# Bayesian Optimized sample-specific Networks Obtained By Omics data (BONOBO)

Enakshi Saha[1,*], Viola Fanfani[1,*], Panagiotis Mandros[1], Marouen Ben-Guebila[1], Jonas Fischer[1], Katherine Hoff-Shutta[1,2], Kimberly Glass[1,2,3], Dawn Lisa DeMeo[2,3], Camila 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

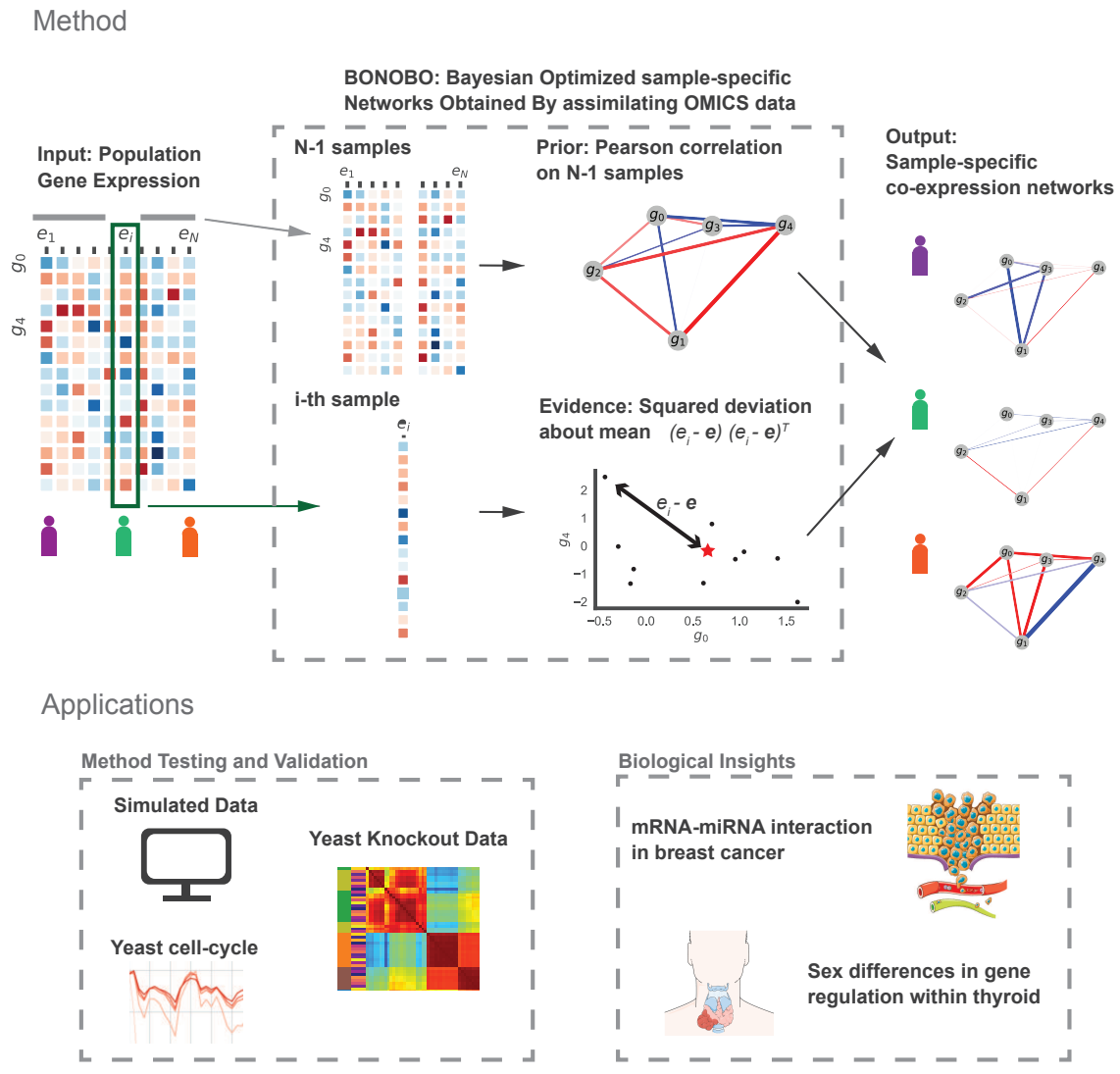
November 3, 2023

**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 coregulated 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. To address these concerns, we introduce BONOBO (Bayesian Optimized Networks Obtained By assimilating Omics data), a scalable Bayesian model for deriving individual sample-specific co-expression networks by recognizing variations in molecular interactions across individuals. For every 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 expression with the prior distribution, BONOBO yields a closed-form solution for the posterior distribution of the sample-specific co-expression matrices, thus making the method extremely scalable. We demonstrate the utility of BONOBO in several contexts, including analyzing gene regulation in yeast transcription factor knockout studies, 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 sample-specific co-expression network inference methods 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

# Graphical Abstract



# 1 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 [1]. 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. While differential expression analysis with conventional tools such as “limma” (or “voom”)[2] enables us to adjust for the effects of these covariates, differences found in transcription levels alone often fail to explain biological differences between the cohorts being compared [3].

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. However, most methods for constructing co-expression networks estimate an aggregate network for the entire population [4, 5, 6], thus failing to capture the heterogeneous, context-specific gene interactions present within individual samples. Trying to overcome these limitations, methods to infer sample-specific co-expression networks have been proposed, such as Single Pearson Correlation Coefficient (SPCC) [7, 8] and Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) [9]. However, these methods produce co-expression matrices that are not positive definite and/or where the estimated correlation values are assigned outside the defined range for correlation measures (for example, [-1,1] for Pearson’s correlation coefficient). This non-positive definiteness can pose significant challenges in downstream analyses, as it violates the fundamental assumptions of correlation networks and can lead to misleading interpretations. Alternatively, other methods designed for personalized characterization of diseases through sample-specific networks [10] and cancer-specific or group-specific networks [11] represent differential networks with respect to an external reference population and hence are susceptible to varying inference depending on the reference sample used.

We develop BONOBO (Bayesian Optimized Networks Obtained By assimilating Omics data), an empirical Bayesian model that derives individual sample-specific co-expression networks (Figure 1), thus facilitating the discovery of differentially co-regulated gene pairs between different conditions and/or phenotypes, while eliminating the effects of confounders. BONOBO derives positive definite co-expression networks from input data alone, without using any 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 network analysis. BONOBO derives a posterior probability distribution for individual correlation matrices, allowing us to test the hypothesis of whether any two pairs of genes have a non-zero correlation, within an individual sample in the data. Based on the results of these hypotheses-testing we 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[12].

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 instance, when comparing aggregate co-expression networks between conditions, such as distinguishing between health and disease or male and female samples, results are frequently confounded by the population’s heterogeneity stemming from nuisance parameters, such as batch effects and/or confounding clinical covariates such as race and 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 also be used as inputs to methods for inferring gene regulatory networks that require a correlation matrix as input, such as PANDA (Passing Messages between Biological Networks to Refine Predicted Interactions) [3], OTTER (Optimize To Estimate Regulation) [13] and EGRET (Estimating the Genetic Regulatory Effect on Transcription factors) [14] to infer sample-specific gene regulatory networks. These methods infer bipartite gene regulatory networks consisting of directed regulatory edges from regulators such as transcription factors (TF) to their target genes, by combining gene co-expression matrices with individual-specific TF-motif and/or chromatin accessibility data.

We demonstrate the advantages of BONOBO using several simulated and real datasets. First, we used simulated data to compare BONOBO’s performance with other state-of-the-art methods. We then used pseudo-bulk gene expression data from knockout experiments in yeast cells and show that BONOBO captures global properties of each yeast strain and also distinguishes 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 find 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 study sex differences in gene regulation within human thyroid. Empirical data indicate that females are three times more likely to develop some types of thyroid conditions over their lifetime than males [15]. Using BONOBO networks as inputs to PANDA, we infer individual-specific gene regulatory networks and compare 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.

BONOBO is available as open-source code in Python through the Network Zoo package (netZooPy v0.9.17;

netzoo.github.io) [16].

# 2 Methods

## 2.1 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 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, $\forall i$.

We assume 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 $j \in \{1, 2, \ldots, N\} \setminus \{i\}$,

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

where $\nu_i \geq g + 1$ denotes the degree 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 similar to the correlation between these same pair of genes across the entire population on average.

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 $0 \leq \delta_i \leq 1$, 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 $S_i \xrightarrow{a.s.} \Sigma$, where $\Sigma$ denotes the population covariance matrix. Thus the hyperparameter $\delta_i$ 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 $\Sigma$. For homogeneous populations we recommend using a smaller value of $\delta_i$, or equivalently, a larger value of $1 - \delta_i$, as this would increase the contribution of the population covariance $\Sigma$ 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 $\delta_i$, 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 $\delta_i$ for every sample.

### 2.1.1 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 $\nu_i$. Hence in order to calibrate $\delta_i$, it suffices to estimate $\nu_i$ for every sample $i$ in the data. The following lemma provides a data-driven approach for calibrating $\nu_i$.

**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 [17].

From (5), summing over $k = 1, \ldots, g$, i.e., 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 $\eta_j^k$ denotes the variance of $k$-th coordinates of $\{x_1, \ldots, x_N\} \setminus x_j$.
2. Estimate $\eta^{(k)} = \frac{1}{N}\sum_{j=1}^{N}\left(\eta_j^k - \frac{1}{N}\sum_{l=1}^{N}\eta_l^k\right)$, 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 $\nu_i$. Thus the estimate of the hyperparameter $\delta_i$ 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 S1.2.5, we illustrate, via simulation experiments, that this data-driven empirical approach of calibrating $\delta_i$ delivers performance on par with the optimal performance achieved by fixing $\delta_i = \delta, \forall i$ to any arbitrary value.

### 2.1.2 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 two pair of genes using the following lemma, which is a direct consequence of the properties of inverse Wishart distribution [17]. 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 $v_{jk}$ would be*

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

*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 $\psi_{jk} = \sqrt{Var(v_{jk})}$, $\sigma_{jk}$ is the posterior mean of $v_{jk}$ and $z_{(1-\alpha/2)}$ denotes the $(1-\alpha/2)$-th quantile of the standard normal distribution.

Equivalently, for every pair of genes $(j,k)$, 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 $\alpha$, provided $2\left(1 - \Phi(\sigma_{jk}/\psi_{jk})\right) \leq \alpha$, where $\Phi$ denotes the cumulative distribution function of the standard normal distribution.

**Remark:** BONOBO derives a dense (complete) network with edges between every pair of genes, where edge weights correspond to $\sigma_{jk}$, 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 $0 < \alpha < 1$.

# 3 Results

## 3.1 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 two other methods for computing sample-specific co-expression: LIONESS [9] and SPCC [7, 8]. We repeated each of the following simulation experiments 100 times and compared BONOBO with LIONESS and SPCC based on the mean sum of squared errors (i.e. the squared Frobenius distance between the true correlation matrix and the estimated correlation matrix) across these 100 iterations. Through samples of varying sizes and different dimensionalities from a homogeneous population we demonstrate that (i) increasing sample size improves the performance of BONOBO at a faster rate than other methods (Supplementary Materials S1.2.1) while (ii) increasing the number of genes deteriorates the performance of all three methods (Supplementary Materials S1.2.2). Next, (iii) we simulated samples from a mixture of two different homogeneous populations (Supplementary Materials S1.2.3) and demonstrated that the performance of BONOBO remains unaffected by the mixing proportion of the two populations. In all these instances of simulation experiments, the mean squared errors for BONOBO were much smaller than those for LIONESS and SPCC. In the next example, (iv) we simulated samples from a mixture of two populations, where one population lost expression for 1% of genes. BONOBO identified this loss of expression (Supplementary Materials S1.2.4) with better accuracy, compared to the two competing methods. Finally (v) we used two simulation examples to illustrate (Supplementary Materials S1.2.5) that the data-driven approach of calibrating hyperparameter $\delta$ described in section 2.1.1, provides comparable estimation accuracy, compared to the optimal performance obtained by choosing an arbitrary fixed value of $\delta$.

Taken together, we have simulated datasets that that resemble many of the scenarios encountered in biological datasets such as different population sizes, single gene knockouts and silencing, mixtures of subpopulations. In all these conditions, BONOBO performs better than both LIONESS and SPCC, thereby demonstrating the efficacy of our method in capturing the "true" correlation patterns between genes, compared to the existing methods for sample-specific coexpression estimation.

## 3.2 BONOBO recovers sample-specific network structure in yeast datasets

Having established the performance characteristics of BONOBO for simulated data, we want to prove that our method is applicable and useful on experimental data. *Saccharomyces Cerevisiae*, i.e. yeast, is a well-studied organism and it has been extensively used to model biological networks [18, 19, 20]. Moreover, perturbation experiments have been key to study the connectivity between biological entities in yeast [21, 22, 23]. We reason that using yeast experiments would allow us to test BONOBO on real data, validating our findings with those in the literature, and testing our method's ability to detect perturbations at the single sample level. First, we applied BONOBO to 48 cell-cycle-synchronized yeast microarray samples [24]. With this smaller dataset we assess the behavior of p-value thresholding in real data and we show that by using BONOBO networks we are able to detect the fluctuations in the cell-cycle transition pathways (see Supplementary Materials S1.3).

Thus, we applied BONOBO to a yeast gene perturbation experiment [25] that includes pseudo-bulked gene expression from $132$ engineered strains that combine genetic and environmental perturbations; $11$ TF knockout (KO) genotypes target the Nitrogen Catabolite Repression (NCR) pathway, the General Amino Acid Control (GAAC) pathway, the Ssy1-Ptr3-Ssy5-sensing (SPS) pathway, and the retrograde pathway. These strains were grown in twelve conditions that included various nitrogen and carbon sources (Supplementary Materials S1.4). While this is originally a scRNA-seq dataset, we created pseudo-bulk expression values for each KO-media combination. This way we can use BONOBO to generate networks that are perturbation-specific and investigate the co-expression changes induced by each KO and media.

Consistent with the results of the original paper, we find that BONOBO's networks in the same growth medium tend to be more similar than those with the same KO (Figure 2A). This makes also logical sense, since different nutrients will perturb larger metabolic pathways, rather than one TF and its gene targets as it is the case for the TF KOs. However, leveraging BONOBO's sample-specific co-expression, we can analyze how much the gene deletion affects each sample's network by looking at the effects of a specific TF KO on the other genes and comparing it to the other samples (Figure S5). It appears that deletion of GCN4 has the strongest effect on the network, visibly changing the correlation patterns of the genes that have the highest edge values with GCN4 (Figure S6). Furthermore, we can pinpoint the effects of GCN4 deletion on each network by reconstructing which edges are the most affected by the perturbation. We then selected the top 100 genes whose edge with GCN4 is the most affected by the perturbation, by means of checking which edges are most different between the GCN4 genotype and the rest of the dataset. As expected, genes that are most perturbed by GCN4 deletion, that targets the GAAC pathway, belong to many pathways related to autophagy such as exosome, phagosome, and autophagy - yeast (Figure

2B). Interestingly, between the top perturbed genes there are known GCN4 targets [26] such as VCX1, MNN10, ACT1,CPA1 ([27, 28], and CCW12 [29] showing that by estimating perturbation-specific co-expression networks BONOBO is able to detect key interactors of the TF. The analysis of yeast experiments recapitulate many of the known properties of yeast cells, showing that BONOBO works well on real data, in addition to simulated datasets. Moreover, BONOBO allows us to investigate the effects of each combination of TF KO and growth media, which would be impossible with conventional population-based correlation measures.

## 3.3 miRNA-gene Interaction in Breast Cancer Subtypes

As evidenced from the simulation experiments and the analysis of gene expression data from perturbed yeast cells, BONOBO successfully recovers sample-specific heterogeneity in the gene co-expression networks. Furthermore, we propose that BONOBO can serve as a useful tool for investigating interaction patterns between multiple omics modalities, while accounting for sample-specific clinical and molecular confounders. MicroRNAs, or miRNAs have been observed to play important parts in RNA silencing by down-regulating the expression of multiple genes and modifications in miRNA levels have been shown to be involved in the development and prognosis of various cancer types [30]. Here, we used individual sample-specific co-expression networks constructed using paired mRNA and miRNA bulk expression data from multiple breast cancer subtypes (GEO accession number GSE19783 [31, 32, 33], preprocessing steps in Supplementary Materials S1.5.1) to understand how correlation between miRNA and genes (or biological pathways) vary across different breast cancer subtypes and whether this interaction between genes and miRNAs have any association with breast cancer survival.

In total we have $101$ sample-specific co-expression networks each of which constitutes of correlations between pairs of genes, pairs of miRNAs and between each gene and each miRNA. From these BONOBO networks we are interested in investigating which biological pathways are most significantly associated with miRNA expression in various breast cancer subtypes. For each network, we account for the association between a gene and all miRNA by summing over all edges connecting the particular gene to the miRNAs and we then applied pathway analysis (Supplementary Materials S1.5.2).

We observed that pathways associated to immune response including Graft vs Host disease, primary immunodeficiency, cytokine-cytokine receptor interaction and natural killer cell mediated cytotoxicity were significantly (at FDR cutoff 0.05) negatively correlated with miRNA expression, across all breast cancer subtypes (Figure S7). Pathways associated to cell adhesion and cell proliferation, such as focal adhesion and ECM receptor interaction were positively correlated with miRNA expression, especially in Basal, Normal-like and Luminal B subtypes, while pathway associated to cell adhesion molecules was significantly (at FDR cutoff 0.05) negatively correlated with miRNA expression in ERBB2, Luminal A and Normal-like breast cancer. These findings align with previous studies [31] that also uncovered significant associations between the expression levels of several miRNAs and biological pathways involved in cell proliferation, cell adhesion and immune response.

Although pathways most correlated (positively or negatively) with miRNA expression were consistent across different breast cancer subtypes, upon closer inspection of the individual mRNA-miRNA edges in the networks, we observe that there is significant difference in the neighborhood of individual genes across different breast cancer subtypes. In particular, comparing individual edges between genes and miRNAs in samples from luminal A and luminal B (Figure 3B) subtypes, we found several genes that were differentially correlated with certain miRNAs in luminal A versus luminal B subtypes. Genes most differentially coexpressed with miRNAs include proto-oncogene *EGFR*; genes involved in immune response such as *PI3KCD and INFGI*; 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, which might be a contributing factor towards diverse clinical presentations observed in both the incidence rates and prognoses of these two breast cancer subtypes. Previous research [34] has also demonstrated that the miRNA dysregulation patterns in luminal A breast cancers differ from those in luminal B breast cancers.

Next we investigated, if the strength of association between miRNA expression and biological pathways have any influence on the survival outcome in various breast cancer subtypes. For every pathway significantly (at 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 every pathway, we fit a Cox proportional hazard model to predict survival, using the pathway scores, while allowing the coefficient of the Cox model to be subtype-specific by including an interaction effect between the pathway score and the breast cancer subtype of every sample.

We observe that a higher correlation between miRNA and pathways associated with immune response such as 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 subtype, while having worse survival among samples from luminal B subtype (Figure 3A). Previous studies [35] have also demonstrated that an 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 mechanisms for tumor suppressive effect of miRNAs in luminal A. In contrast, in luminal B breast cancer, miRNAs may have a role in promoting immune evasion [36], thus leading to more aggressive tumor growth and poorer survival outcome.

Previous studies [37] had identified specific immune gene expression patterns that distinguish luminal A from luminal B subtypes and showed that these distinct immune signatures were associated with a differential ratio between ESR1 and ESR2, a higher value of which was further associated to poorer survival outcome. Our results indicate that subtype-specific regulatory interactions between miRNAs and immune pathways in luminal A versus luminal B breast cancers might be a possible factor contributing towards estrogen receptor mediated survival outcome. Furthermore, we also demonstrated that the genes linked to cell proliferation pathways exhibit distinct patterns of regulation by miRNAs in luminal A versus luminal B breast cancer subtypes. Clinically, this disparity in the regulation of cell proliferation genes might be a crucial factor contributing to the worse prognosis associated with luminal B breast cancer, as it tends to exhibit heightened cellular proliferation [38]. In conclusion, individual-specific heterogeneity in correlation networks between miRNAs and genes can provide valuable insights into the distinct miRNA-gene interaction patterns that distinguish various breast cancer subtypes, which in turn might have significant implications for breast cancer prognosis and personalized therapeutic strategies.

## 3.4 Sample-specific Gene Regulatory Networks Identify Sex Difference in Thyroid

Most thyroid disorders have sex difference in incidence rate with females being significantly more susceptible to be affected by a thyroid condition at some point in their lives, compared to males [39]. Although sex chromosomes, sex hormones, and the immune system [40] have often been cited as possible contributors to this sex difference in thyroid tissue, a system-based analysis exploring the regulatory mechanisms associated to these sex-biased disease manifestation is scarce. We used sample-specific gene regulatory networks constructed from healthy thyroid tissue samples from the Genotype Tissue Expression (GTEx) Project [41] (see Supplementary materials S1.6.1 for all preprocesing steps), to understand how the genes are differentially regulated by transcription factors (TF) in males and females. We combined the sample-specific co-expression networks derived by BONOBO with TF-motif prior information (Supplementary materials S1.6.2) and protein-protein interaction data ( Supplementary materials S1.6.3), using the PANDA network inference algorithm[3] to derive bipartite sample-specific gene regulatory networks that connect transcription factors (TFs) to their target genes (Figure S8).

Based on the differential targeting analysis ( Supplementary materials S1.6.4), we observed that several genes with known relevance in various thyroid cancers and autoimmune conditions are differentially targeted by transcription factors in males and females (Figure S10). The long non-coding RNA *XIST* showed higher targeting in females. Previously, *XIST* had been observed to promote oncogenic activities in papillary thyroid carcinomas (PTC) [42]. Among genes highly targeted in males, tumor suppresor gene *KDM6A* is known to regulate multiple genes involved in immune response, suggesting its potential influence on the risks of developing various autoimmune conditions [43]; Among other genes highly targeted in males, *PCM1* mutation has also been associated with PTC [44]; while mutation in *KMT2C* has been identified as a molecular marker for primary thyroid osteosarcoma [45]. Additionally, overexpression of *SOS1*, which also showed higher targeting in males, have been found to promote cell proliferation and cell apoptosis in PTC cells [46].

Finally, we functionally characterised the genes differentially targeted in male and female samples (see Supplementary materials S1.6.5). We observed that biological pathways associated to immune response such as humoral immune response, B-cell receptor signaling pathway, antigen-receptor mediated signaling and positive regulation of b-cell activation pathway were targeted more in females (Figure 4). Heightened targeting of immune pathways in females may contribute to their increased susceptibility to autoimmune thyroid diseases including Hashimoto's Thyroiditis disease, which often lead to hypothyroidism. On the other hand, pathways associated to cell cycle, cell signaling, metabolic processes and DNA repair were targeted more in males (Figure 4). Disregulation of these pathways have been shown to play integral parts in various thyroid conditions including Graves' Disease and Hashimoto's Thyroiditis [47] and therefore the differential targeting of these pathways in males suggests that they may have a more robust defense against factors that could disrupt thyroid function or lead to the development of these thyroid diseases.

In conclusion, sex-biased differential regulation of key genes and biological pathways might be a contributing factor towards differential risk of various thyroid conditions across both sexes and deciphering these sex-specific gene regulatory patterns through individual-specific gene regulatory networks can aid in developing more effective, personalized interventions for the prevention and treatment of various thyroid diseases.

# 4 Conclusion

Complex human traits and diseases are most often driven by not a single gene but rather by intricate interactions involving multiple genes and regulators. However, the majority of network inference techniques estimate an aggregate network [4, 6] that reflects the average regulatory characteristics of the population, thereby overlooking the diversity within the population that might arise due to various biological (e.g. sex-difference) and/or environmental factors (e.g. carcinogen exposure) [48]. Recognizing the inherent hererogeneity in regulatory processes among individuals, we have introduced BONOBO, a Bayesian parametric model designed to construct personalized sample-specific gene co-expression networks for single samples. These networks enable us to capture population heterogeneity in gene-gene interaction, that are rarely reflected in aggregate co-expression networks constructed by conventional algorithms.

Constructing individual-specific co-expression networks is particularly challenging in bulk expression data as each individual do not have more than one sample. Bayesian statistics enables us to overcome this dearth of individual-level data 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 typically samples in a bulk expression dataset come from a single tissue and from individuals having similar conditions (e.g. primary tumor samples). Integrating this prior information with the individual-level expression data, we estimate posterior distribution of the covarince matrix for each individual in the data. Interestingly, the mean of the posterior distribution turns out to be 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. Thus for every individual, BONOBO compensates for the lack of individual-level information by borrowing stength from all other individuals in the data.

BONOBO is highly scalable as we use 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, thus making BONOBO highly parallelizable, further enhancing 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 where the samples come from a mixture of two different populations with different mean expression and patterns of co-expression, we demonstrate that BONOBO performs better than competing methods even when the underlying assumption of multivariate normality is violated, thus making the method adaptable to a wide range of applications.

In addition to providing point-estimates for each pair of gene-gene correlation, BONOBO provides posterior credible intervals for these individual-specific correlation estimates. These credible regions enable the user to derive sparse co-expression networks by simply pruning out edges between pairs of genes that are uncorrelated with high probability. Using a perturbed yeast cell dataset, we illustrate that these sparse networks not only reflect the sample-specific perturbations globally, but they allow us to investigate the specific neighborhoods that are perturbed by transcription factor KO. We demonstrate that BONOBO can be readily extended to capture interactions between multiple omic data categories to derive individual-specific gene regulatory networks. As an example, we analyze BONOBO networks that capture individual-specific correlation structures between genes and miRNAs in various breast cancer subtypes. Consistent with existing literature [31] we find biological pathways associated to immune response and cell proliferation to be significantly correlated with miRNA expression. Furthermore, through survival analysis, we demonstrate that interactions between miRNAs and immune pathways have varying degrees of prognostic significance between lumina A and luminal B breast cancer subtypes.BONOBO can be combined with existing methods for estimating gene regulatory network, such as PANDA to derive individual-specific bipartite networks with directed edges from transcription factors to target genes. To demonstrate this, we use RNA-seq data from thyroid tissue samples in GTEx and compared the resulting networks between males and females to understand why thyroid conditions are more prevalent among females than males [40]. Our analysis reveals that biological pathways assocaied to cell proliferation, immune response, and various matabolic processes are differentially regulated between males and females, thereby providing a possible mechanism that might contribute towards the observed sex-disparity in the incidence rate of various thyroid conditions.

We also recognize that BONOBO has limitations, some of which could be addressed by future research. Although in this paper we applied BONOBO exclusively on transcriptomics data, the model can be readily adapted to uncover interactions between other omic data types (e.g. proteomics) with little to no modifications, as long as the data can be suitably transformed to resemble a unimodal distribution over a continuous support. 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 promising avenue for future research could involve extending our model to incorporate interactions across a broader range of omics data types, through hierarchical latent variable models and/or

association measures other than Pearson's correlation. Extending BONOBO to other correlation measures, such as the Spearman's rank correlation coefficient, would also allow to overcome Pearson's correlation intrinsic limitations. For instance, Pearson's correlation is heavily influenced by outliers, thereby potentially impacting all individual co-expression networks inferred by BONOBO. Moreover, Pearson's correlation quantifies only the extent of linear association between genes and their molecular regulators. In future research we would like to extend BONOBO to estimate correlation networks based on Chatteejee's correlation coefficient [49] since it is capable of capturing not only linear association but also functional dependence between pairs of genes.

In summary, we derive BONOBO, a Bayesian statistical model for deriving individual-specific co-expression networks, that can be further amalgamated with other network inference methods to infer individual-specific gene regulatory networks. BONOBO can be employed to capture population heterogenerity in interaction patterns involving multiple omics data types, thereby providing a more nuanced understanding of the complex mechanisms of human traits and diseases. Through various real datasets containing multiple omic modalities, we demonstrate that BONOBO can potentially enable network-based disease subtyping and facilitate individualized therapy design in diverse human diseases.

## Code Availability

BONOBO is available through the Network Zoo package (netZooPy v0.9.18; netzoo.github.io).

## Figures

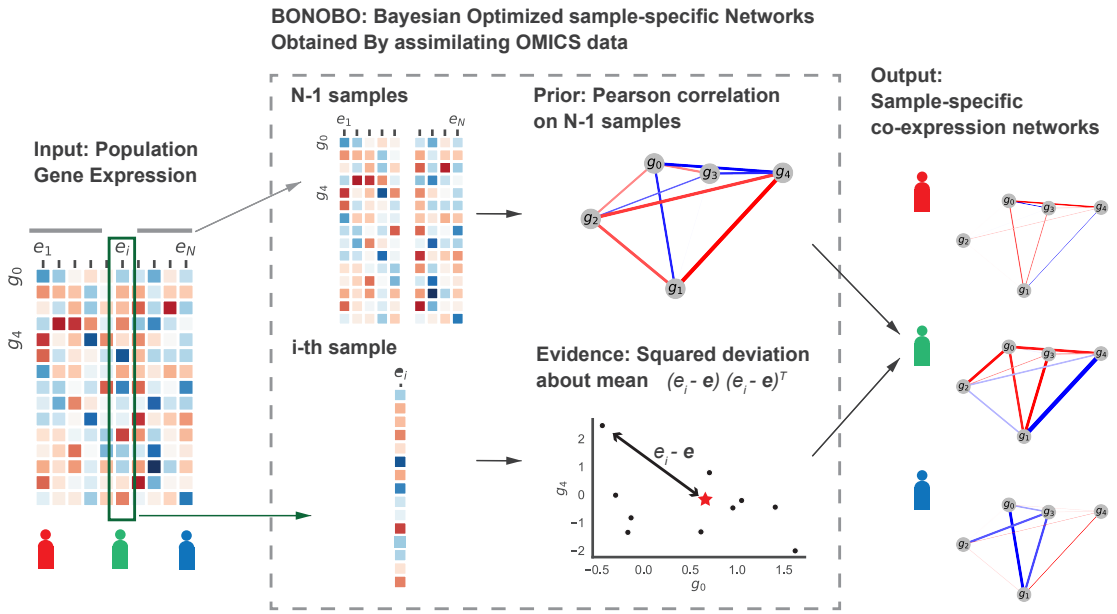



Figure 1: Schematic diagram of BONOBO: BONOBO requires a gene expression matrix as input, from which we would like to extract sample-specific correlation networks. Then, for each of the samples, BONOBO infers the network by using both the Pearson correlation matrix computed on $N-1$ samples and the sample 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.

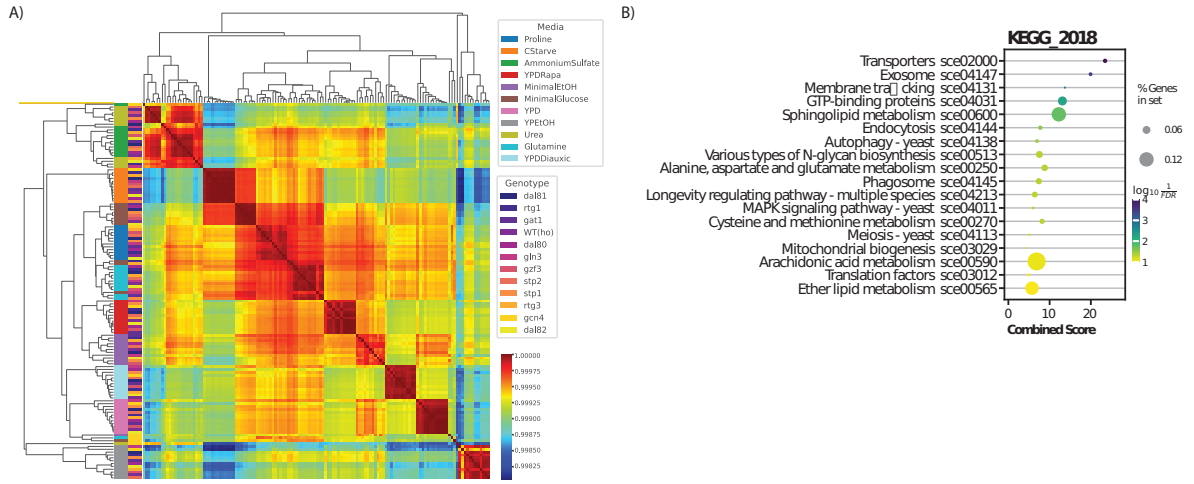



Figure 2: A) Pearson correlation between all sparsified BONOBOs computed on the engineered yeast data. Row labels report both the gene deletion and the growth medium for each sample. As expected, co-expression networks tend to correlate more with the strains subject to the same growth condition, rather than those with the same deletion.B) over-representation analysis of the top 50 genes disrupted by GCN4 perturbation. The enriched terms belong to the KEGG 2018 pathways.

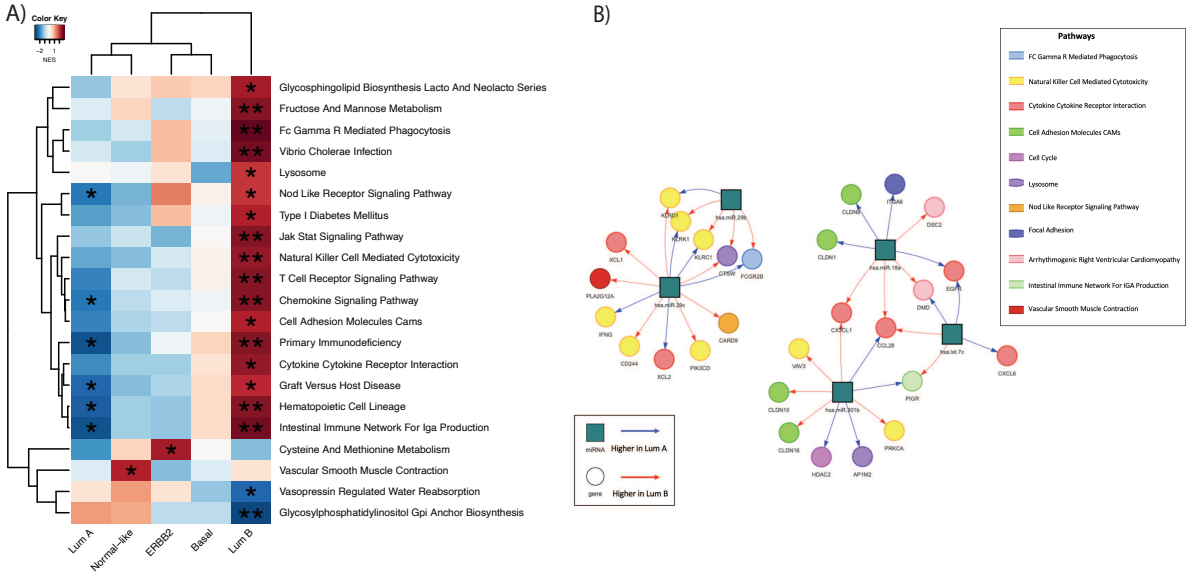



Figure 3: A) Pathways for which the mean correlation with miRNAs is significantly (at significance level 0.05) associated with survival in Cox proportional hazard model: heatmaps are colored by t-statistic of the pathway score in the Cox model; Pathways for which higher correlation with miRNAs is associated to better survival are colored blue and pathways for which higher correlation with miRNAs is associated to worse survival are colored red. B) Pairs of genes and miRNAs that are most differentially coexpressed between Lumina A and Luminal B subtypes: genes nodes are represented by colored circles where the colors correspond to the the biological pathway associated to the gene.

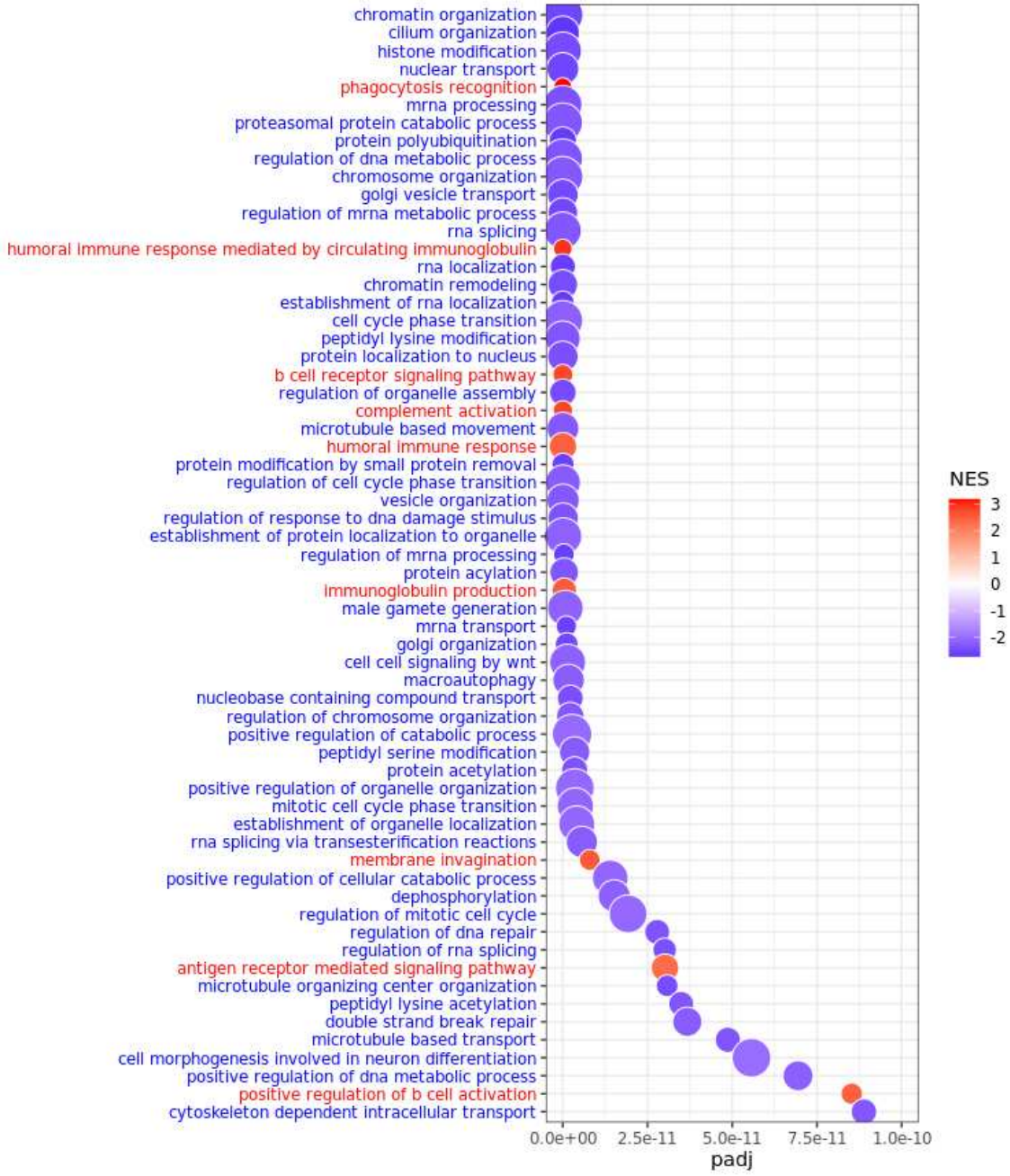



Figure 4: GO biological processes most differentially regulated (at FDR cutoff 1e-10) in males and females in GTEx thyroid samples: pathways highly targeted in males are marked in blue and pathways highly targeted in females are marked in red.

## 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 R01HG125975 and P01HL114501; 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, KG, DLD, CLR 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 CLR; **Writing – Original Draft:** ES and VF; **Writing – Review and Editing:** PM, MBG, JF, KHS, KG, DLD, CLR and JQ; **Visualization:** ES, VF and PM; **Supervision:** JQ, CLR, and DLD; **Funding Acquisition:** JQ, CLR, and DLD

## References

[1] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011.

[2] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[3] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5):e64832, 2013.

[4] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

[5] John A Dawson, Shuyun Ye, and Christina Kendziorski. R/ebcoexpress: an empirical bayesian framework for discovering differential co-expression. *Bioinformatics*, 28(14):1939–1940, 2012.

[6] Gwenaëlle G Lemoine, Marie-Pier Scott-Boyer, Bathilde Ambroise, Olivier Périn, and Arnaud Droit. Gwena: gene co-expression networks analysis and extended modules characterization in a single bioconductor package. *BMC bioinformatics*, 22(1):1–20, 2021.

[7] Xiangtian Yu, Tao Zeng, Xiangdong Wang, Guojun Li, and Luonan Chen. Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. *Journal of translational medicine*, 13:1–13, 2015.

[8] Wanwei Zhang, Tao Zeng, Xiaoping Liu, and Luonan Chen. Diagnosing phenotypes of single-sample individuals by edge biomarkers. *Journal of molecular cell biology*, 7(3):231–241, 2015.

[9] Marieke Lydia Kuijjer, Matthew George Tung, GuoCheng Yuan, John Quackenbush, and Kimberly Glass. Estimating sample-specific regulatory networks. *Iscience*, 14:226–240, 2019.

[10] Xiaoping Liu, Yuetong Wang, Hongbin Ji, Kazuyuki Aihara, and Luonan Chen. Personalized characterization of diseases using sample-specific networks. *Nucleic acids research*, 44(22):e164–e164, 2016.

[11] Wook Lee, De-Shuang Huang, and Kyungsook Han. Constructing cancer patient-specific and group-specific gene networks with multi-omics data. *BMC medical genomics*, 13:1–12, 2020.

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

[13] Deborah Weighill, Marouen Ben Guebila, Camila Lopes-Ramos, Kimberly Glass, John Quackenbush, John Platig, and Rebekka Burkholz. Gene regulatory network inference as relaxed graph matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10263–10272, 2021.

[14] Deborah Weighill, Marouen Ben Guebila, Kimberly Glass, John Quackenbush, and John Platig. Predicting genotype-specific gene regulatory networks. *Genome Research*, 32(3):524–533, 2022.

[15] Margherita Pizzato, Mengmeng Li, Jerome Vignat, Mathieu Laversanne, Deependra Singh, Carlo La Vecchia, and Salvatore Vaccarella. The epidemiological landscape of thyroid cancer worldwide: Globocan estimates for incidence and mortality rates in 2020. *The Lancet Diabetes & Endocrinology*, 10(4):264–272, 2022.

[16] Marouen Ben Guebila, Tian Wang, Camila M Lopes-Ramos, Viola Fanfani, Des Weighill, Rebekka Burkholz, Daniel Schlauch, Joseph N Paulson, Michael Altenbuchinger, Katherine H Shutta, et al. The network zoo: a multilingual package for the inference and analysis of gene regulatory networks. *Genome Biology*, 24(1):45, 2023.

[17] Zhiyong Zhang. A note on wishart and inverse wishart priors for covariance matrix. *Journal of Behavioral Data Science*, 1(2):119–126, 2021.

[18] Steven Hahn and Elton T Young. Transcriptional Regulation in Saccharomyces cerevisiae: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators. *Genetics*, 189(3):705–736. ISSN 1943-2631. doi: 10.1534/genetics.111.127019. URL https://doi.org/10.1534/genetics.111.127019.

[19] Michael Costanzo, Benjamin VanderSluis, Elizabeth N Koch, Anastasia Baryshnikova, Carles Pons, Guihong Tan, Wen Wang, Matej Usaj, Julia Hanchard, Susan D Lee, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):aaf1420, 2016.

[20] Miguel C Teixeira, Pedro T Monteiro, Margarida Palma, Catarina Costa, Cláudia P Godinho, Pedro Pais, Mafalda Cavalheiro, Miguel Antunes, Alexandre Lemos, Tiago Pedreira, and Isabel Sá-Correia. YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. 46(D1):D348–D353. ISSN 0305-1048. doi: 10.1093/nar/gkx842. URL https://doi.org/10.1093/nar/gkx842.

[21] Jüri Reimand, Juan M. Vaquerizas, Annabel E. Todd, Jaak Vilo, and Nicholas M. Luscombe. Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. 38(14):4768–4777. ISSN 0305-1048. doi: 10.1093/nar/gkq232. URL https://doi.org/10.1093/nar/gkq232.

[22] Sean R Hackett, Edward A Baltz, Marc Coram, Bernd J Wranik, Griffin Kim, Adam Baker, Minjie Fan, David G Hendrickson, Marc Berndl, and R Scott McIsaac. Learning causal networks using inducible transcription factors and transcriptome-wide time series. 16(3):e9174. ISSN 1744-4292. doi: 10.15252/msb.20199174. URL https://www.embopress.org/doi/full/10.15252/msb.20199174.

[23] Michael Costanzo, Jing Hou, Vincent Messier, Justin Nelson, Mahfuzur Rahman, Benjamin VanderSluis, Wen Wang, Carles Pons, Catherine Ross, Matej Ušaj, Bryan-Joseph San Luis, Emira Shuteriqi, Elizabeth N. Koch, Patrick Aloy, Chad L. Myers, Charles Boone, and Brenda Andrews. Environmental robustness of the global yeast genetic interaction network. *Science*, 372(6542):eabf8424. doi: 10.1126/science.abf8424. URL https://www.science.org/doi/10.1126/science.abf8424.

[24] Tata Pramila, Wei Wu, Shawna Miles, William Stafford Noble, and Linda L. Breeden. 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. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.1450606. URL http://genesdev.cshlp.org/content/20/16/2266.

[25] Christopher A Jackson, Dayanne M Castro, Giuseppe-Antonio Saldi, Richard Bonneau, and David Gresham. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife*, 9:e51254, 2020. ISSN 2050-084X. doi: 10.7554/eLife.51254. URL https://doi.org/10.7554/eLife.51254.

[26] Edith D Wong, Stuart R Miyasato, Suzi Aleksander, Kalpana Karra, Robert S Nash, Marek S Skrzypek, Shuai Weng, Stacia R Engel, and J Michael Cherry. Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources. *Genetics*, 224(1):iyac191, 01 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyac191. URL https://doi.org/10.1093/genetics/iyac191.

[27] Christopher T. Coey and David J. Clark. A systematic genome-wide account of binding sites for the model transcription factor Gcn4. 32(2):367–377. ISSN 1549-5469. doi: 10.1101/gr.276080.121.

[28] Irem Uluisik, Alaattin Kaya, Dmitri E. Fomenko, Huseyin C. Karakaya, Bradley A. Carlson, Vadim N. Gladyshev, and Ahmet Koc. Boron stress activates the general amino acid control mechanism and inhibits protein synthesis. 6(11):e27772. ISSN 1932-6203. doi: 10.1371/journal.pone.0027772.

[29] Yashpal Rawal, Răzvan V. Chereji, Vishalini Valabhoju, Hongfang Qiu, Josefina Ocampo, David J. Clark, and Alan G. Hinnebusch. Gcn4 Binding in Coding Regions Can Activate Internal and Canonical 5' Promoters in Yeast. 70(2):297–311.e4. ISSN 1097-4164. doi: 10.1016/j.molcel.2018.03.007.

[30] Stefania Oliveto, Marilena Mancino, Nicola Manfrini, and Stefano Biffo. Role of microrna s in translation regulation and cancer. *World journal of biological chemistry*, 8(1):45, 2017.

[31] Espen Enerly, Israel Steinfeld, Kristine Kleivi, Suvi-Katri Leivonen, Miriam R Aure, Hege G Russnes, Jo Anders Rønneberg, Hilde Johnsen, Roy Navon, Einar Rødland, et al. mirna-mrna integrated analysis reveals roles for mirnas in primary breast tumors. *PloS one*, 6(2):e16915, 2011.

[32] Miriam Ragle Aure, Israel Steinfeld, Lars Oliver Baumbusch, Knut Liestøl, Doron Lipson, Sandra Nyberg, Bjørn Naume, Kristine Kleivi Sahlberg, Vessela N Kristensen, Anne-Lise Børresen-Dale, et al. Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PloS one*, 8(1):e53014, 2013.

[33] Vilde D Haakensen, Israel Steinfeld, Radka Saldova, Akram Asadi Shehni, Ilona Kifer, Bjørn Naume, Pauline M Rudd, Anne-Lise Børresen-Dale, and Zohar Yakhini. Serum n-glycan analysis in breast cancer patients–relation to tumour biology and clinical outcome. *Molecular oncology*, 10(1):59–72, 2016.

[34] Vilde D Haakensen, Vegard Nygaard, Liliana Greger, Miriam R Aure, Bastian Fromm, Ida RK Bukholm, Torben Lüders, Suet-Feung Chin, Anna Git, Carlos Caldas, et al. Subtype-specific micro-rna expression signatures in breast cancer progression. *International journal of cancer*, 139(5):1117–1128, 2016.

[35] Raj Pranap Arun, Hannah F Cahill, and Paola Marcato. Breast cancer subtype-specific mirnas: Networks, impacts, and the potential for intervention. *Biomedicines*, 10(3):651, 2022.

[36] Marta Gomarasca, Paola Maroni, Giuseppe Banfi, and Giovanni Lombardi. micrornas 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, 2020.

[37] Bin Zhu, Lap Ah Tse, Difei Wang, Hela Koka, Tongwu Zhang, Mustapha Abubakar, Priscilla Lee, Feng Wang, Cherry Wu, Koon Ho Tsang, et al. Immune gene expression profiling reveals heterogeneity in luminal breast tumors. *Breast Cancer Research*, 21:1–11, 2019.

[38] Felipe Ades, Dimitrios Zardavas, Ivana Bozovic-Spasojevic, Lina Pugliano, Debora Fumagalli, Evandro De Azambuja, Giuseppe Viale, Christos Sotiriou, and Martine Piccart. Luminal b breast cancer: molecular characterization, clinical management, and future perspectives. *Journal of clinical oncology*, 32(25): 2794–2803, 2014.

[39] Mark PJ Vanderpump. The epidemiology of thyroid disease. *British medical bulletin*, 99(1), 2011.

[40] Leila Shobab, Kenneth D Burman, and Leonard Wartofsky. Sex differences in differentiated thyroid cancer. *Thyroid*, 32(3):224–235, 2022.

[41] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[42] Tao Cai, Yan He, and Binyu Peng. Incrna xist stimulates papillary thyroid cancer development through the mir-330-3p/pde5a axis. *Critical Reviews™ in Eukaryotic Gene Expression*, 33(3), 2023.

[43] Yuichiro Itoh, Lisa C Golden, Noriko Itoh, Macy Akiyo Matsukawa, Emily Ren, Vincent Tse, Arthur P Arnold, Rhonda R Voskuhl, et al. The x-linked histone demethylase kdm6a in cd4+ t lymphocytes modulates autoimmunity. *The Journal of clinical investigation*, 129(9):3852–3863, 2019.

[44] Zing Hong Eng, Mardiaty Iryani Abdullah, Khoon Leong Ng, Azlina Abdul Aziz, Nurul Hannis Arba'ie, Nurullainy Mat Rashid, and Sarni Mat Junit. 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, 2023.

[45] Xinpei Wang, Qianqian Wang, Peng Su, Chunyan Chen, Bo Han, and Zhiyan Liu. Kmt2c mutation is a diagnostic molecular marker for primary thyroid osteosarcoma: A case report and literature review. *Frontiers in Medicine*, 9:1030888, 2022.

[46] Renzhu Pang and Shuai Yang. lncrna duxap8 inhibits papillary thyroid carcinoma cell apoptosis via sponging the mir-20b-5p/sos1 axis. *Oncology reports*, 45(5):1–10, 2021.

[47] Haitao Zheng, Jie Xu, Yongli Chu, Wenzhou Jiang, Wenjie Yao, Shaowen Mo, Xicheng Song, and Jin Zhou. 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, 2022.

[48] Enakshi Saha, Marouen Ben-Guebila, Viola Fanfani, Jonas Fischer, Katherine H Shutta, Panagiotis Mandros, Dawn L DeMeo, John Quackenbush, and Camila M Lopes-Ramos. Gene regulatory networks reveal sex difference in lung adenocarcinoma. *bioRxiv*, pages 2023–09, 2023.

[49] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116 (536):2009–2022, 2021.