Supplementary material for "A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity"

Francesca Petralia[1,2,*], Li Wang[1,2,3,*], Jie Peng[4], Arthur Yan[1,2], Jun Zhu[1,2,3], and Pei Wang[1,2]

[1]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, New York, NY 10029, USA
[3]Sema4, a Mount Sinai venture, Stamford, CT
[4]Department of Statistics, University of California, Davis. Davis, CA, USA

Contents

1 EM Algorithm

The latent (unobserved) nature of $\{\{y_g^n\}\}$ and $\{\{z_g^n\}\}$ requires the adoption of the Expectation-Maximization (EM) algorithm. Specifically, the EM algorithm summarizes into the following steps:

- *E-Step* Given the current estimates of the model parameters, i.e., $(\boldsymbol{\Theta}^{(t)}, \boldsymbol{\pi}^{(t)})$, we calculate the expectation of the log-likelihood with respect to the latent variables $(\boldsymbol{Y}, \boldsymbol{Z})$, i.e.,

$$Q^{(t)}(\boldsymbol{\Theta}, \boldsymbol{\pi}) = E\left(\ell(\boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\Theta}, \boldsymbol{\pi}) \,\middle|\, \boldsymbol{X}, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\pi}^{(t)}\right)$$

- *M-step* we find $(\boldsymbol{\Theta}^{(t+1)}, \boldsymbol{\pi}^{(t+1)}, \delta^{(t+1)})$ which are the solution of the following maximization problem

$$\max_{\{\boldsymbol{\Theta}, \boldsymbol{\pi}, \delta\}} Q^{(t)}(\boldsymbol{\Theta}, \boldsymbol{\pi}) + \sum_{n=1}^{N} \ell(h_n; \pi_n, \delta) - P(\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}).$$

with $\ell(.)$ being the log-likelihood function. Given $\boldsymbol{\pi}$, $Q^{(t)}(\boldsymbol{\Theta}, \boldsymbol{\pi})$ is in the form of a penalized Gaussian log-likelihood in $\boldsymbol{\Theta}$ and therefore its maximization can be solved by the graphical lasso algorithm [5]. The conditional densities of $\boldsymbol{\pi}$ and δ are not in closed form and therefore their update needs to be done using numerical optimization. In order to save computational time, we adopt a strategy where first an estimate of $\boldsymbol{\pi}$ and δ is derived by assuming independence across genes.

1.1 Tumor-purity estimation

First, we will obtain stable estimates of $\boldsymbol{\pi}$ and δ by assuming independence across different genes. Therefore, assume that $(y_1^n, \cdots, y_G^n) \overset{iid}{\sim} N(\boldsymbol{\mu}_Y, \sigma_Y \boldsymbol{I})$ and $(z_1^n, \cdots, z_G^n) \overset{iid}{\sim} N(\boldsymbol{\mu}_Z, \sigma_Z \boldsymbol{I})$ with $\boldsymbol{I}$ being the G dimensional identity matrix. Since by assumption

$$x_g^n = \pi_n y_g^n + (1 - \pi_n) z_g^n, \quad g = 1, \cdots, G, \ n = 1, \cdots, N.$$

z_g^n can be written as $z_g^n = \frac{x_g^n - \pi_n y_g^n}{1 - \pi_n}$ and the joint likelihood reduces to:

$$L\left(\boldsymbol{Y}, \{h_n\}_{n=1}^N | \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\pi}\right) = \prod_{n=1}^{N} \prod_{g=1}^{G} \left[\frac{1}{\sqrt{2\pi\sigma_Y}} e^{-\frac{1}{2\sigma_Y}\left(y_g^n - \mu_{Y,g}\right)^2} \frac{1}{\sqrt{2\pi\sigma_Z}} e^{-\frac{1}{2\sigma_Z}\left(\frac{x_g^n - \pi_n y_g^n}{1 - \pi_n} - \mu_{Z,g}\right)^2} \right] \times$$

$$\times \prod_{n=1}^{N} \left[\frac{\Gamma(\alpha_n)\Gamma(\beta_n)}{\Gamma(\delta)} h_n^{\alpha_n - 1} (1 - h_n)^{\beta_n - 1} \right]$$

with $\alpha_n = \pi_n\delta$ and $\beta_n = (1-\pi_n)\delta$. After some algebra, the log-likelihood can be written as

$$\ell(\boldsymbol{Y}, \{h_n\}_{n=1}^N | \boldsymbol{X}, \boldsymbol{\Theta}) = \sum_{g=1}^{G}\left[-\frac{1}{2\sigma_Y}\sum_{n=1}^{N}(y_g^n)^2 + \frac{\mu_{Y,g}}{\sigma_Y}\sum_{n=1}^{N}y_g^n - \frac{n}{2}\log\sigma_Y - \frac{n(\mu_{Y,g})^2}{2\sigma_Y}\right] +$$
$$\sum_{g=1}^{G}\left[-\frac{1}{2\sigma_Z}\sum_{n=1}^{N}(z_g^n)^2 + \frac{\mu_{Z,g}}{\sigma_Z}\sum_{n=1}^{N}z_g^n - \frac{n}{2}\log\sigma_Z - \frac{n(\mu_{Z,g})^2}{2\sigma_Z}\right] +$$
$$\sum_{n=1}^{N}\left[\Gamma(\alpha_n) + \Gamma(\beta_n) - \Gamma(\delta) + (\alpha_n - 1)\log h_n + (\beta_n - 1)\log(1-h_n)\right]$$

where $\mu_{Y,g}$ and $\mu_{Z,g}$ are the gth element of vector $\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_Z$, respectively.

1.1.1 E step

During the *E Step*, we calculate the expectation of the log-likelihood with respect the latent variables $(\boldsymbol{Y}, \boldsymbol{Z})$ given the observed variable $\boldsymbol{X}$. In order to find $E_{(\boldsymbol{Z},\boldsymbol{Y})|\boldsymbol{X}}(\ell(\boldsymbol{Z}, \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta}))$ we need to derive the following quantities:

1. By independence, $E_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_{g=1}^{G}\sum_{n=1}^{N}y_g^n\right) = \sum_{g=1}^{G}\sum_{n=1}^{N}E_{\boldsymbol{Y}|\boldsymbol{X}}\left(y_g^n\right)$. Given that the observed variable $\boldsymbol{X}$ is a linear function of the latent Gaussian variable $\boldsymbol{Y}$ the following holds

$$E_{\boldsymbol{Y}|\boldsymbol{X}}\left(y_g^n\right) = \mu_{Y,g} + \sigma_{X,Y}\sigma_{X,n}^{-1}(x_{n,g} - \mu_{X,n,g})$$

   where $\mu_{X,n,g} = (\pi_n\mu_{Y,g} + (1-\pi_n)\mu_{Z,g})$, $\sigma_{X,n} = \left(\pi_n^2\sigma_Y + (1-\pi_n)^2\sigma_Z\right)$ and $\sigma_{X,Y} = Cov(\boldsymbol{X}, \boldsymbol{Y}) = \pi_n\sigma_Y$.

2. $E_{\boldsymbol{Y}|\boldsymbol{X}}\left(\sum_{g=1}^{G}\sum_{n=1}^{N}(y_g^n)^2\right) = \sum_{g=1}^{G}\sum_{n=1}^{N}E_{\boldsymbol{Y}|\boldsymbol{X}}\left((y_g^n)^2\right)$. Given that the observed variable X is a linear function of the latent Gaussian variable $\boldsymbol{Y}$ the following holds

   $V_{\boldsymbol{Y}|\boldsymbol{X}}\left((y_g^n)^2\right) = \sigma_Y + \sigma_{X,Y}\sigma_X^{-1}\sigma_{X,Y} = \sigma_Y - \pi_n\sigma_Y(\pi_n\sigma_Y + (1-\pi_n)\sigma_Z)^{-1}\pi_n\sigma_Y$

   with $E_{\boldsymbol{Y}|\boldsymbol{X}}\left((y_g^n)^2\right) = V_{\boldsymbol{Y}|\boldsymbol{X}}\left((y_g^n)^2\right) + \left(E_{\boldsymbol{Y}|\boldsymbol{X}}\left(y_g^n\right)\right)^2$

3. Derive $E_{\boldsymbol{Z}|\boldsymbol{X}}\left(\sum_{g=1}^{G}\sum_{n=1}^{N}z_g^n\right)$ and $E_{\boldsymbol{Z}|\boldsymbol{X}}\left(\sum_{g=1}^{G}\sum_{n=1}^{N}(z_g^n)^2\right)$ similarly.

1.1.2 M step

Implement the Expectation Conditional Maximization algorithm (ECM) [8]. Sample each parameter conditioning to others as follows:

1. *For* $n \in \{1, \dots, N\}$,
   derive $\pi_n^{(t)}$ maximizing $\ell\left(\boldsymbol{Z}, \boldsymbol{Y}, \{h_n\}_{n=1}^N | \{\pi_n^*\}, \boldsymbol{\mu}_Y^{(t-1)}, \sigma_Y^{(t-1)}, \sigma_Z^{(t-1)}, \boldsymbol{\mu}_Z^{(t-1)}\right)$
2. Derive $\boldsymbol{\mu}_Y^{(t)}$ maximizing $\ell\left(\boldsymbol{Z}, \boldsymbol{Y}, \{h_n\}_{n=1}^N | \boldsymbol{\mu}_Y^*, \{\pi_n^{(t)}\}, \sigma_Y^{(t-1)}, \sigma_Z^{(t-1)}, \boldsymbol{\mu}_Z^{(t-1)}\right)$
3. Derive $\sigma_Y^{(t)}$ maximizing $\ell\left(\boldsymbol{Z}, \boldsymbol{Y}, \{h_n\}_{n=1}^N | \sigma_Y^*, \boldsymbol{\mu}_Y^{(t)}, \{\pi_n^{(t)}\}, \sigma_Z^{(t-1)}, \boldsymbol{\mu}_Z^{(t-1)}\right)$
4. Derive $\boldsymbol{\mu}_Z^{(t)}$ maximizing $\ell\left(\boldsymbol{Z}, \boldsymbol{Y}, \{h_n\}_{n=1}^N | \boldsymbol{\mu}_Z^*, \{\pi_n^{(t)}\}, \sigma_Z^{(t-1)}, \sigma_Y^{(t)}, \boldsymbol{\mu}_Y^{(t)}\right)$
5. Derive $\sigma_Z^{(t)}$ maximizing $\ell\left(\boldsymbol{Z}, \boldsymbol{Y}, \{h_n\}_{n=1}^N | \sigma_Z^*, \boldsymbol{\mu}_Z^{(t)}, \{\pi_n^{(t)}\}, \sigma_Y^{(t)}, \boldsymbol{\mu}_Y^{(t)}\right)$

1.2 Estimation of concentration matrices

Here, we considered as fixed the tumor purity $\boldsymbol{\pi}$ and derive estimate of parameters in $\boldsymbol{\Theta}$. The log-likelihood can be written as:

$$\ell(\boldsymbol{Y}, \boldsymbol{Z} | \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\pi}) = -\frac{1}{2}|\boldsymbol{\Sigma}_Y| - \sum_{n=1}^{N} \frac{1}{2}(\boldsymbol{Y}^n - \boldsymbol{\mu}_Y)^T \boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{Y}^n - \boldsymbol{\mu}_Y) -$$

$$-\frac{1}{2}|\boldsymbol{\Sigma}_Z| - \sum_{n=1}^{N} \frac{1}{2}(\boldsymbol{Z}^n - \boldsymbol{\mu}_Z)^T \boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{Z}^n - \boldsymbol{\mu}_Z)$$

with $\boldsymbol{Y}^n = (y_1^n, \dots, y_G^n)^T$. After some algebra, we obtain:

$$\ell(\boldsymbol{Y}, \boldsymbol{Z}, | \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\pi}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_Y^{-1}| - \sum_{n=1}^{N} \frac{1}{2}\left[(\boldsymbol{Y}^n - \boldsymbol{\mu}_Y)^T \boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{Y}^n - \boldsymbol{\mu}_Y)\right] +$$

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}_Z^{-1}| - \sum_{n=1}^{N} \frac{1}{2}\left[(\boldsymbol{Z}^n - \boldsymbol{\mu}_Z)^T \boldsymbol{\Sigma}_Z^{-1}(\boldsymbol{Z}^n - \boldsymbol{\mu}_Z)\right]$$

Given that, $z_g^n = \frac{\pi_n y_g^n - x_g^n}{1 - \pi_n}$, the log-likelihood can be written as:

$$\ell(\boldsymbol{Y}, \boldsymbol{Z}, | \boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\pi}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_Y^{-1}| - \sum_{n=1}^{N} \frac{1}{2}\left[(\boldsymbol{Y}^n - \boldsymbol{\mu}_Y)^T \boldsymbol{\Sigma}_Y^{-1}(\boldsymbol{Y}^n - \boldsymbol{\mu}_Y)\right] +$$

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}_Z^{-1}| - \sum_{n=1}^{N} \frac{1}{2}\left[\left(\frac{\pi_n \boldsymbol{Y}^n - \boldsymbol{X}^n}{1 - \pi_n} - \boldsymbol{\mu}_Z\right)^T \boldsymbol{\Sigma}_Z^{-1}\left(\frac{\pi_n \boldsymbol{Y}^n - \boldsymbol{X}^n}{1 - \pi_n} - \boldsymbol{\mu}_Z\right)\right]$$

1.2.1 E step

The expectation step needs to solve $E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[(\boldsymbol{Y}^n-\boldsymbol{a})^T\boldsymbol{C}(\boldsymbol{Y}^n-\boldsymbol{a})\right]$. In particular, let $\boldsymbol{Y}^n|\boldsymbol{X}^n \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}, \boldsymbol{\Sigma}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})$, then

$$E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[(\boldsymbol{Y}^n-\boldsymbol{a})^T\boldsymbol{C}(\boldsymbol{Y}^n-\boldsymbol{a})\right] = \mathrm{tr}\left(E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[(\boldsymbol{Y}^n-\boldsymbol{a})^T\boldsymbol{C}(\boldsymbol{Y}^n-\boldsymbol{a})\right]\right)$$

$$= E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[tr((\boldsymbol{Y}^n-\boldsymbol{a})^T\boldsymbol{C}(\boldsymbol{Y}^n-\boldsymbol{a}))\right]$$

$$= E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[tr((Y^n-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})^T\boldsymbol{C}(Y^n-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}))\right]+E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[tr((a-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})^T\boldsymbol{C}(a-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}))\right]$$

$$= E_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\left[tr(\boldsymbol{C}(\boldsymbol{Y}^n-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})(Y^n-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})^T)\right]+(a-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})^T\boldsymbol{C}(\boldsymbol{a}-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}) =$$

$$= tr(\boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}) + (a-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})^T\boldsymbol{C}(\boldsymbol{a}-\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n})$$

Then, the expected log-likelihood is equal to

$$E_{\boldsymbol{Y}|\boldsymbol{X}}\left[\ell(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\pi})\right] = -\frac{1}{2}\log|\boldsymbol{\Sigma}_Y^{-1}|-\sum_{n=1}^{N}\frac{1}{2}\left[\left(\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}-\boldsymbol{\mu}_Y\right)^T\boldsymbol{\Sigma}_Y^{-1}\left(\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}-\boldsymbol{\mu}_Y\right)\right]$$

$$-\frac{1}{2}\sum_{n=1}^{N}\mathrm{tr}(\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{\Sigma}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}) + N\log|\boldsymbol{\Sigma}_Z^{-1}| - \sum_{n=1}^{N}\mathrm{tr}\left(\frac{\pi_n^2}{(1-\pi_n)^2}\Sigma_{\boldsymbol{Y}^n|\boldsymbol{X}^n}\Sigma_Z^{-1}\right)$$

$$-\sum_{n=1}^{N}\left[(1-\pi_n)^{-2}\left(\boldsymbol{X}^n-\pi\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}-(1-\pi_n)\boldsymbol{\mu}_Z\right)^T\Sigma_Z^{-1}\left(\boldsymbol{X}^n-\pi\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}-(1-\pi_n)\boldsymbol{\mu}_Z\right)\right]$$

where $\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n} = \hat{\boldsymbol{\mu}}_Y+\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{X}^n-\boldsymbol{\mu}_X)$ and $\boldsymbol{\Sigma}_{\boldsymbol{Y}^n|\boldsymbol{X}^n} = \hat{\boldsymbol{\Sigma}}_Y+\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}$ with $\boldsymbol{\Sigma}_{XY} = \hat{\boldsymbol{\Sigma}}_Y$ and $(\hat{\boldsymbol{\Sigma}}_Y, \hat{\boldsymbol{\mu}}_Y)$ being the current estimate of $(\boldsymbol{\Sigma}_Y, \boldsymbol{\mu}_Y)$.

1.2.2 M step

In this step we need to maximize the following penalized expected log-likelihood function:

$$\arg\max_{\boldsymbol{\Theta}} E_{\boldsymbol{Y}|\boldsymbol{X}}\left[\ell(\boldsymbol{Y},\boldsymbol{Z},|\boldsymbol{X},\boldsymbol{\Theta},\boldsymbol{\pi})\right] - \lambda P(\boldsymbol{\Sigma}_Y^{-1},\boldsymbol{\Sigma}_Z^{-1}) \tag{1}$$

with $\boldsymbol{\Theta} = (\boldsymbol{\mu}_Y, \boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Y, \boldsymbol{\Sigma}_Z)$. In particular, the maximization process can be summarized by the following steps:

1. The maximum likelihood estimate of $\boldsymbol{\mu}_Y$ is $\boldsymbol{\mu}_Y^* = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\mu}_{\boldsymbol{Y}^n|\boldsymbol{X}^n}$.
2. The maximum likelihood estimate of $\boldsymbol{\Sigma}_Y$ can be found using function *glasso* available in R Cran.
3. Similarly we can find the maximum likelihood estimates of $\boldsymbol{\mu}_Z$ and $\boldsymbol{\Sigma}_Z$

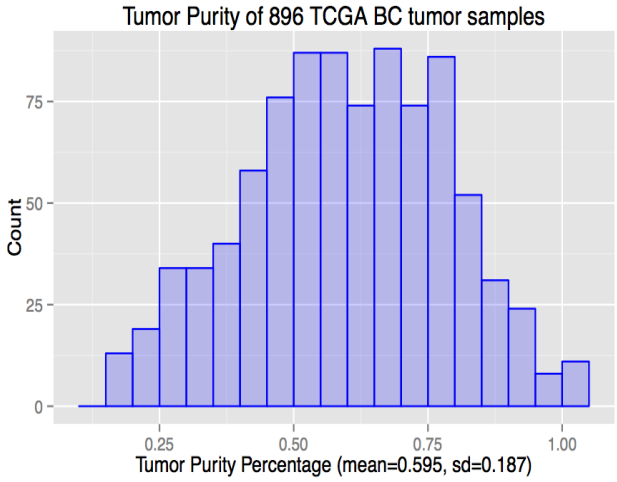

Figure S1: Histogram of tumor-purity for 896 TCGA breast cancer samples. Tumor-purity was estimated using the algorithm ABSOLUTE which infers tumor-purity based on copy number variation data.

2 Synthetic Data

2.1 Tumor purity estimation

2.1.1 Network topology

The total number of nodes was divided into 4 different disjoint sets. Then, for each network, the network topology for each subset of nodes was randomly generated from a power law degree distribution. The two networks were not forced to overlap in any way. This strategy resulted in tumor and normal networks containing 996 each edges with only 19 edges shared across the two networks. Specifically, edges were detected for values in the concentration matrix in absolute value greater than 0.001.

2.1.2 TCGA breast cancer tumor purity

Fig. S1 shows the histogram of tumor purity for 896 TCGA breast cancer samples. Tumor-purity was estimated via ABSOLUTE [3], a well known algorithm which infers tumor-purity from copy number variation data.

2.1.3 Tumor purity under model misspecification

For the synthetic data example in section 4.1, the true and estimated model for the prior estimate of tumor purity h_n were assumed to be the same. In this section we show that good estimation performance can be obtained even in the case of model misspecification. Specifically, h_n was generated considering

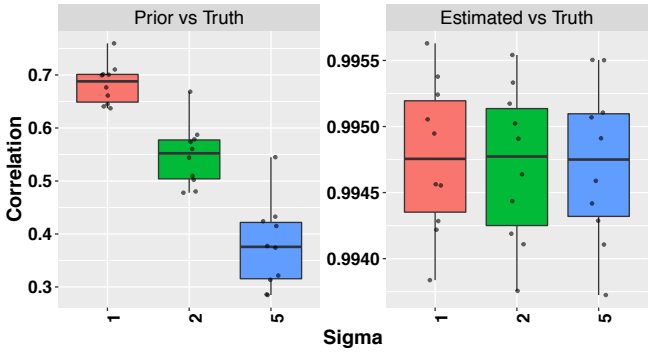

Figure S2: Correlation between prior estimate and true purity (left plot) and correlation between estimated and true purity (right plot) for different values σ.

a logistic regression, i.e., $logit(h_n) = logit(\pi_n) + \epsilon_n$ with ϵ_n being normally distributed with mean zero and variance σ. Given a network topology involving $1,000$ genes, for each value of $\sigma \in \{1, 2, 5\}$, ten independent data sets involving 200 observations each were generated. For each data set, tumor purity was estimated modeling h_n as $h_n \sim Beta(\alpha_n, \beta_n)$ with $\alpha_n = \pi_n\delta$ and $\beta_n = (1 - \pi_n)\delta$. Fig. S2 shows the boxplot of correlation between the prior and the estimated tumor purity with the true tumor-purity over different replicates. As σ increases, the correlation between the prior estimate $\{h_n\}$ and the true purity $\{\pi_n\}$ decreases. However, the estimated tumor purity remains highly correlated with the true tumor-purity for any values of σ.

2.2 Co-expression networks estimation

Given the network topology, we simulated 10 replicates from a Gaussian graphical model with covariance matrices constructed using the technique presented in section 7.1.1 of Danaher et al (2014) [4].

2.2.1 Network topology

Independent networks. The total number of nodes was divided into 4 different disjoint sets. Then, for each network, the network topology for each subset of nodes was randomly generated from a power law distribution. The two networks were not forced to overlap in any way. This strategy resulted in the networks shown in Fig. S3(a) containing 996 edges each with only 19 edges shared across the two networks. Specifically, edges were detected for values in the concentration matrix in absolute value greater than 0.001. The same strategy was adopted in order to generate networks involving 500 nodes. In particular, the tumor and normal networks containing 500 nodes involved 496 edges each with 20 edges shared across the two networks.

Partially overlapping networks. We divided $1,000$ nodes into 4 different disjoint sets. Then, for tumor-network, the network topology for each subset was randomly generated from a power law distribution. For normal-network, two of the components were randomly generated from the power law distribution; while the remaining two components were set equal to two components in tumor-network. Fig. S3(b) shows the topology of the two networks. Tumor-network and Normal-network involved the same number of edges equal to 996 with 514 edges being shared across the two networks.

2.3 Choice of penalty parameters

For the penalty parameters of TSNet (i.e., ρ_y and ρ_z), we considered equally spaced values ranging from 0.18 and 0.33. Then, TSNet model was implemented considering any possible pairwise combination of the two parameters. For the penalty parameter of mixNet, i.e. ρ, we considered equally spaced values ranging from 0.1 to 0.5.

2.4 Star topology

In this section, we evaluate the performance of TSNet and mixNet considering a different type of network topology, i.e., the star topology. Specifically, we consider networks involving $1,000$ nodes which were obtained as the union of four sub-networks involving equally sized disjoint set of genes. Specifically, each sub-network was randomly sampled from the star distribution (the topology of the generated networks is shown in Fig. S4(a)). Based on this topology, we simulated 10 data replicates containing $N = 200$ observations each. Fig. S4(b) shows the average of ROC curves over different replicates for both mixNet and TSNet models. For any false positive rate, TSNet results in higher true positive edges than mixNet.

2.5 Networks estimated via BIC

For TSNet, the penalty parameter was chosen to minimize the BIC defined as $BIC(\rho_y, \rho_z) = -2\ell + (E_{\rho_y}(\mathbf{\Sigma}_y) + E_{\rho_z}(\mathbf{\Sigma}_z))\log(N)$ with $E_{\rho_y}(\mathbf{\Sigma}_y)$ being the number of elements different from zero in the inverse covariance matrix under penalty parameter ρ_y and ℓ being the log-likelihood function. Similarly, mixNet was implemented for different values of the penalty parameters with the best penalty parameter chosen to minimize $BIC(\rho) = -2\ell + E_{\rho}(\mathbf{\Sigma}_x)\log(N)$.

2.6 Comparison with DeMix

In this section, TSNet was compared to a two-step approach where, first, mixed data is deconvoluted into tumor and non-tumor components and, then, networks for both components are estimated using the standard Graphical lasso. In order to deconvolve gene expression from infiltrated tumor tissue, we utilized DeMix

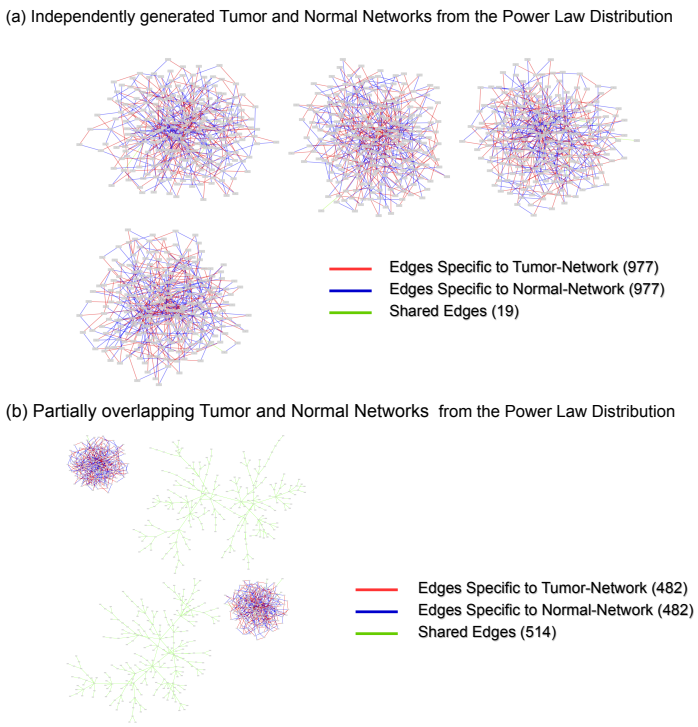

Figure S3: (a) Independently generated tumor and normal networks. For each network, $1,000$ nodes were divided into 4 equally sized non-overlapping sets. Then, for each subset of nodes, a network topology was generated from the star topology. The two networks were independently generated without forcing any overlapping structure. (b) Partially overlapping tumor and normal networks. For each network, $1,000$ nodes were divided into 4 equally sized non-overlapping sets. For the tumor network, a topological structure was generated from a power law distribution for each of the sets of nodes. For the normal network, two of the components were randomly generated from the power law distribution; while the remaining two components were set equal to two of the components in tumor-network.

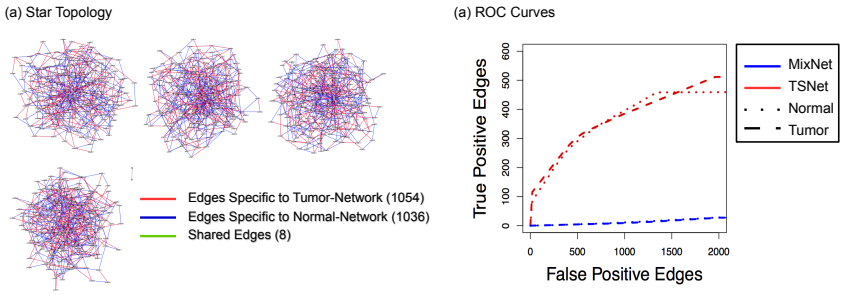

Figure S4: (a) Networks generated from the Star topology. For each network, $1,000$ nodes were divided into 4 equally sized non-overlapping sets. Then, for each subset of nodes. The two networks were independently generated without forcing any overlapping structure. (b) Average of ROC over different replicates for mixNet (blue) and TSNet model (red).

[1] – a well known algorithm for the deconvolution of mixed expression data. DeMix models gene expression of gene g and sample n as

$$x_g^n = \pi_n y_g^n + (1 - \pi_n) z_g^n$$

with $y_g^n \sim \log_2 N(\mu_y, \sigma_y)$ and $z_g^n \sim \log_2 N(\mu_z, \sigma_z)$ with $\log_2 N$ being the log2 Gaussian distribution.

Data generation For this synthetic data scenario, we simulated data from the following model

$$x_g^n = \pi_n y_g^n + (1 - \pi_n) z_g^n$$

with $(z_1^n, \ldots, z_p^n) \sim \log_2 N(\mu_Z, \Sigma_Z)$ and $(y_1^n, \ldots, y_p^n) \sim \log_2 N(\mu_Y, \Sigma_Y)$. In particular, 10 different replicates involving 200 samples each were simulated. In order to deconvolve mixed data into tumor and non-tumor components, DeMix requires some samples with tumor purity equal to zero (i.e., samples from non tumor tissue). For this reason, we set the tumor-purity of a subset of samples equal to zero. In order to assess how an increasing fraction of non-tumor samples might affect the results, we generated data with two different fraction of normal samples, i.e., 10% and 25% of the total number of samples. For the remaining samples, tumor purity was sampled either from a Uniform distribution defined on the interval $[0, 1]$ or a Beta distribution with mean 0.05 and standard deviation 0.04 to mimic the tumor-purity in TCGA Breast cancer (Fig. S1). Specifically, we consider three simulation scenarios. In the first simulation scenario, the number of samples with tumor purity equal to zero was 25% and the tumor purity of the remaining samples was randomly generated from a uniform distribution. For the second synthetic data scenario, we still considered 25% of samples with tumor purity equal to zero, but we generated the tumor purity for the other samples from the Beta distribution. In the final synthetic data scenario, the number of "normal" samples was reduced to 10% and the tumor

purity of other samples was generated from the Beta distribution. Following [1], in all three synthetic data scenarios, mean parameters μ_Z and μ_Y were simulated as follows. A subset of genes (i.e., 25%) was considered to be differentially expressed (DE) between tumor and non-tumor components. For each gene g, the tumor component mean μ_Y^g was sampled from a $N(7, 2)$, while the non-tumor component mean was set equal to $\mu_Z^g = \mu_Y^g + \beta_Z^g$, with $\beta_Z^g \sim N(6, 1.5)$ for DE genes and $\beta_Z^g \sim N(0, 0.2)$ for non-DE genes. The covariance matrices Σ_Z and Σ_Y were randomly generated from a power law distribution. Specifically, we considered networks involving 500 nodes which were obtained as the union of four sub-networks involving different set of genes, with each sub-network being randomly generated from the power law degree distribution.

Implementation Using DeMix, we deconvoluted $\{x_g^n\}$ into gene expression of tumor cells and non-tumor cells, i.e., $\{y_g^n\}$ and $\{z_g^n\}$. Then, the deconvoluted matrices were log2 transformed and the standard graphical lasso was implemented in order to estimate co-expression networks of the two components. On the other hand, TSNet was directly implemented on mixed data $\{x_g^n\}$ after applying genewise standardization (zero mean and unit variance).

Results Fig. S5(a) shows the performance of DeMix and TSNet in recovering the true level of tumor-purity. Specifically, for both methods, we show the correlation between true tumor purity and estimated tumor purity over different replicates. As shown, TSNet outperforms DeMix for all simulation scenarios. Fig. S5(b) shows the mean of ROC over different replicates in estimating tumor and non-tumor networks based on both Demix and TSNet algorithms. As shown, as the number of non-tumor samples increases from 10% to 25% of the total number of samples, both algorithms can estimate non-tumor networks more accurately. However, despite the data was generated from a mixture of log2 Gaussian distributions to favor DeMix model, TSNet outperformed the competitor in estimating both non-tumor and tumor specific networks on all synthetic data scenarios.

2.7 Computational time

Table S1 shows the computational time of TSNet necessary to estimate tumor purity and co-expression networks for different synthetic data examples presented in section 4. Computational time was computed based on an IBM machine with 12 Intel Ivy Bridge (3.5 GHz) cores and 64GB of memory. For tumor purity, Table S1 shows the mean of computational time across 10 different replicates. As shown, the computational time for tumor purity estimation is particularly low for any simulation scenario. For network estimation, we implemented TSNet for different penalty parameters and, then, computed the mean of computational time across different choices of penalty parameters. Then, Table S1 shows the mean across 10 different replicates of those averaged computational times. Although the increased sample size does not affect computational time, an increasing network dimension substantially affects the computational

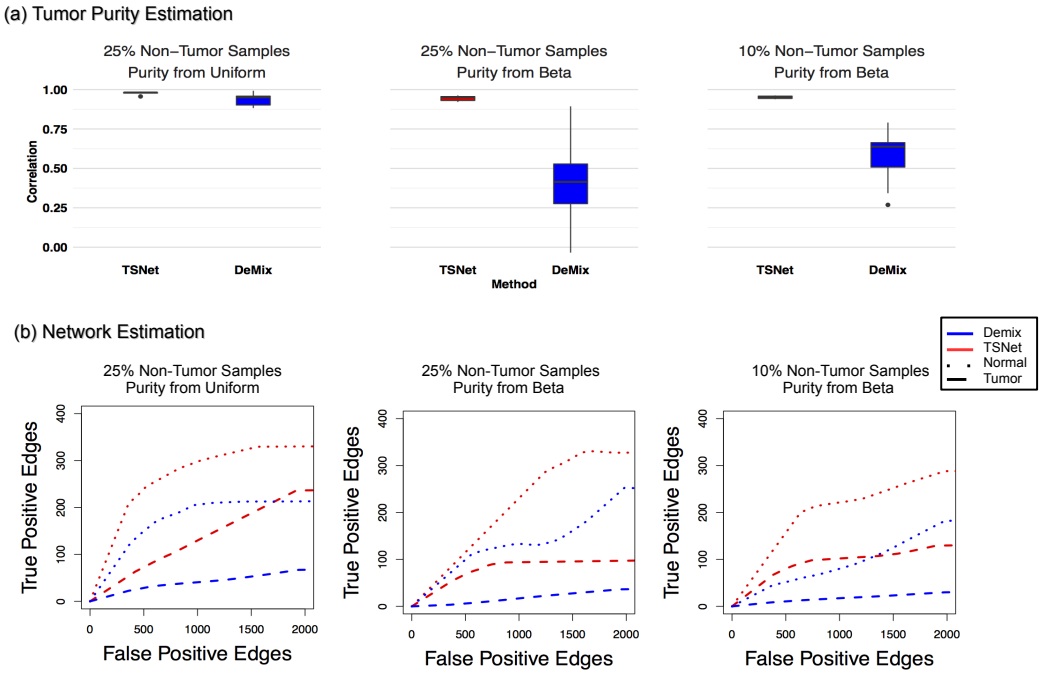

Figure S5: Comparison between DeMix and TSNet (a) Estimation of tumor purity. Correlation between estimated tumor purity and true value over different replicates for different simulation scenarios and different methods (i.e., DeMix and TSNet). (b) Average of ROC over different replicates for different simulation scenarios and different methods.

n	p	Purity	Network
200	500	30	613
200	1,000	39	4,662
400	1,000	61	4,012

Table S1: Computational time in seconds for tumor-purity and network estimation for different synthetic data examples. For each data scenario, we show the mean of computational time over 10 different replicates. Computational time was computed using an IBM node with 12 Intel Ivy Bridge (3.5 GHz) cores and 64GB of memory.

time of the network inference.

3 Real Data

3.1 ABSOLUTE's purity as prior

In this section, we show that similar results were obtained considering tumor-purity from ABSOLUTE [3] as prior. Fig. S6 shows the Pearson's correlation of tumor purities from TSNet, ABSOLUTE [3] and ESTIMATE [10]. This results are very similar to the ones presented in section 4.1, and, therefore, different choices of priors resulted in similar tumor-purity estimates.

3.2 Choice of penalty parameters

For TSNet model, we considered a grid of values for the two penalty parameters (ρ_y, ρ_z) ranging from 0.01 to 1. Then, TSNet model was implemented considering any possible pairwise combination of the two parameters. For mixNet model, we considered values for the penalty parameter ρ ranging from 0.01 to 1. For both TSNet and mixNet, about 600 models have been estimated using different penalty parameters.

3.3 Networks estimated via BIC

TSNet Model: Different penalty parameters resulted in different network dimensions. For TSNet networks, the best model was selected via Bayesian Information Criteria (BIC). Specifically, for TSNet, the penalty parameter was chosen to minimize the BIC defined as $BIC(\rho_y, \rho_z) = -2\ell + (E_{\rho_y}(\mathbf{\Sigma}_y) + E_{\rho_z}(\mathbf{\Sigma}_z)) \log(N)$ with N being the number of observations, $E_{\rho_y}(\mathbf{\Sigma}_y)$ being the number of elements different from zero in the inverse covariance matrix under penalty parameter ρ_y and ρ_z and ℓ the log-likelihood. In particular, an element of the concentration matrix was considered different from zero when greater than 0.001 in absolute value. Following this strategy, we selected the best penalty parameter and we derived the two networks TSNet-tumor and TSNet-normal. In particular,TSNet-

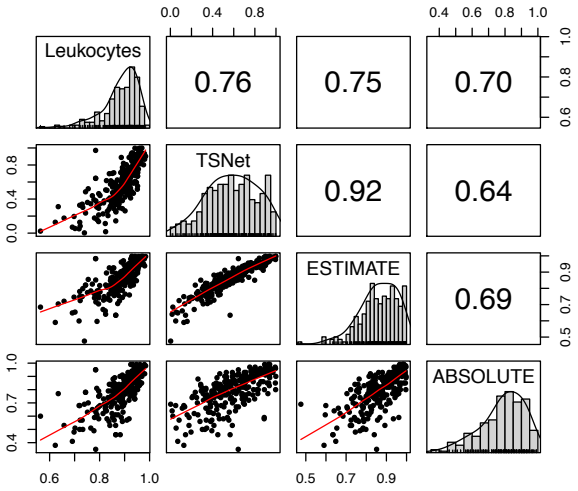

Figure S6: Pearson's correlation of tumor purity from TSNet, ABSOLUTE [3] and ESTIMATE [10] with methylation-based estimates of the fraction of leukocytes in tumor tissue.

tumor and TSNet-normal networks contained 2,625 and 290 edges, respectively.

mixNet Model: Similarly, the best model was selected also for mixNet. mixNet was implemented for different values of the penalty parameters with the best penalty parameter chosen to minimize $BIC(\rho) = -2\ell + E_\rho(\mathbf{\Sigma}_x)\log(N)$, with N being the number of observations, ρ being the penalty parameter, $E_\rho(\mathbf{\Sigma}_x)$ the number of edges obtained using penalty parameter ρ and ℓ the log-likelihood of the data. In particular, the number of edges was calculated by counting the number of elements in the estimated concentration matrix different from zero (i.e., greater than 0.001 in absolute value). The network selected via BIC involved 9,983 edges.

3.4 Consensus networks

As mentioned in section 3.2, models selected via BIC resulted in networks of difficult comparison given their dramatic difference in the number of edges. Therefore, the following technique was adopted to derive consensus networks. For each network type (i.e., TSNet-normal, TSNet-tumor and mixNet), a consensus network was derived based on a series of inferred network models of different dimension ranging from 500 to 3,500 edges (obtained by varying the penalty parameters). Specifically, take TSNet-tumor as an example, we divided the interval $[500, 3500]$ into six equally sized intervals of dimension 500, i.e.,

$[\tau, \tau + 500)$ with $\tau \in \{500, 1000, 1500, 2000, 2500, 3000\}$. Then, for each interval, we considered TSNet-tumor models whose sizes were contained in this interval and, among those networks, we selected the network with the least complexity (the smallest number of nodes). Following this strategy, we derived six different networks, one for each interval. We then derived the final network as the union of edges which were contained in at least 80% of the six networks. Following this strategy, we obtained the final TSNet-tumor, TSNet-normal and mixNet networks which contained 707, 793, and 993 edges, respectively.

3.5 KL Statistics for pathway enrichment of hub-structure

A weighted version of the Kolmogorov-Smirnov statistics [9] was utilized in order to assess if hub-genes in a given network were enriched of a particular pathway. For a given network containing E edges, we first ordered genes in decreasing order of connecting edges, i.e., $\{g_1, \ldots, g_G\}$. Then, for each pathway M containing n_M genes, we computed the following statistics

$$KL(i) = \left[\sum_{j:(g_j \in M)\&(j \leq i)} \frac{w_j}{S_M} - \sum_{j:(g_j \notin M)\&(j \leq i)} \frac{1}{G - n_M} \right]$$

with $S_M = \sum_{j:(g_j \in M)} w_j$, with w_j being the number of connecting edges of gene g_j. Then, for each pathway, we derived a score $S(M)$ which was defined as the maximum deviation from zero of $\{KL(i)\}_{i=1}^{G}$. In order to test the significance of the enrichment and adjust for multiple comparison, we implemented the permutation based technique illustrated in [9] which is summarized in the following steps:

1. Derive $S(M)$ for each pathway M
2. For each permutation ω. Randomly sample a network with E edges, order the genes based on the new network topology and derive $S(M, \omega)$ for each pathway M. This step was repeated for $n(\omega) = 20,000$ permutations.
3. Adjust for the different gene set size by normalizing the score $S(M)$ and $S(M, \omega)$ by dividing them by the mean of $S(M, \omega)$ over different ω which results in the normalized scores $S^*(M)$ and $S^*(M, \omega)$.
4. Compute the FDR for each threshold f_j as follows

$$\text{fdr}(f_j) = \frac{\sum_M \sum_\omega \mathbf{1}_{S^*(M,\omega) > f_j} / n(\omega)}{\sum_M \mathbf{1}_{S^*(M) > f_j}}$$

where $\mathbf{1}_A$ is the indicator function equal to one if A is satisfied and to zero otherwise.

3.6 Enrichment of network topology

Following the strategy illustrated by Zhu et al, (2008) [11], we carried out an enrichment analysis based on the topological structure of each network, i.e., TSNet-tumor, TSNet-normal and mixNet. For the analysis, we considered MSigDB Canonical [7] and Hallmark [6] pathways. In particular, we focused on 442 pathways containing at least 5 of the genes considered in the network analysis. For each network, we carried out a fisher exact test to assess whether the network was enriched of a particular category. Specifically, given a pathway and a network topology, we considered the first order neighborhood of the genes contained in the pathway. From the resulting set of connected nodes, we identified the independent sub-network with the highest number of nodes. Then, considering this sub-network, we carried out a fisher exact test to test whether the pathway of interest was enriched. Based on this test, for each network, we derived p-values for the 442 pathways which were then adjusted for multiple comparison using a Benjamini-adjustment [2].

3.7 Robustness of the method to experimental noise

In order to assess how results might be affected by different levels of experimental noise, we carried out a new simulation experiment in which the ovarian expression data was perturbed by adding different levels of white noises. Specifically, a $(p \times n)$ matrix of white noises was generated by sampling each element independently from a Gaussian distribution with mean zero and standard deviation σ which was then added to the original data matrix. For this experiment, two different levels of noises corresponding to different values of σ (i.e., 0.1 and 0.4) were considered. For simplicity, these two levels of noises will be referred to as Level 1 and Level 2, respectively. Fig. S7(a) shows the comparison between tumor purity estimation based on original and perturbed data for different levels of noise. As shown, the tumor purity estimates from perturbed data is very consistent with that based on the original data with a correlation higher than 0.99 for any level of noise. Then, TSNet was implemented to estimate co-expression networks for the tumor and non-tumor components. For simplicity, we will refer to the network based on original and perturbed data as original network and noisy network, respectively. For sake of comparison, we consider a given network dimension (i.e., 500 nodes) and we compare the edges and hub-gene structure of the original and noisy networks. In particular, as hub-gene, we consider any network node with more than 10 connecting edges. Table S2 shows the number of edges overlapping between original and noisy networks. As shown, the number of overlapping edges is particularly high for the non-tumor network (78% overlapping edges) for Level 1 noise. Although the tumor network experiences a lower overlap between network edges, it has an high percentage of overlapping hub-genes (94% overlapping hubs) no matter the level of noise. This high overlap shows the ability of TSNet to robustly identify hub-genes in the network. Fig. S7(b) shows the scatterplot of the degree of the hub-genes in the original network based on both original and noisy networks. As shown, the

	Edges		Hub-genes	
Noise SD	Non-Tumor	Tumor	Non-Tumor	Tumor
0.10	0.78	0.51	0.69	0.94
0.40	0.52	0.60	0.53	0.94

Table S2: Comparison between network based on original and perturbed ovarian cancer data for different levels of noise based on overlapping edges and hub-gene structure. For each network type (i.e., tumor and non-tumor), the network with 500 nodes is selected. Based on these networks, the table shows (1) the percentage of edges overlapping between original and noisy network and (2) the percentage of hubs genes (i.e., nodes with more than 10 connecting edges) in the original network which are hubs in the noisy network as well.

tumor network is very consistent in the degree distribution of the hub-genes for different levels of noise. On the other hand, the hub-gene structure of the non-tumor network is more affected by increasing levels of noise. Overall, TSNet can recover more than half of the network edges and hub-structure of the original tumor and non-tumor networks for different levels of noise.

References

[1] Jaeil Ahn, Ying Yuan, Giovanni Parmigiani, Milind B Suraokar, Lixia Diao, Ignacio I Wistuba, and Wenyi Wang. Demix: deconvolution for mixed cancer transcriptomes using raw measured data. *Bioinformatics*, 29(15):1865–1871, 2013.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[3] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, et al. Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5):413–421, 2012.

[4] P. Danaher, P. Wang, and D. Witten. The jont graphical lasso for inverse covariance estiamtion across multiple classes. *Journal of the Royal Statistical Society, Series B*, 76(2), 2014.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[6] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6):417–425, 2015.

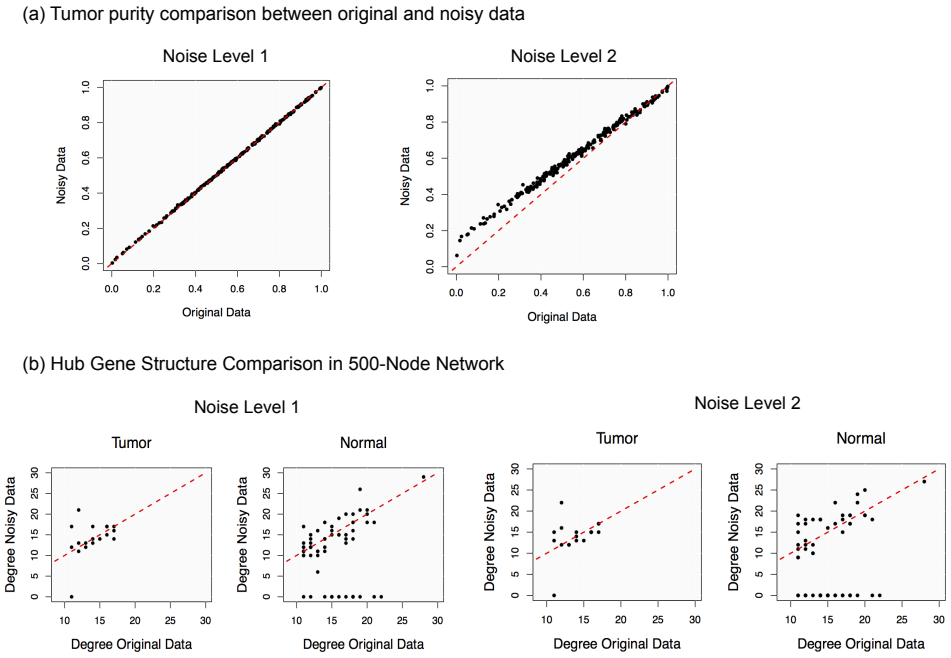

Figure S7: (a) Scatterplot of tumor purity estimated based on original and perturbed data for different levels of noise. (b) Scatterplot of number of connecting edges of hub-genes (i.e., nodes with more than 10 connecting edges) in a 500-node network based on original and noisy data.

[7] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[8] Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[9] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[10] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4, 2013.

[11] Jun Zhu, Bin Zhang, Erin N Smith, Becky Drees, Rachel B Brem, Leonid Kruglyak, Roger E Bumgarner, and Eric E Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, 40(7):854–861, 2008.