Bayesian inference of sample-specific coexpression networks

Enakshi Saha, Viola Fanfani, Panagiotis Mandros, et al.

Genome Res. published online August 12, 2024
Access the most recent version at doi:10.1101/gr.279117.124

Bayesian inference of sample-specific co-expression networks

Enakshi Saha[1,*], Viola Fanfani[1,*], Panagiotis Mandros[1], Marouen Ben-Guebila[1], Jonas Fischer[1], Katherine H. Shutta[1,2], Dawn L. DeMeo[2,3], Camila M. Lopes-Ramos[1,2,3], and John Quackenbush[1,2,4,**]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

[2]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

[3]Department of Medicine, Harvard Medical School, Boston, MA, USA

[4]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

[*]These authors contributed equally.

[**]Corresponding author: johnq@hsph.harvard.edu

July 26, 2024

Abstract

Gene regulatory networks (GRNs) are effective tools for inferring complex interactions between molecules that regulate biological processes and hence can provide insights into drivers of biological systems. Inferring co-expression networks is a critical element of GRN inference, as the correlation between expression patterns may indicate that genes are co-regulated by common factors. However, methods that estimate co-expression networks generally derive an aggregate network representing the mean regulatory properties of the population and so fail to fully capture population heterogeneity. BONOBO (Bayesian Optimized Networks Obtained By assimilating Omics data) is a scalable Bayesian model for deriving individual sample-specific co-expression matrices that recognizes variations in molecular interactions across individuals. For each sample, BONOBO assumes a Gaussian distribution on the log-transformed centered gene expression and a conjugate prior

distribution on the sample-specific co-expression matrix constructed from all other samples in the data. Combining the sample-specific gene co-expression with the prior distribution, BONOBO yields a closed-form solution for the posterior distribution of the sample-specific co-expression matrices, thus allowing the analysis of large datasets. We demonstrate BONOBO's utility in several contexts, including analyzing gene regulation in yeast transcription factor knockout studies, the prognostic significance of miRNA-mRNA interaction in human breast cancer subtypes, and sex differences in gene regulation within human thyroid tissue. We find that BONOBO outperforms other methods that have been used for sample-specific co-expression network inference and provides insight into individual differences in the drivers of biological processes.

***Keywords*—** Gene regulatory network, Co-expression, individual-specific network, Bayesian inference, posterior distribution

Running title: BONOBO.

Introduction

The majority of human traits and diseases are driven not by individual genes, but by networks of genes and proteins interacting with each other. Understanding how genes interact and cooperate under different conditions is a central challenge in deciphering the complexities of cellular processes and their dysregulation in various diseases. Differential expression analysis with conventional tools such as “limma” or “voom” (Ritchie et al. 2015) has allowed us to investigate changes in gene expression attributable to diseases while simultaneously adjusting for the effects of covariates, including age and sex. However, differences in transcription levels alone often fail to explain biological differences between the cohorts being compared (Glass et al. 2013).

Co-expression networks, which represent the coordinated expression patterns of genes across diverse biological samples, can provide insights into processes that are simultaneously activated in different biological states (Becker et al. 2023). However, most methods for constructing co-expression networks estimate an aggregate network for the entire population (Langfelder & Horvath 2008, Dawson et al. 2012, Lemoine et al. 2021), thus failing to capture the heterogeneous, context-specific gene interactions present within individual samples.

Methods to infer sample-specific networks have been proposed to address some of these limitations. This includes Single Pearson Correlation Coefficient (SPCC) (Yu et al. 2015, Zhang et al. 2015), which estimates the significance of sample-specific correlations using a Z-test, Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS)(Kuijjer et al. 2019), which was developed to infer sample-specific gene regulatory networks but has also been applied to Pearson correlation, and sample-specific-weighted correlation network (SWEET)(Chen et al. 2023), which modifies the LIONESS equation to account for size imbalances in dataset subpopulations.

While useful, these methods can produce co-expression matrices that are not positive semidefinite or yield co-expression values outside the defined range for correlation measures (for example, [-1,1] for Pearson correlation coefficient). Because correlation matrices are positive semidefinite by definition, this non-positive semidefiniteness may pose challenges in downstream analyses. For example, if an estimated co-expression matrix is not positive semidefinite, certain linear projections of gene expression, including several principal components (PC), might have negative variance estimates. In more extreme situations, the variance of expression values of certain genes might be negative for some individuals, thus making uncertainty quantification impossible. Alternatively, other methods designed for personalized characterization of diseases through sample-specific networks (Liu et al. 2016) and cancer-specific or group-specific networks (Lee et al. 2020) represent differential networks with respect to an external reference population and hence are susceptible to varying inference depending on the reference used.

We developed BONOBO (Bayesian Optimized Networks Obtained By assimilating Omics data),

an empirical Bayesian model that derives individual sample-specific co-expression networks (Figure 1), facilitating the discovery of gene pairs differentially co-regulated between different conditions and phenotypes, while eliminating the effects of confounders. BONOBO derives positive semidefinite co-expression networks from input data alone, without using external reference datasets. This distinctive feature enables BONOBO to capture correlation structures that remain consistent and comparable across diverse datasets and multiple batches, providing a robust tool for correlation network analysis and meaningful correlation networks for input to gene regulatory network inference tools. BONOBO derives a posterior probability distribution for individual correlation matrices, allowing us to test the hypothesis of whether any pairs of genes have a non-zero correlation within an individual sample in the data. Based on the results of this hypothesis testing, one can infer individual sample-specific sparse co-expression networks by pruning out non-significant edges. This is particularly important for interpretability as empirical data suggests that biological gene networks are sparsely connected (Leclerc 2008).

One of the key strengths of BONOBO lies in its ability to capture the inherent heterogeneity in co-expression patterns among individuals within a population that can be attributed to a range of biological and environmental elements. For example, when comparing aggregate co-expression networks between conditions, such as distinguishing between health and disease or samples from male and female, results are frequently confounded by the study population's heterogeneity stemming from nuisance parameters, such as batch effects or confounding clinical covariates that can include sex or age. BONOBO's individual sample-specific approach explicitly models this heterogeneity, enabling a deeper understanding of the gene networks underlying distinct biological states. In addition, individual sample-specific co-expression networks derived by BONOBO can be used to infer sample-specific gene regulatory networks (GRN), by using the BONOBO networks as an input to methods such as PANDA (Glass et al. 2013), OTTER (Weighill et al. 2021), and EGRET (Weighill et al. 2022).

We demonstrated the advantages of BONOBO using several simulated and real datasets. First, we used simulated data to compare BONOBO's ability to recover "known" sample-specific co-expression matrices with that of other methods for inferring sample-specific networks. We then applied BONOBO to gene expression data from a yeast perturbation experiment and showed that it captures global properties more consistently than other methods and can distinguish the sample-specific effects of single transcription factor knockouts. Next, we examined the interaction between miRNA and mRNA expression in various human breast cancer subtypes using individual-specific co-expression networks derived by BONOBO and found that the correlation patterns between miRNA expression and immune pathways have prognostic significance in Luminal A and Luminal B breast cancer subtypes. In a final application, we analyzed sex differences in gene regulation using RNA-sequencing data from healthy human thyroid tissue. Using BONOBO networks as inputs to PANDA, we inferred individual-specific gene regulatory

networks and compared these between males and females, identifying regulatory differences in immune response, cell proliferation, and metabolic processes, thereby providing a possible mechanism for sex bias in incidence rates of various thyroid conditions such as hypothyroidism and Hashimoto's disease (Pizzato et al. 2022).

BONOBO is available as open-source code in Python through the Network Zoo package (netZooPy v0.10.0; netzoo.github.io) (Ben Guebila et al. 2023).

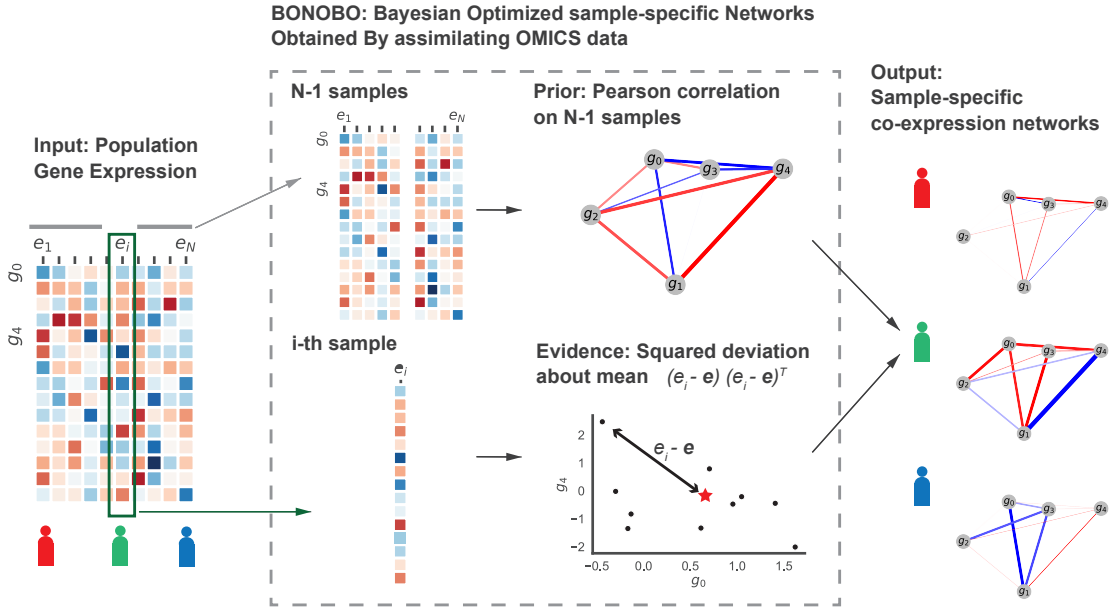


Figure 1: **Schematic diagram of BONOBO**. BONOBO requires a gene expression matrix as input, from which we would like to extract sample-specific correlation networks. Then, for each of the samples, BONOBO infers the network by using both the Pearson correlation matrix computed on $N-1$ samples and the sample-specific squared-deviation about the mean. BONOBO outputs N co-expression networks, one for each sample, and the associated p-values for each of the gene-gene estimated edges.

Methods

BONOBO

Let $x_1, x_2, \ldots, x_N \sim \mathbb{R}^g$ denote the *log-transformed* bulk gene expression values of g genes for N samples. Let us assume that for every sample $i \in \{1, 2, \ldots, N\}$, the centered log expression vector $x_i - \bar{x}$ follows a multivariate normal distribution with mean zero and an unknown sample-specific covariance matrix V_i,

$$x_i - \bar{x} \sim N_g\left(\mathbf{0}_g, V_i\right), \tag{1}$$

where $\mathbf{0}_g \in \mathbb{R}^g$ denotes a vector of all zeros and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ denotes the mean expression across all samples. Our objective is to estimate V_i, the sample-specific covariance matrix of gene expression for the i-th sample, $\forall i$.

It is worth noting that although microarray expression data are often assumed to follow a Gaussian distribution, this assumption is not valid for the expression count data from RNA-sequencing experiments. However, the *log-transformed* expression data is approximately Gaussian and continuous, so in BONOBO we impose a multivariate Gaussian assumption on the *log-transformed* data (see Supplementary Materials S1.8 for a detailed discussion on the validity of this assumption.)

The second assumption we make is that for every sample $i \in \{1, 2, \ldots, N\}$, the sample-specific covariance matrix V_i follows an inverse Wishart prior distribution, given all other samples $j \in \{1, \ldots, N\} \setminus \{i\}$,

$$V_i \sim InvWishart\left((\nu_i - g - 1)S_i, \nu_i\right), \tag{2}$$

where $\nu_i \geq g + 1$ denotes the degrees of freedom and S_i denotes the sample covariance matrix computed from $N - 1$ samples excluding the i-th sample. Under this assumption, the prior mean of the covariance matrix for the i-th sample is $\mathbb{E}[V_i] = S_i$. In other words, we assume that the correlation between any pair of genes for each individual is centered around the correlation between the same pair of genes across the entire population. How much the individual-specific correlation differs from the population-level correlation for every individual, depends on the individual-specific degrees of freedom parameter ν_i. In the following section, we describe a data-driven approach to calibrate this parameter so it reflects how similar a particular sample is to the rest of the population, thereby allowing the individual-specific correlation estimates to deviate from the population-level correlation values accordingly.

The inverse Wishart distribution is a conjugate prior for the covariance matrix of a multivariate normal distribution. Therefore, under the above prior specification, the posterior distribution of the sample-specific covariance matrix V_i also turns out to be an inverse Wishart distribution, as described by the following theorem.

Theorem 1 *Under assumptions* (1) *and* (2)*, the posterior distribution of* V_i *is*

$$V_i | \{x_1, \ldots, x_N\} \sim InvWishart\left((\nu_i - g)\Sigma_i, \nu_i + 1\right), \tag{3}$$

where $\Sigma_i = \frac{(x_i - \bar{x})(x_i - \bar{x})^T + (\nu_i - g - 1)S_i}{\nu_i - g}$ *denotes the posterior mean of* V_i*.*

The proof of the above theorem is given in the appendix S1.1.1.

From (3) we observe that the posterior mean of V_i, the covariance matrix of the i-th sample is a linear combination of the prior mean S_i, which summarizes information from all other samples excluding the

i-th sample and a sample-specific component $(x_i - \bar{x})(x_i - \bar{x})^T$, which summarizes the association between pairwise genes within the i-th sample alone:

$$\Sigma_i = \delta_i (x_i - \bar{x})(x_i - \bar{x})^T + (1 - \delta_i) S_i, \tag{4}$$

where $\delta_i = \frac{1}{\nu_i - g}$. Since $\nu_i - g \geq 1$, we have $0 < \delta_i \leq 1$, which represents the relative contributions of the sample-specific information and the prior information while estimating the posterior mean of V_i.

As sample size n increases, the strong law of large numbers implies $S_i \overset{a.s.}{\to} \Sigma$, where Σ denotes the population covariance matrix. Thus the hyperparameter δ_i quantifies the contribution of the sample-specific information in the posterior mean, while the complement $1 - \delta_i$ quantifies the contribution of the population covariance matrix Σ. For homogeneous populations, we recommend using a smaller value of δ_i, or equivalently, a larger value of $1 - \delta_i$, as this would increase the contribution of the population covariance Σ and give robust estimates of the sample-specific covariance V_i. On the other hand, if the i-th sample is an outlier with respect to the rest of the population, we recommend using a large value of δ_i, thereby decreasing estimation bias. Alternatively, we can set $\delta_i = \delta$, $\forall i$ to some arbitrary value between $(0, 1)$. In the following section, we describe a computationally inexpensive data-driven empirical procedure for calibrating δ_i for every sample.

Fixing prior degrees of freedom

The hyperparameter $\delta_i = \frac{1}{\nu_i - g}$ is a one-to-one function of the prior degrees of freedom ν_i. Hence in order to calibrate δ_i, it suffices to estimate ν_i for every sample i in the data. The following lemma provides a data-driven approach for calibrating ν_i.

Lemma 1 *Under assumption* (2)*, prior variance of the k-th diagonal entry of V_i (denoted by $v_i^{(kk)}$) would be*

$$Var(v_i^{(kk)}) = \frac{2(s_i^{(kk)})^2}{\nu_i - g - 3}, \tag{5}$$

where $\left(s_i^{kk}\right)^2$ denotes the k-th diagonal entry of S_i.

The above lemma is a direct consequence of the properties of inverse Wishart distribution (Zhang 2021).

From (5), summing over $k = 1, \ldots, g$ (over all genes) we get

$$\sum_{k=1}^{g} Var(v_i^{(kk)}) = \frac{2\sum_{k=1}^{g}(s_i^{(kk)})^2}{\nu_i - g - 3} \tag{6}$$

Simplifying the above equation gives us,

$$\nu_i = g + 3 + \frac{2\sum_{k=1}^{g}(s_i^{(kk)})^2}{\sum_{k=1}^{g} Var(v_i^{(kk)})} \tag{7}$$

For every sample i, the right side of (7) is known except for $Var(v_i^{(kk)})$ for $k = 1, \ldots, g$. We can approximate this value from the data as follows:

1. Get N estimates of the variance of the k-th gene by leaving out one sample at a time: $\{\eta_1^k, \eta_2^k, \ldots, \eta_N^k\}$, where η_j^k denotes the variance of k-th coordinates of $\{x_1, \ldots, x_N\} \setminus x_j$.
2. Estimate $\eta^{(k)} = \frac{1}{N}\sum_{j=1}^{N}\left(\eta_j^k - \frac{1}{N}\sum_{l=1}^{N}\eta_l^k\right)^2$, the variance of $\{\eta_1^k, \eta_2^k, \ldots, \eta_N^k\}$.

Replacing $Var(v_i^{(kk)}) = \eta^{(k)}$ on the right-hand side of equation (7) gives us a data-driven estimate of the prior degrees of freedom ν_i. Thus, the estimate of the hyperparameter δ_i becomes

$$\delta_i = \frac{1}{\nu_i - g} = 1/\left[3 + \frac{2\sum_{k=1}^{g}(s_i^{(kk)})^2}{\sum_{k=1}^{g}\eta^{(k)}}\right], \quad \forall i \in \{1, \ldots, N\} \tag{8}$$

In section S1.2.6 of the Supplementary Material, we use simulation experiments to illustrate that this data-driven empirical approach of calibrating δ_i separately for each individual sample i, delivers better or comparable performance than assigning a fixed value of $\delta_i = \delta, \forall i$ for all individuals. This empirical data-driven approach for calibrating the degrees of freedom parameters for the individual-specific co-expression matrices allows BONOBO to more accurately model how individual sample co-expression deviates from the population co-expression, thereby more effectively capturing heterogeneity in gene-gene interaction patterns across the population. For a mathematical explanation of why this data-driven approach is effective, see Supplementary Material S1.2.6.

Hypothesis testing

For every sample i, we can derive a $100(1-\alpha)\%$ posterior credible region for the correlation between any pair of genes as follows: first, we compute the posterior variance of the covariance between any pairs of genes using the following lemma, which is a direct consequence of the properties of inverse Wishart distribution (Zhang 2021). For simplicity, we remove the sample index i.

Lemma 2 *Let v_{jk} denote the covariance between the j-th and the k-th gene. Under assumptions* (1) *and* (2)*, the posterior variance of v_{jk} is*

$$Var(v_{jk}) = \frac{(\nu - g + 1)s_{jk}^2 + (\nu - g - 1)s_{jj}s_{kk}}{(\nu - g)(\nu - g - 3)}, \tag{9}$$

where s_{kk} denotes the k-th diagonal entry of the prior mean S and s_{jk} denotes the (j,k)-th off-diagonal entry (corresponding to the j-th row and the k-th column) of S.

Using the above lemma we can compute an approximate $100(1-\alpha)\%$ posterior credible region for v_{jk} as $\left(\sigma_{jk} - \psi_{jk} z_{(1-\alpha/2)}, \sigma_{jk} + \psi_{jk} z_{(1-\alpha/2)}\right)$, where $\psi_{jk} = \sqrt{Var(v_{jk})}$, σ_{jk} is the posterior mean of v_{jk} and $z_{(1-\alpha/2)}$ denotes the $(1-\alpha/2)$-th quantile of the standard normal distribution.

We use the Central Limit Theorem (CLT) to derive the posterior credible regions. Even though the posterior distribution of each individual-specific co-expression matrix is known, the closed-form expressions of the quantiles of Inverse Wishart distribution are theoretically intractable, and unlike many other distributions, such as Gaussian, no software implementation is available for Inverse Wishart quantiles. To construct a posterior credible region for each individual, we need to simulate a large number of samples from the corresponding Inverse Wishart posterior, which is computationally expensive. Moreover, this credible region constructed from simulated observations would give us the joint credible region for the correlation between all pairs of genes and not the marginal credible region for each pair of genes, which we are interested in estimating as they are biologically more interpretable. Therefore we use the CLT to derive a computationally inexpensive credible region.

In low dimensional settings, Bayesian credible regions can be used as frequentist confidence intervals, as justified in Section 10.2 of (Van der Vaart 2000). Thus we can reject the null hypothesis $H_0 : v_{jk} = 0$, in favor of the alternative hypothesis $H_1 : v_{jk} \neq 0$ at significance level α, provided $2\left(1 - \Phi(\sigma_{jk}/\psi_{jk})\right) \leq \alpha$, where Φ denotes the cumulative distribution function of the standard normal distribution. In practice, BONOBO derives a dense (complete) network with edges between every pair of genes, where edge weights correspond to σ_{jk}, the posterior mean of the covariance between genes j and k. We can generate a sparse covariance network by simply pruning out edges for which the $100(1-\alpha)\%$ posterior credible regions contain zero for a suitable value of $0 < \alpha < 1$.

Computing co-expression from covariance

For every sample i, BONOBO first computes a sample-specific covariance matrix V_i following the above-mentioned procedure. These sample-specific covariance matrices can be subsequently converted to the corresponding sample-specific correlation (generally co-expression) matrices as follows. For each pair of genes, we can compute the sample-specific co-expression simply by dividing the covariance between the two genes by the product of the standard deviations of these same two genes. Formally speaking, for sample i, the correlation between genes j and k would be $\frac{v_{jk}^{(i)}}{\sqrt{v_{jj}^{(i)} v_{kk}^{(i)}}}$, where $v_{jk}^{(i)}$ denotes the (j,k)-th entry of the covariance matrix V_i.

Results

Simulated data and comparison with other methods

Although we recognize that simulated gene regulatory networks may not capture the full complexity of the gene expression and the effects of regulation, simulation is an important tool as it provides a measure of "ground truth" against which various methods can be rigorously benchmarked and compared. We performed five simulation experiments and compared BONOBO with three other methods for computing sample-specific co-expression: LIONESS (here used only in the LIONESS::Pearson configuration), SPCC, and SWEET. We repeated each of the simulation experiments 100 times and compared BONOBO with LIONESS::Pearson, SPCC, and SWEET based on the mean sum of squared errors (the squared Frobenius norm of the difference between the true correlation matrix and the estimated correlation matrix) across these 100 iterations.

First, we examined how the performance of BONOBO compares with the competing methods for varying sample sizes, based on 100-dimensional gene expression data generated from a homogeneous population (Supplementary Materials S1.2.1). We found that as sample size increases from 10 through 1000, the mean squared error (MSE) declines for BONOBO, LIONESS::Pearson, and SWEET, and remains unchanged for SPCC (Figure 2A). For any sample size, BONOBO outperforms all competing methods, although the performance improvement over SWEET is much smaller compared to the performance gains over LIONESS::Pearson and SPCC. However, upon closer inspection (Figure S1 top left) one can see that BONOBO not only provides a steady decrease in MSE over SWEET across all sample sizes, but as sample size increases, BONOBO exhibits increasingly better performance compared to SWEET by providing up to 40% decrease in MSE for sample size = 1000.

Next, we simulated samples from a mixture of two different homogeneous populations (Supplementary Materials S1.2.2) and demonstrated that the performance of BONOBO, along with SWEET and LIONESS::Pearson, declines as the proportion of samples from the smaller sub-population increases, meaning that the overall population becomes more diverse (Figure 2B). For any given mixture proportions of subpopulations, the MSEs for BONOBO were much smaller than those for LIONESS::Pearson and SPCC, while SWEET performed slightly worse than BONOBO (Figure S1 top right).

Next, we compared the performance of BONOBO to the other methods by keeping sample sizes fixed at $n = 100$ and varying the number of genes from 100 to 1500. We examined two different situations, the first using samples drawn from a homogeneous population (Supplementary Materials S1.2.3) and the second in which samples are drawn with a $20 : 80$ mix from two homogeneous sub-populations (Supplementary Materials S1.2.4) One can see that the performance of all four methods declines as the number of genes increases (Figure 2C-D). However, for any fixed number of genes, BONOBO

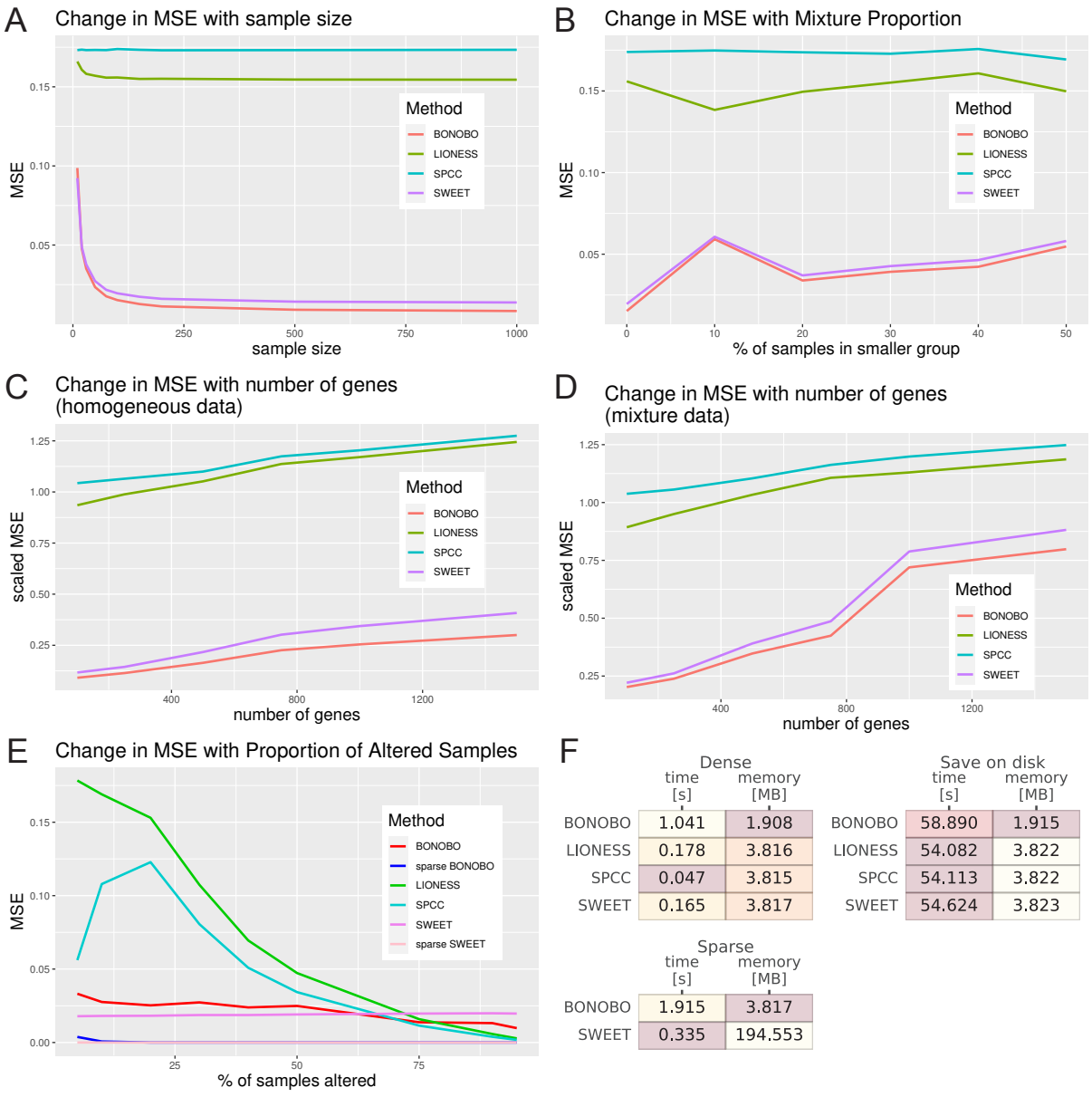


Figure 2: Legend in next page

Figure 2: **Comparisons between BONOBO and other methods for the estimation of sample-specific co-expression on simulated data**. Performance of BONOBO, LIONESS::Pearson, SPCC, and SWEET on simulated data: (A) simulated data from homogeneous population: Change in MSE with respect to sample size. (B) Simulated data from a mixture of two populations: Change in MSE with respect to the percentage of samples in the smaller population. (C) Simulated data from a homogeneous population: change in scaled MSE (MSE divided by the variance) with respect to the number of genes. (D) Simulated data from a mixture of two populations, where 20% samples come from one population, and the remaining 80% samples come from another population with distinct mean and covariance of gene expression: change in scaled MSE with respect to the number of genes. (E) Simulated data where some samples lost expression of 1% genes: a change in MSE with respect to the proportion of altered samples. BONOBO outperforms all three competing methods by providing lower MSE across varying sample sizes, varying mixture proportions of sub-populations, and varying the number of genes. (F) Computational resources required to run each method: We report average time and memory consumption measured over 10 datasets, with 500 genes and 100 samples, darker colors correspond to better performance/lower resources. We first compare resources to infer "Dense" networks, the standard implementation of each method, and those needed to infer "Sparse" BONOBO and SWEET. Then, we compare the resources required to infer and save to disk each network, "Save on disk," which is a more realistic scenario when dealing with larger datasets. BONOBO consistently performs better in terms of required memory, while being slower than the others. However, we can see that when networks are saved on disk, which is a time-consuming task, BONOBO's disadvantage is greatly reduced.

has a smaller MSE compared to the other methods. Further, as the number of genes is increased, the difference between the MSE of BONOBO and the MSE of the closest competitor, SWEET, grows (Figure S1 bottom right) such that BONOBO's performance as we approach genome-scale analyses becomes substantial.

In the next example, we simulated samples from a mixture of two populations, where one population lost expression for 1% of genes. The sparse version of BONOBO identified this loss of expression (Supplementary Materials S1.2.5) with better accuracy compared to BONOBO without sparsity, as well as other non-sparse competitors including SWEET, LIONESS::Pearson and SPCC, while both sparse SWEET and sparse BONOBO had comparable performance (Figure 2E).

Additionally, we used two simulation examples to illustrate (Supplementary Materials S1.2.6) that the data-driven approach of calibrating hyperparameters δ_i for each sample i described in section entitled "Fixing Prior Degrees of Freedom," provides a better estimation accuracy compared to the performance obtained by choosing a fixed value of $\delta_i = \delta, \forall i$ for all samples.

We also compared the computational performance of the four methods. We generated 100 simulated datasets with 500 genes and 100 samples each, applied BONOBO, LIONESS::Pearson, SPCC, and SWEET to each dataset (see Supplementary Materials S1.3), and measured how much time and memory each method required on average (Figure 2F). We found BONOBO to be more memory effi-

cient than the other methods, as it only needs to store one gene-by-gene matrix in memory. For the sparse case, BONOBO must store another gene-by-gene p-value matrix, effectively doubling the memory requirement. SWEET uses an even more inefficient strategy, simultaneously keeping all networks in memory. When considering the speed of execution, BONOBO was slower than the other methods, but this is somewhat mitigated by the fact that saving each network on disk, which is necessary when dealing with real data, adds a considerable time over-head to the computation (additional 50 seconds), hence reducing the speed advantage of SWEET, LIONESS::Pearson, and SPCC. For completeness, we repeated the same analyses for varying numbers of genes and samples (Supplementary Figure S3, S4, S5) confirming the trend.

The simulated datasets we analyzed resemble many scenarios that are encountered in the analysis of biological datasets, including different population sizes, gene knockouts and silencing, and mixtures of subpopulations. In all these conditions, BONOBO performs as well as, or better than, LIONESS::Pearson, SPCC, and SWEET. Further, although the relative performance advantage of BONOBO in some conditions is minor, and its run time is greater than the other methods, BONOBO is more memory-efficient, making it scalable for use with genome-wide correlation networks in situations where memory limitations would prevent use of the other methods.

BONOBO recovers sample-specific network structure in yeast datasets

Having assessed the performance of BONOBO relative to other methods for inferring single-sample correlation networks using simulated data, we then explored its performance in analyzing correlation networks in real-world data. *Saccharomyces cerevisiae* (brewer's yeast) has been extensively used to explore gene function and to explore various network models (Hahn & Young 2011, Costanzo et al. 2016, Teixeira et al. 2018, Reimand et al. 2010, Hackett et al. 2020, Costanzo et al. 2021), and so we chose expression data from two experiments to test how well BONOBO could find biologically relevant changes in gene expression correlation patterns. As a baseline test, we used BONOBO to estimate correlation networks using microarray expression data from synchronized yeast cells at 48 time points (Pramila et al. 2006) and found that BONOBO was able to accurately identify both cyclic fluctuations in cell cycle stage transition pathways and to identify biologically meaningful changes in functionally relevant groups of genes (see Supplementary Materials S1.4).

We then turned our attention to a yeast gene perturbation experiment in which eleven transcription factors were knocked out (11 KO strains) and each was grown in twelve different growth media (Jackson et al. 2020). The 11 TF knockout (KO) targets include known regulators of the Nitrogen Catabolite Repression (NCR) pathway, the General Amino Acid Control (GAAC) pathway, the Ssy1-Ptr3-Ssy5-sensing (SPS) pathway, and the retrograde pathway; the twelve different growth media represent a number of

nitrogen and carbon sources (Supplementary Materials S1.5). Because the original data came from a single-cell RNA-seq analysis we created pseudo-bulked expression data for the 132 KO/media combinations. Jackson and colleagues reported that the transcriptional effects of changing growth medium (and carbon or nitrogen source) had a greater effect than did the individual TF knockouts. This is expected since the nutrients included or excluded from each growth medium will specifically activate genome-wide metabolic pathways; while knocking out one TF is likely to produce localized changes among that TF's gene targets.

Consistent with expectations, we found that BONOBO's networks were far more similar for yeast in the same growth medium irrespective of TF KO than for strains with the same KO irrespective of growth media (see Figure S9). Going beyond this qualitative observation, we can use this hypothesis, that is same-media networks should be more similar to each other than the other networks, as "ground truth" to assess the relative performance of BONOBO and other single-sample correlation network estimation methods. To do this, we used multiple quantitative methods, including clustering, correlation, and distance metrics, to assess the global similarities between sample-specific networks (Supplementary Materials S1.5.1).

As in the simulations, we computed both sparse and dense networks, and then computed pairwise correlations and distances between all the networks derived for each sample (Figure 3A) and used a number of clustering methods (*k*-means, spectral, and agglomeration clustering) to determine how well correlation matrices inferred using each method reflected the growth media effect (Figure 3B-C). We found that BONOBO more consistently detected these growth media similarity patterns than did the other methods and that the sparsification statistical framework we described provides more manageable networks while maintaining strong performance (Figure 3, S11, and S12).

None of the single-sample correlation network methods were able to pick out the weak similarities caused by TF KOs as they were masked by the signal from the growth media (Figures S10, S11 and S12). However, we reasoned that if we controlled the analysis for the effect of the growth media, we could discern the effects of TF KO. Looking at each media individually, we tested whether genetically perturbed samples showed stronger perturbations in the edges connected to the knocked-out TF. For each growth medium and each TF, we selected only the edges with the highest weights and tested whether these edges in the wild-type were significantly different compared to the perturbed samples using Kolmogorov-Smirnov tests. We summarized these results by computing the F1 score; where for each TF KO, we computed the "true positive" (TP) rate as the proportion of edges connected to that specific TF that are significantly different between wild-type and the perturbed samples with that specific TF knocked out (Section S1.5, Figure S16). We found that although all methods were able to detect KO effects with generally comparable performance, BONOBO networks were consistently better at detecting changes resulting from the knockout of the TFs *GCN4, GAT1, GLN3, and RTG3*.

It is worth noting that the BONOBO model explicitly assumes that log-transformed gene expression values have a Gaussian distribution. However, after using the Kolmogorov-Smirnov test, we were unable to reject the Gaussian distribution null hypothesis for only 66.61% of genes (Figure S25 right panel). Nevertheless, BONOBO performed at least comparably to, and sometimes even better than the other methods in all the tests we performed using yeast data.

Finally, we wanted to understand what BONOBO's sample-specific co-expression networks can reveal about how a TF KO affects the overall network (Figure S17). This is a challenging task since we have shown that the effect of changing growth medium is greater in most instances than knocking out a TF. The one possible exception in this dataset is the deletion of *GCN4,* which produced easily detectable changes in the correlation patterns of the genes that are most strongly correlated with *GCN4* across samples (Figure S18). We selected the 100 genes whose correlation edge weights differed the most between the *GCN4* knockout and the rest of the dataset. As expected, genes that were most perturbed by *GCN4* deletion, which targets the GAAC pathway, belong to many pathways related to autophagy, such as exosome, phagosome, and autophagy - yeast (Figure S19). Among the most significantly perturbed genes were known *GCN4* targets (Wong et al. 2023), including *VCX1, MNN10, ACT1, CPA1* (Coey & Clark 2022, Uluisik et al. 2011), and *CCW12* (Rawal et al. 2018), demonstrating that perturbation-specific co-expression networks estimated using BONOBO can be used to detect key regulatory targets of the TF.

miRNA-gene interaction in breast cancer subtypes

We also wanted to explore the application of BONOBO to multi-omic datasets while accounting for sample-specific clinical and molecular confounders. MicroRNAs (miRNAs) play an important regulatory role through RNA silencing, resulting in the down-regulation of their target genes; changes in miRNA expression levels and their downstream effects have been shown to play a role in the cancer development and are predictive of outcomes in some cancers (Oliveto et al. 2017). We calculated co-expression networks using paired mRNA and miRNA expression data from 101 breast cancer samples representing the five canonical molecular subtypes (GEO accession number GSE19783 (Enerly et al. 2011, Aure et al. 2013, Haakensen, Steinfeld, Saldova, Shehni, Kifer, Naume, Rudd, Borresen-Dale & Yakhini 2016), pre-processing steps in Supplementary Materials S1.6.1). We then used BONOBO to infer sample-specific co-expression networks to explore how correlation between miRNA and gene expression vary between breast cancer subtypes (normal-like, Luminal A, Luminal B, ERBB2/HER2-positive, and Basal) and whether patterns of co-expression can be used to predict survival outcomes.

BONOBO produced 101 sample-specific co-expression networks, each of which includes correlations between pairs of genes, pairs of miRNAs, and between each gene and each miRNA. We compared

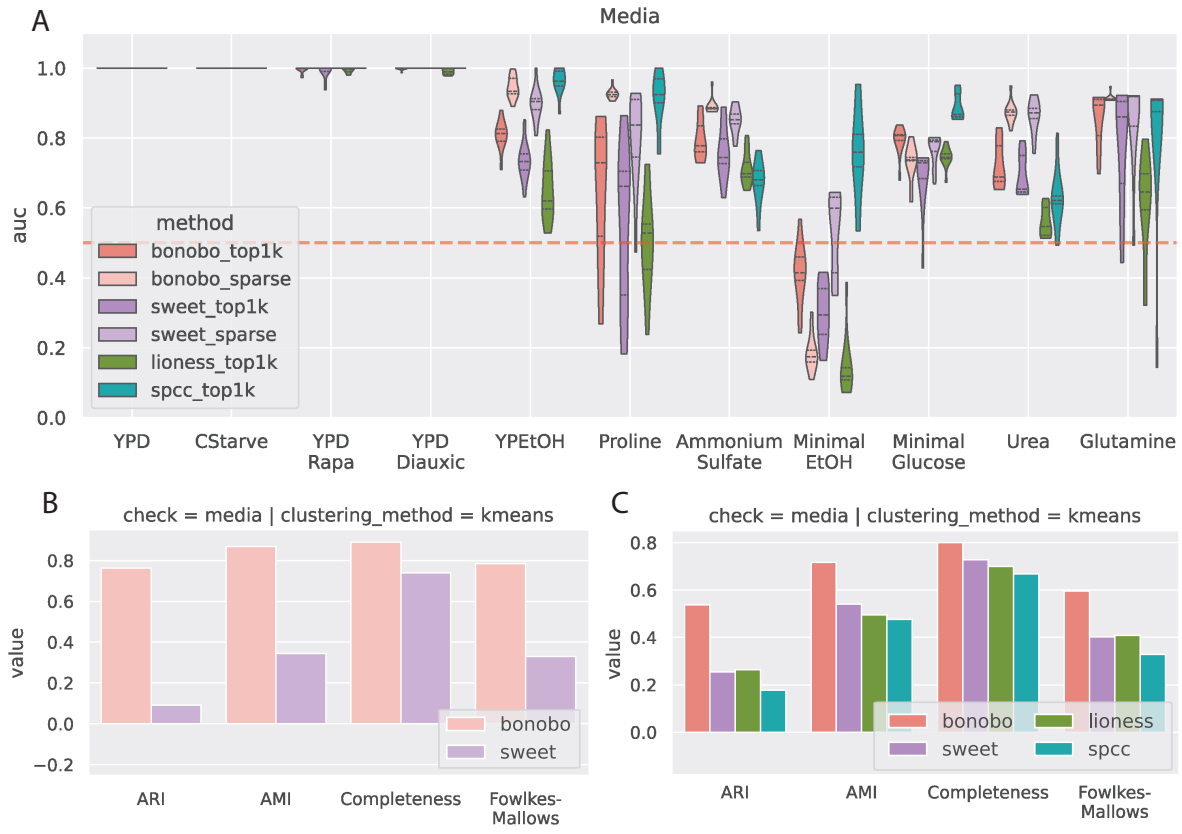


Figure 3: **Comparison between methods for detecting media perturbations in yeast experiments**. We used only the edges returned by the sparsified versions of SWEET and BONOBO and the 1000 strongest edges, on average, for each method (top1k labels). A) Network similarity between perturbed yeast samples. For each sample, we computed Pearson correlation coefficients with all the other samples–which is always positive here–and estimated the ROC curve using a binary label that is 1 for the samples grown in the same media. For each comparison, we computed the AUROC (y-axis), grouping the samples by the growth medium (x-axis), and we compared the results between different methods. $auc = 1$ means that samples in the same media have the highest correlation values compared to samples in different media, $auc = 0.5$ (red dashed line) is the "random" performance, which means that samples with the same label are not more similar to each other than the rest. B) Clustering performance on sparse BONOBO and SWEET. For all networks, we used *k*-means clustering and evaluated how well the resulting clusters captured growth media similarity using four different metrics: adjusted rand score (ARI), adjusted mutual info score (AMI), completeness, Fowlkes-Mallows Score. Greater values indicate that the clusters group samples grown in the same media. C) Clustering performance on top1k networks using the same protocol as in B. Sparse networks, both for SWEET and BONOBO outperform the naive "top1k" networks, with BONOBO outperforming SWEET. While for some growth media, SPCC has a greater auc than BONOBO, BONOBO outperforms SPCC in the clustering test.

the miRNA-gene edges between subtypes. For each network, we ranked the genes by their "miRNA-specific degree" and performed pre-ranked gene set enrichment analysis (Supplementary Materials S1.6.2). A gene's "miRNA-specific degree" is calculated by summing all co-expression values between that gene and each of the miRNAs. For instance, if for a particular sample, $\{\rho_{g_j m_1}, \rho_{g_j m_2}, \ldots, \rho_{g_j m_K}\}$ represent the correlations between gene g_j with miRNAs $m_1, \ldots, m_K$ respectively, then the total correlation between gene g_j with miRNAs is computed as $\sum_{k=1}^{K} \rho_{g_j m_k}$. By ranking genes in this way and performing pathway enrichment analysis, we can identify those biological pathways that are either significantly positively or negatively correlated with miRNA expression.

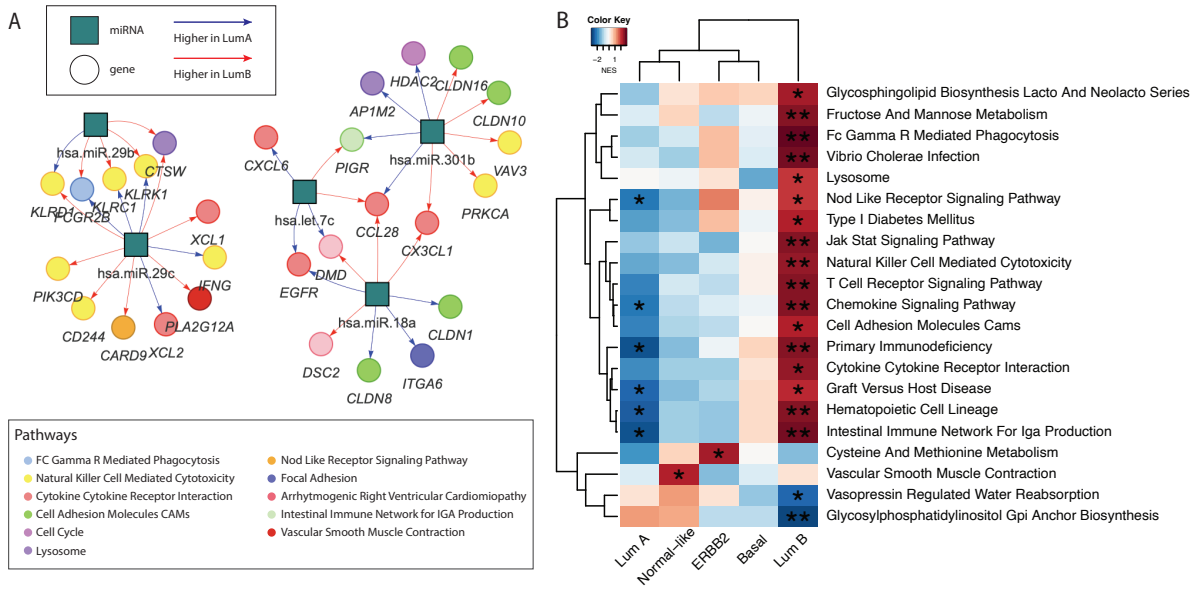


Figure 4: **BONOBO networks reveal distinct miRNA-mRNA co-expression patterns in different breast cancer subtypes.** A) Pairs of genes and miRNAs that are most differentially co-expressed between Luminal A and Luminal B subtypes: gene nodes are represented by colored circles where the colors correspond to the biological pathway associated with that gene; miRNAs are represented by teal squares. Edges are colored red if they have higher weights in Luminal A compared to Luminal B and blue if its miRNA-specific degree is greater in Luminal B than Luminal A. B) Pathways for which the mean correlation with miRNAs is significantly (at significance level 0.05) associated with survival in Cox proportional hazard model: heatmaps are colored by t-statistic of the pathway score in the Cox model. Rows correspond to pathways and columns represent the breast cancer subtypes. Pathways for which a higher correlation with miRNAs is associated with better survival are colored blue, and pathways for which a higher correlation with miRNAs is associated with worse survival are colored red.

We found that pathways associated with immune response, including Graft vs Host disease, primary immunodeficiency, cytokine-cytokine receptor interaction, and natural killer cell-mediated cytotoxicity, were significantly negatively correlated (FDR cutoff 0.05) with miRNA expression across all breast cancer subtypes (Figure S20). Pathways associated with cell adhesion and cell proliferation, such as

focal adhesion and ECM receptor interaction, were positively correlated with miRNA expression, especially in the Basal, Normal-like, and Luminal B subtypes, while pathways associated with cell adhesion molecules were significantly negatively correlated (FDR cutoff 0.05) with miRNA expression in ERBB2, Luminal A, and Normal-like breast cancer. These findings are consistent with previous studies (Enerly et al. 2011) that reported significant associations between the expression levels of several miRNAs and biological pathways involved in cell proliferation, cell adhesion, and immune response.

Although the pathways most strongly correlated (positively or negatively) with miRNA expression were consistent between subtypes, when we more closely inspected the individual miRNA-mRNA edges in the networks, we found that there are significant differences between subtypes in the neighborhood of individual genes. For example, when we compared individual correlation edges between genes and miRNAs in Luminal A and Luminal B breast cancer (Figure 4A), we found several genes that were differentially correlated with miRNA expression between the subtypes. Genes most differentially co-expressed with miRNAs include the proto-oncogene *EGFR*; genes involved in immune response such as *PI3KCD and IFNG*; genes involved in cell-cell adhesion and cell proliferation such as *CLDN1, CLDN8, CLDN10, CLDN16* etc. These differences between miRNA-gene interactions highlight the distinct molecular landscapes of Luminal A and Luminal B (Haakensen, Nygaard, Greger, Aure, Fromm, Bukholm, Lüders, Chin, Git, Caldas et al. 2016).

We next investigated if the strength of association between miRNA expression and biological pathways is predictive of survival in each of the subtypes. For every pathway significantly (FDR cutoff 0.05) correlated (positively or negatively) with miRNA expression, we computed a pathway score, defined by the mean of the total correlation between all miRNAs and each gene in that pathway. Then, for each pathway, we fit a Cox proportional hazard model, allowing the coefficients of the Cox model to be subtype-specific. We found that a high correlation between miRNA and immune-related pathways, including Chemokine signaling pathway, primary immunodeficiency, Graft vs. Host disease, Hematopoietic cell lineage, and intestinal immune network for IgA production, was associated with better survival among samples from Luminal A, but worse survival among samples from Luminal B subtype (Figure 4B). Previous studies (Arun et al. 2022) have reported that increased expression of miRNAs is associated with tumor suppression among Luminal A breast cancers, thus leading to slower tumor growth and improved prognosis. Our analysis indicates that an up-regulation of immune pathways by miRNAs might provide a possible mechanism to explain the tumor-suppressive effect of miRNAs in Luminal A. In Luminal B breast cancer, miRNAs may have a role in regulating immune evasion processes (Gomarasca et al. 2020), resulting in more aggressive tumor growth and poorer survival outcomes.

Previous studies (Zhu et al. 2019) have identified specific immune gene expression patterns that distinguish Luminal A from Luminal B breast cancer and showed that these distinct immune signatures were associated with a differential ratio between ESR1 and ESR2, a higher value of which has been

associated with poorer survival outcomes. Our results indicate that subtype-specific regulatory interactions between miRNAs and immune pathways in Luminal A breast cancers might contribute to estrogen receptor-mediated survival outcomes. We also found that the genes linked to cell proliferation pathways exhibit distinct patterns of regulation by miRNAs in Luminal A when compared to Luminal B breast cancer. Clinically, this disparity in the regulation of cell proliferation genes might be a crucial factor contributing to poorer prognosis in Luminal B breast cancer patients (Ades et al. 2014).

Sample-specific gene regulatory networks identify sex difference in thyroid

Many thyroid disorders exhibit sex-based differences in incidence, with females being significantly more susceptible to thyroid conditions than age-matched males (Vanderpump 2011). Although sex chromosomes, sex hormones, and the immune system (Shobab et al. 2022) have often been suggested as possible contributors to this sex difference in thyroid tissue, there are few published systems-based analyses exploring the molecular mechanisms driving the sex differences.

To fill this gap in our understanding of thyroid diseases, we used data from the Genotype-Tissue Expression (GTEx) Project (Lonsdale et al. 2013) (see Supplementary materials S1.7.1 for all preprocessing steps), to calculate individual-specific gene-gene correlation matrices using BONOBO. These co-expression matrices were subsequently combined with TF-motif prior regulatory networks (Supplementary materials S1.7.2), and TF protein-protein interaction data (Supplementary materials S1.7.3) as input to PANDA (Glass et al. 2013) to estimate 653 sample-specific gene regulatory networks (linking TFs to target genes) in healthy research subjects (Figure S21). We performed differential targeting analysis comparing male and female networks (Supplementary materials S1.7.4) and identified several genes with known relevance to thyroid cancers and autoimmune conditions (Figure S23).

For example, the long non-coding RNA *XIST* was more highly targeted in females. Previously, *XIST* has been reported to promote oncogenic activity in papillary thyroid carcinomas (Cai et al. 2023). Among genes highly targeted in males, the tumor suppressor gene *KDM6A* is known to regulate multiple genes involved in immune response, suggesting its potential influence on the risks of developing various autoimmune conditions (Itoh et al. 2019). Among the other genes more highly targeted in males are *PCM1*, *KMT2C*, and *SOS1*. A mutation in *PCM1* has been associated with papillary thyroid carcinoma (Eng et al. 2023), a mutation in the *KMT2C* gene has been identified as a molecular marker for primary thyroid osteosarcoma (Wang et al. 2022), and over-expression of *SOS1* has been found to promote cell proliferation and cell apoptosis in papillary thyroid carcinoma cells (Pang & Yang 2021).

Using genes differentially targeted between the sexes that were ranked by limma t-statistics, we performed gene set enrichment analysis using Gene Ontology Biological Processes (GOBP) annotations (see Supplementary Materials S1.7.5). Using genes more highly targeted in females, we found

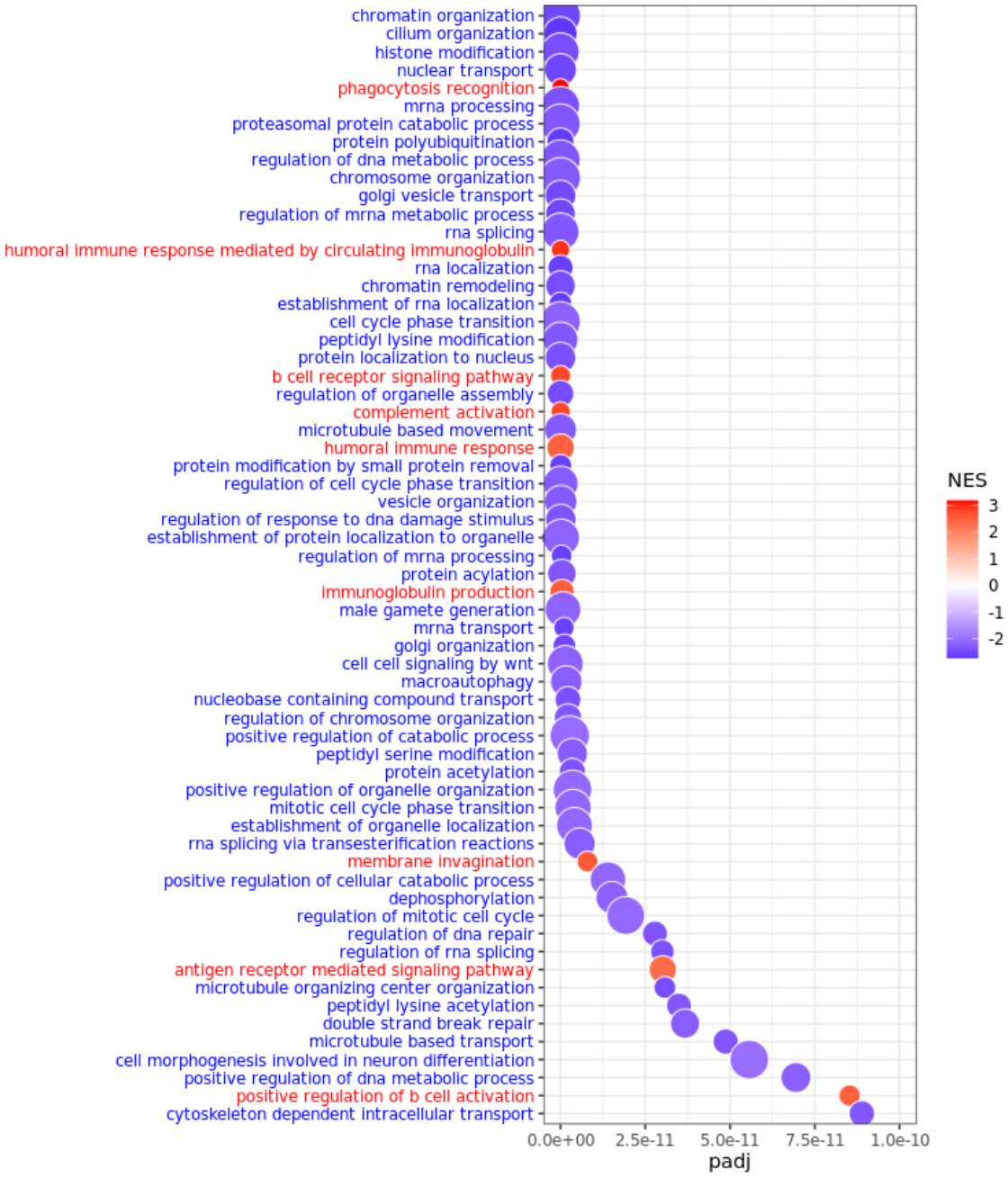


Figure 5: **Differential regulation between males and females in healthy thyroid samples**. GO biological processes most differentially regulated (at FDR cutoff 1×10^{-10}) in males and females in GTEx thyroid samples. Pathways are ordered from top to bottom in ascending order of adjusted p-values (padj) such that pathways represented at the top are most differentially targeted by transcription factors in males and females. Pathways highly targeted in males are marked blue and pathways highly targeted in females are marked red.

enrichment for biological processes associated with immune response, such as humoral immune response, B cell receptor signaling pathway, antigen-receptor mediated signaling, and positive regulation of the B cell activation pathway (Figure 5); higher relative targeting of immune pathways in females may contribute to their increased susceptibility to autoimmune thyroid diseases, including Hashimoto's Thyroiditis disease, which often leads to hypothyroidism.

For the genes more highly targeted in males, we found enrichment for processes associated with cell cycle, cell signaling, metabolic processes, and DNA repair (Figure 5); dysregulation of these pathways has been shown to play functional roles in various thyroid conditions, including Graves' Disease and Hashimoto's Thyroiditis (Zheng et al. 2022). While validation experiments are needed to determine the direction of these effects, the differential targeting of these pathways in males suggests that they may have a more robust defense against factors that could disrupt normal thyroid functioning.

To emphasize the advantage of using individual-specific networks as opposed to population-level networks for sex-difference analyses, we performed a gene set enrichment analysis based on sex-specific population-level gene regulatory networks constructed separately for males and females using PANDA. Comparing sex-biased biological pathways discovered by individual-specific BONOBO networks and pathways discovered by population-level sex-specific gene regulatory networks, we found twelve pathways, all associated with immune response (Table S2.2) for which the apparent direction of sex bias was different in the two methods. Examining individual-specific networks, we found that these pathways are highly targeted by TFs in females compared to males, whereas the comparison of sex-specific aggregate networks found the same pathways to be more highly targeted in males.

To understand the reason for this difference for the 12 functional classes, we looked at the average pathway-targeting score (defined as the mean of edge weights connecting TFs to all genes in these 12 pathways) and discovered that males have more outliers with higher targeting scores than females (Figure S24). It is the presence of these outliers that led to the conclusion that these processes are more highly targeted in males, when looking at aggregate networks, even though in reality the targeting scores in females are significantly higher than those in males (p-value of Wilcoxon rank sum test = 0.016).

In fact, greater targeting of immune-related pathways in females has been reported previously in other studies involving individual-specific regulatory networks (Lopes-Ramos et al. 2020, Saha et al. 2023). Biologically, it makes sense that females would have a higher targeting of immune-related pathways since greater targeting could make immune processes in females being more susceptible to disruption, giving rise to more frequent immune-related ailments in thyroid tissue such as Hashimoto's disease and other autoimmune conditions.

Overall, individual-specific gene regulatory networks derived by combining BONOBO with PANDA identify biological pathways that are differentially regulated by sex and are not discernible from the

analysis of population-level sex-specific networks. This sex-biased differential regulation of key genes and pathways might contribute to the differential risk of various thyroid conditions between biological males and females.

Discussion

Complex human traits and diseases are most often driven by not a single gene but rather by intricate interactions involving multiple genes and the transcription factors that regulate them. Most network inference methods estimate an aggregate network (Langfelder & Horvath 2008, Lemoine et al. 2021) that reflects the average characteristics of the population and ignores the diversity that exists in gene expression and regulation (Saha et al. 2023). Other methods that have been used to infer individual-specific co-expression networks (Kuijjer et al. 2019, Chen et al. 2023) can produce extreme edge weights that are difficult to interpret as correlations. In that sense, BONOBO fills an important gap in methods by estimating true correlation matrices that are positive semidefinite by definition (in the univariate setting, this translates to expression of individual genes having non-negative variance). Indeed, for the yeast KO experiment, SWEET, LIONESS::Pearson, and SPCC networks all have negative eigenvalues (Figure S8), representing negative variances of the principal components (PC) of gene expression. In contrast, since BONOBO estimated positive semi-definite co-expression matrices for every individual, any linear combination of genes, not just PCs, is ensured to have non-negative variances. Moreover, matrix operations, including matrix inversion and Cholesky decomposition, are interpretable and numerically stable only for positive semi-definite matrices. Thus, unlike other methods, co-expression matrices inferred by BONOBO can be inverted to compute individual-specific partial correlation networks, which are more effective in discerning direct and indirect interactions between pairs of genes, compared to correlation networks (Shutta et al. 2023). Furthermore, it has been shown that using correctly defined distances on a Riemannian manifold between the covariances from scRNA-seq data improves the discovery of differential gene programs (Ozbay et al. 2023). This approach, which leverages Riemannian geometry of positive semidefinite matrices, could be extended to detect differences between sample-specific networks derived by BONOBO and not any other existing competing methods. Lastly, because BONOBO is a probabilistic model, one can generate synthetic samples for every individual, which can further be used for sample-specific inference in other contexts.

Constructing individual-specific co-expression networks is challenging as each individual typically has only a single sample. Bayesian statistics enables us to overcome this problem by incorporating prior information derived from other individuals in the same dataset. We assume that individual-level covariance matrices come from an inverse Wishart prior whose mean is equal to the sample covariance matrix computed from all other individuals in the given dataset. This assumption is based on the fact

that samples in a bulk expression dataset typically come from a single tissue and from individuals having similar conditions. Integrating this prior information with the individual-level expression data, BONOBO can estimate the posterior distribution of the covariance matrix for each individual in the data. The mean of the posterior distribution is a weighted average of the deviation of the individual expression from the mean expression and the estimated population covariance from all the other individuals in the dataset.

BONOBO is also highly scalable due to the use of a conjugate prior distribution over individual-specific covariance matrices, which enables us to derive a closed-form expression of the posterior distribution, thus eliminating the need for running computationally expensive Markov chain Monte Carlo. In addition, the posterior distribution of the covariance matrix for each individual is computed separately, without any influence of the posterior distribution of other individuals. These properties, inherent to BONOBO, make the method highly parallelizable and enhance computational efficacy.

BONOBO is based on minimal assumptions and only one tuning parameter that can be efficiently calibrated using a data-driven approach. BONOBO assumes that for every individual in the data, the log-transformed expression values of genes follow multivariate Gaussian distribution with a covariance matrix unique to every sample. Through simulated examples in which the study population is a mix of two different populations with different mean expression and patterns of co-expression, we show that BONOBO outperforms other methods in estimating sample-specific co-expression even when the underlying assumption of multivariate normality is violated, thus making the method adaptable to a wide range of applications.

In developing BONOBO, we also assumed that the covariance matrix for a particular individual follows an inverse-Wishart distribution, which is the multivariate analog of inverse-Gamma, the most widely used prior distribution for the univariate variance parameter. The prior mean of the individual's covariance is assumed to be equal to the population covariance matrix estimated using all other samples in the data, which is an unbiased estimate of the population covariance. In BONOBO, the prior degrees of freedom, which characterize how much the individual covariance might vary from the population covariance, is estimated from the data empirically to capture how different the individual of interest is from the rest of the population; if that individual is an outlier, their covariance is allowed to deviate further from the population covariance. If the individual is more similar to the rest of the population, their covariance is assumed to deviate less from the population covariance. Consequently, the assumption of an inverse-Wishart prior distribution on the individual-specific covariance matrices is not too stringent for practical purposes.

It is important to note that benchmarking single-sample networks is particularly challenging (De Marzio et al. 2023). Since our goal is to estimate sample-specific co-expression, we simulated multiple gene expression samples for every individual from a multivariate normal distribution and used these samples to calculate individual-specific co-expression matrices. We showed that, for these data, the MSE of

the sample-specific co-expression networks estimated by BONOBO was much lower than that of the other methods. Further, as the dimensionality of the estimation problem increased, BONOBO provided increasingly better performance than its closest competitor SWEET. One potential reason could be that BONOBO assumes that individual-specific co-expression matrices follow an Inverse-Wishart distribution with a prior mean equal to the co-expression computed from all other samples in the dataset, implying that individual-specific co-expression is *a priori* assumed to be centered around the population's aggregate co-expression pattern. Intuitively, this is a reasonable assumption as a sample-specific co-expression network should only deviate slightly from an aggregate co-expression network.

However, assuming that the sample-specific correlation is centered around the aggregate correlation may also impact our results by reducing the overall variability of the estimated networks. When we applied BONOBO to the yeast data both for the cell cycle and the TF knockout perturbation experiments, we found high values of correlation (with pairwise correlation > 0.99 in most cases) across all samples. This could present problems when using BONOBO to detect outlying samples or when it is used in highly heterogeneous populations. In these situations, other methods, such as LIONESS::Pearson, might be more appropriate for identifying samples that are distant from the aggregate estimate.

Methodologically, this situation can be mitigated by imposing a non-informative prior on the individual-specific co-expression matrix, such as the Lewandowski-Kurowicka-Joe distribution with parameter 1 ($LKJ(1)$) (Lewandowski et al. 2009), which puts uniform probability over the space of all correlation matrices. However, a completely non-informative prior would not allow BONOBO to borrow information from other samples. A possible trade-off would be a mixture prior distribution $\omega LKJ(1) + (1 - \omega)InvWishart(V_i, \nu_i)$, where $InvWishart((\nu - g - 1)S_i, \nu_i)$ is the prior used by BONOBO and $0 < \omega < 1$ is a mixture weight that captures how far the outlier is from the center of the population. This prior specification is not conjugate to multivariate normal covariance, and therefore the posterior distribution would not have a closed-form expression, increasing the computational complexity of the estimation algorithm.

We also note that BONOBO has some intrinsic limitations. Although we applied BONOBO exclusively to transcriptomics data, the model could be adapted to other omic data types (such as epigenomics or proteomics) with minor modifications—provided the data can be suitably transformed to resemble a unimodal distribution. However, it is important to note that BONOBO is not applicable for omics modalities characterized by binary or categorical data, such as mutation profiles. A potential area for future research could involve extending our model to incorporate interactions across a broader range of omics data types through hierarchical latent variable models or association measures other than Pearson correlation. Extending BONOBO to other correlation measures, such as Spearman's rank correlation coefficient, would also allow us to overcome the limitations intrinsic to Pearson correlation, such as the fact that it is known to be heavily influenced by outliers. Extending BONOBO to other measures, such as Chatteerjee's correlation coefficient (Chatterjee 2021), could allow BONOBO to estimate

networks based on non-linear associations between the expression of genes. .

Finally, we recognize that in its current implementation, BONOBO does not control for false discovery rate (FDR) when computing the p-values for specific edges in individual co-expression networks. Theoretically, by including an edge with p-value bounded above by α, BONOBO includes an expected $\alpha g(g-1)/2$ false positive edges in the network, where g denotes the number of genes. We believe the inclusion of false positive edges might be problematic although to what extent BONOBO might be prone to false discovery is hard to verify due to the absence of ground truth for real data applications. In the next release of BONOBO, we plan to implement a Benjamini-Hochberg procedure for FDR correction and give users the option to choose between raw p-values and adjusted p-values.

Software availability

BONOBO is available through the Network Zoo package (netZooPy v0.10.0; netzoo.github.io), specific code can be found inside the `https://github.com/netZoo/netZooPy/tree/master/netZooPy/bonobo` folder. Data and code to reproduce the simulation experiments and real data analysis are available at Harvard Dataverse (doi: https://doi.org/10.7910/DVN/CAVAAB). We have also included a copy of the code in the supplemental material of this manuscript. A tutorial on generating and analyzing BONOBO networks on yeast is available on Netbooks (version 2.3.4) (Ben Guebila et al. 2022): `https://netbooks.networkmedicine.org/` under the title "Estimating single-sample co-expression networks for yeast genetic screens using BONOBO". Finally, BONOBO networks for the breast cancer miRNA application, and single sample PANDAs for thyroid samples are accessible on GRAND (version 1.6) (`https://grand.networkmedicine.org/` (Ben Guebila et al. 2021))

Competing Interest Statement

The authors declare no competing interests.

Acknowledgements

This work was supported by grants from the National Institutes of Health: ES, CMLR, MBG, VF, JF, KHS, PM, and JQ were supported by R35CA220523; MBG and JQ were also supported by U24CA231846; JQ received additional support from P50CA127003; JQ and DLD were supported by R01HG011393; KHS and DLD were supported by P01HL114501; DLD is also supported by K24HL171900; KHS was

supported by T32HL007427; CMLR was supported by K01HL166376; CMLR and ES were also supported by the American Lung Association grant LCD-821824.

Author Contributions

Conceptualization: ES, VF, PM, MBG, JF, KHS, DLD, CMLR and JQ; **Methodology:** ES and VF; **Formal Analysis:** ES, VF and PM; **Investigation:** ES and VF; **Resources:** JQ and MBG; **Data Curation:** ES, VF and CMLR; **Writing – Original Draft:** ES and VF; **Writing – Review and Editing:** PM, MBG, JF, KHS, DLD, CMLR and JQ; **Visualization:** ES, VF and PM; **Supervision:** JQ, CMLR, and DLD; **Funding Acquisition:** JQ, CMLR, and DLD

References

Ades, F., Zardavas, D., Bozovic-Spasojevic, I., Pugliano, L., Fumagalli, D., De Azambuja, E., Viale, G., Sotiriou, C. & Piccart, M. (2014), 'Luminal b breast cancer: molecular characterization, clinical management, and future perspectives', *Journal of clinical oncology* **32**(25), 2794–2803.

Arun, R. P., Cahill, H. F. & Marcato, P. (2022), 'Breast cancer subtype-specific mirnas: Networks, impacts, and the potential for intervention', *Biomedicines* **10**(3), 651.

Aure, M. R., Steinfeld, I., Baumbusch, L. O., Liestøl, K., Lipson, D., Nyberg, S., Naume, B., Sahlberg, K. K., Kristensen, V. N., Børresen-Dale, A.-L. et al. (2013), 'Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data', *PloS one* **8**(1), e53014.

Becker, M., Nassar, H., Espinosa, C., Stelzer, I. A., Feyaerts, D., Berson, E., Bidoki, N. H., Chang, A. L., Saarunya, G., Culos, A. et al. (2023), 'Large-scale correlation network construction for unraveling the coordination of complex biological systems', *Nature Computational Science* pp. 1–14.

Ben Guebila, M., Lopes-Ramos, C. M., Weighill, D., Sonawane, A. R., Burkholz, R., Shamsaei, B., Platig, J., Glass, K., Kuijjer, M. L. & Quackenbush, J. (2021), 'GRAND: a database of gene regulatory network models across human conditions', *Nucleic Acids Research* **50**(D1), D610–D621.
URL: *https://doi.org/10.1093/nar/gkab778*

Ben Guebila, M., Wang, T., Lopes-Ramos, C. M., Fanfani, V., Weighill, D., Burkholz, R., Schlauch, D., Paulson, J. N., Altenbuchinger, M., Shutta, K. H. et al. (2023), 'The network zoo: a multilingual package for the inference and analysis of gene regulatory networks', *Genome Biology* **24**(1), 45.

Ben Guebila, M., Weighill, D., Lopes-Ramos, C. M., Burkholz, R., Pop, R. T., Palepu, K., Shapoval, M., Fagny, M., Schlauch, D., Glass, K., Altenbuchinger, M., Kuijjer, M. L., Platig, J. & Quackenbush, J. (2022), 'An online notebook resource for reproducible inference, analysis and publication of gene regulatory networks', *Nature Methods* **19**(5), 511–513.

Cai, T., He, Y. & Peng, B. (2023), 'Incrna xist stimulates papillary thyroid cancer development through the mir-330-3p/pde5a axis', *Critical Reviews™ in Eukaryotic Gene Expression* **33**(3).

Chatterjee, S. (2021), 'A new coefficient of correlation', *Journal of the American Statistical Association* **116**(536), 2009–2022.

Chen, H.-H., Hsueh, C.-W., Lee, C.-H., Hao, T.-Y., Tu, T.-Y., Chang, L.-Y., Lee, J.-C. & Lin, C.-Y. (2023), 'Sweet: a single-sample network inference method for deciphering individual features in disease', *Briefings in Bioinformatics* **24**(2), bbad032.

Coey, C. T. & Clark, D. J. (2022), 'A systematic genome-wide account of binding sites for the model transcription factor Gcn4', **32**(2), 367–377.

Costanzo, M., Hou, J., Messier, V., Nelson, J., Rahman, M., VanderSluis, B., Wang, W., Pons, C., Ross, C., Ušaj, M., San Luis, B.-J., Shuteriqi, E., Koch, E. N., Aloy, P., Myers, C. L., Boone, C. & Andrews, B. (2021), 'Environmental robustness of the global yeast genetic interaction network', *Science* **372**(6542), eabf8424.
URL: *https://www.science.org/doi/10.1126/science.abf8424*

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D. et al. (2016), 'A global genetic interaction network maps a wiring diagram of cellular function', *Science* **353**(6306), aaf1420.

Dawson, J. A., Ye, S. & Kendziorski, C. (2012), 'R/ebcoexpress: an empirical bayesian framework for discovering differential co-expression', *Bioinformatics* **28**(14), 1939–1940.

De Marzio, M., Glass, K. & Kuijjer, M. L. (2023), 'Single-sample network modeling on omics data', *BMC biology* **21**(1), 296.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.-K., Aure, M. R., Russnes, H. G., Rønneberg, J. A., Johnsen, H., Navon, R., Rødland, E. et al. (2011), 'mirna-mrna integrated analysis reveals roles for mirnas in primary breast tumors', *PloS one* **6**(2), e16915.

Eng, Z. H., Abdullah, M. I., Ng, K. L., Abdul Aziz, A., Arba'ie, N. H., Mat Rashid, N. & Mat Junit, S. (2023), 'Whole-exome sequencing and bioinformatic analyses revealed differences in gene mutation profiles

in papillary thyroid cancer patients with and without benign thyroid goitre background', *Frontiers in Endocrinology* **13**, 1039494.

Glass, K., Huttenhower, C., Quackenbush, J. & Yuan, G.-C. (2013), 'Passing messages between biological networks to refine predicted interactions', *PloS one* **8**(5), e64832.

Gomarasca, M., Maroni, P., Banfi, G. & Lombardi, G. (2020), 'microrna s in the antitumor immune response and in bone metastasis of breast cancer: from biological mechanisms to therapeutics', *International Journal of Molecular Sciences* **21**(8), 2805.

Haakensen, V. D., Nygaard, V., Greger, L., Aure, M. R., Fromm, B., Bukholm, I. R., Lüders, T., Chin, S.-F., Git, A., Caldas, C. et al. (2016), 'Subtype-specific micro-rna expression signatures in breast cancer progression', *International journal of cancer* **139**(5), 1117–1128.

Haakensen, V., Steinfeld, I., Saldova, R., Shehni, A. A., Kifer, I., Naume, B., Rudd, P. M., Borresen-Dale, A.-L. & Yakhini, Z. (2016), 'Serum n-glycan analysis in breast cancer patients–relation to tumour biology and clinical outcome', *Molecular oncology* **10**(1), 59–72.

Hackett, S. R., Baltz, E. A., Coram, M., Wranik, B. J., Kim, G., Baker, A., Fan, M., Hendrickson, D. G., Berndl, M. & McIsaac, R. S. (2020), 'Learning causal networks using inducible transcription factors and transcriptome-wide time series', **16**(3), e9174.
URL: *https://www.embopress.org/doi/full/10.15252/msb.20199174*

Hahn, S. & Young, E. T. (2011), 'Transcriptional Regulation in Saccharomyces cerevisiae: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators', *Genetics* **189**(3), 705–736.
URL: *https://doi.org/10.1534/genetics.111.127019*

Itoh, Y., Golden, L. C., Itoh, N., Matsukawa, M. A., Ren, E., Tse, V., Arnold, A. P., Voskuhl, R. R. et al. (2019), 'The x-linked histone demethylase kdm6a in cd4+ t lymphocytes modulates autoimmunity', *The Journal of clinical investigation* **129**(9), 3852–3863.

Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R. & Gresham, D. (2020), 'Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments', *eLife* **9**, e51254.
URL: *https://doi.org/10.7554/eLife.51254*

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J. & Glass, K. (2019), 'Estimating sample-specific regulatory networks', *Iscience* **14**, 226–240.

Langfelder, P. & Horvath, S. (2008), 'Wgcna: an r package for weighted correlation network analysis', *BMC bioinformatics* **9**(1), 1–13.

Leclerc, R. D. (2008), 'Survival of the sparsest: robust gene networks are parsimonious', *Molecular systems biology* **4**(1), 213.

Lee, W., Huang, D.-S. & Han, K. (2020), 'Constructing cancer patient-specific and group-specific gene networks with multi-omics data', *BMC medical genomics* **13**, 1–12.

Lemoine, G. G., Scott-Boyer, M.-P., Ambroise, B., Périn, O. & Droit, A. (2021), 'Gwena: gene co-expression networks analysis and extended modules characterization in a single bioconductor package', *BMC bioinformatics* **22**(1), 1–20.

Lewandowski, D., Kurowicka, D. & Joe, H. (2009), 'Generating random correlation matrices based on vines and extended onion method', *Journal of multivariate analysis* **100**(9), 1989–2001.

Liu, X., Wang, Y., Ji, H., Aihara, K. & Chen, L. (2016), 'Personalized characterization of diseases using sample-specific networks', *Nucleic acids research* **44**(22), e164–e164.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. et al. (2013), 'The genotype-tissue expression (gtex) project', *Nature genetics* **45**(6), 580–585.

Lopes-Ramos, C. M., Chen, C.-Y., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., Platig, J., Glass, K., Quackenbush, J. & DeMeo, D. L. (2020), 'Sex differences in gene expression and regulatory networks across 29 human tissues', *Cell reports* **31**(12).

Oliveto, S., Mancino, M., Manfrini, N. & Biffo, S. (2017), 'Role of microrna s in translation regulation and cancer', *World journal of biological chemistry* **8**(1), 45.

Ozbay, S., Parekh, A. & Singh, R. (2023), 'Navigating the manifold of single-cell gene coexpression to discover interpretable gene programs', *bioRxiv* pp. 2023–11.

Pang, R. & Yang, S. (2021), 'lncrna duxap8 inhibits papillary thyroid carcinoma cell apoptosis via sponging the mir-20b-5p/sos1 axis', *Oncology reports* **45**(5), 1–10.

Pizzato, M., Li, M., Vignat, J., Laversanne, M., Singh, D., La Vecchia, C. & Vaccarella, S. (2022), 'The epidemiological landscape of thyroid cancer worldwide: Globocan estimates for incidence and mortality rates in 2020', *The Lancet Diabetes & Endocrinology* **10**(4), 264–272.

Pramila, T., Wu, W., Miles, S., Noble, W. S. & Breeden, L. L. (2006), 'The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle', **20**(16), 2266–2278.
URL: *http://genesdev.cshlp.org/content/20/16/2266*

Rawal, Y., Chereji, R. V., Valabhoju, V., Qiu, H., Ocampo, J., Clark, D. J. & Hinnebusch, A. G. (2018), 'Gcn4 Binding in Coding Regions Can Activate Internal and Canonical 5' Promoters in Yeast', **70**(2), 297–311.e4.

Reimand, J., Vaquerizas, J. M., Todd, A. E., Vilo, J. & Luscombe, N. M. (2010), 'Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets', **38**(14), 4768–4777.
URL: *https://doi.org/10.1093/nar/gkq232*

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. & Smyth, G. K. (2015), 'limma powers differential expression analyses for rna-sequencing and microarray studies', *Nucleic acids research* **43**(7), e47–e47.

Saha, E., Ben-Guebila, M., Fanfani, V., Fischer, J., Shutta, K. H., Mandros, P., DeMeo, D. L., Quackenbush, J. & Lopes-Ramos, C. M. (2023), 'Gene regulatory networks reveal sex difference in lung adenocarcinoma', *bioRxiv* .

Shobab, L., Burman, K. D. & Wartofsky, L. (2022), 'Sex differences in differentiated thyroid cancer', *Thyroid* **32**(3), 224–235.

Shutta, K. H., Weighill, D., Burkholz, R., Guebila, M. B., DeMeo, D. L., Zacharias, H. U., Quackenbush, J. & Altenbuchinger, M. (2023), 'Dragon: determining regulatory associations using graphical models on multi-omic networks', *Nucleic Acids Research* **51**(3), e15–e15.

Teixeira, M. C., Monteiro, P. T., Palma, M., Costa, C., Godinho, C. P., Pais, P., Cavalheiro, M., Antunes, M., Lemos, A., Pedreira, T. & Sá-Correia, I. (2018), 'YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae', **46**(D1), D348–D353.
URL: *https://doi.org/10.1093/nar/gkx842*

Uluisik, I., Kaya, A., Fomenko, D. E., Karakaya, H. C., Carlson, B. A., Gladyshev, V. N. & Koc, A. (2011), 'Boron stress activates the general amino acid control mechanism and inhibits protein synthesis', **6**(11), e27772.

Van der Vaart, A. W. (2000), *Asymptotic statistics*, Vol. 3, Cambridge university press.

Vanderpump, M. P. (2011), 'The epidemiology of thyroid disease.', *British medical bulletin* **99**(1).

Wang, X., Wang, Q., Su, P., Chen, C., Han, B. & Liu, Z. (2022), 'Kmt2c mutation is a diagnostic molecular marker for primary thyroid osteosarcoma: A case report and literature review', *Frontiers in Medicine* **9**, 1030888.

Weighill, D., Guebila, M. B., Glass, K., Quackenbush, J. & Platig, J. (2022), 'Predicting genotype-specific gene regulatory networks', *Genome Research* **32**(3), 524–533.

Weighill, D., Guebila, M. B., Lopes-Ramos, C., Glass, K., Quackenbush, J., Platig, J. & Burkholz, R. (2021), Gene regulatory network inference as relaxed graph matching, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 35, pp. 10263–10272.

Wong, E. D., Miyasato, S. R., Aleksander, S., Karra, K., Nash, R. S., Skrzypek, M. S., Weng, S., Engel, S. R. & Cherry, J. M. (2023), 'Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources', *Genetics* **224**(1), iyac191.
URL: *https://doi.org/10.1093/genetics/iyac191*

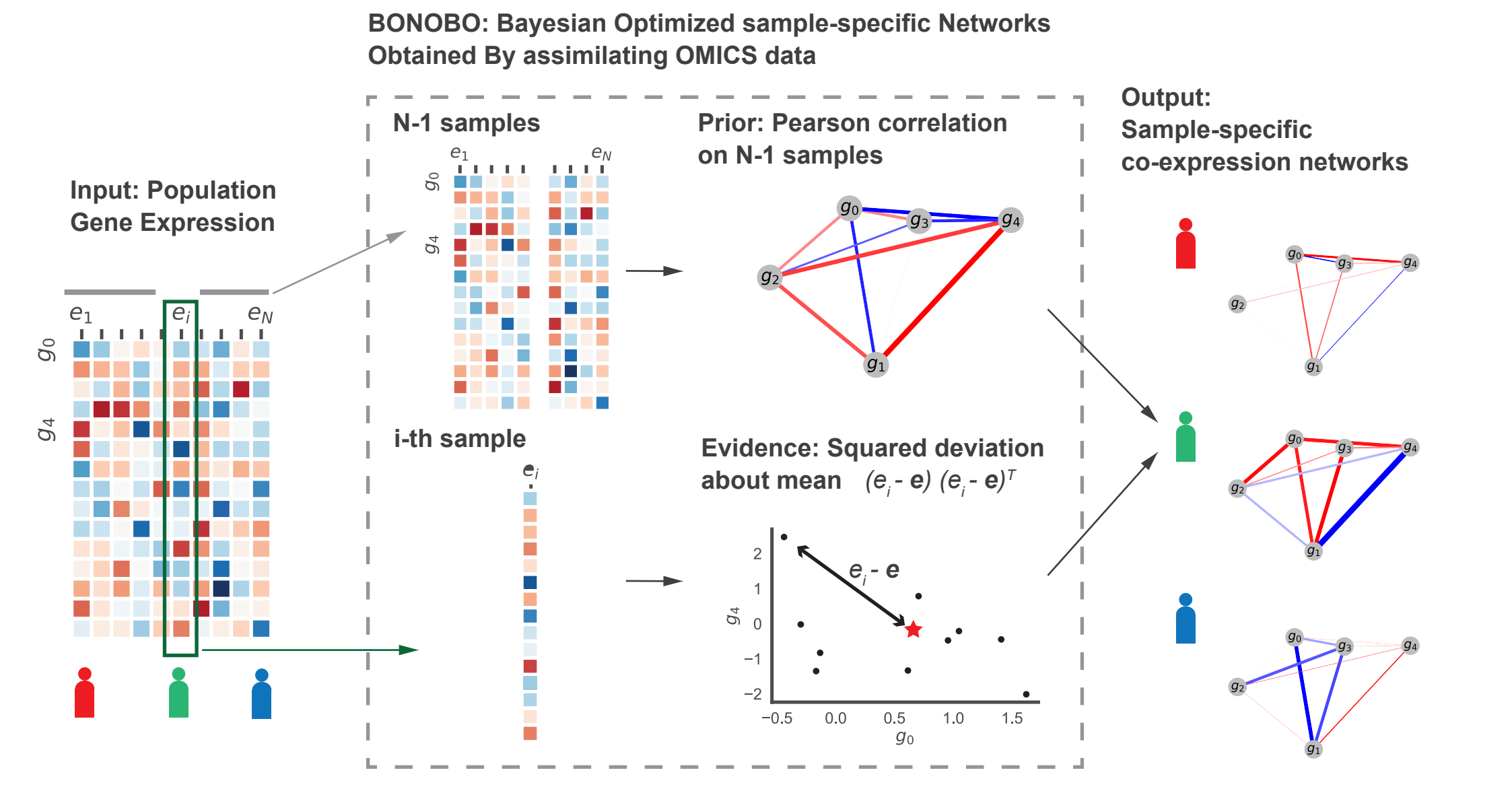
Yu, X., Zeng, T., Wang, X., Li, G. & Chen, L. (2015), 'Unravelling personalized dysfunctional gene network of complex diseases based on differential network model', *Journal of translational medicine* **13**, 1–13.

Zhang, W., Zeng, T., Liu, X. & Chen, L. (2015), 'Diagnosing phenotypes of single-sample individuals by edge biomarkers', *Journal of molecular cell biology* **7**(3), 231–241.

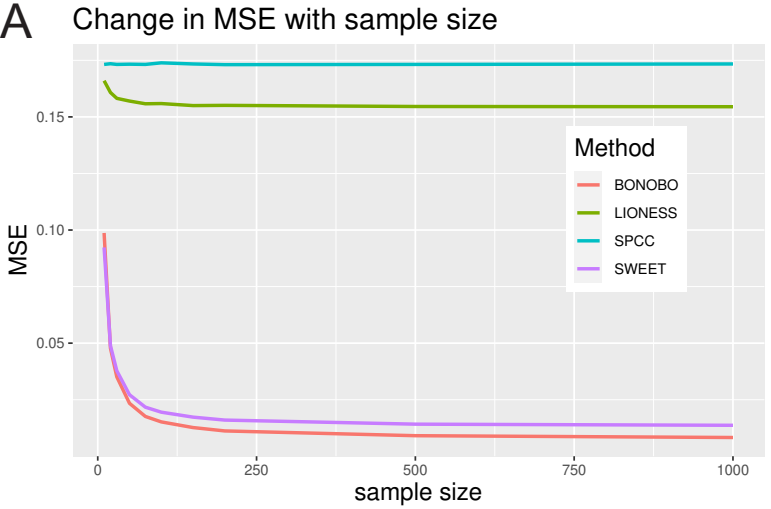
Zhang, Z. (2021), 'A note on wishart and inverse wishart priors for covariance matrix', *Journal of Behavioral Data Science* **1**(2), 119–126.

Zheng, H., Xu, J., Chu, Y., Jiang, W., Yao, W., Mo, S., Song, X. & Zhou, J. (2022), 'A global regulatory network for dysregulated gene expression and abnormal metabolic signaling in immune cells in the microenvironment of graves' disease and hashimoto's thyroiditis', *Frontiers in Immunology* **13**, 879824.

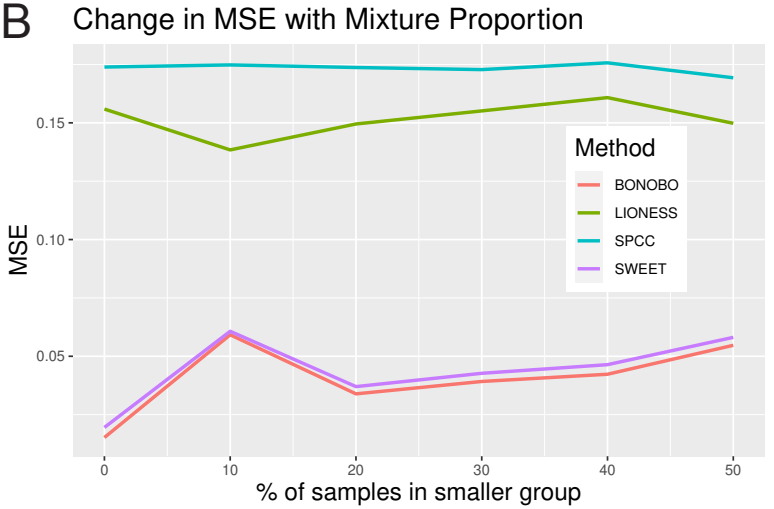
Zhu, B., Tse, L. A., Wang, D., Koka, H., Zhang, T., Abubakar, M., Lee, P., Wang, F., Wu, C., Tsang, K. H. et al. (2019), 'Immune gene expression profiling reveals heterogeneity in luminal breast tumors', *Breast Cancer Research* **21**, 1–11.


BONOBO: Bayesian Optimized sample-specific Networks
Obtained By assimilating OMICS data
Input: Population
Gene Expression
N-1 samples
i-th sample
Prior: Pearson correlation
on N-1 samples
Evidence: Squared deviation
about mean $(e_i - e)\,(e_i - e)^T$
Output:
Sample-specific
co-expression networks

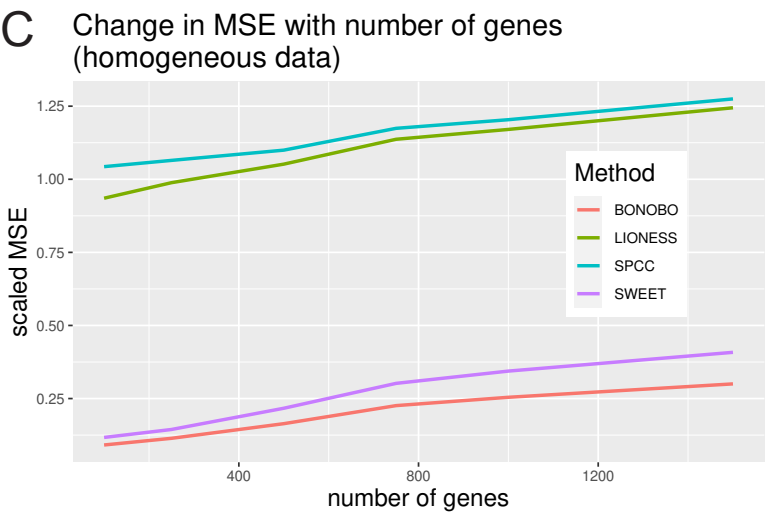
A



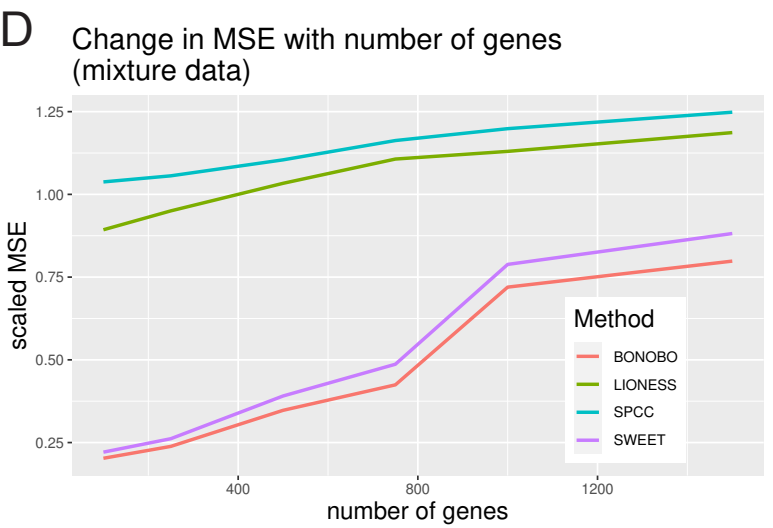
B



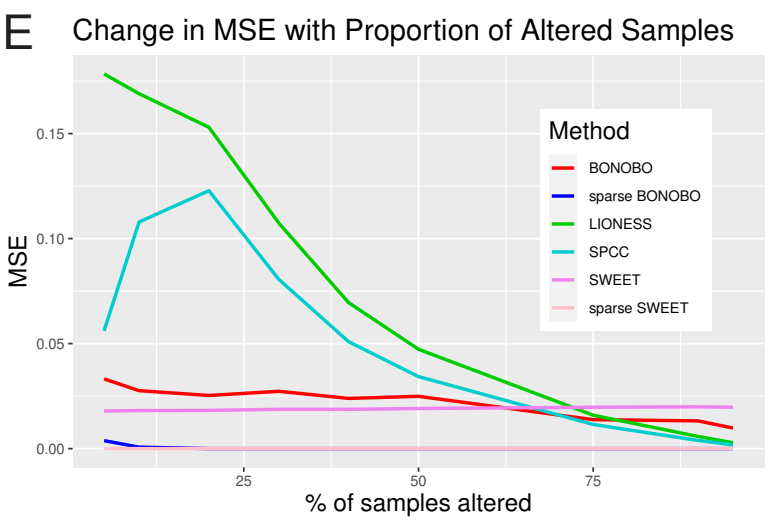
C



D



E



F

Dense

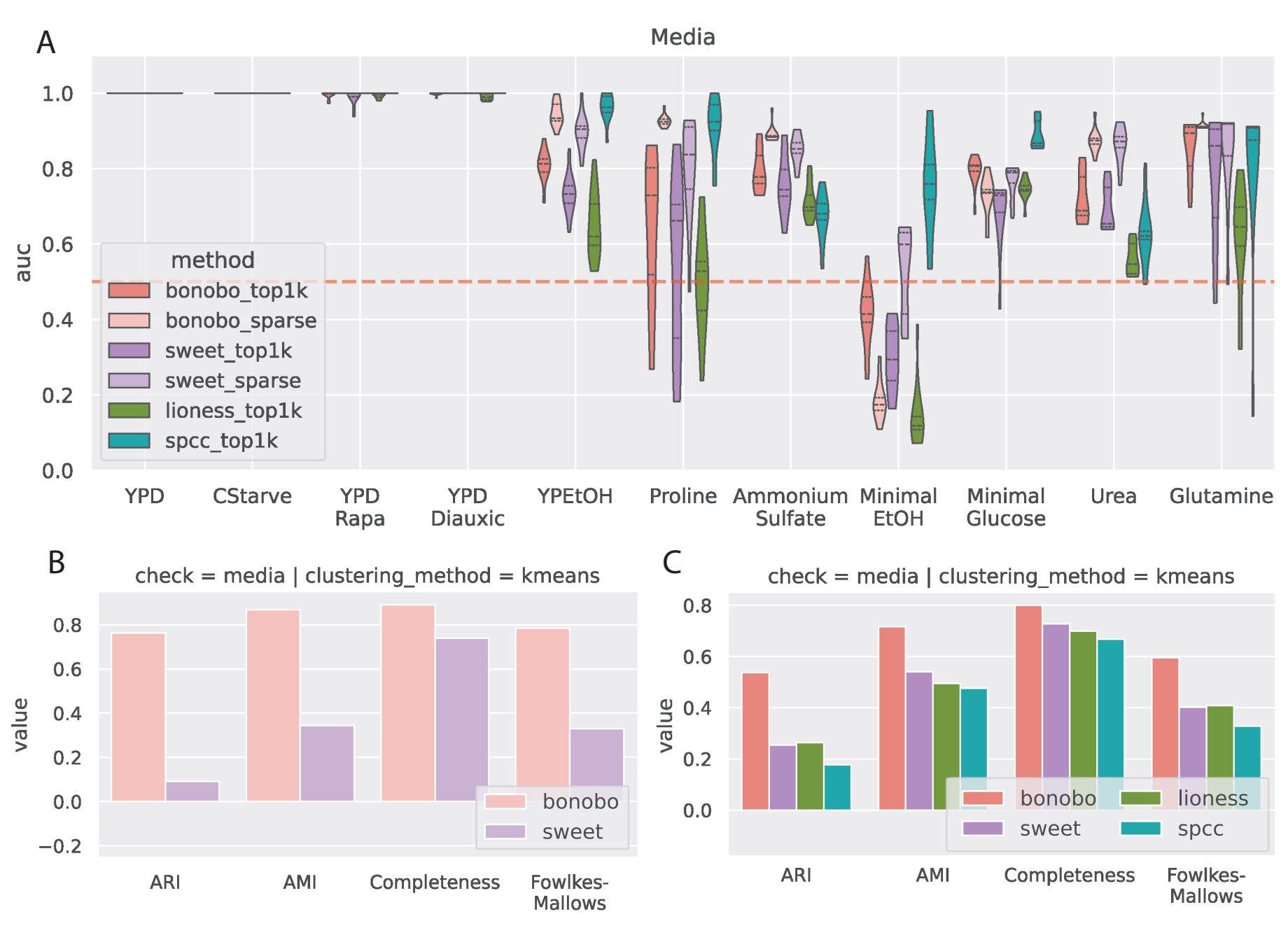
	time [s]	memory [MB]
BONOBO	1.041	1.908
LIONESS	0.178	3.816
SPCC	0.047	3.815
SWEET	0.165	3.817

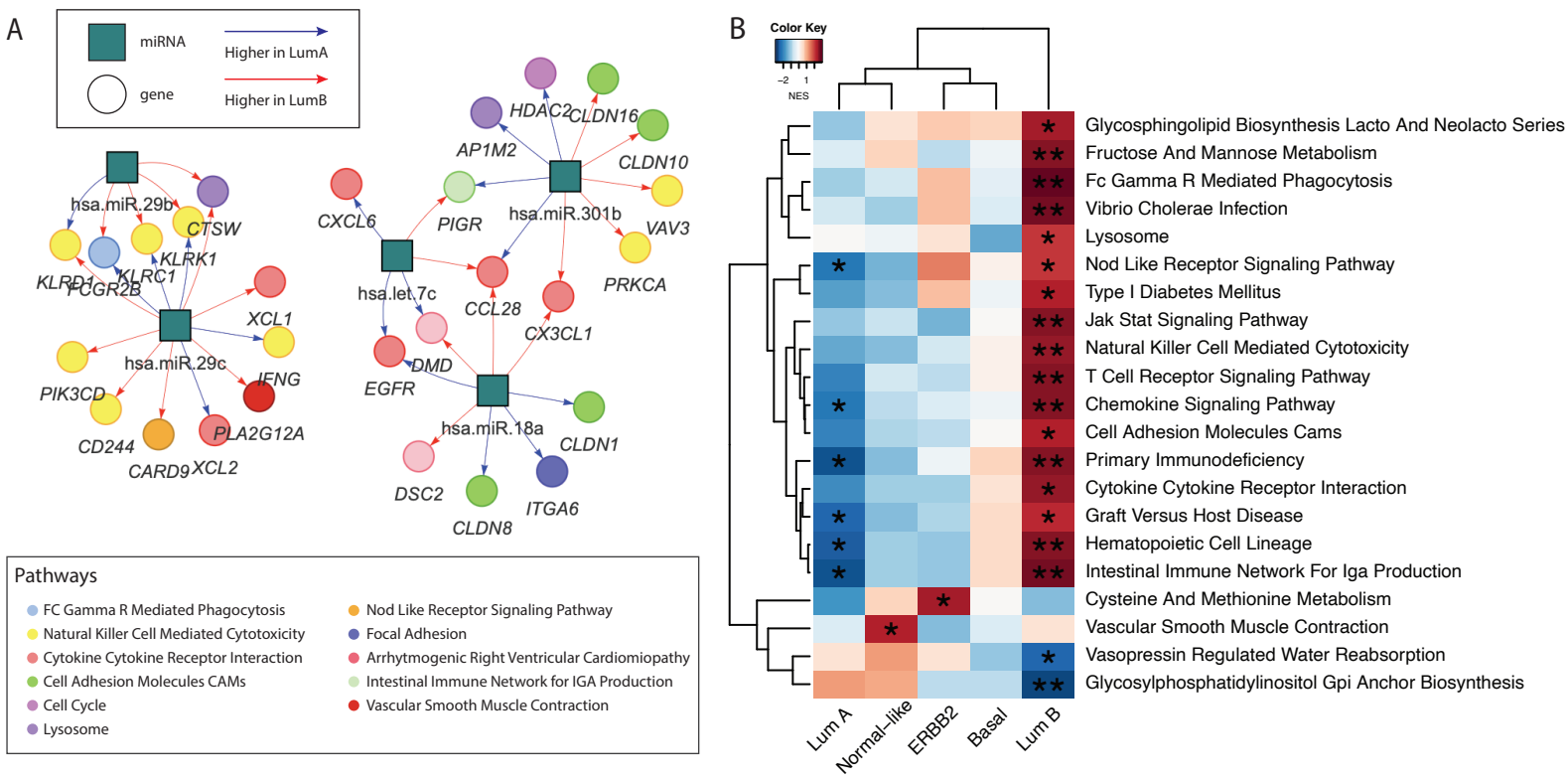
Save on disk

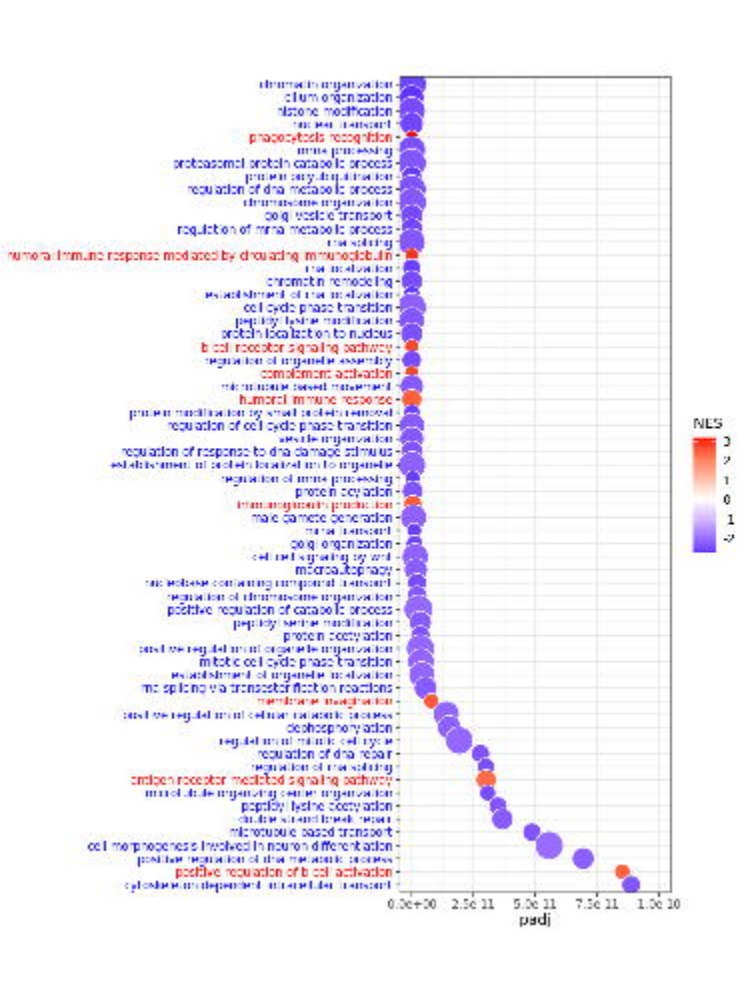
	time [s]	memory [MB]
BONOBO	58.890	1.915
LIONESS	54.082	3.822
SPCC	54.113	3.822
SWEET	54.624	3.823

Sparse

	time [s]	memory [MB]
BONOBO	1.915	3.817
SWEET	0.335	194.553


A
Media
auc
1.0
0.8
0.6
0.4
0.2
0.0
method
bonobo_top1k
bonobo_sparse
sweet_top1k
sweet_sparse
lioness_top1k
spcc_top1k
YPD
CStarve
YPD Rapa
YPD Diauxic
YPEtOH
Proline
Ammonium Sulfate
Minimal EtOH
Minimal Glucose
Urea
Glutamine
B
check = media | clustering_method = kmeans
value
0.8
0.6
0.4
0.2
0.0
−0.2
bonobo
sweet
ARI
AMI
Completeness
Fowlkes-Mallows
C
check = media | clustering_method = kmeans
value
0.8
0.6
0.4
0.2
0.0
bonobo
sweet
lioness
spcc
ARI
AMI
Completeness
Fowlkes-Mallows


A
miRNA
gene
Higher in LumA
Higher in LumB
hsa.miR.29b
CTSW
KLRK1
KLRC1
KLRD1
FCGR2B
XCL1
hsa.miR.29c
IFNG
PIK3CD
PLA2G12A
CD244
CARD9
XCL2
HDAC2
CLDN16
AP1M2
CLDN10
CXCL6
PIGR
hsa.miR.301b
VAV3
hsa.let.7c
CCL28
PRKCA
CX3CL1
DMD
EGFR
hsa.miR.18a
CLDN1
DSC2
ITGA6
CLDN8
Pathways
FC Gamma R Mediated Phagocytosis
Natural Killer Cell Mediated Cytotoxicity
Cytokine Cytokine Receptor Interaction
Cell Adhesion Molecules CAMs
Cell Cycle
Lysosome
Nod Like Receptor Signaling Pathway
Focal Adhesion
Arrhytmogenic Right Ventricular Cardiomiopathy
Intestinal Immune Network for IGA Production
Vascular Smooth Muscle Contraction
B
Color Key
−2
1
NES
Glycosphingolipid Biosynthesis Lacto And Neolacto Series
Fructose And Mannose Metabolism
Fc Gamma R Mediated Phagocytosis
Vibrio Cholerae Infection
Lysosome
Nod Like Receptor Signaling Pathway
Type I Diabetes Mellitus
Jak Stat Signaling Pathway
Natural Killer Cell Mediated Cytotoxicity
T Cell Receptor Signaling Pathway
Chemokine Signaling Pathway
Cell Adhesion Molecules Cams
Primary Immunodeficiency
Cytokine Cytokine Receptor Interaction
Graft Versus Host Disease
Hematopoietic Cell Lineage
Intestinal Immune Network For Iga Production
Cysteine And Methionine Metabolism
Vascular Smooth Muscle Contraction
Vasopressin Regulated Water Reabsorption
Glycosylphosphatidylinositol Gpi Anchor Biosynthesis
Lum A
Normal−like
ERBB2
Basal
Lum B


chromatin organization
cilium organization
histone modification
nuclear transport
phagocytosis recognition
mrna processing
proteasomal protein catabolic process
protein ubiquitination
regulation of dna metabolic process
chromosome organization
golgi vesicle transport
regulation of mrna metabolic process
rna splicing
humoral immune response mediated by circulating immunoglobulin
rna localization
chromatin remodeling
establishment of rna localization
cell cycle phase transition
peptidyl lysine modification
protein localization to nucleus
b cell receptor signaling pathway
regulation of organelle assembly
complement activation
microtubule based movement
humoral immune response
protein modification by small protein removal
regulation of cell cycle phase transition
vesicle organization
regulation of response to dna damage stimulus
establishment of protein localization to organelle
regulation of mrna processing
protein acylation
innate immune response
male gamete generation
nuclear transport
golgi organization
cell cell signaling by wnt
mitochondrion
rna localization
regulation of chromosome organization
positive regulation of catabolic process
peptidyl serine modification
protein acylation
positive regulation of organelle organization
mitotic cell cycle phase transition
establishment of organelle localization
rna splicing via transesterification reactions
membrane invagination
positive regulation of cellular catabolic process
autophagy
regulation of mitotic cell cycle
regulation of dna repair
regulation of rna splicing
cytokine receptor mediated signaling pathway
microtubule organizing center organization
peptidyl lysine acylation
double strand break repair
microtubule based transport
cell morphogenesis involved in neuron differentiation
positive regulation of dna metabolic process
positive regulation of b cell activation
cytoskeleton dependent intracellular transport
0.0e+00
2.5e 11
5.0e 11
7.5e 11
1.0e 10
padj
NES
2
1
0
1
2