Supplementary Information

Supplementary Note 1: Conditions for and Proof of Convergence of Principal Components

Lemma: *Let* $\mathbf{X}$ *be a matrix of expression data with signal both due to artifacts* $\mathbf{A}$*, and due to a genuine network of linear expression relationships. Then under the conditions below and provided that the node degree distribution of the network follows a power-law, the principal components of* X *consistently estimate a linear space spanning the artifacts* A *and not the network structure.*

Proof:
Decompose a gene expression matrix with n samples and m genes $\mathbf{X}_{m\times n} = (\mathbf{x}_1, ..., \mathbf{x}_m)^t$ as follows:

$$\mathbf{X} = \boldsymbol{\mu} \times \mathbf{1} + \mathbf{\Gamma_A A} + \mathbf{\Gamma_N N} + \mathbf{U}$$

where,

- $\boldsymbol{\mu} = (\mu_1, ..., \mu_m)^t$ is an m dimensional column vector with $\mu_i := E[\mathbf{x}_i]$, $i = 1, ..., m$ and $\mathbf{1}$ is an n dimensional row vector of 1's.
- There are L artifacts or confounders ($L < n$), forming an $L \times n$ matrix A with an associated coefficient matrix Γ_A.
- N is an $m \times n$ matrix of expression data without any network structure, with associated $m \times m$ coefficient vector Γ_N. Features i and k are share an edge if γ^N_{ik} or γ^N_{ki} are nonzero. This represents a linear relationship between the expression levels of genes. To avoid circularity, the diagonal entries of Γ_N are set to zero.
- $\mathbf{U}$ is an $m \times n$ matrix of pairwise independent mean zero random noise

Based on our previous work [1], we make the following additional assumptions about the behavior of the data in the experiment.

1. The number of non-zero entries in the network Γ_N follows a power-law distribution with an exponential coefficient $2 < \alpha < 3$ [2]. As we point out in the main text power-law degree distributions have been observed in gene expression networks, for example yeast co-expression networks [3, 4] and *Caenorhabditis elegans* [5], and the preferential attachment model characteristic of scale-free networks has been explained by gene duplication [6, 7, 8]. Further, network inference algorithms such as WGCNA also employ this assumption.

2. The entries in the artifact and network coefficient, pre-network expression data, and independent noise matrices have bounded fourth moment:

$$\begin{aligned} 0 < E\left[\left(\gamma_{A_{i,j}}\right)^4\right] &\leq B_{\gamma_A} \\ 0 < E\left[\left(\gamma_{N_{i,j}}\right)^4\right] &\leq B_{\gamma_N} \\ 0 < E\left[\left(N_{i,j}\right)^4\right] &\leq B_N \\ 0 < E\left[\left(u_{i,j}\right)^4\right] &\leq B_U. \end{aligned}$$

   Therefore, by Liyapunov's inequality, there exist (finite) bounds B'_{γ_A}, B'_{γ_N}, B'_N, and B'_U, on the variances:

$$\begin{aligned} 0 < \mathrm{Var}\left(\gamma_{\mathrm{A_{i,j}}}\right) = E\left[\left(\gamma_{A_{i,j}}\right)^2\right] &\leq B'_{\gamma_A} \\ 0 < \mathrm{Var}\left(\gamma_{\mathrm{N_{i,j}}}\right) = E\left[\left(\gamma_{N_{i,j}}\right)^2\right] &\leq B'_{\gamma_N} \\ 0 < \mathrm{Var}\left(\mathrm{N_{i,j}}\right) = E\left[\left(N_{i,j}\right)^2\right] &\leq B'_N \\ 0 < \mathrm{Var}\left(\mathrm{u_{i,j}}\right) = E\left[\left(u_{i,j}\right)^2\right] &\leq B'_U. \end{aligned}$$

   This is true for most common distributions used to model gene expression data or a suitably transformed version.

3. There exists a positive definite matrix Δ for which the following hold:

   (a) $\lim_{m\to\infty} \|\frac{1}{m} A^t \Gamma_A^t \Gamma_A A - A^t \Delta A\|_F = 0$

   (b) $A^t \Delta A$ has eigenvalues $\lambda_1 > .... > \lambda_L > \lambda_{L+1} = ... = \lambda_n = 0$

   This assumption means that the batch effects and other artifacts are sufficiently widespread as to affect a fixed and non-negligable percentage of the genes in the data set.

Additionally, we assume without loss of generality, that expression levels of each gene in $\mathbf{X}$ is centered.

4. $\boldsymbol{\mu} = \vec{0}$.

5. The expression data in the absence of any network structure, N, has mean $E[N] = \vec{0}$ where $\vec{0}$ is an m-dimensional column vector. Further, in the absence of network structure, the genes are pairwise independent. Therefore, by Assumption 2 the entries of N converge almost surely to zero.

Based on this model, we show that the principal components of the matrix $\mathbf{X}$ estimate the artifacts and are not corrupted by the signal from the network terms.
The eigen-vectors of the matrix $\frac{1}{m}\mathbf{X}^T\mathbf{X}$ are equal to the right singular vectors of the matrix $\mathbf{X}$. Given

observed data $\mathbf{X}$, the empirical variance-covariance matrix of the data $\hat{\Sigma}$ takes the form:

$$\begin{aligned}
\hat{\Sigma} &= \frac{1}{m}\mathbf{X}^t\mathbf{X} \\
&= \frac{1}{m}\left(\Gamma_A A + \Gamma_N N + \mathbf{U}\right)^t\left(\Gamma_A A + \Gamma_N N + \mathbf{U}\right) \\
&= \frac{1}{m}\left(A^t\Gamma_A^t + N^t\Gamma_N^t + \mathbf{U}^t\right)\left(\Gamma_A A + \Gamma_N N + \mathbf{U}\right) \\
&= \frac{1}{m}(A^t\Gamma_A^t\Gamma_A A + A^t\Gamma_A^t\Gamma_N N + A^t\Gamma_A^t\mathbf{U} + N^t\Gamma_N^t\Gamma_A A + N^t\Gamma_N^t\Gamma_N N + N^t\Gamma_N^t\mathbf{U} + \\
&\quad \mathbf{U}^t\Gamma_A A + \mathbf{U}^t\Gamma_N N + \mathbf{U}^t\mathbf{U}) \\
&= \frac{1}{m}\left(A^t\Gamma_A^t\Gamma_A A + A^t\Gamma_A^t\Gamma_N N + N^t\Gamma_N^t\Gamma_A A + N^t\Gamma_N^t\Gamma_N N\right) + \\
&\quad \frac{1}{m}\left(A^t\Gamma_A^t\mathbf{U} + N^t\Gamma_N^t\mathbf{U} + \mathbf{U}^t\Gamma_A A + \mathbf{U}^t\Gamma_N N + \mathbf{U}^t\mathbf{U}\right) \\
&= \frac{1}{m}A^t\Gamma_A^t\Gamma_A A + \frac{1}{m}A^t\Gamma_A^t\Gamma_N N + \frac{1}{m}N^t\Gamma_N^t\Gamma_A A + \frac{1}{m}N^t\Gamma_N^t\Gamma_N N + \\
&\quad \frac{1}{m}A^t\Gamma_A^t\mathbf{U} + \frac{1}{m}N^t\Gamma_N^t\mathbf{U} + \frac{1}{m}\mathbf{U}^t\Gamma_A A + \frac{1}{m}\mathbf{U}^t\Gamma_N N + \frac{1}{m}\mathbf{U}^t\mathbf{U}
\end{aligned}$$

We will show that as the number of features grows, the empirical variance-covariance matrix, after centering by an estimate of the background variation, converges to the same thing as if there were no network structure:

$$\tilde{\mathbf{X}}_{unstr} := \Gamma_A A + \mathbf{U}.$$

Then we can show that the principal components of the confounded matrix are consistent estimators of the confounding variables.
Therefore, we will show that, holding the number of observations n fixed, there exits an $n \times n$ matrix $\mathcal{L}$ so that:

$$\lim_{m\to\infty}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^t\tilde{\mathbf{X}}^{\text{unstr}} - \hat{\sigma}^2_{ave}\mathbf{I} = \mathcal{L}$$

$$\lim_{m\to\infty}\frac{1}{m}\mathbf{X}^t\mathbf{X} - \hat{\sigma}^2_{ave}\mathbf{I} = \mathcal{L}$$

where, borrowing the notation from Leek 2011, we let $V_L(\mathbf{X}) = \{v_1(\mathbf{X}), ..., v_L(\mathbf{X})\}$ be a matrix of the first L right singular vectors of $\mathbf{X}$ and $\hat{\Gamma}_L$ the least squares estimates from regressing $\mathbf{X}$ on $V_L(\mathbf{X})$. Then, we define:

$$\sigma^2_{ave} := \frac{1}{m(n-L)}\|\mathbf{X} - \hat{\Gamma}_L V_L(\mathbf{X})\|_F,$$

where we estimate L using a permutation approach through the 'num.sv' function in the *sva* package.

To determine $\mathcal{L}$, we write:

$$\begin{aligned}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^t\tilde{\mathbf{X}}^{\text{unstr}}-\hat{\sigma}^2_{ave}\mathbf{I} &= \frac{1}{m}\left(\Gamma_A A+U\right)^t\left(\Gamma_A A+U\right)-\hat{\sigma}^2_{ave}\mathbf{I}\\ &= \frac{1}{m}\left(A^t\Gamma_A^t+U^t\right)\left(\Gamma_A A+U\right)-\hat{\sigma}^2_{ave}\mathbf{I}\\ &= \frac{1}{m}A^t\Gamma_A^t\Gamma_A A+\frac{1}{m}A^t\Gamma_A^t U+\frac{1}{m}U^t\Gamma_A A+\frac{1}{m}^t U-\hat{\sigma}^2_{ave}\mathbf{I}\end{aligned}$$

Letting m grow,

$$\begin{aligned}&\lim_{m\to\infty}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^t\tilde{\mathbf{X}}^{\text{unstr}}-\hat{\sigma}^2_{ave}\mathbf{I}\\ &\quad=\lim_{m\to\infty}\frac{1}{m}\Gamma_A^t\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t U+\lim_{m\to\infty}\frac{1}{m}U^t\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}U^tU-\hat{\sigma}^2_{ave}\mathbf{I}\\ &\quad=A^t\Delta A+\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t U+\lim_{m\to\infty}\frac{1}{m}U^t\Gamma_A A+\lim_{m\to\infty}\frac{1}{m}U^tU-\hat{\sigma}^2_{ave}\mathbf{I}\end{aligned}$$

Leek 2011 shows that the terms $\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t U+\lim_{m\to\infty}\frac{1}{m}U^t\Gamma_A A$ both converge almost surely to zero by the Kolmogorov Strong Law of Large Numbers (KSLLN). Further, Leek 2011 uses KSLLN to show that the off diagonal elements of $\frac{1}{m}U^tU$ converge almost surely to zero, while the diagonals converge almost surely to $\hat{\sigma}^2_{ave}$. Therefore,

$$\lim_{m\to\infty}\frac{1}{m}\left(\tilde{\mathbf{X}}^{\text{unstr}}\right)^t\tilde{\mathbf{X}}^{\text{unstr}}-\hat{\sigma}^2_{ave}\mathbf{I}=A^t\Delta A,$$

and

$$\mathcal{L}=A^t\Delta A.$$

The limit of the empirical variance-covariance matrix is as follows:

$$
\begin{aligned}
&\lim_{m\to\infty}\frac{1}{m}\mathbf{X}^t\mathbf{X}-\hat{\sigma}^2_{ave}\mathbf{I}\\
&=\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_AA+\lim_{m\to\infty}\frac{1}{m}\Gamma_AA^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_AA+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_NN+-\hat{\sigma}^2_{ave}\mathbf{I}\\
&\quad\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\mathbf{U}+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\mathbf{U}+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_AA+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\mathbf{U}-\hat{\sigma}^2_{ave}\mathbf{I}\\
&=\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_AA+\lim_{m\to\infty}\frac{1}{m}\Gamma_AA^tU+\lim_{m\to\infty}\frac{1}{m}U^t\Gamma_AA+\lim_{m\to\infty}\frac{1}{m}U^tU-\hat{\sigma}^2_{ave}\mathbf{I}+\\
&\quad\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_AA+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\mathbf{U}\\
&=A^t\Delta A+\lim_{m\to\infty}\frac{1}{m}U^tU-\hat{\sigma}^2_{ave}\mathbf{I}+\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_AA+\\
&\quad\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_NN+\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\mathbf{U}\\
&=A^t\Delta A+\underbrace{\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_NN}_{(1)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_AA}_{(2)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_NN}_{(3)}+\\
&\quad\underbrace{\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_NN}_{(4)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\mathbf{U}}_{(5)}
\end{aligned}
$$

We consider the convergence of (1) through (5) separately:

1. $$\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_NN=\lim_{m\to\infty}A^t\frac{1}{m}\Gamma_A^t\Gamma_NN$$

   We first consider $Q:=\frac{1}{m}\Gamma_A^t\Gamma_N$, an $L\times m$ matrix with entries indexed by $l\in\{1,...,L\},k\in\{1,...,m\}$:

$$
\begin{aligned}
q_{lk}&=Q_{l,k}\\
&=\frac{1}{m}\sum_{j=1}^{m}\Gamma_{A_{j,l}}\Gamma_{N_{j,k}}\\
&=\frac{1}{m}\sum_{j=1}^{m}\gamma_{A_{j,l}}\gamma_{Nj,k}\\
&=\frac{1}{m}\sum_{\{j:\gamma_{N_{j,k}}\neq0\}}\gamma_{A_{j,l}}\gamma_{N_{j,k}}+\frac{1}{m}\sum_{\{j:\gamma_{N_{j,k}}=0\}}\gamma_{A_{j,l}}\gamma_{N_{j,k}}\\
&=\frac{1}{m}\sum_{\{j:\gamma_{N_{j,k}}\neq0\}}\gamma_{A_{j,l}}\gamma_{N_{j,k}}+\frac{1}{m}\sum_{\{j:\gamma_{N_{j,k}}=0\}}\gamma_{A_{j,l}}\times0\\
&=\frac{1}{m}\sum_{\{j:\gamma_{N_{j,k}}\neq0\}}\gamma_{A_{j,l}}\gamma_{N_{j,k}}
\end{aligned}
$$

Suppose that there are $0 \leq d \leq m$ indices j for which $\gamma_{N\,j,k} \neq 0$, so that there are d terms $\gamma_{A_{j,l}}\gamma_{N_{j,k}}$ in the summation contributing to q_{lk}. We can re-index these terms as $\gamma_{A_{j',l}}\gamma_{N_{j',k}}, j' = 1, ..., d$.

For any fixed k, whenever $\gamma_{N_{j,k}}$, necessarily genes k and j share an edge. Therefore, given d non-zero coefficients $\gamma_{N_{j,k}}$, gene k has at least degree d. However, [9] show that for scale free networks following a power-law degree distribution $p_k \sim k^{\alpha-1}$, as assumed in our framework, the maximum degree of a vertex in the network follows $k_{\max} \sim m^{\frac{1}{\alpha-1}}$, and $d \leq m^{\frac{1}{\alpha-1}}$. Therefore, we can write each element as:

$$\begin{aligned}
q_{lk} &= \frac{1}{m} \sum_{\{j:\gamma_{N_{j,k}} \neq 0\}} \gamma_{A_{j,l}}\gamma_{N_{j,k}} \\
&= \frac{1}{m} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}} \\
&= \frac{d}{m}\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}} \\
&\leq \frac{m^{\frac{1}{\alpha-1}}}{m}\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}} \\
&= m^{-1}m^{\frac{1}{\alpha-1}}\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}} \\
&= m^{\frac{2-\alpha}{\alpha-1}}\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}} \\
&= m^{\frac{-(\alpha-2)}{\alpha-1}}\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}} \\
&= \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d} \sum_{j'=1}^{d} \gamma_{A_{j',l}}\gamma_{N_{j',k}}
\end{aligned}$$

By Assumption 1 ($2 < \alpha < 3$), so that $\frac{\alpha-1}{\alpha-2} > 1$ and $\lim_{m\to\infty} \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} = 0$.

Now, consider the expectation of the terms inside of the summation. For any j', applying the Cauchy-Schwarz inequality to $|\gamma_{A_{j',l}}|$ and $|\gamma_{N_{j',k}}|$

$$\begin{aligned}
E\left[|\gamma_{A_{j',l}}\gamma_{N_{j',k}}|\right] &\leq \sqrt{E\left[|\gamma_{A_{j',l}}|^2\right] E\left[|\gamma_{N_{j',k}}|^2\right]} \\
&= \sqrt{E\left[(\gamma_{A_{j',l}})^2\right] E\left[(\gamma_{N_{j',k}})^2\right]} \\
&\leq \sqrt{B'_{\gamma_A} \times B'_{\gamma_N}} \quad \text{By Assumption 2} \\
&= B^* \quad \text{where we define the bound } \; B^* := \sqrt{B'_{\gamma_A} \times B'_{\gamma_N}},
\end{aligned}$$

and

$$-\infty < -B^* \leq E\left[\gamma_{A_{j',l}}\gamma_{N_{j',k}}\right] \leq B^* < \infty,$$

and by the Strong Law of Large Numbers,

$$\frac{1}{d}\sum_{j'=1}^{d}\gamma_{A_{j',l}}\gamma_{N_{j',k}} \xrightarrow{a.s.} E\left[\gamma_{A_{j',l}}\gamma_{N_{j',k}}\right],$$

therefore, for each l, k:

$$q_{lk} \leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d}\sum_{j'=1}^{d}\gamma_{A_{j',l}}\gamma_{N_{j',k}} \xrightarrow{a.s.} 0$$

s

and

$$Q \xrightarrow{a.s.} 0.$$

Recall, the matrix of artifacts A is $L \times n$ dimensional, so that it is fixed with respect to m, and, as shown in Assumption 5, $N \xrightarrow{a.s.} 0$, so that by Slutsky's Theorem:

$$\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_N N = 0$$

2.

$$\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_A A$$

By symmetry, the same argument as in (1) holds, and

$$\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_A A = 0$$

3.

$$\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_N N = \lim_{m\to\infty}N^t\frac{1}{m}\Gamma_N^t\Gamma_N N$$

We will first consider $P := \frac{1}{m}\Gamma_N^t\Gamma_N$, an $m \times m$ matrix with entries indexed by $l, k \in \{1, ..., m\}$:

$$\begin{aligned} p_{lk} &= P_{l,k} \\ &= \frac{1}{m}\sum_{j=1}^{m}\Gamma_{N_{j,l}}\Gamma_{N_{j,k}} \\ &= \frac{1}{m}\sum_{j=1}^{m}\gamma_{N_{j,l}}\gamma_{N\,j,k} \end{aligned}$$

We will consider the diagonal and off-diagonal entries of P separately. The diagonal entries

$(k = l)$ take the form:

$$\begin{aligned} p_{ll} &= \frac{1}{m}\sum_{j=1}^{m}\gamma_{N_{j,l}}\gamma_{N\,j,l} \\ &= \frac{1}{m}\sum_{j=1}^{m}\gamma^2_{N_{j,l}} \\ &= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0\}}\gamma^2_{N_{j,l}} + \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}=0\}}\gamma^2_{N_{j,l}} \\ &= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0\}}\gamma^2_{N_{j,l}} \end{aligned}$$

Now, whenever $\gamma_{N_{j,l}} \neq 0$, by definition, genes j and l share an edge, so that d', the number of j such that $\gamma_{N_{j,l}} \neq 0$ is equal to the degree of vertex l. Following the argument from the proof of (1) , $d' \leq m^{\frac{1}{\alpha-1}}, 2 < \alpha < 3$ and:

$$\begin{aligned} p_{ll} &= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0\}}\gamma^2_{N_{j,l}} \\ &= \frac{1}{m}\sum_{j'=1}^{d'}\gamma^2_{N_{j',l}} \\ &= \frac{d'}{m}\frac{1}{d'}\sum_{j'=1}^{d'}\gamma^2_{N_{j',l}} \\ &\leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d'}\sum_{j'=1}^{d'}\gamma^2_{N_{j',l}} \end{aligned}$$

Again, by Assumption 1

$$\lim_{m\to\infty}\frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} = 0.$$

Further, by Assumption 2

$$E\left[\gamma^4_{N_{j',l}}\right] \leq B_{\gamma_N},$$

so that applying the Strong Law of Large Numbers,

$$\frac{1}{d}\sum_{j'=1}^{d'}\gamma^2_{N_{j',l}} \xrightarrow{a.s.} E\left[\gamma^2_{N_{j',l}}\right] \leq B'_{\gamma_N},$$

and for each l:

$$0 \leq p_{ll} \leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d'}\sum_{j'=1}^{d'}\gamma^2_{N_{j',l}} \xrightarrow{a.s.} 0.$$

We now consider the off-diagonal entries($k \neq l$):

$$\begin{aligned}
p_{lk} &= P_{l,k} \\
&= \frac{1}{m}\sum_{j=1}^{m} \Gamma_{N_{j,l}}\Gamma_{N_{j,k}} \\
&= \frac{1}{m}\sum_{j=1}^{m} \gamma_{N_{j,l}}\gamma_{N\,j,k} \\
&= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0 \text{ and } \gamma_{\mathrm{N_{j,k}}}\neq 0\}} \gamma_{N_{j,l}}\gamma_{N\,j,k} + \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}=0 \text{ or } \gamma_{\mathrm{N_{j,k}}}=0\}} \gamma_{N_{j,l}}\gamma_{N\,j,k} \\
&= \frac{1}{m}\sum_{\{j:\gamma_{N_{j,l}}\neq 0 \text{ and } \gamma_{\mathrm{N_{j,k}}}\neq 0\}} \gamma_{N_{j,l}}\gamma_{N\,j,k}
\end{aligned}$$

If both $\gamma_{N_{j,l}} \neq 0$ and $\gamma_{N_{j,k}} \neq 0$ then gene j shares and edge with both genes l and k, so that d', the number of j such that $\gamma_{N_{j,l}} \neq 0$ and $\gamma_{N_{j,k}} \neq 0$ will be bounded by the maximum of the degrees of vertices l and k. The same argument as used for the diagonal entries then follows:

$$p_{lk} \leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d''}\sum_{j'=1}^{d'} \gamma_{N_{j',l}}\gamma_{N_{j',k}},$$

and

$$\lim_{m\to\infty} \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}} = 0.$$

Further, for any j', by Assumption 2 and the Cauchy-Schwarz inequality to $|\gamma_{N_{j',l}}|$ and $|\gamma_{N_{j',k}}|$

$$\begin{aligned}
E\left[|\gamma_{N_{j',l}}\gamma_{N_{j',k}}|\right] &\leq \sqrt{E\left[|\gamma_{N_{j',l}}|^2\right] E\left[|\gamma_{N_{j',k}}|^2\right]} \\
&= \sqrt{E\left[(\gamma_{N_{j',l}})^2\right] E\left[(\gamma_{N_{j',k}})^2\right]} \\
&\leq \sqrt{B'_{\gamma_N} \times B'_{\gamma_N}}
\end{aligned}$$

and

$$-\infty < -(B'_{\gamma_N})^2 \leq E\left[\gamma_{N_{j',l}}\gamma_{N_{j',k}}\right](B'_{\gamma_N})^2 < \infty,$$

and by the Strong Law of Large Numbers,

$$\frac{1}{d''}\sum_{j'=1}^{d''} \gamma_{N_{j',l}}\gamma_{N_{j',k}} \xrightarrow{a.s.} E\left[\gamma_{N_{j',l}}\gamma_{N_{j',k}}\right],$$

therefore, for each $l \neq k$:

$$p_{lk} \leq \frac{1}{m^{\frac{\alpha-1}{\alpha-2}}}\frac{1}{d''}\sum_{j'=1}^{d''} \gamma_{N_{j',l}}\gamma_{N_{j',k}} \xrightarrow{a.s.} 0.$$

Therefore, both the diagonal and off-diagonal entries in P converge to zero, and

$$P \xrightarrow{a.s.} 0.$$

As shown in Assumption 5, $N \xrightarrow{a.s.} 0$, so that by Slutsky's Theorem:

$$\lim_{m\to\infty} \frac{1}{m} N^t \Gamma_N^t \Gamma_N N = 0$$

4.
$$\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_N N$$
This term converges almost surely to zero by the KSLLN since $E[U]=0$ and Γ_N and U have bounded fourth moments.

5.
$$\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\mathbf{U}$$
This term converges almost surely to zero by the KSLLN since $E[U]=0$ and Γ_N and U have bounded fourth moments.

Therefore, all of the terms (1)-(5) converge almost surely to zero and the limit of the empirical variance-covariance matrix is

$$\lim_{m\to\infty}\frac{1}{m}\mathbf{X}^t\mathbf{X}-\hat{\sigma}^2_{ave}\mathbf{I}=\mathbf{A^t\Delta A}+\underbrace{\lim_{m\to\infty}\frac{1}{m}A^t\Gamma_A^t\Gamma_N N}_{(1)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_A A}_{(2)}+$$
$$\underbrace{\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\Gamma_N N}_{(3)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}\mathbf{U}^t\Gamma_N N}_{(4)}+\underbrace{\lim_{m\to\infty}\frac{1}{m}N^t\Gamma_N^t\mathbf{U}}_{(5)}=A^t\Delta A=\mathcal{L}$$

The principal components of this matrix consistently estimate the space spanend by the confounding artifacts as we have previously demonstrated [1].

Supplementary Note 2: Online Methods

All analyses was performed using R and scripts are available on github at:
`https://github.com/leekgroup/networks_correction`

2.1 Simulation example

We construct a true underlying network with eight nodes that represent genes and three edges that represent conditional dependencies between the genes. Next, we simulate 10,000 observations from a multivariate normal distribution that encode the conditional dependencies corresponding to three edges as non-zero entries in the precision matrix (Figure 1a). Then, to introduce confounding in the data, we simulate a sample specific term from a standard normal distribution, and add a scalar multiple of that to genes 2 through 6 (Figure 1d). Finally, to correct the data, we regress out the first principal component from the confounded data (Figure 1g). We used graphical lasso to reconstruct networks using the three versions of the data. The code for this simulation example and network reconstruction can be found at: `https://github.com/leekgroup/networks_correction/blob/master/publication_rmd/simulation_example_fig1/figure1.Rmd`

2.2 Co-expression network reconstruction

To evaluate our correction method and it's effect on reconstruction of co-expression networks, we used two methods to infer the structure of gene co-expression networks: a) weighted gene co-expression networks (WGCNA)[10] and b) graphical lasso[11]

2.2.1 WGCNA

Weighted Gene Co-expression Network Analysis (WGCNA) identifies relationships between genes through a power transform of the Pearson correlation coefficient. The algorithm first computes the pairwise correlation between genes i and j:

$$a_{ij} = |corr(x_i, x_j)|^{\beta} \tag{1}$$

where $\beta \geq 1$. β is the soft-thresholding parameter and it's value is selected such that the networks obtained are scale free. This is assesed by setting a scale-free topology fit R^2 at least 0.85, between $log(p(k))$ and $log(k)$ where $p(k)$ is the fraction of nodes with at least k neighbors[12, 10, 13, 14].Next, the adjacency matrix between genes is transformed into a topological overlap matrix (TOM) [12]. The TOM is then input into an average linkage hierarchical clustering algorithm to identify network modules – defined as 'groups of nodes with high topological overlap and indicates high levels of co-expression [12].

2.2.2 Graphical Lasso

Given our gene expression data contains N multivariate gaussian observations each of dimension p, i.e. for each observation, we have expression measurements for p genes, graphical lasso estimates the structure of the co-expression network over genes by maximizing L_1-penalized log likelihood of a multivariate gaussian with mean μ and covariance Σ over Θ:

$$\log\det\Theta - \text{trace}(\text{S}\Theta) - \lambda||\Theta||_1 \tag{2}$$

Here S is the empirical covariance matrix and $\Theta = \Sigma^{-1}$ is the inverse covariance matrix. The L_1 penalty on Θ induces and controls the amount of sparsity in the solution [11]. Hence, if an entry $\Theta_{i,j}$ is 0, then variable i is conditionally independent of variable j given other variables. We used 'QUIC' package[15] in R to infer co-expression network structure with graphical lasso.

2.3 Determining sample specific estimate of GC bias

Studies have shown that GC content of genes have significant impact on sequencing read coverage in DNA-seq and RNA-seq experiments. This eventually introduces sample specific biases in expression quantification. To quantify the effect of GC bias, using transcript level fasta files from Gencode v25 we first computed the GC% of each transcript by:

$$GC\%(T) = \frac{(\#G + \#C)}{(\#A + \#T + \#G + \#C)}$$

We summarized GC content of genes, by averaging over all transcripts belonging to the gene. Suppose k transcripts were transcribed from gene G_i then,

$$GC\%(G_i) = \frac{\sum_{j=1}^{k} GC\%(T_j)}{k}$$

Next using a linear model, we obtain sample specific estimates of GC content of genes:

$$E_i = \mu + \beta_i \times G$$

where, E_i is the vector of expression values of all genes in sample i, G is the GC content for each gene and β_i is the estimate of GC bias for sample i.

2.4 Network reconstruction using GTEx data

We used gene expression RNAseq data from five tissues in the GTEx project [16] that included Adipose Subcutaneous, Lung, Muscle Skeletal, Thyroid and Whole Blood. In each tissue we filtered for genes that had scaled expression (counts scaled by the total coverage of the sample) of at least 0.1 in $\geq 25\%$ of total number of observations. Next, we $\log_2$ transformed the scaled gene expression data, and performed following steps to select the most variable 5000 genes across five tissues:

(a) Select genes expressed in all five tissues.

(b) For each tissue, assign a rank to each gene by variance, such that the most variable gene is ranked first and least variable gene is ranked in last.

(c) Using the ranked list of genes from five tissues, assign an average rank to each gene across five tissues.

(d) Select top 5000 genes from the average rank list for network inference with WGCNA and graphical lasso.

2.4.1 Network inference with WGCNA

For learning the structure of networks using WGCNA, for each tissue, we performed the following steps for data correction and network inference:

- Raw uncorrected expression data:
    1. We transformed the expression of each gene to a Gaussian distribution by projecting the expression of each gene to the quantiles of a standard normal.
    2. Next, to reconstruct unsigned weighted co-expression networks with WGCNA, we identified lowest power for which scale-free fit R^2 between $\log(p(k))$ and $\log(k)$ exceeds 0.85. Here $p(k)$ is the fraction of node in the network with at least k neighbors
    3. After that we used the 'blockwisemodules' function in the *WGCNA* CRAN package to perform co-expression module detection at varying cut-heights of hierarchical dendrogram ranging from 0.9 to 1.0.
- Residuals from RIN/Exonic Rate/ GC bias:
    1. First using a linear model, we regressed the RNA integrity number (RIN), exonic rate or sample specific estimate of GC bias on the expression data and computed the residuals.
    2. Next we transformed the expression of each gene to Gaussian distribution by projecting the expression of each gene to the quantiles of a standard normal.
    3. Then for module detection, we repeated step 2-3 from raw uncorrected expression data
- Residuals from principal components:
    1. Using a permutation based approach, we determined the number of principal components to correct the data for with the 'num.sv' function in the Bioconductor package *sva*.(Supplementary Table 1)

2. Next we normalized the expression data such that expression of each gene was centered at mean and had unit standard deviation. After that, we performed singular value decomposition of the normalized expression data.
3. Using a linear model we regressed the top s principal components (s as determined in step 1) on the expression data and computed the residuals.
4. We then transformed the expression of each gene to Gaussian distribution (as described earlier).
5. For module detection, we repeated step 2-3 from raw expression data

For networks that were reconstructed with WGCNA, we considered all genes in the same module to be a fully-connected subgraph.

2.4.2 Network inference with graphical lasso

As described above, we use a permutation based approach to determine the number of principal components to include [17] in our linear model for correction of expression data. The steps for data correction and network inference for each tissue were as follows:

- Raw uncorrected expression data:
  1. We transformed the expression of each gene to Gaussian distribution (as described earlier).
  2. Next using the transformed data, we computed the gene covariance matrix and used graphical lasso for co-expression network reconstruction. For this we used the 'QUIC' function in the *QUIC*[15] R package and inferred the networks with penalization parameters λ ranging from 0.3 to 1.0.
- Residuals from RIN/Exonic rate/ GC bias
  1. Using a linear model, we regressed the RNA integrity number (RIN), exonic rate, or GC bias on the expression data and computed the residuals.
  2. To reconstruct networks, we repeated steps 1-2 from raw uncorrected expression data
- Residuals from principal components:
  1. We used the 'num.sv' function in the *sva* Bioconductor package to determine the number of principal components to correct for.
  2. Next we scaled the gene expression data such that expression of each gene was centered at mean and had a unit standard deviation. Next we performed singular value decomposition of the scaled expression data.
  3. Using a linear model, we regressed the top s principal components (as determined in step 1) on the scaled expression data and then computed the residuals.
  4. For co-expression network reconstruction, we repeated steps 1-2 from raw uncorrected expression data.

2.4.3 Network evaluation

Since the underlying network structure is generally unknown, we used a) genes known to be functional in the same pathways and b) known transcription factors and their targets as ground truth to assess these networks.

- ***Canonical pathway databases:*** We downloaded the latest pathway information (2016) from KEGG, Biocarta and Pathway Interaction Database from Enrichr[18, 19], that were also annotated as canonical pathways by MSigDB [20]. The number of pathways/genesets in each of these databases were:
  - KEGG - 293
  - Biocarta - 237
  - Reactome - 1530
  - Pathway Interaction Database - 209

  We obtained a total of 714616 unique real connections from these databases. Any pair of genes that have at least one pathway in common were assumed as true connection. An edge that was observed between a pair of genes in the inferred network (from WGCNA or graphical lasso) and was also present in the list of *real connections* was called as a true positive (TP). We defined false positive (FP) to be an edge that was observed between a pair of genes in the inferred network, however was absent in the list of *real connections*. False negatives (FN) were the edges that were missing in the inferred network but were present in the list of *real connections*. Using this definition of true positive, false positive and false negative, we compute *precision* and *recall* for the networks inferred by WGCNA and graphical lasso with different forms of corrected and uncorrected data.

- ***Shared true positives:*** We obtained a refined list of *real connections* described above by restricting to pairs of genes that were present in at least two pathway databases. By doing this, we reduced to a more confident set of 49559 unique real connections.

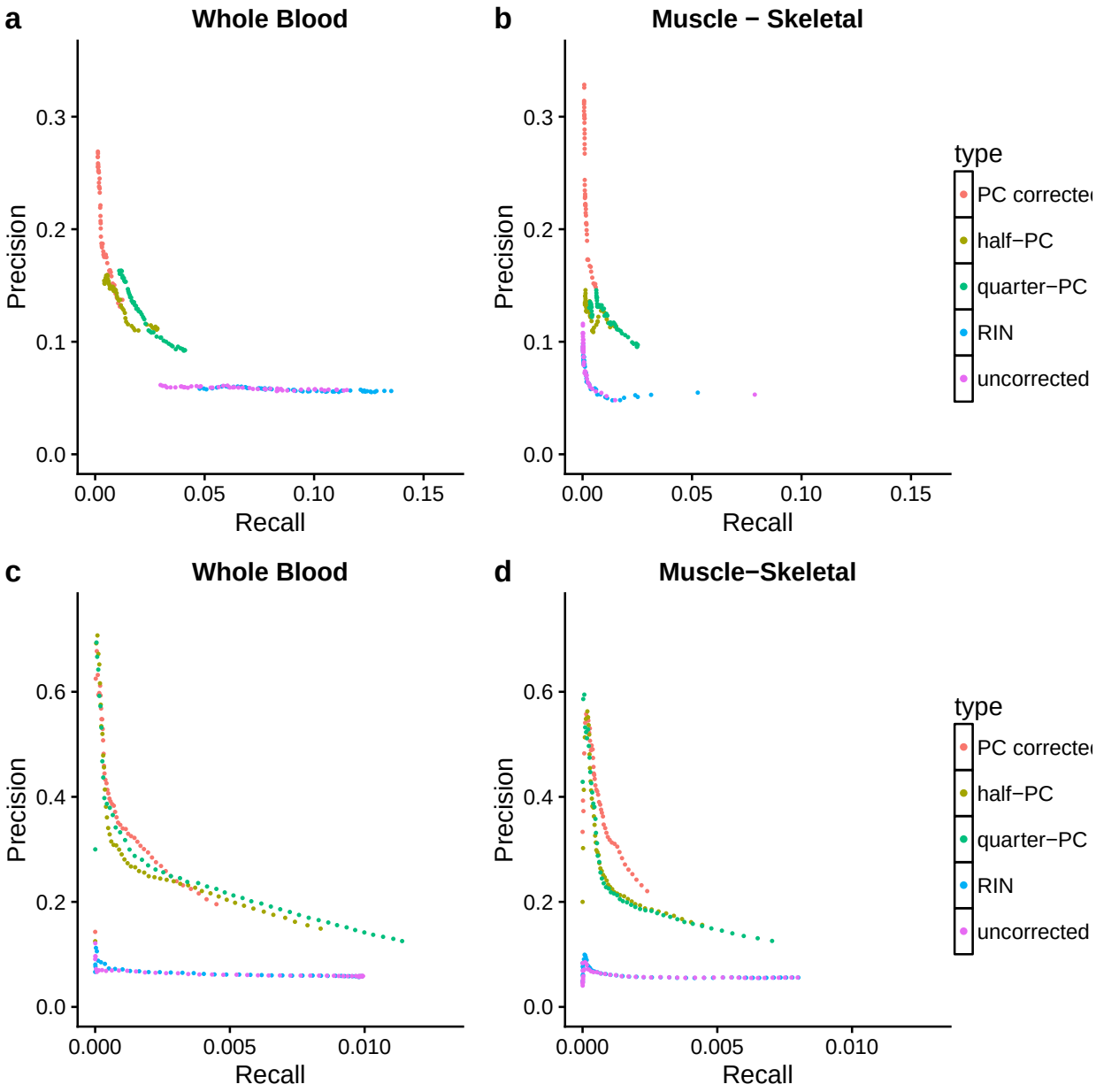
- ***Transcription factor targets:*** We downloaded the list of known transcription factor target genes (ChEA 2016) from Enrichr[18, 19]. Connection between a each transcription factor with it's targets were called as true connections. Edges inferred from co-expression networks that were also true connections were called as true positives. False positives and false negatives were defined as described above.

All TP, FP and FN were computed with genes restricted to the most variable 5000 genes that were used for reconstructing co-expression networks. Using the above mentioned definitions of TP, FP and FN, we compute precision and recall as given below:

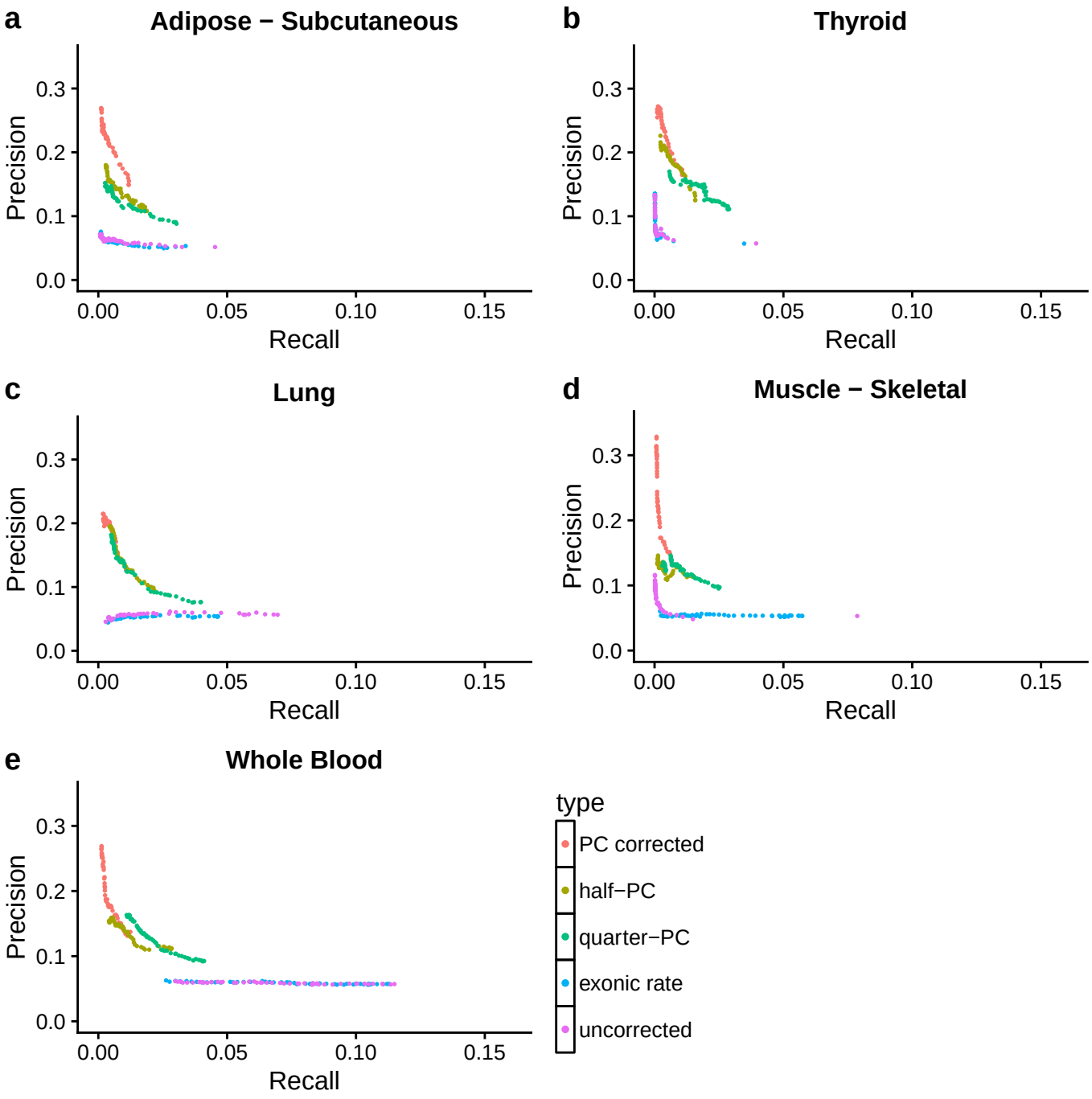
$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

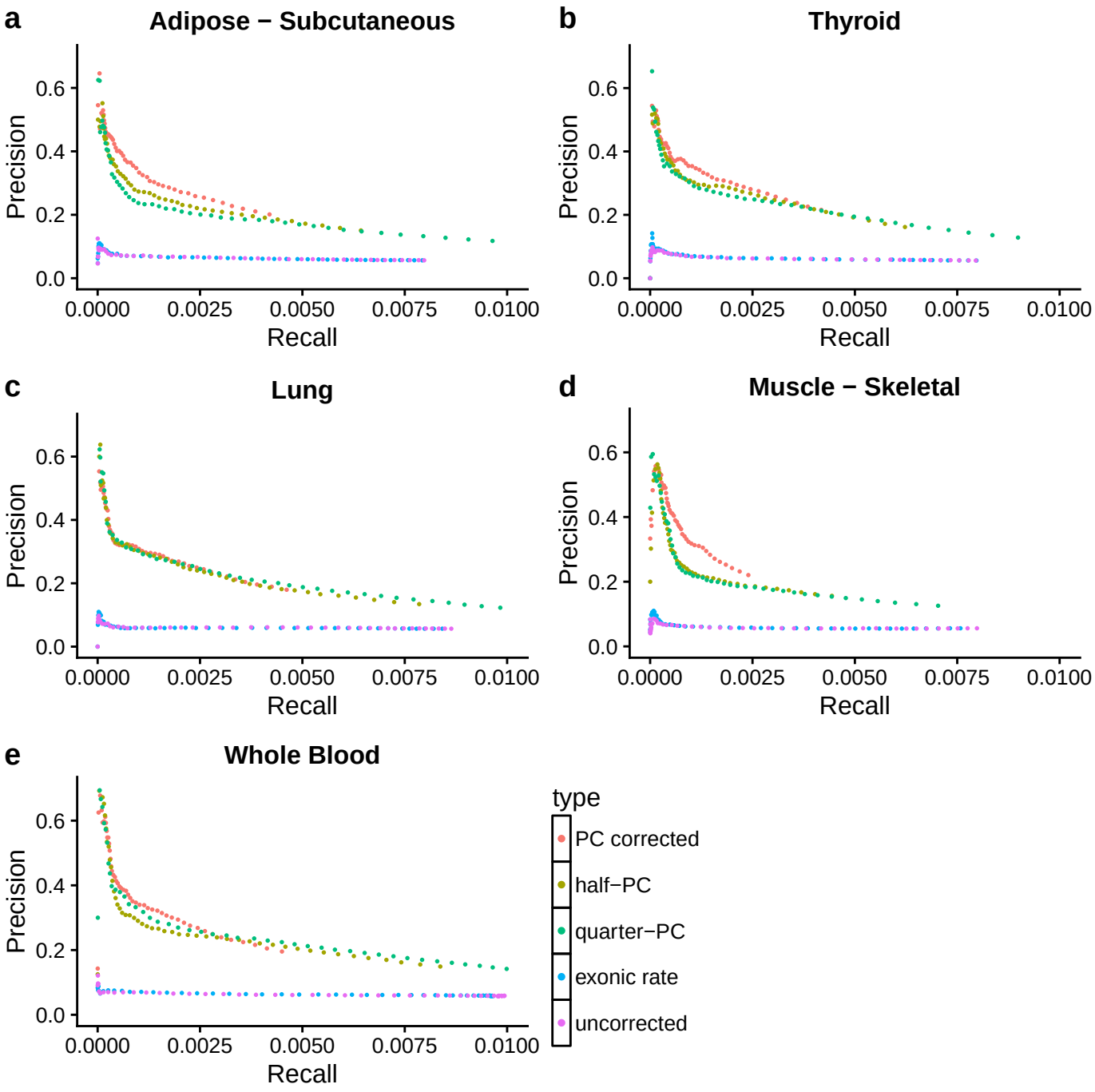
Supplementary Note 3: Supplementary Figures and Tables



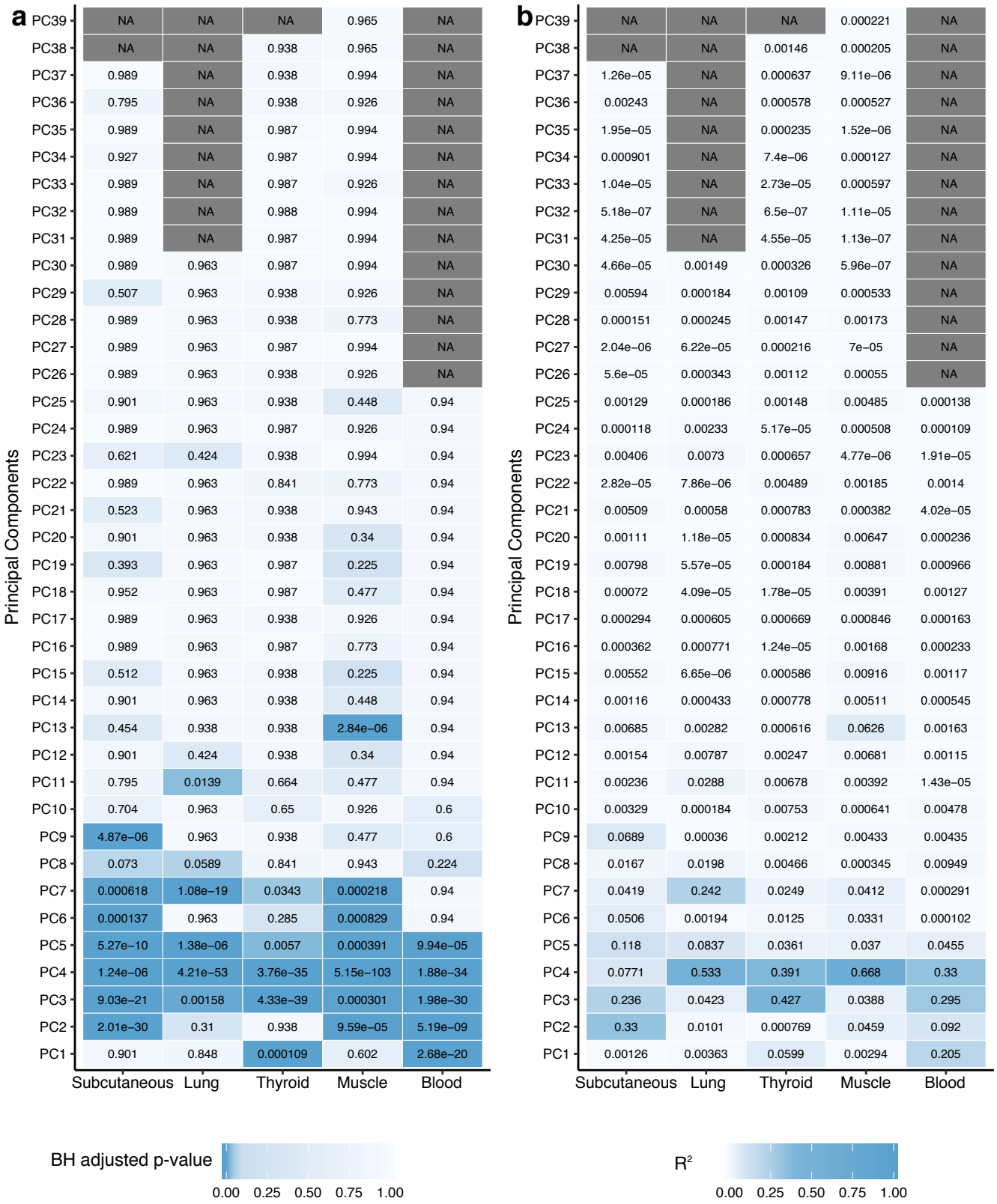
Supplementary Figure 1: Precision and recall curves of (a-b) WGCNA networks obtained at a varying cut-heights. (c-d) networks obtained from graphical lasso with varying size of penalty parameter. Each point corresponds to the network obtained at a specific cut-height in WGCNA or penalty parameter in graphical lasso



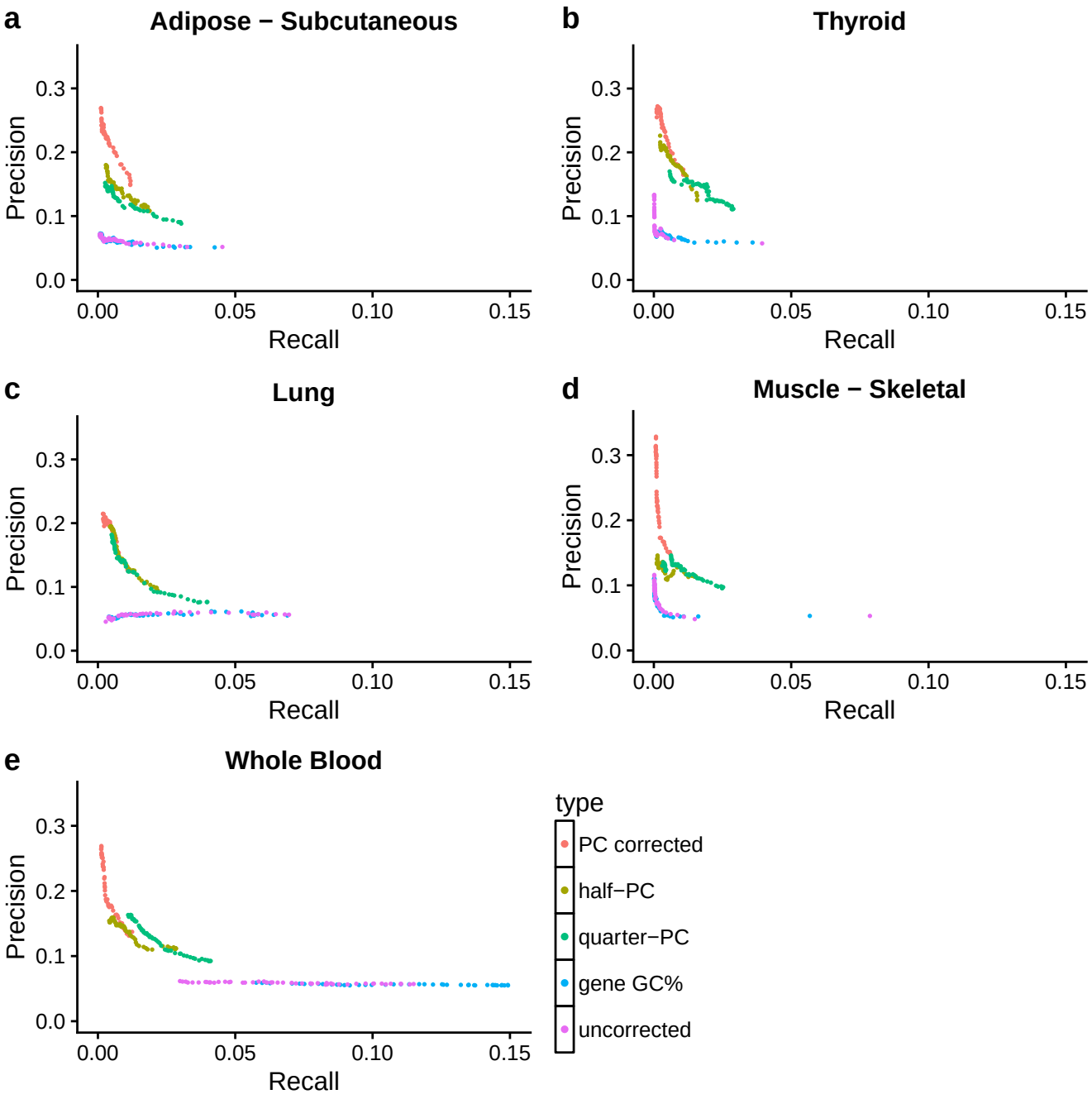
Supplementary Figure 2: Precision-Recall curves of WGCNA modules using canonical pathway databases. WGCNA networks obtained at a varying cut-heights. Each color corresponds to the correction approach, and each point corresponds to the network obtained at a specific cut-height. Exonic rate is the known confounder used in this figure.



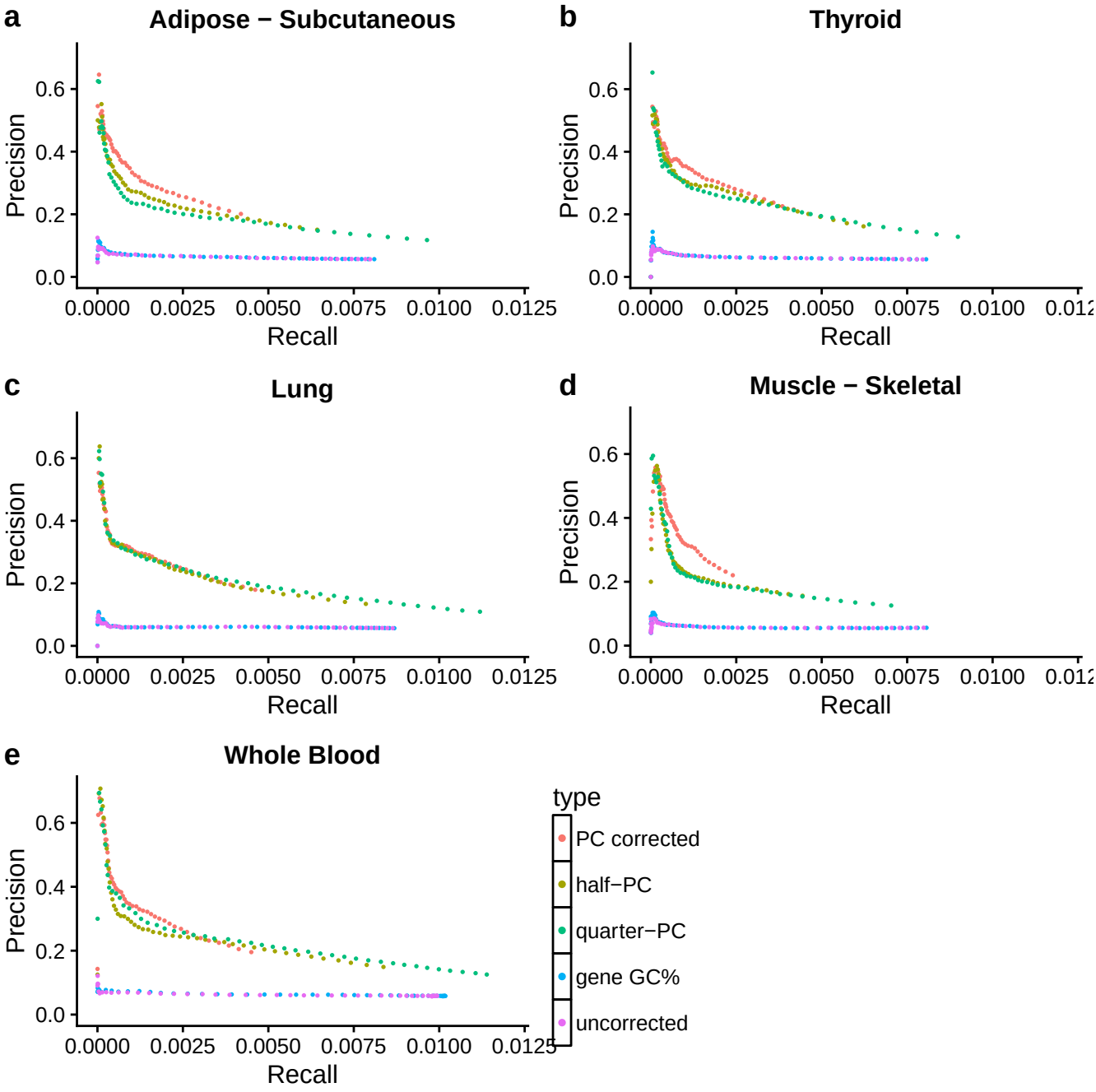
Supplementary Figure 3: Precision-Recall curves of networks inferred with graphical lasso using canonical pathway databases. Each color corresponds to the correction approach, and each point in the figure corresponds to precision and recall of networks obtained at a specific L1 penalty parameter value (penalty parameter ranges from 0.3 to 1.0). Exonic rate is the known confounder used in this figure



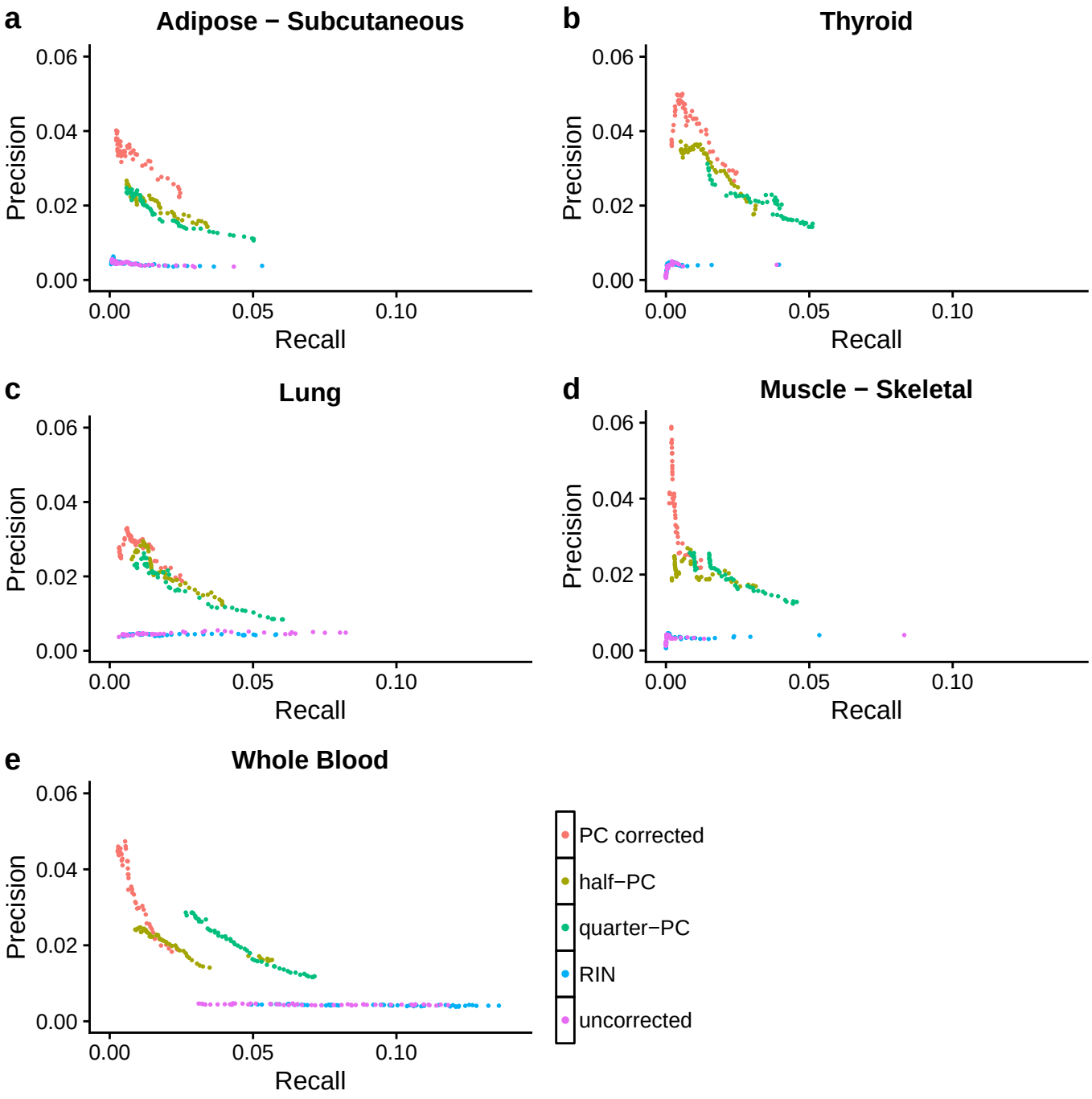
Supplementary Figure 4: Principal component loadings of gene expression are significantly associated with estimates of sample specific GC bias. Association was tested using a linear model. Panel (a) shows BH adjusted p-values and (b) shows R-squared



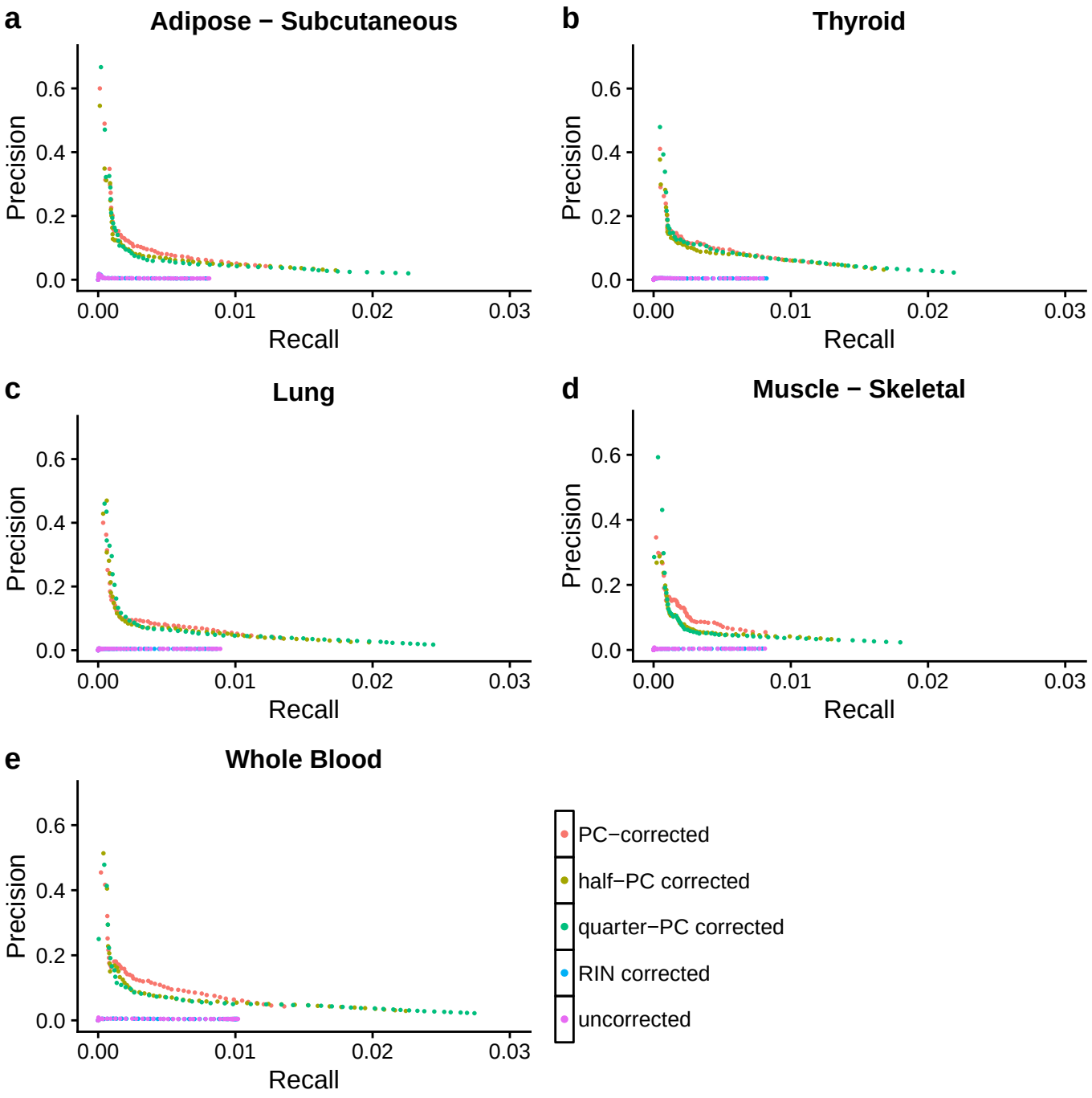
Supplementary Figure 5: Precision-Recall curves of WGCNA modules using canonical pathway databases. WGCNA networks obtained at a varying cut-heights. Each color corresponds to the correction approach, and each point corresponds to the network obtained at a specific cut-height. GC bias is the known confounder used in this figure.



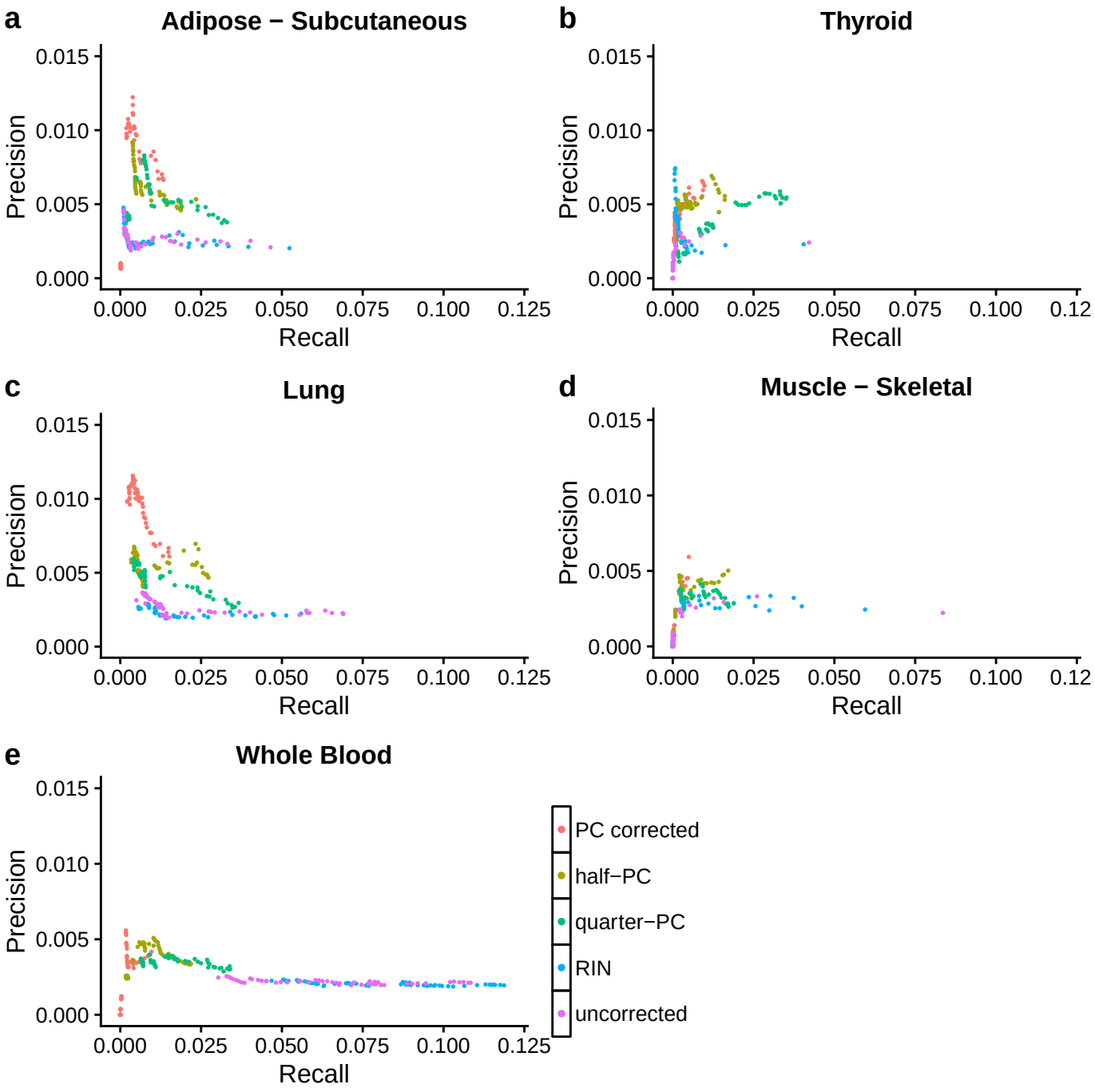
Supplementary Figure 6: Precision-Recall curves of networks inferred with graphical lasso using canonical pathway databases. Each color corresponds to the correction approach, and each point in the figure corresponds to precision and recall of networks obtained at a specific L1 penalty parameter value (penalty parameter ranges from 0.3 to 1.0). GC bias is the known confounder used in this figure.



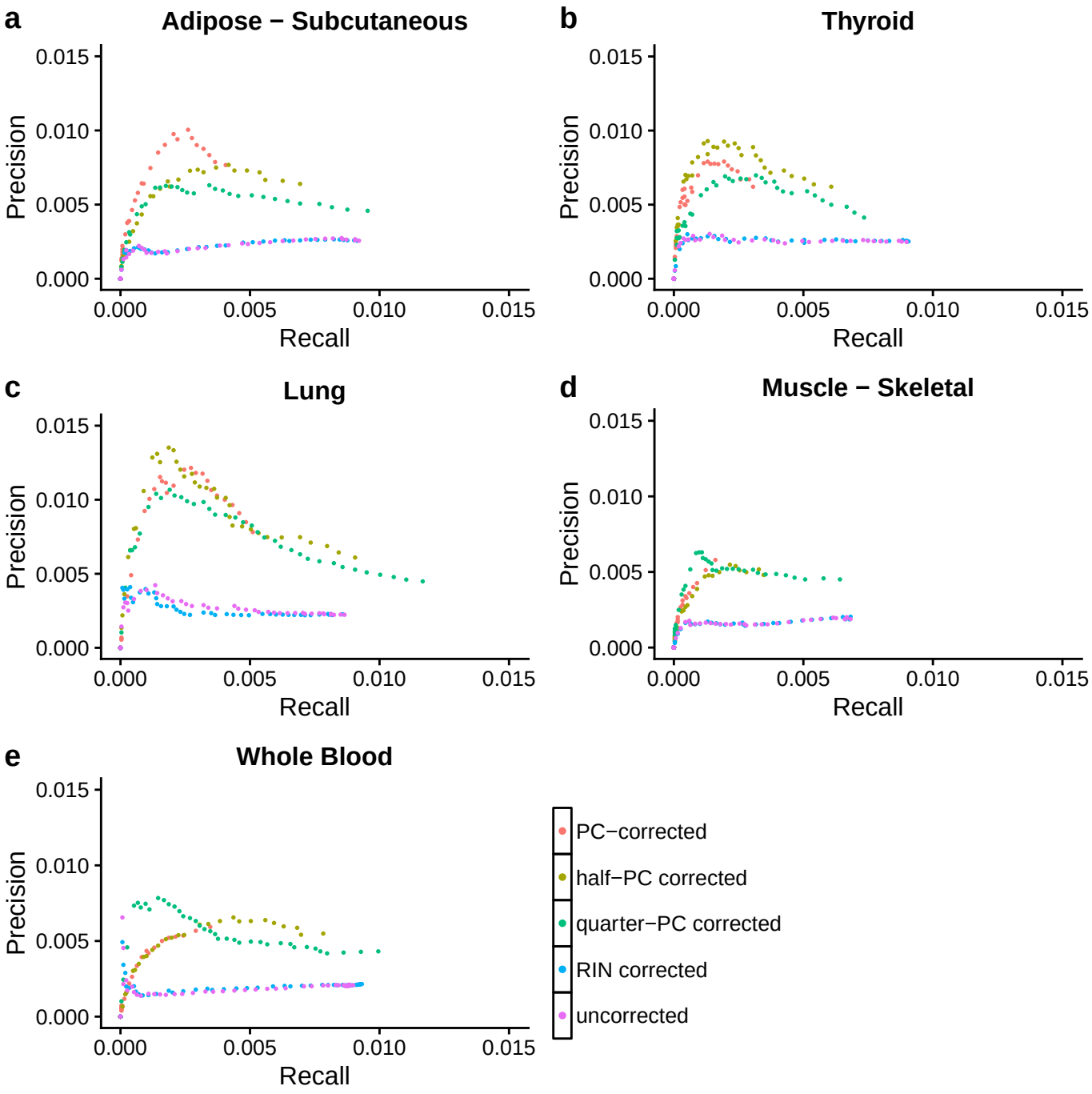
Supplementary Figure 7: Precision-Recall curves of WGCNA modules using list of shared true positives (Supplementary Note 2.3.3). Precision and recall curves of WGCNA networks obtained at a varying cut-heights. Each color corresponds to the correction approach, and each point corresponds to the network obtained at a specific cut-height



Supplementary Figure 8: Precision-Recall curves of networks inferred with graphical lasso using list of shared true positives (Supplementary Note 2.3.3). Each color corresponds to the correction approach, and each point in the figure corresponds to precision and recall of networks obtained at a specific L1 penalty parameter value (penalty parameter ranges from 0.3 to 1.0)



Supplementary Figure 9: Precision-Recall curves of WGCNA modules using ChEA2016 (transcription factor targets). Precision and recall curves of WGCNA networks obtained at a varying cut-heights. Each color corresponds to the correction approach, and each point corresponds to the network obtained at a specific cut-height.



Supplementary Figure 10: Precision-Recall curves of networks inferred with graphical lasso using ChEA2016 (transcription factor targets). Each color corresponds to the correction approach, and each point in the figure corresponds to precision and recall of networks obtained at a specific L1 penalty parameter value (penalty parameter ranges from 0.3 to 1.0).

	Total # of PCs removed
Adipose Subcutaneous	37
Lung	30
Thyroid	38
Skeletal Muscle	39
Whole Blood	25

Supplementary Table 1: Number of principal components removed

References

[1] Leek, J. T. Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics* **67**, 344–352 (2011).

[2] Barabási, A.-L., Albert, R. & Jeong, H. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications* **281**, 69–77 (2000).

[3] Van Noort, V., Snel, B. & Huynen, M. A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports* **5**, 280–284 (2004).

[4] Carlson, M. R. *et al.* Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC genomics* **7**, 1 (2006).

[5] Kim, S. K. *et al.* A gene expression map for caenorhabditis elegans. *Science* **293**, 2087–2092 (2001).

[6] Rzhetsky, A. & Gomez, S. M. Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. *Bioinformatics* **17**, 988–996 (2001).

[7] Bhan, A., Galas, D. J. & Dewey, T. G. A duplication growth model of gene expression networks. *Bioinformatics* **18**, 1486–1493 (2002).

[8] Jordan, I. K., Mariño-Ramírez, L., Wolf, Y. I. & Koonin, E. V. Conservation and coevolution in the scale-free human gene coexpression network. *Molecular biology and evolution* **21**, 2058–2070 (2004).

[9] Newman, M. E. The structure and function of complex networks. *SIAM review* **45**, 167–256 (2003).

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

[11] Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).

[12] Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4** (2005).

[13] Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS comput biol* **4**, e1000117 (2008).

[14] Sedaghat, N., Saegusa, T., Randolph, T. & Shojaie, A. Comparative study of computational methods for reconstructing genetic networks of cancer-related pathways. *Cancer informatics* **13**, 55 (2014).

[15] Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. & Ravikumar, P. Quic: quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research* **15**, 2911–2947 (2014).

[16] Consortium, G. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017). URL `http://dx.doi.org/10.1038/nature24277`.

[17] Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

[18] Chen, E. Y. *et al.* Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics* **14**, 128 (2013).

[19] Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90–W97 (2016).

[20] Liberzon, A. *et al.* Molecular signatures database (msigdb) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).