraints, it is possible that many of these components explain similar variation in the observations; for this quantification we are assuming this PVE is disjoint and normalizing across all component-wise PVEs. We also calculated the total PVE explained by each component category by summing the total PVE for all components jointly in each category. The distribution of the number of genes contained in each loading and the PVE by component and across SD, SS, DS, DD categories show that SD and SS components made up the vast majority of recovered components, and that, on average, SD components accounted for a larger PVE than SS components (Fig 3a). The number of components that fell into the SS, SD, DS, DD categories accounted for 64.1%, 35.7%, 0.1%, 0.1%, respectively, of the total number of components. In the same order, components in the four categories accounted for 24.7%, 74.8%, 0.3% and 0.1% of the total PVE.

We selected components from the fitted BicMix model to identify ER+ and ER- specific gene co-expression networks (see Methods). Moreover, to recover gene co-expression networks that are differentially expressed across ER+ and ER- samples, we identified components corresponding to factors that had a significant difference in the mean rank of the factor value between the ER+ and ER- samples based on a Wilcoxon signed-rank test ($p \leq \frac{0.05}{53814} = 9.29 \times 10^{-7}$). Across our components, we found 996 components unique to ER+ samples, 135 components unique to ER- samples, and 17,051 components differential across ER+ and ER- samples. Interestingly, we note that differential ER status represents the bulk of the SD components recovered across all of the runs. We contrasted the correlation of all of the observed covariates with a representative sample of the SD factors (S1 Fig), and we found that ER status correlated well with many of these SD factors. Tumor grade also correlated well with many of the SD factors; however, tumor grade is somewhat anti-correlated with ER status (S2 Fig), so we chose to focus on ER status because of its clinical utility.

The precision matrices of the subsets of components corresponding to the three network types were constructed and edges among these genes were tested using our method for extracting gene co-expression networks from the fitted biclustering model (see Methods) [63]. For the ER- specific network, we recovered a total of 15 genes and 10 edges that are replicated ≥ 2 times; for the ER+ specific network, we recovered 621 genes and 760 edges that are replicated ≥ 2 times (S3 and S4 Figs; S1 and S2 Tables). We recovered more genes for ER+ than ER- specific networks because there are many more ER+ samples than ER- samples. We found one node and no edges shared across the ER+ and ER- specific networks.

For the co-expression network differential across ER+ and ER- samples, we recovered 432 genes and 8,156 edges that were replicated ≥ 15 times across the 900 runs (Fig 4; S3 Table). We hypothesized that, in these differential networks, the 432 genes may be divided into two subgroups: a group of genes that is up-regulated in the ER+ samples and down-regulated in the ER- samples, and a group of genes that is down-regulated in the ER+ samples and up-regulated

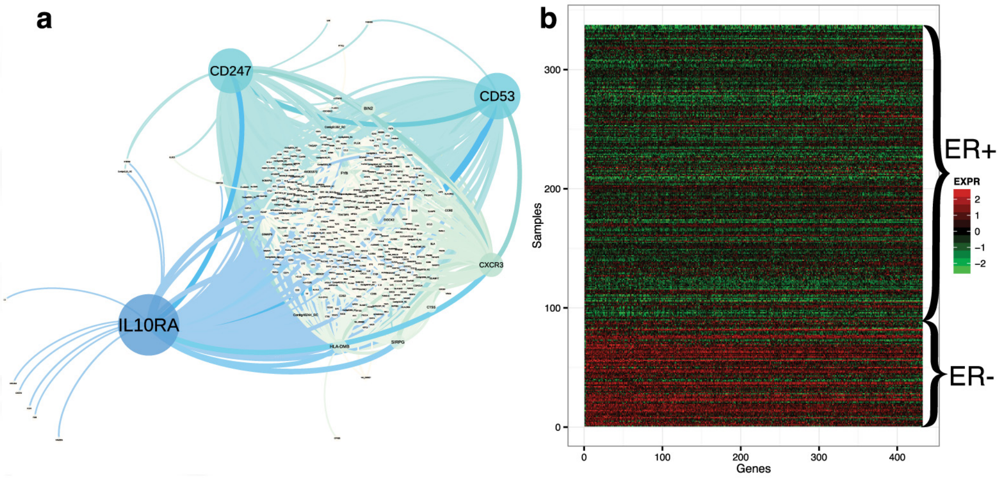

Fig 4. Differential ER-status gene co-expression network and gene expression for ER differential genes. Panel a: differential ER gene co-expression network, where node size corresponds to betweenness centrality, which quantifies the number of shortest paths between all pairs of nodes in the network in which the gene is included. Panel b: gene expression levels for 432 genes in the ER-status differential co-expression network.

doi:10.1371/journal.pcbi.1004791.g004

in the ER- samples. To explore this hypothesis, we quantified differential expression levels for the 432 genes in the differential gene co-expression network (Fig 4b). We found 430 genes that are up-regulated in ER- samples and down-regulated in ER+ samples. In contrast, we found two genes that are down-regulated in ER- samples and up-regulated in ER+ samples (Fig 4b, first two columns). This bias is likely due to the unbalanced number of ER+ and ER- samples. The genes in the ER-status specific co-expression networks also show a pattern of differential expression (S5 Fig), although not as strong, because gene correlation is differential across networks as opposed to coordinated differential expression levels across a subset of genes.

In the ER-status differential network, we found that many of the annotated genes play critical roles in breast cancer development. For genes that are up-regulated in ER- and down-regulated in ER+ patients:

- *CD53*, the gene that encodes the leukocyte surface antigen *CD53* protein, and *IL10RA*, the gene that encodes the receptor for interleukin 10, belong to a set of three genes (including *DMFS*) that are predictive of outcome specifically in ER tumors [73];
- *CD247* encodes the CD3 zeta chain protein; a previous study has shown that the down-regulation of CD3 zeta expression plays an important role in breast cancer progression [74, 75];
- *CXCR3* encodes the chemokine (*C-X3-C* motif) ligand protein; studies have shown that *CXCR3* deficiency speeds tumor progression [76], and that *CXCR3* isoforms have divergent roles in promoting cancer stem-like cell survival and metastasis [77].

For genes that are up-regulated in ER+ and down-regulated in ER- patients:

- *SFRP2* encodes the secreted frizzled-related protein, which inhibits the growth of triple-negative breast tumors [78];
- *COL12A1* encodes the *alpha chain of type XII collagen*, which has been found to be a regulatory target of *miR-26b* and is predictive of breast cancer recurrence [79].

Cardiovascular and Pharmacogenetic (CAP) RNA expression study data. We next applied our biclustering model to a gene expression study with 10,195 expressed genes measured in 480 human lymphoblastoid cell lines (LCLs) [45]. In these data, there are 221 female samples, 259 male samples, 64 smokers, and 416 non-smokers. Age and BMI are also available for each sample. The data were processed according to previous work [45, 60]; In particular, we removed genes with probes on the gene expression array that aligned to multiple regions of the genome using a BLAST analysis against human genome reference hg19, leaving 8,718 expressed genes. No known covariates nor PCs were controlled in these data before projecting each gene expression level to the quantiles of a standard normal distribution (S6 Fig).

We set $K = 400$ and ran EM from 1000 starting points with $a = b = 0.5$, $c = 1$, $d = 0.5$, $e = f = 1$, $\nu = 1$ and $\alpha = \beta = 1$ [60]. After removing runs that did not converge, we recovered a total of 865 runs. On average there were 93 factors across runs, of which on average 21 were dense factors and dense loadings (where both the loading and the factor mixture component posterior probability of being non-sparse ≥ 0.9).

Sex-differential networks. To construct a sex differential network, we selected factors that had differential mean values across sex. We recovered 68 genes corresponding to 78 edges that were replicated ≥ 5 times across runs (Fig 5a). The expected number of edges to replicate in this experiment under the null hypothesis is approximately 0.8 (Eq 89). Many of the genes that were identified to have differential co-expression partners across sex play important roles in sex determination or sex specific regulatory activities.

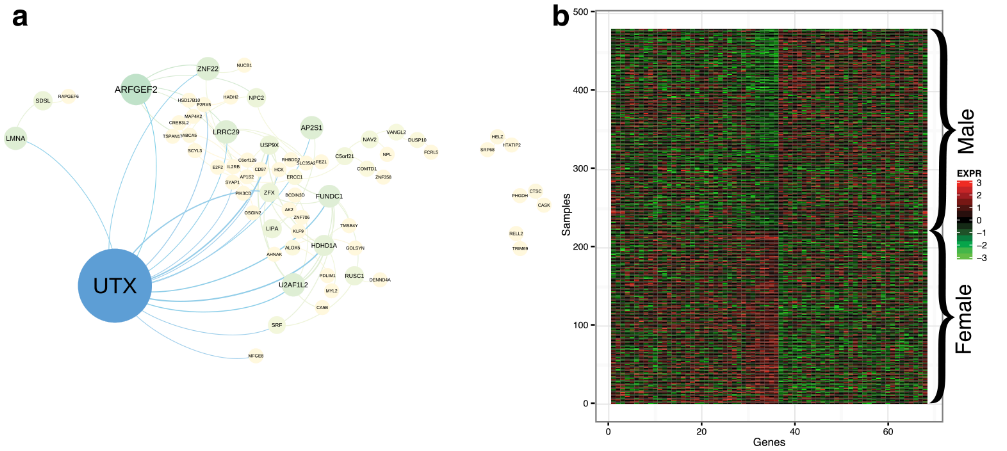

Fig 5. Sex differential gene co-expression network and gene expression levels for sex differential genes. Panel a: differential sex gene co-expression network, where node size corresponds to betweenness centrality. Panel b: gene expression levels for 61 genes in the sex differential gene co-expression network.

doi:10.1371/journal.pcbi.1004791.g005

- The hub gene, *UTX*, regulates somatic and germ cell epigenetic reprogramming [80]; more recently it has been shown to be a sex specific tumor suppressor in T-cell acute lymphoblastic leukemia [81];
- *ZFX*, a gene on the human X chromosome that is structurally similar to the Y chromosome gene *ZFY*, has been used for sex identification in many species [82];
- *USP9X*, an X-linked gene, is differentially expressed across sexes in the adult mouse brain, and is hypothesized to play a role in differential neural development in women with XO Turner syndrome [83].

A heatmap of the gene expression levels of these genes across male and female samples shows patterns of differential expression as in the genes highlighted in the breast cancer ER-status differential co-expression network (Fig 5b).

Sex specific and smoking status specific networks. To construct sex specific networks and smoking status specific networks, we identified components with non-zero factor values for either only male or only female samples, or only smokers or only non-smokers, respectively. Using these components, we recovered 57 genes with 176 edges specific to males and 38 genes with 80 edges specific to females, where both networks included edges that were recovered ≥ 10 times across runs (Fig 6). The expected number of edges to replicate in this experiment under the null hypothesis is approximately 2.3×10^{-7} (males) and 9.4×10^{-7} (females; Eq 89).

For the smoking status specific networks, we identified 31 genes with 36 edges specific to smokers that were recovered ≥ 5 times across runs (S7a Fig); we identified 294 genes with 2,272 edges specific to non-smokers that were recovered ≥ 10 times across runs (S7b Fig). The expected number of edges to replicate in this experiment under the null hypothesis is approximately 4.6×10^{-7} (smokers) and 1.3×10^{-6} (non-smokers; Eq 89). The decrease in power to detect smoker specific edges compared to non-smoker specific edges is likely due to the imbalanced number of smokers (64) and non-smokers (416) in this study.

Genotype-Tissue Expression (GTEx) study. We applied BicMix to a subset of the pilot Genotype-Tissue Expression (GTEx) project data [46]. Our goal for these data was to identify gene co-expression networks that were specific to a single tissue while controlling for networks that appeared across all tissues and other confounding effects. The subset of gene expression

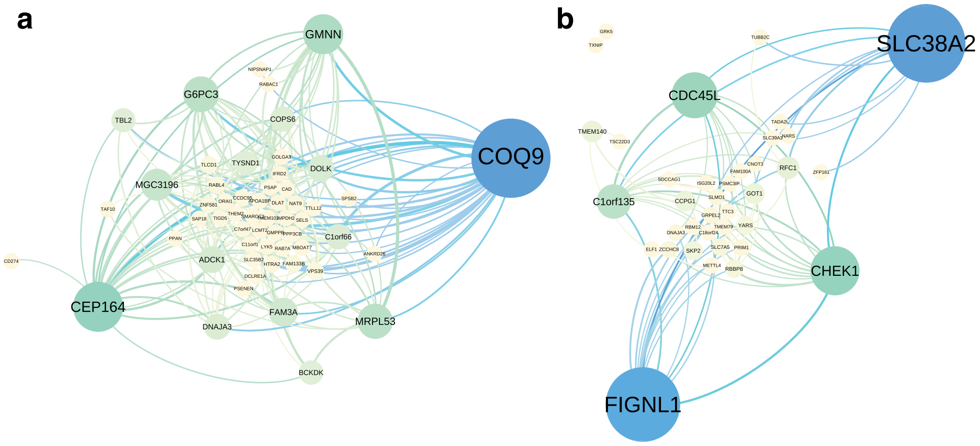

Fig 6. Sex specific gene co-expression networks in the CAP gene expression data. Panel a: Gene co-expression network specific to males. Panel b: Gene co-expression network specific to females. Node size and color correspond to betweenness centrality.

doi:10.1371/journal.pcbi.1004791.g006

data that we used contained $p = 20{,}850$ genes measured in $n = 446$ samples across four tissues: adipose ($n_f = 103$), artery ($n_a = 118$), lung ($n_l = 122$), and skin ($n_s = 106$). We preprocessed these RNA-sequencing data as described in Methods. We concatenated these data sample-wise to create a single response matrix $\mathbf{Y} \in \Re^{p \times (n_f + n_a + n_l + n_s)}$.

We set $K = 300$ and ran EM from 1000 starting points with $a = b = 0.5$, $c = 1$, $d = 0.5$, $e = f = 1$, $v = 1$ and $\alpha = \beta = 1$ [60]. After eliminating runs that did not converge, we used a total of 958 runs. In these runs, we recovered on average 119 factors, approximately eight of which were dense in both factors and loadings (S8 Fig).

Comparison with biclustering results from comparative methods. To determine whether or not the existing biclustering methods also recover similarly informative biclusters, we ran Fabia, CC, Bimax, Spectral, and Plaid on the GTEx data. To control for confounding effects, we removed the effects of five principal components of the gene expression matrix separately within each tissue so as to maintain the tissue specific effects for all methods except Fabia and BicMix. We found that both Plaid and Fabia were able to separate the tissues in the sample space, as was principal components analysis (PCA; S9 Fig). Starting from 300 biclusters, Plaid reduced the number of biclusters to four, with each bicluster containing samples from a subset of one tissue; adipose samples, however, were not identified uniquely (S10 Fig).

Results from Fabia depended on the specified number of biclusters: when the number of clusters is set to the number of tissues (four), the biclusters distinguished the four tissues (S11 Fig). However, with larger numbers of clusters specified, the separation of the samples by tissue became increasingly obfuscated until, with 20 clusters, there were no tissue specific patterns across clusters (S12 Fig). Compared to results from Plaid and Fabia, the other methods produced uninterpretable results: with 300 clusters, CC found one bicluster that contained all genes and one sample; Bimax failed to recover any biclusters in these data; Spectral found two biclusters, each with zero genes. BicMix, as described above, identified the four tissues in separate clusters as with Fabia (four factors) and PCA (S13 Fig).

Tissue specific networks. In the BicMix results, we identified all of the bicluster factors and loadings that appeared uniquely in one of the four tissues. We collected genes and edges that were recovered ≥ 10 times across 1000 runs. We recovered 152 genes with 379 edges for adipose tissue, 167 genes with 966 edges for artery tissue, 98 genes with 199 edges for lung tissue, and 70 genes with 171 edges for skin tissue (Fig 7; S4–S7 Tables). The expected number of

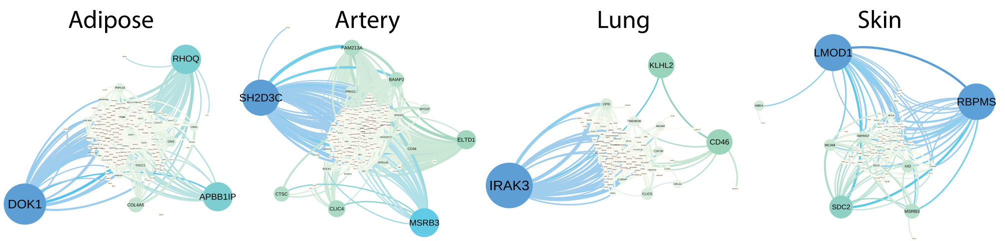

Fig 7. Tissue specific gene co-expression networks in the GTEx pilot data. Adipose: gene co-expression network for adipose. Artery: gene co-expression network for artery. Lung: gene co-expression network for lung. Skin: gene co-expression network for skin. Node size and color correspond to betweenness centrality.

doi:10.1371/journal.pcbi.1004791.g007

edges to replicate in this experiment under the null hypothesis is approximately 9.3×10^{-6} (adipose), -2.8×10^{-6} (artery), 8.2×10^{-6} (adipose), 1.7×10^{-6} (skin; Eq 89). Note that the negative expectation is due to the approximation in Eq (89).

To study the biological significance of the tissue specific gene clusters, we applied the functional gene annotation tool, DAVID (version 1.1) [84] to perform a gene ontology enrichment analysis on the recovered tissue specific gene clusters from a subset of 100 randomly chosen runs. Using a false discovery rate (FDR) threshold of 0.05, we found 262 (adipose), 265 (artery), 672 (lung), and 560 (skin) *molecular function* Gene Ontology (GO) categories enriched in clusters specific to each of the four tissues (S8–S11 Tables). In particular, the adipose specific gene clusters were most enriched for *lipid biosynthetic process* (FDR $\leq 2.8 \times 10^{-4}$), *lipid metabolism* (FDR $\leq 1.7 \times 10^{-2}$), and *insulin signaling pathway* (FDR ≤ 0.046); Interestingly, artery specific functions also appeared in adipose enriched terms; it has been shown that angiogenesis and vascular functions modulate obesity, adipose metabolism, and insulin sensitivity [85]. The artery specific gene clusters were most enriched for *vascular smooth muscle contraction* (FDR $\leq 9 \times 10^{-3}$), *circulatory system process* (FDR ≤ 0.03), and *ventricular cardiac muscle cell development* (FDR ≤ 0.049). The lung specific gene clusters were enriched for *respiratory burst* (FDR $\leq 5 \times 10^{-3}$) and *Toll-like-receptor* (FDR $\leq 1.8 \times 10^{-2}$), where *Toll-like-receptor* has a role in lung disease [86]. The skin specific gene clusters were most enriched for *melanogenesis* (FDR $\leq 1.9 \times 10^{-2}$), the process of formation of pigmentation, typically of the skin, and *epidermolysis bullosa (EB)* (FDR ≤ 0.03), an inherited connective tissue disease causing blisters in the skin and mucosal membranes. We also observed that many GO terms were enriched in multiple tissues, suggesting that a similar set of biological functions are achieved via different, cell specific mechanisms across tissues.

In the tissue specific networks, we found that the genes with larger betweenness centrality were known to play important roles in tissue specific functions. For adipose tissue, we found that the hub genes play important roles in adipose metabolism and show significant associations with obesity related traits. For example, *DOK1*, the gene that encodes *Docking Protein 1*, was shown to mediate high-fat diet-induced adipocyte hypertrophy and obesity through modulation of *PPAR*-gamma phosphorylation [87]. *APBB1IP*, which encodes the *Rap1-GTP*-interacting adaptor, has been shown to be involved in regulating metabolism and protecting against obesity [88]. *RHOQ*, also known as *TC10*, alters its gene expression levels in visceral adipose tissue of rats that are given a high-fat diet [89]. Genes with smaller betweenness centrality play important adipose specific functions as well; for example, gene expression levels of *PNPLA3* are affected by diet-induced obesity [90], and *LGPAT1* was shown to influence BMI and percent body fat in Native Americans [91].

In the artery specific network, we found that the genes with larger betweenness centrality play important roles in *angiogenesis* and *vasculogenesis*. For example, *ELTD1*, also known as epidermal growth factor (*EGF*), regulates angiogenesis [92]. *CLIC4*, a gene that encodes the chloride intracellular channel protein 4, regulates vasculogenesis through endothelial tube formation; abnormal *CLIC4* expression may play a role in pulmonary arterial hypertension pathology [93]. *CD36* encodes the thrombospondin receptor, and polymorphisms in this gene have significant associations with coronary artery heart disease risk [94]. *BAIAP2* encodes the brain specific angiogenesis inhibitor 1 *BAI1*, and is protective of angiogenesis [95].

In the lung specific network, the genes with larger betweenness centrality play important roles in maintaining the proper function of the lungs or are associated with lung and respiratory diseases. For example, *IRAK3* encodes a member of the interleukin-1 receptor-associated kinase protein family; hospital patients on mechanical ventilation with injury have differential expression of *TLR4* and *IRAK3* [96]. *CD46* encodes a regulatory protein for which higher expression in the lungs of ex-smokers appears to reduce inflammation and protect individuals from emphysema and chronic obstructive pulmonary disease [97, 98]. *IL18RAP* encodes the interleukin 18 receptor accessory protein. A recent study found that genetic variants in *IL1RL1* and *IL18R1* were significantly associated with bronchial hyperresponsiveness [99].

In the skin specific network, we found genes with larger betweenness centrality played critical biological roles in the skin, and were often related to skin tumor growth or carcinoma. For example, *RBPMS* encodes the RNA binding protein with multiple splicing; *RBPMS* is one of six genes that were shared among the top up-regulated genes both in dedifferentiated carcinoma and in carcinoma with loss of 13q [100]. *MCAM* encodes the melanoma cell adhesion molecule; increased expression of *MCAM* in human melanoma cells leads to increased tumor growth and metastasis [101, 102]. Finally, *MSRB* encodes the methionine sulfoxide reductase, which is important for antioxidant repair in human skin [103]. We note that it is possible in the current methodological framework to recover gene networks specific to subsets of the context specific samples; here it is possible that a subset of these skin samples include pre-cancerous cells, which will lead to recovering skin specific networks with edges corresponding to genes that are co-expressed only in pre-cancerous skin tissue.

Tissue specific trans-eQTLs. To validate the tissue specific networks identified in the GTEx data, we performed a trans-eQTL analysis on the genes in these networks as follows. An expression QTL (eQTL) is, in this context, a single nucleotide polymorphism (SNP) that regulates expression levels of a specific gene. Transcription regulation may happen in an allele specific way (cis-eQTL) or may be mediated by a non-allele specific process (trans-eQTL). Given a gene A, a SNP Q that has been identified as a cis-eQTL for gene A, and a target gene B that is a neighbor of A in the tissue specific network, we tested for association between Q and B in gene expression values derived from samples with matched primary tissue. Using a list of eQTLs identified in previous work [104], we performed a pairwise univariate regression analysis of these eQTLs informed by the tissue specific networks. The same univariate analysis was performed on the permuted gene expression data to find p values under the null distribution of no trans-association.

Using our adipose specific networks, we found 40 cis-eQTLs for genes *RHOQ* and *PARVA12*, which were trans-eQTLs for many different target genes in adipose tissue samples (FDR ≤ 0.2). Two of these genes with trans-eQTLs specific to adipose tissue were *TK2* and *EHD2*, which have been associated with the abnormal development of adipose tissues and adipokine levels in mice [105] and lead to specific changes in white adipose tissue of growth hormone receptor-null mice [106], respectively. In artery, we found 121 cis-eQTLs for nine genes that act as trans-eQTLs to multiple target genes. The artery specific trans-eQTL target genes include *CCM2L*, involved in controlling vascular stability and growth [107], and *ARHGEF15*, which affects retinal angiogenesis in endothelial cells [108]. In lung, we found 19 cis-eQTLs for seven genes that are trans-eQTLs for seven target genes including *IRAK3*, described above, and *SIGLEC5*, which has been shown regulate acute pulmonary neutrophil inflammation [109]. In skin samples, we found 13 cis-eQTLs targeting three genes that are trans-eQTLs for two target genes that are unique to skin samples. One of these skin specific trans-eQTLs affects *SLC11A1*, which has been associated with susceptibility to Buruli ulcers [110]. (Full list of tissue specific trans-eQTLs in S12–S15 Tables.)

Discussion

In this work, we developed a statistical approach to biclustering based on a Bayesian sparse latent factor model. We included a two-component mixture distribution to allow both sparse and dense representations of the features or samples, which captures heterogeneous sources of structured variation within the gene expression data. We used the regularized covariance matrix estimated from the latent factor model to build a Gaussian graphical model with the features represented as nodes in the undirected network. By extracting covariance matrices corresponding to subsets of components, we were able to identify gene co-expression networks that were shared across all samples, unique to a subset of samples, or differential across sample subsets.

We applied our methodology to breast tumor tissue gene expression samples and recovered co-expression networks that are differential across ER+ and ER- tumor types. We applied our methodology to gene expression data from the CAP project and recovered sex-differential, sex specific, and smoking status specific gene co-expression networks. We applied our methodology to the GTEx gene expression pilot data and recovered tissue specific networks for four tissues, which we validated by identifying tissue specific trans-eQTLs.

Factor analysis methods, including the biclustering approach presented here but extending to many other well-studied models, are statistical tools developed for exploratory analyses of the data. In this work, we have exploited the latent structure in both the factor and the loading matrix to estimate the covariance matrix that is specific to sample subsets. We considered tissue type and tumor types, but these methods can be used for any observed binary, categorical, integer, or continuous covariate (e.g., case-control status, batch, sex, age, EBV load). We showed in the Results that the recovered latent structure has substantial context specific biological meaning.

Our results on the GTEx data show that a number of genes were identified as part of multiple tissue specific networks. While individual genes may overlap across networks, the interactions of those genes did not. Genes that co-occurred in multiple tissue specific networks are good candidates to test for differential function across tissues. We also used this approach to study sexual dimorphism, extracting gene networks specific to one sex or differential across the sexes. We showed the potential of this approach to improve statistical power to identify sex specific trans-eQTLs.

In this version of BicMix, extracting a covariance matrix specific to a subset of the samples was performed *post hoc*: the linear projection to the latent space was performed in a mostly unsupervised way, although our three layer sparsity inducing priors add additional structure above SFA-type approaches. As described in the Results, there were multiple categories of gene-interactions that we recovered. These categories included: gene interactions that existed across context, gene interactions that were unique to specific contexts, and gene interactions that were present across contexts but differentially interact in different contexts. However, this approach currently does not use known context to directly inform the projection.

Indirectly, we saw that the sparsity structure on the samples allowed small subsets of the samples to inform projection, but this still relied on a post hoc interpretation of those sample subsets to recover specific network types. Correlation between contexts, such as tissue and batch, or smoking status and sex, would confound these results; here we checked for these correlation among observed covariates (S2 and S6 Figs) and also validated our results using literature searches and trans-eQTL recovery. Furthermore, it may be the case that, for a sample context of interest (e.g., age, sex), there is insufficient signal that is uniquely attributable to those samples (e.g., female) to identify a covariance matrix corresponding to the values of interest in this unsupervised framework. We are currently extending this approach so that the linear projection is explicitly informed by the context of interest.

Methods

Bayesian biclustering model for BicMix

We consider the following factor analysis model:

$$\mathbf{Y} = \mathbf{\Lambda X} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{Y} \in \Re^{p \times n}$ is the matrix of observed variables, $\mathbf{\Lambda} \in \Re^{p \times K}$ is the loading matrix, $\mathbf{X} \in \Re^{K \times n}$ is the factor matrix, and $\epsilon \in \Re^{p \times n}$ is the residual error matrix for p genes and n samples. We assumed $\epsilon_{\cdot,i} \sim \mathcal{N}(0, \Psi)$, where $\Psi = \text{diag}(\psi_1, \ldots, \psi_p)$.

In previous work [60, 62], a three parameter beta ($\mathcal{TPB}$) [61] prior was used to model the variance of $\mathbf{\Lambda}$. The three parameter distribution has the form

$$f(x : a, b, \phi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \phi^b x^{b-1} (1-x)^{a-1} \{1 + (\phi - 1)x\}^{-(a+b)}, \tag{5}$$

for $x \in (0, 1)$, $a > 0$, $b > 0$ and $\phi > 0$.

We used $\mathcal{TPB}$ to induce flexible shrinkage to both $\mathbf{\Lambda}$ and $\mathbf{X}$. Specifically, we included three layers of shrinkage—global, factor specific, and local—for both the factors and the loadings. Next we describe the sparsity-inducing structure for $\mathbf{\Lambda}$ and $\mathbf{X}$.

Hierarchical structure on $\mathbf{\Lambda}$. The hierarchical structure for $\mathbf{\Lambda}$ is written as

$$\varrho \sim \mathcal{TPB}(e, f, v), \tag{6}$$

$$\zeta_k \sim \mathcal{TPB}\left(c, d, \frac{1}{\varrho} - 1\right) \tag{7}$$

$$\varphi_{j,k} \sim \mathcal{TPB}\left(a, b, \frac{1}{\zeta_k} - 1\right), \tag{8}$$

$$\Lambda_{j,k} \sim \mathcal{N}\left(0, \frac{1}{\varphi_{j,k}} - 1\right). \tag{9}$$

We used the fact that

$$\varphi \sim \mathcal{TPB}(a, b, v) \Leftrightarrow \frac{\theta}{v} \sim Be'(a, b) \Leftrightarrow \theta \sim \mathcal{G}a(a, \delta) \text{ and } \delta \sim \mathcal{G}a(b, v), \tag{10}$$

where $Be'(a, b)$ and $\mathcal{G}a$ indicate an inverse beta and a gamma distribution, respectively. Making the substitution $\eta = \frac{1}{\varrho} - 1,\ \phi_k = \frac{1}{\zeta_k} - 1,\ \theta_{j,k} = \frac{1}{\varphi_{j,k}} - 1$, we get the equivalent hierarchical structure [61]:

$$\gamma \sim \mathcal{G}a(f, v), \tag{11}$$

$$\eta \sim \mathcal{G}a(e, \gamma), \tag{12}$$

$$\tau_k \sim \mathcal{G}a(d, \eta), \tag{13}$$

$$\phi_k \sim \mathcal{G}a(c, \tau_k), \tag{14}$$

$$\delta_{j,k} \sim \mathcal{G}a(b, \phi_k), \tag{15}$$

$$\theta_{j,k} \sim \mathcal{G}a(a, \delta_{j,k}), \tag{16}$$

$$\Lambda_{j,k} \sim \mathcal{N}(0, \theta_{j,k}). \tag{17}$$

We applied a two-component mixture model to jointly model possibly dense confounding effects by letting $\theta_{j,k}$ be generated from a mixture of sparse and dense components:

$$\theta_{j,k} \sim \pi \mathcal{G}a(a, \delta_{j,k}) + (1 - \pi)\delta(\phi_k), \tag{18}$$

where the hidden variable z_k, which indicates whether or not loading k is sparse (1) or dense (0), is generated from the following beta-Bernoulli distribution:

$$\pi | \alpha, \beta \sim Be(\alpha, \beta) \tag{19}$$

$$z_k | \pi \sim \text{Bern}(\pi), \ \ k = \{1, \cdots, K\}. \tag{20}$$

Hierarchical structure for X. The hierarchical structure inducing sparsity in **X**, which is structurally identical to that for Λ, is:

$$\varphi \sim \mathcal{G}a(f_X, \xi), \tag{21}$$

$$\chi \sim \mathcal{G}a(e_X, \varphi), \tag{22}$$

$$\kappa_k \sim \mathcal{G}a(d_X, \chi), \tag{23}$$

$$\omega_k \sim \mathcal{G}a(c_X, \kappa_k) \tag{24}$$

$$\rho_{k,i} \sim \mathcal{G}a(b_X, \omega_k), \tag{25}$$

$$\sigma_{k,i} \sim \pi \mathcal{G}a(a_X, \rho_{k,i}) + (1 - \pi)\delta(\omega_k) \tag{26}$$

$$x_{k,i} \sim \mathcal{N}(0, \sigma_{k,i}), \tag{27}$$

with $\sigma_{k,i}$ generated from a two component mixture. Here, the hidden variable o_k, which indicates whether or not the factor is sparse (1) or dense (0), has the following beta-Bernoulli distribution:

$$\pi_X|\alpha_X, \beta_X \sim Be(\alpha_X, \beta_X) \tag{28}$$

$$o_k|\pi_X \sim \text{Bern}(\pi_X), \;\; k = \{1, \cdots, K\}. \tag{29}$$

Variational expectation maximization

Because of the large dimension of matrices to which we apply BicMix, we used approximate methods for parameter estimation. In particular, we used variational expectation maximization (VEM) to estimate values for latent variables and parameters directly from the data in an approximate way. Extending previous work [60], the posterior probability $\mathcal{P} = p(\mathbf{\Lambda}, \mathbf{X}, \mathbf{z}, \mathbf{o}, \mathbf{\Theta}|\mathbf{Y})$ is written as:

$$\begin{aligned} \mathcal{P} &\propto p(\mathbf{Y}|\mathbf{\Lambda}, \mathbf{X})p(\mathbf{\Lambda}|\mathbf{z}, \mathbf{\Theta}_{\mathbf{\Lambda}})p(\mathbf{X}|\mathbf{o}, \mathbf{\Theta}_{\mathbf{X}})p(\mathbf{z}|\mathbf{\Theta}_{\mathbf{\Lambda}})p(\mathbf{o}|\mathbf{\Theta}_{\mathbf{X}})p(\mathbf{\Theta}_{\Lambda})p(\mathbf{\Theta}_{X}) \\ &\propto p(\mathbf{Y}|\mathbf{\Lambda}, \mathbf{X})\mathcal{P}(\mathbf{\Lambda})\mathcal{P}(\mathbf{X}), \end{aligned} \tag{30}$$

where we used $\mathbf{\Theta}_{\Lambda}$ and $\mathbf{\Theta}_X$ to denote the set of parameters related to $\mathbf{\Lambda}$ and $\mathbf{X}$, respectively. Then,

$$\begin{aligned} \mathcal{P}(\mathbf{\Lambda}) &= p(\mathbf{\Lambda}|\mathbf{z}, \mathbf{\Theta}_{\mathbf{\Lambda}})p(\mathbf{z}|\mathbf{\Theta}_{\mathbf{\Lambda}})p(\mathbf{\Theta}_{\Lambda}) \\ &= \left[\prod_{j=1}^{p}\prod_{k=1}^{K}\mathcal{N}(\Lambda_{j,k}|\theta_{j,k})\mathcal{G}a(\theta_{j,k}|a, \delta_{j,k})\mathcal{G}a(\delta_{j,k}|b, \phi_k)\right]^{\mathbb{1}_{z_k=1}} \\ &\times \left[\prod_{j=1}^{p}\prod_{k=1}^{K}\mathcal{N}(\Lambda_{j,k}|\phi_k)\right]^{\mathbb{1}_{z_k=0}} \left[\prod_{k=1}^{K}\mathcal{B}ern(z_k|\pi)\right]\mathcal{B}eta(\pi|\alpha, \beta) \\ &\times \left[\prod_{k=1}^{K}\mathcal{G}a(\phi_k|c, \tau_k)\mathcal{G}a(\tau_k|d, \eta)\right]\mathcal{G}a(\eta|e, \gamma)\mathcal{G}a(\gamma|f, \nu) \end{aligned} \tag{31}$$

and

$$\begin{aligned} \mathcal{P}(\mathbf{X}) &= p(\mathbf{X}|\mathbf{o}, \mathbf{\Theta}_{\mathbf{X}})p(\mathbf{o}|\mathbf{\Theta}_{\mathbf{X}})p(\mathbf{\Theta}_{X}) \\ &= \left[\prod_{k=1}^{K}\prod_{i=1}^{n}\mathcal{N}(x_{k,i}|\sigma_{k,i})\mathcal{G}a(\sigma_{k,i}|a_X, \rho_{k,i})\mathcal{G}a(\rho_{k,i}|b_X, \omega_k)\right]^{\mathbb{1}_{o_k=1}} \\ &\times \left[\prod_{k=1}^{K}\prod_{i=1}^{n}\mathcal{N}(x_{k,i}|\omega_k)\right]^{\mathbb{1}_{o_k=0}} \left[\prod_{k=1}^{K}\mathcal{B}ern(o_k|\pi_X)\right]\mathcal{B}eta(\pi_X|\alpha_X, \beta_X) \\ &\times \left[\prod_{k=1}^{K}\mathcal{G}a(\omega_k|c_X, \kappa_k)\mathcal{G}a(\kappa_k|d_X, \chi)\right]\mathcal{G}a(\chi|e_X, \varphi)\mathcal{G}a(\varphi|f_X, \xi) \end{aligned} \tag{32}$$

To derive the variational EM algorithm, we expanded the posterior probability (Eq (30)) and wrote the expected complete log likelihood for parameters related to

$\boldsymbol{\Lambda}$: $Q(\boldsymbol{\Theta}_\Lambda) = \langle \ell_c(\boldsymbol{\Theta}_\Lambda,\boldsymbol{\Lambda}|\mathbf{z},\mathbf{X},\mathbf{Y})\rangle$ as:

$$\begin{aligned}
Q(\boldsymbol{\Theta}_\Lambda) \quad & \propto \sum_{j=1}^{p}\sum_{i=1}^{n}\Big\langle \log p(y_{j,i}|\boldsymbol{\Lambda},\mathbf{X},\boldsymbol{\Theta}_\Lambda,\mathbf{z})\Big\rangle \\
& + \sum_{j=1}^{p}\sum_{k=1}^{K}\Big\langle p(z_k|\boldsymbol{\Theta}_\Lambda)\log p(\Lambda_{j,k}|\boldsymbol{\Theta}_\Lambda, z_k)\Big\rangle + \log p(\boldsymbol{\Theta}_\Lambda) \\
& \propto -\frac{p}{2}\ln|\boldsymbol{\Psi}| - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left(y_{j,i}-\sum_{k=1}^{K}\Lambda_{j,k}\langle x_{k,i}\rangle\right)^2}{2\psi_{j,j}} + \sum_{j=1}^{p}\sum_{k=1}^{K}\langle 1-z_k\rangle\left\{-\frac{1}{2}\ln\phi_k - \frac{\Lambda_{j,k}^2}{2\phi_k}\right\} \\
& + \sum_{j=1}^{p}\sum_{k=1}^{K}\langle z_k\rangle\left\{-\frac{1}{2}\ln\theta_{j,k} - \frac{\Lambda_{j,k}^2}{2\theta_{j,k}} + a\ln\delta_{j,k} + (a-1)\ln\theta_{j,k} - \delta_{j,k}\theta_{j,k}\right\} \\
& + \sum_{j=1}^{p}\sum_{k=1}^{K}\langle z_k\rangle\left\{b\ln\phi_k + (b-1)\ln\delta_{j,k} - \phi_k\delta_{j,k}\right\} + \sum_{k=1}^{K}\{\langle z_k\rangle\ln\pi + (1-\langle z_k\rangle)\ln(1-\pi)\} \\
& + \sum_{k=1}^{K}\{c\ln\tau_k + (c-1)\ln\phi_k - \tau_k\phi_k + d\ln\eta + (d-1)\ln\tau_k - \eta\tau_k\} \\
& + e\ln\gamma + (e-1)\ln\eta - \gamma\eta + f\ln v + (f-1)\ln\gamma - v\gamma + \alpha\ln\pi + \beta\ln(1-\pi),
\end{aligned} \tag{33}$$

where we used $\langle X\rangle$ to represent the expected value of X.

Similarly, the expected complete log likelihood for parameters related to $\mathbf{X}$ takes the following form:

$$\begin{aligned}
Q(\boldsymbol{\Theta}_X) \quad & \propto \sum_{j=1}^{p}\sum_{i=1}^{n}\Big\langle \log p(y_{j,i}|\boldsymbol{\Lambda},\mathbf{X},\boldsymbol{\Theta}_X,\mathbf{O})\Big\rangle \\
& + \sum_{k=1}^{K}\sum_{i=1}^{n}\Big\langle p(o_k|\boldsymbol{\Theta}_X)\log p(x_{k,i}|\boldsymbol{\Theta}_X, O_k)\Big\rangle + \log p(\boldsymbol{\Theta}_X) \\
& \propto -\frac{p}{2}\ln|\boldsymbol{\Psi}| - \sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\left(y_{j,i}-\sum_{k=1}^{K}\Lambda_{j,k}\langle x_{k,i}\rangle\right)^2}{2\psi_{j,j}} + \sum_{k=1}^{K}\sum_{i=1}^{n}\langle 1-o_k\rangle\left\{-\frac{1}{2}\ln\omega_k - \frac{\langle x_{k,i}^2\rangle}{2\omega_k}\right\} \\
& + \sum_{k=1}^{K}\sum_{i=1}^{n}\langle o_k\rangle\left\{-\frac{1}{2}\ln\sigma_{k,i} - \frac{\langle x_{k,i}^2\rangle}{2\sigma_{k,i}} + a_X\ln\rho_{k,i} + (a_X-1)\ln\sigma_{k,i} - \rho_{k,i}\sigma_{k,i}\right\} \\
& + \sum_{k=1}^{K}\sum_{i=1}^{n}\langle o_k\rangle\left\{b_X\ln\omega_k + (b_X-1)\ln\rho_{k,i} - \omega_k\rho_{k,i}\right\} + \sum_{k=1}^{K}\{\langle o_k\rangle\ln\pi_X + (1-\langle o_k\rangle)\ln(1-\pi_X)\} \\
& + \sum_{k=1}^{K}\{c_X\ln\kappa_k + (c_X-1)\ln\omega_k - \kappa_k\omega_k + d_X\ln\chi + (d_X-1)\ln\kappa_k - \chi\kappa_k\} \\
& + e_X\ln\varphi + (e_X-1)\ln\chi - \varphi\chi + f_X\ln\xi + (f_X-1)\ln\varphi - \xi\varphi + \alpha_X\ln\pi_X + \beta_X\ln(1-\pi_X).
\end{aligned} \tag{34}$$

To simplify the calculation, we assumed that the joint distribution $p(o_k,x_{k,i})$ factorizes as $p(o_k)p(x_{k,i})$, implying that corresponding factors' and loadings' sparsity statuses are independent.

We computed the MAP estimates for the parameters that encourage sparsity in the Λ matrix, $\hat{\boldsymbol{\Theta}}_\Lambda = argmax_{\boldsymbol{\Theta}_\Lambda} Q(\boldsymbol{\Theta}_\Lambda)$. Specifically, we solved equation $\frac{\partial Q(\boldsymbol{\Theta}_\Lambda)}{\partial \boldsymbol{\Theta}_\Lambda} = 0$ to find the closed form MAP estimates. The MAP estimate for the jth row of $\mathbf{\Lambda}$, $\Lambda_{j,\cdot}$, in matrix form, is:

$$\hat{\Lambda}_{j,\cdot} = Y_{j,\cdot}\psi_{j,j}^{-1}\mathbf{X}^T\left(\left\langle \mathbf{X}\psi_{j,j}^{-1}\mathbf{X}^T\right\rangle + \mathbf{Z}\mathbf{\Theta}_j^{-1} + (\mathbf{I} - \langle\mathbf{Z}\rangle)\mathbf{\Phi}^{-1}\right)^{-1}, \tag{35}$$

where

$$\mathbf{\Theta}_j = \begin{pmatrix} \theta_{j,1} & 0 & \cdots & 0 \\ 0 & \theta_{j,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{j,K} \end{pmatrix}, \quad \mathbf{\Phi} = \begin{pmatrix} \phi_1 & 0 & \cdots & 0 \\ 0 & \phi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \phi_K \end{pmatrix}, \tag{36}$$

and

$$\mathbf{Z} = \begin{pmatrix} \langle z_1\rangle & 0 & \cdots & 0 \\ 0 & \langle z_2\rangle & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle z_K\rangle \end{pmatrix} \tag{37}$$

and $\mathbf{I}$ is the identity matrix.

We computed the expected value of $\mathbf{X}$, $\langle\mathbf{X}\rangle$, which has the following form:

$$\left\langle \mathbf{X}_{\cdot,i}\right\rangle = (\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda} + \langle O\rangle\Sigma_i^{-1} + (\mathbf{I} - \langle O\rangle)\mathbf{\Omega}^{-1})^{-1}\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}Y_{\cdot,i}, \tag{38}$$

where

$$\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_{1,i} & 0 & \cdots & 0 \\ 0 & \sigma_{2,i} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{K,i} \end{pmatrix}, \quad \mathbf{\Omega} = \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_K \end{pmatrix}, \tag{39}$$

and

$$\mathbf{O} = \begin{pmatrix} \langle o_1\rangle & 0 & \cdots & 0 \\ 0 & \langle o_2\rangle & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle o_K\rangle. \end{pmatrix} \tag{40}$$

We computed the expected value of $\mathbf{X}\psi_{j,j}^{-1}\mathbf{X}^T$:

$$\left\langle \mathbf{X}\psi_{j,j}^{-1}\mathbf{X}^T\right\rangle = \psi_{j,j}^{-1}\left(\langle\mathbf{X}\rangle\langle\mathbf{X}\rangle^T + \mathbf{\Sigma}_X\right) \tag{41}$$

where Σ_X denotes the covariance matrix of $\mathbf{X}$.

Parameter $\theta_{j,k}$ has a generalized inverse Gaussian conditional probability [60, 61], and its MAP estimate is:

$$\hat{\theta}_{j,k} = \frac{2a - 3 + \sqrt{(2a-3)^2 + 8\Lambda_{j,k}^2\delta_{j,k}}}{4\delta_{j,k}}. \tag{42}$$

Similarly, the MAP estimate for $\sigma_{k,i}$ has the following closed form:

$$\hat{\sigma}_{k,i} = \frac{2a_X - 3 + \sqrt{(2a_X-3)^2 + 8\langle x_{k,i}^2\rangle\rho_{k,i}}}{4\rho_{k,i}}. \tag{43}$$

The MAP estimate for $\delta_{j,k}$ is:

$$\hat{\delta}_{j,k} = \frac{a + b - 1}{\theta_{j,k} + \phi_k}. \tag{44}$$

Correspondingly,

$$\hat{\rho}_{k,i} = \frac{a_X + b_X - 1}{\sigma_{k,i} + \omega_k}. \tag{45}$$

Parameter ϕ_k generates both sparse and dense components, and its MAP estimate takes the form:

$$\hat{\phi}_k = \frac{H + \sqrt{H^2 + MT}}{M}, \tag{46}$$

where

$$H = pb\langle z_k\rangle + c - 1 - \frac{p}{2}(1 - \langle z_k\rangle) \tag{47}$$

$$M = 2\left(\langle z_k\rangle \sum_{j=1}^{p} \delta_{j,k} + \tau_k\right) \tag{48}$$

$$T = \sum_{j=1}^{p} \Lambda_{j,k}^2. \tag{49}$$

Correspondingly,

$$\hat{\omega}_k = \frac{H_X + \sqrt{H_X^2 + M_X T_X}}{M_X}, \tag{50}$$

where

$$H_X = nb_X\langle o_k\rangle + c_X - 1 - \frac{n}{2}(1 - \langle o_k\rangle) \tag{51}$$

$$M_X = 2\left(\langle o_k\rangle \sum_{i=1}^{n} \rho_{k,i} + \kappa_k\right) \tag{52}$$

$$T_X = \sum_{i=1}^{n} \left\langle x_{k,i}^2 \right\rangle. \tag{53}$$

The following parameters have similar updates to $\delta_{j,k}$, with simple closed forms because of the conjugacy of the distributions:

$$\hat{\tau}_k = \frac{c + d - 1}{\phi_k + \eta} \tag{54}$$

$$\hat{\eta} = \frac{Kd + e - 1}{\gamma + \sum_k \tau_k} \tag{55}$$

$$\hat{\gamma} = \frac{e + f - 1}{\eta + v} \tag{56}$$

The corresponding parameters related to **X** have similar forms:

$$\hat{\kappa}_k = \frac{c_X + d_X - 1}{\omega_k + \chi} \tag{57}$$

$$\hat{\chi} = \frac{Kd_X + e_X - 1}{\varphi + \sum_k \kappa_k} \tag{58}$$

$$\hat{\varphi} = \frac{e_X + f_X - 1}{\chi + \xi}. \tag{59}$$

The expected value of $z_k|\mathbf{\Theta}$ is computed as:

$$\begin{aligned} \langle z_k|\mathbf{\Theta}_\Lambda\rangle &= p(z_k = 1|\mathbf{\Theta}_\Lambda) \\ &= \frac{\pi \prod_{j=1}^{p} \mathcal{N}(\Lambda_{j,k}|\theta_{j,k})\mathcal{G}a(\theta_{j,k}|a, \delta_{j,k})\mathcal{G}a(\delta_{j,k}|b, \phi_k)}{(1-\pi)(\prod_{j=1}^{p} \mathcal{N}(\Lambda_{j,k}|\phi_k)) + \pi \prod_{j=1}^{p} \mathcal{N}(\Lambda_{j,k}|\theta_{j,k})\mathcal{G}a(\theta_{j,k}|a, \delta_{j,k})\mathcal{G}a(\delta_{j,k}|b, \phi_k)} \end{aligned} \tag{60}$$

The prior on the indicator variable for sparse and dense components, π, has a beta distribution, and its geometric mean is the following:

$$\langle \ln \pi\rangle = \psi\left(\sum_{k=1}^{K} \langle z_k\rangle + \alpha\right) - \psi(K + \alpha + \beta) \tag{61}$$

where ψ is the digamma function.

The corresponding parameters related to **X** are

$$\begin{aligned}\langle o_k|\boldsymbol{\Theta}_X\rangle &= p(o_k=1|\boldsymbol{\Theta}_X)\\ &= \frac{\pi\prod_{i=1}^{n}\mathcal{N}(x_{k,i}|\sigma_{k,i})\mathcal{G}a(\sigma_{k,i}|a_X,\rho_{k,i})\mathcal{G}a(\rho_{k,i}|b_X,\omega_k)}{(1-\pi)(\prod_{i=1}^{n}\mathcal{N}(x_{k,i}|\omega_k))+\pi\prod_{i=1}^{n}\mathcal{N}(x_{k,i}|\sigma_{k,i})\mathcal{G}a(\sigma_{k,i}|a_X,\rho_{k,i})\mathcal{G}a(\rho_{k,i}|b_X,\omega_k)}.\end{aligned} \tag{62}$$

$$\langle\,\ln\pi_X\rangle=\psi\left(\sum_{k=1}^{K}o_k+\alpha_X\right)-\psi(K+\alpha_X+\beta_X) \tag{63}$$

Assuming that the residual precision has a conjugate (gamma) prior, $\frac{1}{\psi_{j,j}}\sim\mathcal{G}a(1,1)$, then we have

$$\boldsymbol{\Psi}=\mathrm{diag}\left(\frac{\mathbf{YY}^{\mathrm{T}}-\mathbf{2Y}\langle\mathbf{X}^{\mathrm{T}}\rangle\boldsymbol{\Lambda}^{\mathrm{T}}+\boldsymbol{\Lambda}\langle\mathbf{XX}^{\mathrm{T}}\rangle\boldsymbol{\Lambda}^{\mathrm{T}}+2\mathbf{I}}{n+2}\right). \tag{64}$$

In estimating the factors, we invert a $K \times K$ matrix for each sample, and, in estimating the loading matrix, we invert a $K \times K$ matrix for reach gene, so the main source of computational complexity in our algorithm is $\mathcal{O}((n+p)K^3)$. To summarize the description above, we write the complete VEM algorithm for parameter updates:

Algorithm 1: Variational expectation maximization for BicMix

```
Data: $Y$, $K$, $n\_itr$, $a$, $b$, $c$, $d$, $e$, $f$, $a_X$, $b_X$, $c_X$, $d_X$, $e_X$, $f_X$, $\alpha$, $\beta$, $\alpha_X$, $\beta_X$
Initialize starting values:
Sample $\eta$, $\gamma$, $\chi$, $\varphi \leftarrow \mathcal{G}a(1,1)$
Sample $\pi \leftarrow \mathcal{B}eta(\alpha,\beta)$, $\pi_X \leftarrow \mathcal{B}eta(\alpha_X,\beta_X)$,
for $j \leftarrow 1$ to $p$ do
  Sample $\psi_{j,j} \leftarrow \mathcal{G}a(1,1)$
for $k \leftarrow 1$ to $K$ do
  Sample $z_k \leftarrow \mathcal{B}ern(\pi)$, $o_k \leftarrow \mathcal{B}ern(\pi_X)$
  Sample $\phi_k$, $\tau_k$, $\omega_k$, $\kappa_k \leftarrow \mathcal{G}a(1,1)$
  for $j \leftarrow 1$ to $p$ do
    Sample $\Lambda_{j,k} \leftarrow \mathcal{N}(0,1)$, Sample $\theta_{j,k}$, $\delta_{j,k} \leftarrow \mathcal{G}a(1,1)$
  for $i \leftarrow 1$ to $n$ do
    Sample $x_{k,i} \leftarrow \mathcal{N}(0,1)$, Sample $\sigma_{k,i}$, $\rho_{k,i} \leftarrow \mathcal{N}(0,1)$
for $t \leftarrow 1$ to $n\_itr$ do
  for $j \leftarrow 1$ to $p$ do
    Update $\Lambda_{j,\cdot} \leftarrow$ Eq (35)
  for $k \leftarrow 1$ to $K$ do
    Update $\theta_{j,k} \leftarrow$ Eq (42), $\delta_{j,k} \leftarrow$ Eq (44)
  for $k \leftarrow 1$ to $K$ do
    Update $\phi_k \leftarrow$ Eq (46), $\tau_k \leftarrow$ Eq (54), $z_k \leftarrow$ Eq (60)
  Update $\eta \leftarrow$ Eq (55), $\gamma \leftarrow$ Eq (56), $\pi \leftarrow$ Eq (61)
  for $i \leftarrow 1$ to $n$ do
    Update $X_{\cdot,i} \leftarrow$ Eq (38)
    for $k \leftarrow 1$ to $K$ do
      Update $\sigma_{k,i} \leftarrow$ Eq (43), $\rho_{k,i} \leftarrow$ Eq (45)
  for $k \leftarrow 1$ to $K$ do
    Update $\omega_k \leftarrow$ Eq (50), $\kappa_k \leftarrow$ Eq (57), $o_k \leftarrow$ Eq (62)
  Update $\chi \leftarrow$ Eq (58), $\varphi \leftarrow$ Eq (59), $\pi_X \leftarrow$ Eq (63)
  for $j \leftarrow 1$ to $p$ do
    Update $\psi_{j,j} \leftarrow$ Eq (64)
Output $\boldsymbol{\Lambda}$, X, z, o
```

Across hundreds of runs, we found that the number of recovered factors was stable within each application with low variance. As a rule of thumb, we initialized the number of latent factors to $K = 2 \times \min(n, p)$, and, if we find that the number of factors recovered is not reduced by approximately a third, we increased the number of initial factors until we saw this substantial reduction in K.

MCMC: Conditional distributions of BicMix parameters

We derive below the conditional distributions that capture the MCMC approach that we implemented for BicMix. In our manuscript, we used MCMC to compute the warm start parameter settings in the simulations.

We updated the loading matrix $\mathbf{\Lambda}$ one row at a time, where each row consists of values across the K components. The jth row of the loading matrix, $\Lambda_{j,\cdot}$, has the following posterior distribution,

$$\Lambda_{j,\cdot}|Y_{j,\cdot}\mathbf{X}, \mathbf{\Theta}_i, \psi_{j,j} \sim \mathcal{N}\left(Y_{j,\cdot}\psi_{j,j}^{-1}\mathbf{X}^T\left(\mathbf{X}\psi_{j,j}^{-1}\mathbf{X}^T + \mathbf{V}_j^{-1}\right)^{-1}, \mathbf{X}\psi_{j,j}^{-1}\mathbf{X}^T + \mathbf{V}_j^{-1}\right) \tag{65}$$

where $\mathbf{V}_j$ is a $K \times K$ diagonal matrix. If we used $V_{j,k,k}$ to denote the (k, k)th element for $\mathbf{V}_j$, then we sampled $V_{j,k,k}$ and its related parameters as follows:

$$V_{j,k,k} = \begin{cases} \theta_{j,k} & \text{if } z_k = 1; \\ \phi_k & \text{if } z_k = 0. \end{cases} \tag{66}$$

Similarly, each column of $\mathbf{X}$ consists of values across the K components; the ith column of the factor matrix, $X_{\cdot,i}$, has the following posterior distribution,

$$X_{\cdot,i}|Y_{\cdot,i}, \mathbf{\Lambda}, \mathbf{\Sigma}_i, \mathbf{\Psi} \sim \mathcal{N}\left((\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda} + \mathbf{W}_i^{-1})^{-1}\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}Y_{\cdot,i}, \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda} + \mathbf{W}_i^{-1}\right), \tag{67}$$

where $\mathbf{W}_i$ is a $K \times K$ diagonal matrix. If $W_{i,k,k}$ denotes the (k, k)th element of W_i, then we sampled the value of $W_{i,k,k}$ as follows:

$$W_{i,k,k} = \begin{cases} \sigma_{k,i} & \text{if } o_k = 1; \\ \omega_k & \text{if } o_k = 0. \end{cases} \tag{68}$$

We sampled values for the parameters conditional on sparse and dense state as follows. If $z_k = 1$,

$$\theta_{j,k}|\Lambda_{j,k}, \delta_{j,k} \sim \mathcal{GIG}\left(a - \frac{1}{2}, 2\delta_{j,k}, \Lambda_{j,k}^2\right) \tag{69}$$

$$\delta_{j,k}|\theta_{j,k}, \phi_k \sim \mathcal{G}a(a + b, \theta_{j,k} + \phi_k) \tag{70}$$

$$\phi_k|\delta_{j,k}, \tau_k \sim \mathcal{G}a\left(pb + c, \sum_{j=1}^{p}\delta_{j,k} + \tau_k\right). \tag{71}$$

If $z_k = 0$,

$$\phi_k|\tau_k, \Lambda_{j,k} \sim \mathcal{GIG}\left(c - \frac{p}{2}, 2\tau_k, \sum_{j=1}^{p}\Lambda_{j,k}^2\right). \tag{72}$$

Correspondingly, the following parameters related to **X** were sampled as follows. If $o_k = 1$

$$\sigma_{k,i} | x_{k,i}, \rho_{k,i} \sim \mathcal{GIG}\left(a_X - \frac{1}{2}, 2\rho_{k,i}, x_{k,i}^2\right) \tag{73}$$

$$\rho_{k,i} | \sigma_{k,i}, \omega_k \sim \mathcal{G}a(a_X + b_X, \rho_{k,i} + \omega_k) \tag{74}$$

$$\omega_k | \rho_{k,i}, \omega_k \sim \mathcal{G}a\left(nb_X + c_X, \sum_{i=1}^{n} \rho_{k,i} + \omega_k\right). \tag{75}$$

If $o_k = 0$

$$\omega_k | \kappa_k, x_{k,i} \sim \mathcal{GIG}\left(c_X - \frac{n}{2}, 2\kappa_k, \sum_{i=1}^{n} x_{k,i}^2\right). \tag{76}$$

The following parameters are not sparse or dense component specific; they each have a gamma conditional distribution because of conjugacy:

$$\tau_k | \phi_k, \eta \sim \mathcal{G}a(c + d, \phi_k + \eta) \tag{77}$$

$$\eta | \gamma, \tau_k \sim \mathcal{G}a\left(Kd + e, \gamma + \sum_{k=1}^{K} \tau_k\right) \tag{78}$$

$$\gamma | \eta, v \sim \mathcal{G}a(e + f, \eta + v). \tag{79}$$

Parameters related to **X** were sampled as,

$$\kappa_k \sim \mathcal{G}a(c_X + d_X, \omega_k + \chi) \tag{80}$$

$$\chi \sim \mathcal{G}a\left(Kd_X + e_X, \varphi + \sum_{k} \kappa_k\right) \tag{81}$$

$$\varphi \sim \mathcal{G}a(e_X + f_X, \chi + \xi). \tag{82}$$

The conditional probability for z_k has a Bernoulli distribution:

$$p(z_k = 1 | \boldsymbol{\Theta}_\Lambda) = \frac{\pi \prod_{j=1}^{p} \mathcal{N}(\Lambda_{j,k} | \theta_{j,k}) \mathcal{G}a(\theta_{j,k} | a, \delta_{j,k}) \mathcal{G}a(\delta_{j,k} | b, \phi_k)}{(1-\pi)(\prod_{j=1}^{p} \mathcal{N}(\Lambda_{j,k} | \phi_k)) + \pi \prod_{j=1}^{p} \mathcal{N}(\Lambda_{j,k} | \theta_{j,k}) \mathcal{G}a(\theta_{j,k} | a, \delta_{j,k}) \mathcal{G}a(\delta_{j,k} | b, \phi_k)}.$$

Let $p_z = p(z_k = 1 | \boldsymbol{\Theta}_\Lambda)$; then the conditional probability for z_k is

$$z_k | p_z \sim \mathcal{B}ern(p_z). \tag{83}$$

The mixing proportion π has a beta conditional probability:

$$\pi | \alpha, \beta, z_k \sim \mathcal{B}eta\left(\alpha + \sum_{k=1}^{K} \mathbb{1}_{z_k=1}, K - \sum_{k=1}^{K} \mathbb{1}_{z_k=0} + \beta\right) \tag{84}$$

where $\mathbb{1}$ is the indicator function.

Similarly for **X**, the conditional probability for o_k has a Bernoulli distribution:

$$p(o_k = 1|\boldsymbol{\Theta}_X) = \frac{\pi \prod_{i=1}^{n} \mathcal{N}(x_{k,i}|\sigma_{k,i}) \mathcal{G}a(\sigma_{k,i}|a_X, \rho_{k,i}) \mathcal{G}a(\rho_{k,i}|b_X, \omega_k)}{(1-\pi)(\prod_{i=1}^{n} \mathcal{N}(x_{k,i}|\omega_k)) + \pi \prod_{i=1}^{n} \mathcal{N}(x_{k,i}|\sigma_{k,i}) \mathcal{G}a(\sigma_{k,i}|a_X, \rho_{k,i}) \mathcal{G}a(\rho_{k,i}|b_X, \omega_k)}. \quad (85)$$

Let $p_X = p(o_k = 1|\boldsymbol{\Theta}_X)$; then the conditional probability for o_k is

$$o_k|p_X \sim \mathcal{B}ern(p_X). \quad (86)$$

The mixing proportion π has a beta conditional probability:

$$\pi_X|\alpha_X, \beta_X, O_k \sim \mathcal{B}eta\left(\alpha_X + \sum_{k=1}^{K} \mathbb{1}_{o_k=1}, K - \sum_{k=1}^{K} \mathbb{1}_{o_k=0} + \beta_X\right) \quad (87)$$

where $\mathbb{1}$ is the indicator function.

Finally, we have,

$$\psi_{j,j} \sim \mathcal{IG}\left(\frac{n}{2} + 1, \frac{\sum_{i=1}^{n} \left(y_{j,i} - \sum_{k=1}^{K} \Lambda_{j,k} x_{k,i}\right)^2}{2} + 1\right). \quad (88)$$

We implemented the following MCMC algorithm for sampling the parameters of the BicMix model.

Algorithm 2: MCMC algorithm for BicMix

```
Data: $Y$ $p \times n$ gene expression matrix, $K$, $n\_itr$
Initialize parameters
for $t \leftarrow 1$ to $n\_itr$ do
  for $j \leftarrow 1$ to $p$ do
    for $k \leftarrow 1$ to $K$ do
      Sample $V_{j,k,k}$ according to Eq (66)
    Sample $\Lambda_{j,\cdot}$ according to Eq (65)
  for $k \leftarrow 1$ to $K$ do
    if $z_k = 1$ then
      Sample $\phi_k$ according to Eq (71)
      for $j \leftarrow 1$ to $p$ do
        Sample $\theta_{j,k}$ according to Eq (69), $\delta_{j,k}$ according to Eq (70),
    if $z_k = 0$ then
      Sample $\phi_k$ according to Eq (72)
  for $k \leftarrow 1$ to $K$ do
    Sample $\tau_k$ according to Eq (77), $z_k$ with Eq (83)
  Sample $\eta$ according to Eq (78), $\gamma$ with Eq (79), $\pi$ with Eq (84)
  for $i \leftarrow 1$ to $n$ do
    for $k \leftarrow 1$ to $K$ do
      Sample $W_{i,k,k}$ according to Eq (68)
    Sample $X_{\cdot,i}$ according to Eq (67)
  for $k \leftarrow 1$ to $K$ do
    if $o_k = 1$ then
      Sample $\omega_k$ according to Eq (75)
      for $i \leftarrow 1$ to $n$ do
        Sample $\sigma_{k,i}$ according to Eq (73), $\rho_{k,i}$ according to Eq (74),
    if $o_k = 0$ then
      Sample $\omega_k$ according to Eq (76)
  for $k \leftarrow 1$ to $K$ do
```

```
      Sample $\kappa_k$ according to Eq (80), $o_k$ according to Eq (86)
  Sample $\chi$ according to Eq (81), $\varphi$ with Eq (82), $\pi_X$ with Eq (87)
  for $j \leftarrow 1$ to $p$ do
    Sample $\psi_{j,j}$ according to Eq (88)
Output $\mathbf{\Lambda}, \mathbf{X}, \mathbf{z}, \mathbf{o}$
```

Data processing and comparative methods

Processing the breast cancer gene expression data. The breast cancer data set is maintained by Dana-Farber Cancer Institute at Harvard University, and is available through their R package: breastCancerNKI version 1.3.1 [70]. We removed genes with $>$ 10% missing values. We imputed the remaining missing values using the R package `impute` (version 1.36.0) [71, 111]. We projected the gene expression levels of each gene to the quantiles of a standard normal. There were 24,158 genes remaining in the data set after filtering.

Processing the CAP gene expression data. The Cardiovascular and Pharmacogenetics (CAP) gene expression data were generated from the Krauss Lab at the Children's Hospital Oakland Research Institute and is publicly available at Gene Expression Omnimbus (GEO), GSE36868. We used expression levels from 8,718 expressed genes measured in a sample of 480 human immortalized blood cell lines (LCLs) [45]. The data were processed according to previous work [45, 60], including filtering unexpressed genes; however, neither known covariates nor PCs were controlled for before quantile normalization. We also removed genes with probes on the gene expression array that aligned to multiple regions of the genome using a BLAST analysis and human reference genome hg19 in order to remove gene pairs that appeared well correlated due to a co-hybridization effect. We did this because pairs of genes that were artefactually correlated due to co-hybridization were incorrectly identified by BicMix as co-regulated genes, and we wanted to control this artifact.

Processing the GTEx gene expression data. We downloaded from dbGaP the Genotype-Tissue Expression (GTEx) project v4 pilot data [46]. The subset of gene expression data that we used contained p = 20,134 genes measured in n = 446 samples across four tissues: adipose (n_f = 103), artery (n_a = 118), lung (n_l = 122) and skin (n_s = 106). We preprocessed these RNA-seq data as in earlier work and recapitulated here [104].

We trimmed the RNA-seq reads from the GTEx pilot data v4 [46] using `Trimmomatic` (v.0.30) [112]. We also trimmed the adapter sequences and overrepresented contaminant sequences that are identified by `FastQC` (v.0.10.1) [113] with 2 seed mismatches and a simple clip threshold of 20. For all reads, we trimmed the leading and trailing nucleotides (low quality or Ns) until a canonical base was encountered with quality greater than 3. For adaptive quality trimming, we scanned the reads with a 4-base sliding window and trimmed when the average quality per base dropped below 20. We discarded sequences shorter than 30 nucleotides. We were left with 3.65 B read pairs in adipose, 4.02 B read pairs in artery, 4.63 B read pairs in lung, 3.63 B read pairs in skin after trimming.

Before mapping the RNA-seq reads to human reference genome assembly GRCh37.p13, we prepared the genome with `STAR aligner genomeGenerate` mode by setting the splice junction database (`sjdbGTFfile`) to GENCODE v.19, and setting the splice junction database overhang (`sjdbOverhang`) to 75 bp. We used default settings for the rest of the parameters. We ran `STAR aligner alignReads` using default settings except setting `outFilterMultimapNmax` to 1. We mapped 14.52 B read pairs (91%) uniquely, and we discarded the 1.2 B read pairs (7.5%) that mapped to multiple loci.

We converted the mapped reads to read counts (one read pair per read count) using the software `featureCounts` [114] on settings that discounted multi-overlapping and chimeric reads. We also required both ends of a read pair to be mapped to the same gene. After filtering all genes with > 25% non-zero values across all tissues, we were left with 20,134 genes. We normalized the read abundance by gene length, GC-content, and library size using the `Bioconductor` R package `cqn` [115], and the resulting gene expression values were mapped to the quantiles of the standard normal distribution. Note that we did not correct for confounding factors nor known covariates before quantile normalization. We concatenated these gene expression data sample-wise to create a single response matrix $\mathbf{Y} \in \Re^{p\times(n_f + n_a + n_l + n_s)}$.

Simulation comparison

Using these simulated data, we compared BicMix to five other methods: Fabia, Plaid, CC, Bimax, and Spectral biclustering. We ran these methods using the following settings.

For Sim1, we set the number of components to the correct values, and ran each method as follows.

- We ran Fabia (version 2.10.2) using its default parameter settings.
- We ran Fabia-truth using default parameter settings. We set the sparsity threshold in Fabia to the number (from 100 quantiles of the uniform distribution over [0.1, 5]) that produced the closest match in the recovered matrices to the number of non-zero elements in the simulated data.
- We ran Plaid (implemented in the R package `biclust` [116] version 1.0.2) using `background` = `TRUE` to capture the noise, maximum layers were set to 10, number of iterations to find starting values was set to 10, and the number of iterations to find the layers was set to 100.
- We ran CC (implemented in the R package `biclust` [116] version 1.0.2) by setting the maximum accepted score `delta` = `1.5` and the scaling factor `alpha` = `1.0`.
- We ran Bimax (implemented in the R package `biclust` [116] version 1.0.2) by setting the minimum number of rows and columns to 2.
- We ran Spectral biclustering (implemented in the R package `biclust` [116] version 1.0.2) by setting the normalization method to `bistochastization`, the number of eigenvalues for constructing the biclusters was set to 10, and the minimum number of rows and columns for the biclusters were set to 2.

For Sim2, we corrected the simulated data for the dense components by controlling for five PCs in the simulated gene expression data, and we ran the methods on the residual matrix as in Sim1, setting the number of components to 10.

Redundancy of components

We calculated a simple statistic to check the redundancy of the multiple components recovered across multiple runs as follows. For every component in all runs, we counted the number of genes with non-zero values, denoted as n_g, and the number of samples with non-zero values, denoted as n_s for each component. We then grouped the components that share the same n_g and n_s. For each pair of components in the same group, we counted how many components have non-zero values for the same genes and the same samples (i.e., the ℓ_0 norm between components, or the Manhattan distance). Redundant components corresponded to pairs of factors and loadings for which the Manhattan distance is zero.

Algorithm for identifying gene co-expression networks from BicMix

We write out the algorithm we used to build the gene co-expression networks using the fitted BicMix model. Note that the sparsity-inducing prior on the covariance matrix of the factors increases the difficulty of computing the gene-wise covariance matrix relative to the common identity matrix covariance in the prior of the factors; however, all of the elements necessary to compute an estimate of the factor covariance matrix have been explicitly quantified in the VEM algorithm already.

Algorithm 3: Algorithm to construct gene co-expression network

```
Data: $p \times K$ loading matrix and $K \times n$ factor matrix; $\Psi$; net_type, rep, c, (d).
for $i \leftarrow 1$ to n_runs do
  for $k_1 \leftarrow 1$ to $K - 1$ do
    for $k_2 \leftarrow K_1 + 1$ to $K$ do
      if $cor(\Lambda_{k_1}, \Lambda_{k_2}) \times cor(\mathbf{X}_{k_1}, X_{k_2}) > 0.5$ then
        discard run
  if net_type = subset specific then
    Add component to $A$ when non-zero factors are only in context $c$
  if net_type = subset differential then
    Compute Wilcoxon signed-rank test for non-zero factor values across contexts $c$, $d$ Add component to $A$ when $p < 1 \times 10^{-10}$
  Construct the covariance matrix for $\mathbf{X}$ as $\mathbf{\Sigma} \leftarrow \langle \mathbf{X}\mathbf{X}^{\mathbf{T}} \rangle - \langle \mathbf{X} \rangle \langle \mathbf{X} \rangle^{\mathbf{T}}$ (Eqs 41 and 38)
  Calculate the variance for the residual as $\mathbf{\Psi} \leftarrow$ Eq (64)
  Construct the precision matrix for subset $A$ as $\Delta^i = \left(\mathbf{\Lambda}^i_A \Sigma^i_{A,A} \mathbf{\Lambda}^i_A T + \Psi\right)^{-1}$
  Run GeneNet [63] on $\mathbf{\Delta}^i$ to test significance of edges
  Store edges with probability of presence $\geq 0.8$
Count number of times each edge is found across all runs
Keep edges that are found $\geq$ rep times
Output the nodes and edges
Draw graph using Gephi [117]
```

Quantifying the expected number of edges using ensemble method

We computed the (approximate) expected number of edges to appear r or more times at random using our ensemble method as follows. Let R represent the total number of runs, r represent the threshold for number of runs an edge must appear, E represent the total number of possible edges, and $\hat{e}$ represent the average number of edges per run. Note that this is an approximate expectation because we are using $\hat{e}$ instead of the true number of edges recovered in each run; this expectation has a noticeable impact on the result when the variance in edges per run is large (this was the case only in the breast cancer data set), but otherwise lead to a reasonable approximation.

We computed the probability of a single edge occurring in at least r networks as follows:

$$
\begin{aligned}
Pr(|e_i| \geq r) &= 1 - \sum_{j=1}^{r-1} Pr(|e_i| = j) \\
Pr(|e_i| = j) &= \binom{R}{j} \left(\frac{\hat{e}}{E}\right)^j \left(1 - \frac{\hat{e}}{E}\right)^{R-j}.
\end{aligned}
$$

Then the expected number of edges that will occur r or more times at random was approximated as follows, assuming edge independence:

$$E_r[|e|] = E \cdot Pr(|e_i| \geq r). \tag{89}$$

Supporting Information

S1 Fig. Correlation between the sparse loadings (genes), dense factors (samples), and experimental covariates in the breast cancer data. The x-axis represents 30 recovered factors; the y-axis represents the observed covariates; darker blue and red represent large magnitude correlations, whereas white represents no correlation.
(PDF)

S2 Fig. Correlation among the known covariates in breast cancer data. The x- and y-axes represents the observed covariates; darker blue and red represent large magnitude correlations, whereas white represents no correlation.
(PDF)

S3 Fig. Gene co-expression network specific to ER- samples. Node size corresponds to betweenness centrality.
(PDF)

S4 Fig. Gene co-expression network specific to ER+ samples. Node size corresponds to betweenness centrality.
(PDF)

S5 Fig. Heatmap of genes identified in the ER+ specific and ER- specific networks. There is evidence of differential expression levels across the two sample types for these genes that are in the ER+ and ER- specific networks.
(PDF)

S6 Fig. Correlation among the known covariates in the CAP data. The x- and y-axes represents the observed covariates; darker blue and red represent large magnitude correlations, whereas white represents no correlation.
(PDF)

S7 Fig. Smoking status specific gene co-expression networks in the CAP data. Panel a: Gene co-expression network specific to smokers. Panel b: Gene co-expression network specific to non-smokers.
(PDF)

S8 Fig. Distribution of the number of genes in loadings and the number of samples in the factors for the GTEx data. Both the sparse and dense loadings and factors are shown. Left: number of genes for all loadings; Middle: number of samples for all factors; Right: a summary of the PVE explained by the components across all runs, where upper bound and the lower bound of the ribbon correspond to the maximum and minimum PVE, and the solid line correspond to the median. The components are inversely sorted by the median.
(PDF)

S9 Fig. Principal components analysis applied to the GTEx pilot data. PC1 effectively separates the four tissue types.
(PDF)

S10 Fig. Plaid applied to the GTEx pilot data. Each sample point is plotted with jitter to denote the density of each tissue in each included or excluded component. The four factors each capture variation in one (or a subset of one) of the tissues reasonably well, except for adipose.
(PDF)

S11 Fig. Fabia with four components applied to the GTEx pilot data. Across factors 1, 2, and 3, the four tissue types are effectively separated.
(PDF)

S12 Fig. Fabia with twenty components applied to the GTEx pilot data. With 20 factors, Fabia is no longer able to separate the four tissue types because of limited sparsity.
(PDF)

S13 Fig. BicMix applied to the GTEx pilot data. The substantial sparsity induced in BicMix is illustrated in these panels. Note that BicMix separates all four tissues in the first four factors.
(PDF)

S1 Table. Genes that are specific to ER- patients. The importance of the genes are ordered by their betweenness centralities.
(TEX)

S2 Table. Genes that are specific to ER+ patients. The importance of the genes are ordered by their betweenness centralities.
(TEX)

S3 Table. Genes that are differential between ER+ and ER- patients. The importance of the genes are ordered by their betweenness centralities.
(TEX)

S4 Table. Genes that are in the adipose-specific network. The genes are ordered by their betweenness centralities.
(TEX)

S5 Table. Genes that are in the artery-specific network. The genes are ordered by their betweenness centralities.
(TEX)

S6 Table. Genes that are in the lung-specific network. The genes are ordered by their betweenness centralities.
(TEX)

S7 Table. Genes that are in the skin-specific network. The genes are ordered by their betweenness centralities.
(TEX)

S8 Table. Gene Ontology enrichment analysis across recovered factors for adipose tissue. All term enrichments are for a cutoff of FDR ≤ 0.05. Gene names are suppressed, but the number of genes in each factor is shown.
(TEX)

S9 Table. Gene Ontology enrichment analysis across recovered factors for artery tissue. All term enrichments are for a cutoff of FDR ≤ 0.05. Gene names are suppressed, but the number of genes in each factor is shown.
(TEX)

S10 Table. Gene Ontology enrichment analysis across recovered factors for lung tissue. All term enrichments are for a cutoff of FDR ≤ 0.05. Gene names are suppressed, but the number of genes in each factor is shown.
(TEX)

S11 Table. Gene Ontology enrichment analysis across recovered factors for skin tissue. All term enrichments are for a cutoff of FDR ≤ 0.05. Gene names are suppressed, but the number of genes in each factor is shown.
(TEX)

S12 Table. eQTLs for adipose tissue. Associations are listed in the order of the significance of p values.
(TEX)

S13 Table. eQTLs for artery tissue. Associations are listed in the order of the significance of p values.
(TEX)

S14 Table. eQTLs for lung tissue. Associations are listed in the order of the significance of p values.
(TEX)

S15 Table. eQTLs for skin tissue. Associations are listed in the order of the significance of p values.
(TEX)

Acknowledgments

The authors would like to acknowledge conversations with Sebastian Seung and Eleazar Eskin that contributed to this manuscript.

Author Contributions

Conceived and designed the experiments: BEE CG. Performed the experiments: CG. Analyzed the data: CG BEE. Contributed reagents/materials/analysis tools: ICM SZ. Wrote the paper: BEE CG CDB.

References

1. Hung JH, Whitfield TW, Yang TH, Hu Z, Weng Z, et al. (2010) Identification of functional modules that correlate with phenotypic difference: The influence of network topology. Genome Biology 11: R23. doi: 10.1186/gb-2010-11-2-r23 PMID: 20187943
2. Parkkinen JA, Kaski S (2010) Searching for functional gene modules with interaction component models. BMC Systems Biology 4: 1–11. doi: 10.1186/1752-0509-4-4
3. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95: 14863–14868. doi: 10.1073/pnas.95.25.14863
4. Jiang D, Tang C, Zhang A (2004) Cluster analysis for gene expression data: A survey. IEEE Transactions on Knowledge and Data Engineering 16: 1370–1386. doi: 10.1109/TKDE.2004.68
5. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, et al. (2006) GenePattern 2.0. Nature Genetics 38: 500–501. doi: 10.1038/ng0506-500 PMID: 16642009
6. de Souto MC, Costa IG, Araujo DSd, Ludermir TB, Schliep A (2008) Clustering cancer gene expression data: a comparative study. BMC Bioinformatics 9: 497. doi: 10.1186/1471-2105-9-497 PMID: 19038021

7. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. Journal of Computational Biology 7: 601–620. doi: 10.1089/106652700750050961 PMID: 11108481

8. Davidich MI, Bornholdt S (2008) Boolean network model predicts cell cycle sequence of fission yeast. PLoS ONE 3: e1672. doi: 10.1371/journal.pone.0001672 PMID: 18301750

9. MacNeil LT, Walhout AJM (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Research 21: 645–657. doi: 10.1101/gr.097378.109 PMID: 21324878

10. Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology 9: 770–780. doi: 10.1038/nrm2503 PMID: 18797474

11. Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4. doi: 10.2202/1544-6115.1128 PMID: 16646834

12. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, et al. (2006) The Inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biology 7: R36. doi: 10.1186/gb-2006-7-5-r36 PMID: 16686963

13. Ruan J, Dean AK, Zhang W (2010) A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Systems Biology 4: 8. doi: 10.1186/1752-0509-4-8 PMID: 20122284

14. Glass K, Huttenhower C, Quackenbush J, Yuan GC (2013) Passing messages between biological networks to refine predicted interactions. PLoS ONE 8: e64832. doi: 10.1371/journal.pone.0064832 PMID: 23741402

15. Engelhardt B, Stephens M (2010) Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. PLoS Genetics 6: e1001117. doi: 10.1371/journal.pgen.1001117 PMID: 20862358

16. Carvalho CM, Lucas JE, Wang Q, Chang J, Nevins JR, et al. (2008) High-dimensional sparse factor modelling: Applications in gene expression genomics. Journal of the American Statistical Association 103: 1438–1456. doi: 10.1198/016214508000000869 PMID: 21218139

17. West M (2003) Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Statistics 7: 723–732.

18. Bhattacharya A, Dunson DB (2011) Sparse Bayesian infinite factor models. Biometrika 98: 291–306. doi: 10.1093/biomet/asr013 PMID: 23049129

19. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848–853. doi: 10.1126/science.1136678 PMID: 17289997

20. Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genetics 9: e1003649. doi: 10.1371/journal.pgen.1003649 PMID: 23935528

21. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10: 515–534. doi: 10.1093/biostatistics/kxp008 PMID: 19377034

22. Srivastava S, Engelhardt BE, Dunson DB (2014) Expandable factor analysis. arXiv preprint arXiv:14071158: 1–32.

23. Cheng Y, Church GM (2000) Biclustering of expression data. Proceedings of the International Conference on Intelligent Systems for Molecular Biology 8: 93–103.

24. Ben-Dor A, Chor B, Karp R, Yakhini Z (2003) Discovering local structure in gene expression data: The order-preserving submatrix problem. Journal of Computational Biology 10: 373–384. doi: 10.1089/10665270360688075 PMID: 12935334

25. Murali TM, Kasif S (2003) Extracting conserved gene expression motifs from gene expression data. Proceedings of the Pacific Symposium on Biocomputing: 77–88.

26. Li G, Ma Q, Tang H, Paterson AH, Xu Y (2009) QUBIC: A qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Research 37: e101–e101. doi: 10.1093/nar/gkp491 PMID: 19509312

27. Prelic A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22: 1122–1129. doi: 10.1093/bioinformatics/btl060 PMID: 16500941

28. Bergmann S, Ihmels J, Barkai N (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. Physical review E, Statistical, nonlinear, and soft matter physics 67: 031902. doi: 10.1103/PhysRevE.67.031902 PMID: 12689096

29. Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, et al. (2009) Detailing regulatory networks through large scale data integration. Bioinformatics 25: 3267–3274. doi: 10.1093/bioinformatics/btp588 PMID: 19825796

30. Lazzeroni L, Owen A (2000) Plaid models for gene expression data. Statistica Sinica 12: 61–86.

31. Gu J, Liu JS (2008) Bayesian biclustering of gene expression data. BMC Genomics 9: S4. doi: 10.1186/1471-2164-9-S1-S4 PMID: 18366617

32. Bozdag D, Parvin JD, Catalyurek UV (2009) A biclustering method to discover co-regulated genes using diverse gene expression datasets. In: Rajasekaran S, editor, Bioinformatics and Computational Biology, Springer Berlin Heidelberg, number 5462 in Lecture Notes in Computer Science. pp. 151–163.

33. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, et al. (2010) FABIA: Factor analysis for bicluster acquisition. Bioinformatics 26: 1520–1527. doi: 10.1093/bioinformatics/btq227 PMID: 20418340

34. Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral biclustering of microarray data: Coclustering genes and conditions. Genome Research 13: 703–716. doi: 10.1101/gr.648603 PMID: 12671006

35. Aguilar-Ruiz JS (2005) Shifting and scaling patterns from gene expression data. Bioinformatics 21: 3840–3845. doi: 10.1093/bioinformatics/bti641 PMID: 16144809

36. Storey JD, Akey JM, Biswas S, Leek JT (2007) On the design and analysis of gene expression studies in human populations. Nature Genetics 39: 808–809. doi: 10.1038/ng0707-808

37. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11: 733–739. doi: 10.1038/nrg2825 PMID: 20838408

38. Leek J, Storey J (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genetics 3: e161–1735. doi: 10.1371/journal.pgen.0030161

39. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772. doi: 10.1038/nature08872 PMID: 20220758

40. Stegle O, Parts L, Durbin R, Winn J (2010) A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. PLoS Computational Biology 6: e1000770. doi: 10.1371/journal.pcbi.1000770 PMID: 20463871

41. Listgarten J, Kadie C, Schadt E, Heckerman D (2010) Correction for hidden confounders in the genetic analysis of gene expression. Proceedings of the National Academy of Sciences of the United States of America 107: 16465–16470. doi: 10.1073/pnas.1002425107 PMID: 20810919

42. Pierson E, Koller D, Battle A, Mostafavi S, Consortium G, et al. (2015) Sharing and specificity of co-expression networks across 35 human tissues. PLOS Computational Biology 11: e1004220. doi: 10.1371/journal.pcbi.1004220 PMID: 25970446

43. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536. doi: 10.1038/415530a

44. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine 347: 1999–2009. doi: 10.1056/NEJMoa021967 PMID: 12490681

45. Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. PLoS Genetics 9: e1003649+. doi: 10.1371/journal.pgen.1003649 PMID: 23935528

46. Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348: 648–660. doi: 10.1126/science.1262110

47. Hartigan JA (1972) Direct clustering of a data matrix. Journal of the American Statistical Association 67: 123–129. doi: 10.1080/01621459.1972.10481214

48. Van Mechelen I, Bock HH, De Boeck P (2004) Two-mode clustering methods: A structured overview. Statistical Methods in Medical Research 13: 363–394. doi: 10.1191/0962280204sm373ra PMID: 15516031

49. Patrikainen A, Meila M (2006) Comparing subspace clusterings. IEEE Transactions on Knowledge and Data Engineering 18: 902–916. doi: 10.1109/TKDE.2006.106

50. Kriegel HP, Kröger P, Zimek A (2009) Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans Knowl Discov Data 3: 1–58. doi: 10.1145/1497577.1497578

51. Yoon S, Benini L, De Micheli G (2007) Co-clustering: a versatile tool for data analysis in biomedical informatics. IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society 11: 493–494. doi: 10.1109/TITB.2007.897575

52. Busygin S, Prokopyev O, Pardalos PM (2008) Biclustering in data mining. Computers & Operations Research 35: 2964–2987. doi: 10.1016/j.cor.2007.01.005

53. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: A survey. IEEE/ACM Transactions in Computational Biology and Bioinformatics 1: 24–45. doi: 10.1109/TCBB.2004.2

54. Madeira SC, Oliveira AL (2009) A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. Algorithms for Molecular Biology 4: 8. doi: 10.1186/1748-7188-4-8 PMID: 19497096

55. Turner H, Bailey T, Krzanowski W (2005) Improved biclustering of microarray data demonstrated through systematic performance tests. Computational Statistics & Data Analysis 48: 235–254. doi: 10.1016/j.csda.2004.02.003

56. Santamaría R, Quintales L, Therón R (2007) Methods to bicluster validation and comparison in microarray data. In: Yin H, Tino P, Corchado E, Byrne W, Yao X, editors, Intelligent Data Engineering and Automated Learning, Springer Berlin Heidelberg, number 4881 in Lecture Notes in Computer Science. pp. 780–789.

57. Neng Fan NB (2009) Recent advances of data biclustering with application in computational neuroscience: 105–132.

58. de Castro P, de Franga F, Ferreira H, Von Zuben F (2007) Evaluating the performance of a biclustering algorithm applied to collaborative filtering: A comparative analysis. In: Proceedings of the 7th International Conference on Hybrid Intelligent Systems. pp. 65–70.

59. Eren K, Deveci M, Kücüktunc O, Catalyürek UV (2012) A comparative analysis of biclustering algorithms for gene expression data. Briefings in Bioinformatics: 32.

60. Gao C, Brown CD, Engelhardt BE (2013) A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. ArXiv preprint arXiv:13104792.

61. Armagan A, Dunson DB, Clyde M (2011) Generalized beta mixtures of Gaussians. In: Proceedings of Neural Information Processing Systems. pp. 523–531.

62. Gao C, Engelhardt B (2012) A sparse factor analysis model for high dimensional latent spaces. NIPS: Workshop on Analysis Operator Learning vs Dictionary Learning: Fraternal Twins in Sparse Modeling.

63. Schäfer J, Strimmer K (2005) An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics 21: 754–764. doi: 10.1093/bioinformatics/bti062 PMID: 15479708

64. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical Applications in Genetics and Molecular Biology 4. PMID: 16646851

65. Breiman L (1996) Bagging predictors. Machine Learning 24: 123–140. doi: 10.1023/A:1018054314350

66. Carvalho CM, Polson NG, Scott JG (2010) The horseshoe estimator for sparse signals. Biometrika 97: 465–480. doi: 10.1093/biomet/asq017

67. Strawderman WE (1971) Proper Bayes minimax estimators of the multivariate normal mean. The Annals of Mathematical Statistics 42: 385–388. doi: 10.1214/aoms/1177693528

68. Berger J (1980) A robust generalized Bayes estimator and confidence region for a multivariate normal mean. The Annals of Statistics 8: 716–761. doi: 10.1214/aos/1176345068

69. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536. doi: 10.1038/415530a

70. Schroeder M, Haibe-Kains B, Culhane A, Sotiriou C, Bontempi G, et al. (2011) breastCancerNKI: Gene expression dataset. R package version 1.0.6.

71. Hastie T, Tibshirani R, Sherlock G, Eisen M, Brown P, et al. (1999) Imputing missing data for gene expression arrays. Technical report.

72. Zhang MH, Man HT, Zhao XD, Dong N, Ma SL (2014) Estrogen receptor-positive breast cancer molecular signatures and therapeutic potentials. Biomedical Reports 2: 41–52. doi: 10.3892/br.2013.187 PMID: 24649067

73. Hu Y, Wu G, Rusch M, Lukes L, Buetow KH, et al. (2012) Integrated cross-species transcriptional network analysis of metastatic susceptibility. Proceedings of the National Academy of Sciences 109: 3184–3189. doi: 10.1073/pnas.1117872109

74. Schüle J, Bergkvist L, Håkansson L, Gustafsson B, Håkansson A (2002) Down-regulation of the *CD3-ζ* chain in sentinel node biopsies from breast cancer patients. Breast Cancer Research and Treatment 74: 33–40. doi: 10.1023/A:1016009913699 PMID: 12150450

75. Yu B, Zhang W (2011) Down-regulation of *CD3-ζ* is a breast cancer biomarker associated with immune suppression. Cell Biology International 35: 165–169. doi: 10.1042/CBI20100346 PMID: 20883209

76. Oghumu S, Varikuti S, Terrazas C, Kotov D, Nasser MW, et al. (2014) *CXCR3* deficiency enhances tumor progression by promoting macrophage M2 polarization in a murine breast cancer model. Immunology 143: 109–119. doi: 10.1111/imm.12293 PMID: 24679047

77. Li Y, Reader JC, Ma X, Kundu N, Kochel T, et al. (2014) Divergent roles of *CXCR3* isoforms in promoting cancer stem-like cell survival and metastasis. Breast Cancer Research and Treatment: 1–13.

78. King TD, Suto MJ, Li Y (2012) The wnt/*β*-catenin signaling pathway: A potential therapeutic target in the treatment of triple negative breast cancer. Journal of Cellular Biochemistry 113: 13–18. doi: 10.1002/jcb.23350 PMID: 21898546

79. Verghese ET, Drury R, Green CA, Holliday DL, Lu X, et al. (2013) MiR-26b is down-regulated in carcinoma-associated fibroblasts from ER-positive breast cancers leading to enhanced cell migration and invasion. The Journal of Pathology 231: 388–399. doi: 10.1002/path.4248 PMID: 23939832

80. Mansour AA, Gafni O, Weinberger L, Zviran A, Ayyash M, et al. (2012) The H3K27 demethylase *UTX* regulates somatic and germ cell epigenetic reprogramming. Nature 488: 409–413. doi: 10.1038/nature11272 PMID: 22801502

81. Van der Meulen J, Sanghvi V, Mavrakis K, Durinck K, Fang F, et al. (2015) The H3K27me3 demethylase *UTX* is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. Blood 125: 13–21. doi: 10.1182/blood-2014-05-577270 PMID: 25320243

82. Aasen E, Medrano J (1990) Amplification of the *ZFY* and *ZFX* genes for sex identification in humans, cattle, sheep and goats. Biotechnology 8: 1279–1281. doi: 10.1038/nbt1290-1279 PMID: 1369448

83. Xu J, Taya S, Kaibuchi K, Arnold A (2005) Sexually dimorphic expression of *USP9X* is related to sex chromosome complement in adult mouse brain. The European Journal of Neuroscience 21: 3017–3022. doi: 10.1111/j.1460-9568.2005.04134.x PMID: 15978012

84. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology 4: P3. doi: 10.1186/gb-2003-4-5-p3 PMID: 12734009

85. Cao Y (2013) Angiogenesis and vascular functions in modulation of obesity, adipose metabolism, and insulin sensitivity. Cell Metabolism 18: 478–489. doi: 10.1016/j.cmet.2013.08.008 PMID: 24035587

86. Basu S, Fenton MJ (2004) Toll-like receptors: Function and roles in lung disease. American Journal of Physiology 286: L887–L892. PMID: 15064235

87. Hosooka T, Noguchi T, Kotani K, Nakamura T, Sakaue H, et al. (2008) *DOK1* mediates high-fat diet–induced adipocyte hypertrophy and obesity through modulation of *PPAR-γ* phosphorylation. Nature Medicine 14: 188–193. doi: 10.1038/nm1706 PMID: 18204460

88. Yeung F, Ramírez CM, Mateos-Gomez PA, Pinzaru A, Ceccarini G, et al. (2013) Non-telomeric role for *RAP1* in regulating metabolism and protecting against obesity. Cell Reports 3: 1847–1856. doi: 10.1016/j.celrep.2013.05.032 PMID: 23791522

89. Jun HS, Hwang K, Kim Y, Park T (2008) High-fat diet alters *PP2A*, *TC10*, and *CIP4* expression in visceral adipose tissue of rats. Obesity 16: 1226–1231. doi: 10.1038/oby.2008.220 PMID: 18388891

90. Oliver P, Caimari A, Díaz-Rúa R, Palou A (2012) Diet-induced obesity affects expression of adiponutrin/*PNPLA3* and adipose triglyceride lipase, two members of the same family. International Journal of Obesity 36: 225–232. doi: 10.1038/ijo.2011.92 PMID: 21556044

91. Traurig MT, Orczewska JI, Ortiz DJ, Bian L, Marinelarena AM, et al. (2013) Evidence for a role of *LPGAT1* in influencing BMI and percent body fat in Native Americans. Obesity 21: 193–202. doi: 10.1002/oby.20243 PMID: 23505186

92. Masiero M, Simões FC, Han HD, Snell C, Peterkin T, et al. (2013) A core human primary tumor angiogenesis signature identifies the endothelial orphan receptor *ELTD1* as a key regulator of angiogenesis. Cancer Cell 24: 229–241. doi: 10.1016/j.ccr.2013.06.004 PMID: 23871637

93. Wojciak-Stothard B, Abdul-Salam VB, Lao KH, Tsang H, Irwin DC, et al. (2014) Aberrant chloride intracellular channel 4 expression contributes to endothelial dysfunction in pulmonary arterial hypertension. Circulation 129: 1770–1780. doi: 10.1161/CIRCULATIONAHA.113.006797 PMID: 24503951

94. Zhang Y, Ling Z, Deng S, Du H, Yin Y, et al. (2014) Associations between *CD36* gene polymorphisms and susceptibility to coronary artery heart disease. Brazilian Journal of Medical and Biological Research 47: 895–903. doi: 10.1590/1414-431X20143825 PMID: 25118627

95. Koh JT, Lee ZH, Ahn KY, Kim JK, Bae CS, et al. (2001) Characterization of mouse brain-specific angiogenesis inhibitor 1 (*BAI1*) and phytanoyl-CoA alpha-hydroxylase-associated protein 1, a novel *BAI1*-binding protein. Molecular Brain Research 87: 223–237. doi: 10.1016/S0169-328X(01)00004-3 PMID: 11245925

96. Villar J, Cabrera NE, Casula M, Flores C, Valladares F, et al. (2010) Mechanical ventilation modulates *TLR4* and *IRAK-3* in a non-infectious, ventilator-induced lung injury model. Respiratory Research 11: 27. doi: 10.1186/1465-9921-11-27 PMID: 20199666

97. Grumelli S, Lu B, Peterson L, Maeno T, Gerard C (2011) *CD46* protects against chronic obstructive pulmonary disease. PLoS ONE 6: e18785. doi: 10.1371/journal.pone.0018785 PMID: 21573156

98. Burdorf L, Stoddard T, Zhang T, Rybak E, Riner A, et al. (2014) Expression of human *CD46* modulates inflammation associated with *GalTKO* lung xenograft injury. American Journal of Transplantation 14: 1084–1095. doi: 10.1111/ajt.12673 PMID: 24698431

99. Reijmerink NE, Postma DS, Koppelman GH (2010) The candidate gene approach in asthma: What happens with the neighbours? European Journal of Human Genetics 18: 17. doi: 10.1038/ejhg.2009.128 PMID: 19654613

100. Skawran B, Steinemann D, Becker T, Buurman R, Flik J, et al. (2008) Loss of 13q is associated with genes involved in cell cycle and proliferation in dedifferentiated hepatocellular carcinoma. Modern Pathology 21: 1479–1489. doi: 10.1038/modpathol.2008.147 PMID: 18820673

101. Xie S, Luca M, Huang S, Gutman M, Reich R, et al. (1997) Expression of *MCAM/MUC18* by human melanoma cells leads to increased tumor growth and metastasis. Cancer Research 57: 2295–2303. PMID: 9187135

102. Mills L, Tellez C, Huang S, Baker C, McCarty M, et al. (2002) Fully human antibodies to *MCAM/MUC18* inhibit tumor growth and metastasis of human melanoma. Cancer Research 62: 5106–5114. PMID: 12208768

103. Taungjaruwinai WM, Bhawan J, Keady M, Thiele JJ (2009) Differential expression of the antioxidant repair enzyme methionine sulfoxide reductase (*MSRA* and *MSRB*) in human skin. The American Journal of Dermatopathology 31: 427–431. doi: 10.1097/DAD.0b013e3181882c21 PMID: 19542914

104. McDowell I, Pai A, Guo C, Vockley C, Brown C, et al. (2014) Identification of long intergenic non-coding RNA eQTLs in four primary tissues reveals association with obesity-related traits. In Review.

105. Villarroya J, Dorado B, Vila MR, Garcia-Arumí E, Domingo P, et al. (2011) Thymidine kinase 2 deficiency-induced mitochondrial DNA depletion causes abnormal development of adipose tissues and adipokine levels in mice. PLoS ONE 6: e29691. doi: 10.1371/journal.pone.0029691 PMID: 22216345

106. Sackmann-Sala L, Berryman DE, Lubbers ER, Zhang H, Vesel CB, et al. (2014) Age-related and depot-specific changes in white adipose tissue of growth hormone receptor-null mice. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 69: 34–43. doi: 10.1093/gerona/glt110

107. Zheng X, Xu C, Smith AO, Stratman AN, Zou Z, et al. (2012) Dynamic regulation of the cerebral cavernous malformation pathway controls vascular stability and growth. Developmental Cell 23: 342–355. doi: 10.1016/j.devcel.2012.06.004 PMID: 22898778

108. Kusuhara S, Fukushima Y, Fukuhara S, Jakt LM, Okada M, et al. (2012) *ARHGEF15* promotes retinal angiogenesis by mediating *VEGF*-induced *CDC42* activation and potentiating *RHOJ* inactivation in endothelial cells. PLoS ONE 7: e45858. doi: 10.1371/journal.pone.0045858 PMID: 23029280

109. McMillan SJ, Sharma RS, McKenzie EJ, Richards HE, Zhang J, et al. (2013) *Siglec-E* is a negative regulator of acute pulmonary neutrophil inflammation and suppresses *CD11b* *β*2-integrin–dependent signaling. Blood 121: 2084–2094. doi: 10.1182/blood-2012-08-449983 PMID: 23315163

110. Stienstra Y, Van Der Werf T, Oosterom E, Nolte I, Van der Graaf W, et al. (2006) Susceptibility to Buruli ulcer is associated with the *SLC11A1* (*NRAMP1*) D543N polymorphism. Genes and Immunity 7: 185–189. doi: 10.1038/sj.gene.6364281 PMID: 16395392

111. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, et al. (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520–525. doi: 10.1093/bioinformatics/17.6.520 PMID: 11395428

112. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. doi: 10.1093/bioinformatics/btu170 PMID: 24695404

113. Andrews S (2012). http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

114. Liao Y, Smyth GK, Shi W (2014) featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30: 923–930. doi: 10.1093/bioinformatics/btt656 PMID: 24227677

115. Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13: 204–216. doi: 10.1093/biostatistics/kxr054 PMID: 22285995

116. Kaiser S, Santamaria R, Theron R, Quintales L, Leisch F (2009) biclust: Bicluster algorithms. R package version 07 2.

117. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks.