Research Report
Institute of Bioinformatics, Johannes Kepler University Linz

FABIA: Factor Analysis for Bicluster Acquisition
— *supplementary material* —

Sepp Hochreiter[1], Ulrich Bodenhofer[1], Martin Heusel[1], Andreas Mayr[1], Andreas Mitterecker[1], Adetayo Kasim[3], Tatsiana Khamiakova[3], Suzy Van Sanden[3], Dan Lin[3], Willem Talloen[4], Luc Bijnens[4], Hinrich W. H. Göhlmann[4], Ziv Shkedy[3], and Djork-Arné Clevert[1,2]

[1]Institute of Bioinformatics, Johannes Kepler University, Linz, Austria
[2]Dept. of Nephrology and Internal Intensive Care, Charité, Berlin, Germany
[3]Inst. for Biostatistics and Statistical Bioinformatics, Hasselt University, Belgium
[4]Johnson & Johnson Pharmaceutical Research & Development, a Division of Janssen Pharmaceutica, Beerse, Belgium

Contents

List of Figures

List of Tables

S1 Introduction

This document contains supplementary information for the paper "FABIA: Factor Analysis for Bicluster Acquisition". Its purpose is to provide

1. further explanations to the method,
2. mathematical details and derivations,
3. data preprocessing recommendations,
4. more detailed description of the evaluation procedure,
5. details on the settings used for the different biclustering methods in the comparative analysis, and
6. additional experimental results especially on the data set of Li *et al.* (2009) and for benchmark data created according to an additive model,
7. an extensive biological interpretation of biclustering results obtained by FABIA on the three gene expression data sets.

The document is structured analogous to the main paper. Some (sub)sections have been split in order to further structure the supplementary material.

S2 The FABIA Model

Depiction of a Factor Analysis Model

Figure S1 shows a simple factor analysis model with two factors.

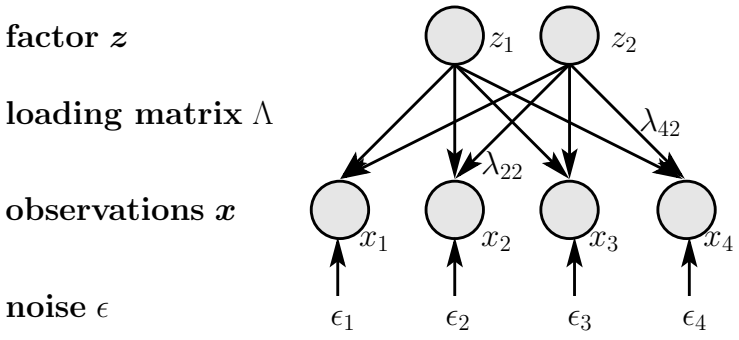

Figure S1: Factor analysis model with two factors and four observations.

The Product of Laplacians is Very Sparse

In our model, we have the products $\boldsymbol{\lambda}_i \; \boldsymbol{z}_i$, where both $\boldsymbol{\lambda}_i$ and $\boldsymbol{z}_i$ are Laplacian. Here we describe the distribution resulting from the product of two Laplacian variables.

The product of variables according to independent Laplace distributions is distributed proportionally to

$$\frac{1}{\alpha_1\,\alpha_2} K_0\left(2\,\sqrt{\frac{x}{\alpha_1\,\alpha_2}}\right)\,,$$

where K_0 is 0-th order modified Bessel function of the second kind and α_1 and α_2 are the scale parameters of the Laplacians (Bithas *et al.*, 2007). For $x > 1$ we have $K_0(x) \approx \sqrt{\frac{\pi}{2\,x}}\,e^{-x}$, which gives

$$\frac{1}{\alpha_1\,\alpha_2}\sqrt{\frac{\pi\,\sqrt{\alpha_1\,\alpha_2}}{4\,\sqrt{x}}}\,\exp\left(-\,2\,\sqrt{\frac{x}{\alpha_1\,\alpha_2}}\right)$$

for the distribution of the product of the variables. For large x, this distribution is governed by $\exp\left(-\,a\,\sqrt{x}\right)$, therefore, the tails are heavier than that of the Laplace distribution.

However, the noise term in our model is Gaussian, which reduces the sparseness of the data generated by the model. The sparseness of the model is between Gaussian (only noise) and the 0-th order modified Bessel function (no noise), depending on the signal-to-noise ratio.

In summary, our model produces data which are of about the same sparseness as Laplacian distributed data.

S3 Model Selection

S3.1 Variational Approach for Sparse Factors

We obtain the following lower bound on the likelihood:

$$\begin{aligned}
\log\,p(\boldsymbol{x}) \;\geq\; \log\,p(\boldsymbol{x}|\boldsymbol{\xi}) \;&=\; \int Q(\tilde{\boldsymbol{z}})\,\log\,p(\boldsymbol{x}|\boldsymbol{\xi})\,d\tilde{\boldsymbol{z}} \\
&=\; \int Q(\tilde{\boldsymbol{z}})\,\log\frac{Q(\tilde{\boldsymbol{z}})}{p(\tilde{\boldsymbol{z}}\mid\boldsymbol{x},\boldsymbol{\xi})}\,d\tilde{\boldsymbol{z}} - \int Q(\tilde{\boldsymbol{z}})\,\log\frac{Q(\tilde{\boldsymbol{z}})}{p(\tilde{\boldsymbol{z}},\boldsymbol{x}\mid\boldsymbol{\xi})}\,d\tilde{\boldsymbol{z}} \\
&\geq\; \int Q(\tilde{\boldsymbol{z}})\,\log p(\tilde{\boldsymbol{z}},\boldsymbol{x}\mid\boldsymbol{\xi})\,d\tilde{\boldsymbol{z}}
\end{aligned}$$

We used

$$p(\boldsymbol{x}\mid\boldsymbol{\xi}) \;=\; \frac{p(\tilde{\boldsymbol{z}},\boldsymbol{x}\mid\boldsymbol{\xi})}{p(\tilde{\boldsymbol{z}}\mid\boldsymbol{x},\boldsymbol{\xi})}\,.$$

Like for the standard EM algorithm, we now set, for each sample $\boldsymbol{x}_j$,

$$Q(\tilde{\boldsymbol{z}}_j) \;=\; p\big(\tilde{\boldsymbol{z}}_j\mid\boldsymbol{x}_j,\boldsymbol{\Lambda}^{\mathrm{old}},\boldsymbol{\Psi}^{\mathrm{old}}\big)\,.$$

For evaluating

$$p(\tilde{\boldsymbol{z}}_j,\boldsymbol{x}_j\mid\boldsymbol{\xi}_j) \;=\; p(\boldsymbol{x}_j\mid\tilde{\boldsymbol{z}}_j)\,p(\tilde{\boldsymbol{z}}_j\mid\boldsymbol{\xi}_j)\,,$$

we need the prior $p(\tilde{\boldsymbol{z}}_j)$. According to Girolami (2001) and Palmer *et al.* (2006),

$$p(\tilde{\boldsymbol{z}}_j) \;=\; \Big(\tfrac{1}{\sqrt{2}}\Big)^p \prod_{i=1}^{p} \exp(-\sqrt{2}\,|z_{ij}|) \;=\; \arg\max_{\boldsymbol{\xi}_j} p(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{\xi}_j)$$
$$= \arg\max_{\boldsymbol{\xi}_j} \Big(\tfrac{1}{\sqrt{2}}\Big)^p \prod_{i=1}^{p} \phi(\xi_{ij})\, \mathcal{N}\left(\tilde{z}_{ij}\;,\; |\xi_{ij}|\right)\;,$$

where

$$\phi(\xi_{ij}) \;=\; \exp\big(-\tfrac{1}{2}\,|\xi_{ij}|\big)\,\sqrt{2\,\pi\,|\xi_{ij}|}\,.$$

Maximizing the lower bound on the likelihood with respect to $\boldsymbol{\xi}_j$ (for the j-th sample $\boldsymbol{x}_j$) gives

$$\xi_{ij} \;=\; \sqrt{\int p(\boldsymbol{z}_j \mid \boldsymbol{x}_j)\; z_{ij}^2\; d\tilde{z}_j}\,.$$

For computing ξ_{ij}^2, we can use

$$p(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) \;\geq\; p(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j, \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\xi}^{\text{old}}) \;=\; \frac{p(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j, \boldsymbol{\Lambda}, \boldsymbol{\Psi})\; p(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{\xi}_j^{\text{old}})}{\int p(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j, \boldsymbol{\Lambda}, \boldsymbol{\Psi})\; p(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{\xi}_j^{\text{old}})\; d\tilde{\boldsymbol{z}}_j}\;,$$

where we know both the variational $p(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{\xi}_j^{\text{old}})$ and the Gaussian

$$p(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) \;=\; \mathcal{N}\left(\boldsymbol{\Lambda}\;\tilde{\boldsymbol{z}}_j\;,\; \boldsymbol{\Psi}\right)\;.$$

Using $\boldsymbol{\Xi}_j = \text{diag}(\boldsymbol{\xi}_j)$, the variational prior is the following multi-modal Gaussian:

$$\Big(\tfrac{1}{\sqrt{2}}\Big)^p\; \phi(\boldsymbol{\Xi}_j)\, \mathcal{N}\left(\tilde{\boldsymbol{z}}_j\;,\; \boldsymbol{\Xi}_j\right)$$

The posterior of $\tilde{\boldsymbol{z}}_j$ is now basically a product of Gaussians for which we can compute the conditional expectations analytically.

The conditional mean is

$$E(\boldsymbol{z}_j \mid \boldsymbol{x}_j) \;=\; \Big(\big(\boldsymbol{\Lambda}^{\text{old}}\big)^T\,\big(\boldsymbol{\Psi}^{\text{old}}\big)^{-1}\,\boldsymbol{\Lambda}^{\text{old}} \;+\; \big(\boldsymbol{\Xi}_j^{\text{old}}\big)^{-1}\Big)^{-1}\big(\boldsymbol{\Lambda}^{\text{old}}\big)^T\,\big(\boldsymbol{\Psi}^{\text{old}}\big)^{-1}\,\boldsymbol{x}_j$$

and the conditional covariance is

$$E\big(\tilde{\boldsymbol{z}}_j\;\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\big) \;=\; \Big(\big(\boldsymbol{\Lambda}^{\text{old}}\big)^T\,\big(\boldsymbol{\Psi}^{\text{old}}\big)^{-1}\,\boldsymbol{\Lambda}^{\text{old}} \;+\; \big(\boldsymbol{\Xi}_j^{\text{old}}\big)^{-1}\Big)^{-1} \;+\; E(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j)\,E(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j)^T\,.$$

The update for $\boldsymbol{\xi}_j$ is now

$$\boldsymbol{\xi}_j \;=\; \text{diag}\Big(\sqrt{\text{E}\big(\tilde{\boldsymbol{z}}_j\;\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\big)}\Big)\,.$$

S3.2 New Update Rules for Sparse Loadings

The update rule for FABIA is derived from Eq. (S1) and Eq. (S2) below.

For FABIAS we defined sparseness as

$$\mathrm{sp}(\boldsymbol{\lambda}_i) \;=\; \frac{\sqrt{n} \;-\; \sum_{k=1}^{n} |\lambda_{ki}| \;/\; \sum_{k=1}^{n} \lambda_{ki}^2}{\sqrt{n} \;-\; 1} \,,$$

and have an update rule given desired sparseness of spL:

$$\boldsymbol{\Lambda}^{\mathrm{new}} \;=\; \mathrm{proj}\left(\frac{\frac{1}{l} \sum_{j=1}^{l} \boldsymbol{x}_j \, \mathrm{E}\left(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j\right)^T}{\frac{1}{l} \sum_{j=1}^{l} \mathrm{E}(\tilde{\boldsymbol{z}}_j \; \tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j)}, \mathrm{spL} \right) \,.$$

Here we mean that given

$$\boldsymbol{\Lambda}^{\mathrm{temp}} \;=\; \frac{\frac{1}{l} \sum_{j=1}^{l} \boldsymbol{x}_j \, \mathrm{E}\left(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j\right)^T}{\frac{1}{l} \sum_{j=1}^{l} \mathrm{E}(\tilde{\boldsymbol{z}}_j \; \tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j)} \,,$$

the actual update is

$$\begin{aligned} \forall_i : \;\; \boldsymbol{\lambda}_i^{\mathrm{new}} \;&=\; \mathrm{proj}\left(\boldsymbol{\lambda}_i^{\mathrm{temp}}, \mathrm{spL}\right) \\ &=\; \min_{\boldsymbol{\lambda}} \left\| \boldsymbol{\lambda} \;-\; \boldsymbol{\lambda}_i^{\mathrm{temp}} \right\|^2 \qquad \text{s.t.} \quad \mathrm{sp}(\boldsymbol{\lambda}) \;=\; \mathrm{spL} \quad \text{and} \quad \|\boldsymbol{\lambda}\| \;=\; 1 \,. \end{aligned}$$

The length of the vector $\boldsymbol{\lambda}$ is kept at 1 by $\|\boldsymbol{\lambda}\| = 1$, but another constant length c is possible giving $\|\boldsymbol{\lambda}\|^2 = c$.

S3.3 Extremely Sparse Priors

Some gene expression data sets are sparser than Laplacian. In estimating DNA copy numbers with Affymetrix SNP 6 arrays, we observed a kurtosis larger than 30. Since noise reduces the sparseness, we need extremely sparse priors to address the characteristics of distributions with such heavy tails.

S3.3.1 Extremely Sparse Priors on Loadings

The derivatives of the negative "log"-densities of the distributions listed in Section 3.3 of the main paper are given as follows:

Generalized Gaussians:

$$\frac{\partial \left(-\ln p(z)\right)}{\partial z} \;\propto\; \beta |z|^{\beta \,-\, 1}, \qquad \text{where } 0 \;<\; \beta \;\leq\; 1.$$

Jeffrey's prior:

$$\frac{\partial \left(-\ln p(z)\right)}{\partial z} \;\propto\; \frac{1}{|z|}$$

Improper prior:

$$\frac{\partial\,(-\ln p(z))}{\partial z} \;\propto\; \beta|z|^{-\beta\,-\,1}, \qquad \text{where } 0 \;<\; \beta.$$

Let us denote the negative exponent of $|z|$ in the derivatives by spl:

$$\mathrm{spl} \;=\; \begin{cases} 1 \;-\; \beta & \text{for generalized Gaussian} \\ 1 & \text{for Jeffrey's prior} \\ 1 \;+\; \beta & \text{for the improper prior} \end{cases}$$

All $\mathrm{spl} \;\geq\; 0$ are possible, where $\mathrm{spl} \;=\; 0$ corresponds to the Laplace prior and a larger spl represents sparser priors.

For the M-step of the EM algorithm on the posterior on the parameters, we have to solve

$$\boldsymbol{\Lambda}^{\mathrm{new}}\;\frac{1}{l}\sum_{j=1}^{l}\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\;\tilde{\boldsymbol{z}}_j^T\mid\boldsymbol{x}_j\right)\;-\;\frac{1}{l}\;\sum_{j=1}^{l}\boldsymbol{x}_j\;\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\mid\boldsymbol{x}_j\right)^T\;+\;\frac{\alpha}{l}\;\boldsymbol{\Psi}\;\boldsymbol{\Lambda}^{-\mathrm{spl}}\;=\;\mathbf{0}\,, \tag{S1}$$

where we define $\boldsymbol{\Lambda}^{-\mathrm{spl}} \;=\; |\boldsymbol{\Lambda}|^{-\mathrm{spl}}\;\mathrm{sign}\,(\boldsymbol{\Lambda})$ with element-wise operations (absolute value, sign, exponentiation, multiplication). This results in the following solution:

$$\begin{aligned}\boldsymbol{\Lambda}^{\mathrm{new}} \;&=\; \Big(\frac{1}{l}\;\sum_{j=1}^{l}\boldsymbol{x}_j\;\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\mid\boldsymbol{x}_j\right)^T\;-\;\frac{\alpha}{l}\;\boldsymbol{\Psi}\;\boldsymbol{\Lambda}^{-\mathrm{spl}}\Big)\;\Big(\frac{1}{l}\sum_{j=1}^{l}\mathrm{E}\big(\tilde{\boldsymbol{z}}_j\;\tilde{\boldsymbol{z}}_j^T\mid\boldsymbol{x}_j\big)\Big)^{-1} \\ &=\; \underbrace{\Big(\sum_{j=1}^{l}\boldsymbol{x}_j\;\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\mid\boldsymbol{x}_j\right)^T\Big)\;\Big(\sum_{j=1}^{l}\mathrm{E}\big(\tilde{\boldsymbol{z}}_j\;\tilde{\boldsymbol{z}}_j^T\mid\boldsymbol{x}_j\big)\Big)^{-1}}_{=\boldsymbol{\Lambda}^{\mathrm{tmp}}}\;-\;\underbrace{\big(\alpha\;\boldsymbol{\Psi}\;\boldsymbol{\Lambda}^{-\mathrm{spl}}\big)\;\Big(\sum_{j=1}^{l}\mathrm{E}\big(\tilde{\boldsymbol{z}}_j\;\tilde{\boldsymbol{z}}_j^T\mid\boldsymbol{x}_j\big)\Big)^{-1}}_{=\boldsymbol{\Lambda}^{\mathrm{pr}}}\end{aligned} \tag{S2}$$

Without the prior, the solution would be $\boldsymbol{\Lambda}^{\mathrm{tmp}}$; correspondingly, $\boldsymbol{\Lambda}^{\mathrm{pr}}$ is the contribution of the prior. The goal is to produce a sparse solution. However, especially for an α that is not sufficiently small, the prior's contribution $\boldsymbol{\Lambda}^{\mathrm{pr}}$ may "overshoot" zero. Therefore, we propose the following truncation update (for all $k = 1, \ldots, n$ and all $i = 1, \ldots, p$):

$$\boldsymbol{\lambda}_{ki}^{\mathrm{new}} \;=\; \begin{cases} \boldsymbol{\lambda}_{ki}^{\mathrm{tmp}} \;-\; \boldsymbol{\lambda}_{ki}^{\mathrm{pr}} & \text{if } \mathrm{sign}(\boldsymbol{\lambda}_{ki}^{\mathrm{tmp}} \;-\; \boldsymbol{\lambda}_{ki}^{\mathrm{pr}}) \;=\; \mathrm{sign}(\boldsymbol{\lambda}_{ki}^{\mathrm{tmp}}) \\ 0 & \text{otherwise} \end{cases}$$

S3.3.2 Extremely Sparse Priors on Factors

We use the same priors as for the loadings. We want to represent the priors through a convex variational form according to Palmer *et al.* (2006). We have to show that

$$g(z) \;=\; -\;\ln p(\sqrt{z})$$

is increasing and concave for $z > 0$. To this end, we consider first and second order derivatives of our three extremely sparse priors:

Generalized Gaussians: corresponds to $g(z) = z^{\beta/2}$ with $0 < \beta \leq 1$, leading to

$$\begin{aligned} \frac{\partial g(z)}{\partial z} &= \frac{\beta}{2}|z|^{\beta/2\,-\,1} > 0\,, \\ \frac{\partial^2 g(z)}{\partial z^2} &= -\frac{\beta}{2}\,(1\,-\,\beta/2)\,|z|^{\beta/2\,-\,2} < 0\,. \end{aligned}$$

Jeffrey's prior: corresponds to $g(z) = \frac{1}{2}\,\ln|z|$, yielding

$$\frac{\partial g(z)}{\partial z} = \frac{1}{2}\,\frac{1}{|z|} > 0 \qquad \text{and} \qquad \frac{\partial^2 g(z)}{\partial z^2} = -\frac{1}{2}\,\frac{1}{|z|^2} < 0\,.$$

Improper prior: corresponds to $g(z) = -\,z^{-\,\beta/2}$ with $\beta > 0$, resulting in

$$\begin{aligned} \frac{\partial g(z)}{\partial z} &= \frac{\beta}{2}\,|z|^{-\beta/2\,-\,1} > 0\,, \\ \frac{\partial^2 g(z)}{\partial z^2} &= -\frac{\beta}{2}\,(1\,+\,\beta/2)\,|z|^{-\,\beta/2\,-\,2} < 0\,. \end{aligned}$$

We can summarize that, for all three priors, g is increasing and concave for $z > 0$, which allows us to represent the priors through a convex variational form.

According to Palmer *et al.* (2006), therefore, the update for the variational parameter $\boldsymbol{\xi}_j$ is

$$\boldsymbol{\xi}_j = 2\,\frac{\partial g}{\partial \tilde{\boldsymbol{z}}}\left(\operatorname{diag}\left(\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\,\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\right)\right)\right)\,,$$

which results in the following updates:

Generalized Gaussians:

$$\boldsymbol{\xi}_j = \beta\,\operatorname{diag}\left(\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\,\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\right)\right)^{\beta/2\,-\,1}$$

Jeffrey's prior:

$$\boldsymbol{\xi}_j = \operatorname{diag}\left(\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\,\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\right)\right)^{-1}$$

Improper prior:

$$\boldsymbol{\xi}_j = \beta\,\operatorname{diag}\left(\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\,\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\right)\right)^{-\beta/2\,-\,1}$$

We denote the negative exponent of $E\big(\tilde{\boldsymbol{z}}_j\,\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\big)$ in the update by spz:

$$\mathrm{spz} = \begin{cases} 1\,-\,\beta/2 & \text{for generalized Gaussian} \\ 1 & \text{for Jeffrey's prior} \\ 1\,+\,\beta/2 & \text{for improper prior} \end{cases}$$

This definition covers all $\mathrm{spz} \geq 1/2$. The smallest $\mathrm{spz} = 1/2$ ($\beta = 1$) represents the Laplace prior and $\mathrm{spz} > 1/2$ leads to sparser priors; the update of the variational variable can then we written in the following general form:

$$\boldsymbol{\xi}_j \propto \operatorname{diag}\left(\mathrm{E}\left(\tilde{\boldsymbol{z}}_j\,\tilde{\boldsymbol{z}}_j^T \mid \boldsymbol{x}_j\right)^{\mathrm{spz}}\right)$$

S3.4 Data Preprocessing and Initialization

The data $\boldsymbol{x}$ may be centered to zero mean, to zero median, or to zero mode. The latter two centerings are supposed to result in sparser raw data. We recommend median centering because it is both suited for giving sparse data and robust. In a second step, the data can be normalized.

Centering

Our prior distribution $p(\tilde{\boldsymbol{z}})$ is assumed to be unimodal, symmetric, decorrelated and to have zero mean. The additive noise is assumed to be Gaussian with mean 0. Thus, the model distribution $p(\boldsymbol{x})$ of the observations $\boldsymbol{x}$ is unimodal and symmetric. The symmetry follows from

$$\begin{aligned} p(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}) \;&\sim\; \mathcal{N}(\boldsymbol{x} - \boldsymbol{\Lambda}\tilde{\boldsymbol{z}}\,,\; \boldsymbol{\Psi}) \\ p(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}) \;&=\; p(-\boldsymbol{x} \mid -\tilde{\boldsymbol{z}}) \end{aligned}$$

and

$$\begin{aligned} p(\boldsymbol{x}) \;&=\; \int p(\boldsymbol{x} \mid \tilde{\boldsymbol{z}})\; p(\tilde{\boldsymbol{z}}) d\tilde{\boldsymbol{z}} \\ &= \int p(-\boldsymbol{x} \mid -\tilde{\boldsymbol{z}})\; p(-\tilde{\boldsymbol{z}}) d\tilde{\boldsymbol{z}} \;=\; p(-\boldsymbol{x})\;, \end{aligned}$$

where, in the second equality, the change of the $d\tilde{\boldsymbol{z}}$ to $d(-\tilde{\boldsymbol{z}})$ and the change of integration limits introduce both a "$-$"-sign which cancels. The unimodality follows from (Dharmadhikari and Jogdeo, 1976, Theorem 3.4).

For the first and third moment of $p(\boldsymbol{x})$, we obtain immediately

$$\mathrm{E}(\boldsymbol{x}) \;=\; \mathbf{0} \quad \text{and} \quad \mathrm{E}(\boldsymbol{x}^3) \;=\; \mathbf{0}\;,$$

where the latter means that all third moments are zero. The second moment of $p(\boldsymbol{x})$ is

$$\begin{aligned} \mathrm{E}(\boldsymbol{x}\,\boldsymbol{x}^T) \;&=\; \boldsymbol{\Lambda}\mathrm{E}(\boldsymbol{z}\boldsymbol{z}^T)\boldsymbol{\Lambda}^T \;+\; \boldsymbol{\Lambda}\mathrm{E}(\boldsymbol{z})\mathrm{E}(\boldsymbol{\epsilon}^T) \;+\; \mathrm{E}(\boldsymbol{z})\mathrm{E}(\boldsymbol{\epsilon})\boldsymbol{\Lambda}^T \;+ \mathrm{E}(\boldsymbol{\epsilon}\,\boldsymbol{\epsilon}^T) \\ &=\; \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \;+\; \mathrm{diag}(\sigma_k^2) \;=\; \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \;+\; \boldsymbol{\Psi}\;. \end{aligned}$$

Note that, for a unimodal distribution of a random variable x with $\mathrm{E}(x) = 0$, $\mathrm{E}(x^2) = 1$, and $\mathrm{E}(x^3) = \tau$, the mode is $-\frac{1}{2}\tau$ and the median $-\frac{1}{6}\tau$ (Haldane, 1942; Hall, 1980). For $p(x)$ we see that $\tau = 0$. *The mean, the mode, and the median are all zero for our model.*

If the real underlying distribution or the empirical values are skewed, then mean, median, and mode differ from each other. In this case, we can either set the mean, the median, or the mode to zero. For the latter, we either add $\frac{1}{2}\tau$ to x giving $\mathrm{E}(x) = \frac{1}{2}\tau$ to move the mode to zero, or we subtract $3 \cdot \mathrm{median}$ from each x using $\mathrm{median} = -\frac{1}{6}\tau$.

We have the following centering methods:

- *Zero mean centering* — not sparse.
- *Zero median centering* — sparser than mean and robust.
- *Zero mode centering* — sparser but less robust.

Since sparseness is one of the key features of our model, we want to set the mode to zero. However, the estimation of τ is not robust, therefore we prefer zero median centering because the median is closer to the mode than the mean.

Normalization

If the correlation of weak signals is of interest too, we recommend to normalize the data. Our R package supports the following two methods:

1. Row-wise division by standard deviation (scaling to unit variance)
2. Row-wise division by the difference of 75% and 25% quantile

Initialization

The simplest strategy is to initialize $\mathbf{\Lambda}$ randomly while ensuring that

$$\mathbf{\Psi} = \mathrm{diag}\Big(\mathrm{covar}(\boldsymbol{x}) - \mathbf{\Lambda}\mathbf{\Lambda}^T\Big) \geq \delta > 0 \; .$$

Random initialization with the same range is justified after normalization of the components to unit second moment. The variational parameter vectors $\boldsymbol{\xi}_j$ are initialized by vectors of ones.

S4 Information Content of Biclusters

A highly desired property for biclustering algorithms is to rank the extracted biclusters analogously to principal components which are ranked according to the data variance they explain. We rank biclusters according to the information they contain about the data.

S4.1 Information of the Latent Variables on Observations

We measure the information in biclusters through the mutual information between $\tilde{\boldsymbol{z}}$ and $\boldsymbol{x}$, that is, how much information about $\boldsymbol{x}$ is contained in $\tilde{\boldsymbol{z}}$. This idea is the basis of the I/NI calls in (Talloen *et al.*, 2007).

The entropy of a multivariate Gaussian $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ with density

$$p(\boldsymbol{x}) \;=\; \frac{1}{(2\,\pi)^{n/2}\;|\mathbf{\Sigma}|^{1/2}} \exp\big(-\;\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\,\mathbf{\Sigma}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})\big)$$

is

$$\mathrm{H}\left(\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})\right) \;=\; \ln\Big(\sqrt{(2\,\pi\,e)^n\;|\mathbf{\Sigma}|}\Big) \; .$$

Mutual information of $\boldsymbol{x}$ and $\tilde{\boldsymbol{z}}$ is defined as

$$\mathrm{I}(\boldsymbol{x}; \tilde{\boldsymbol{z}}) \;=\; \mathrm{H}(\boldsymbol{x}) \;-\; \mathrm{H}(\boldsymbol{x} \mid \tilde{\boldsymbol{z}}) \; .$$

In our model, we have

$$\boldsymbol{x}_j \;\sim\; \mathcal{N}\left(\mathbf{0}\;,\; \mathbf{\Psi} \;+\; \mathbf{\Lambda}\;\mathbf{\Xi}_j\;\mathbf{\Lambda}^T\right) \qquad \text{and} \qquad \boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j \;\sim\; \mathcal{N}\left(\mathbf{\Lambda}\,\tilde{\boldsymbol{z}}_j\;,\; \mathbf{\Psi}\right) \; .$$

Thus, the mutual information is

$$\begin{aligned}
\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) &= \mathrm{H}(\boldsymbol{x}_j) \;-\; \mathrm{H}(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j) \\
&= \ln\left(\sqrt{(2\,\pi\,e)^n\;|\boldsymbol{\Psi} \;+\; \boldsymbol{\Lambda}\;\boldsymbol{\Xi}_j\;\boldsymbol{\Lambda}^T|}\right) \;-\; \ln\left(\sqrt{(2\,\pi\,e)^n\;|\boldsymbol{\Psi}|}\right) \\
&= \frac{1}{2}\;\ln\left|\left(\boldsymbol{\Psi} \;+\; \boldsymbol{\Lambda}\;\boldsymbol{\Xi}_j\;\boldsymbol{\Lambda}^T\right)\;\boldsymbol{\Psi}^{-1}\right| \\
&= \frac{1}{2}\;\ln\left|\boldsymbol{I}_n \;+\; \boldsymbol{\Lambda}\;\boldsymbol{\Xi}_j\;\boldsymbol{\Lambda}^T\;\boldsymbol{\Psi}^{-1}\right| \\
&= \frac{1}{2}\;\ln\left|\boldsymbol{I}_p \;+\; \boldsymbol{\Lambda}^T\;\boldsymbol{\Psi}^{-1}\;\boldsymbol{\Lambda}\;\boldsymbol{\Xi}_j\right| \\
&= \frac{1}{2}\;\ln\left|\boldsymbol{I}_p \;+\; \boldsymbol{\Xi}_j\;\boldsymbol{\Lambda}^T\;\boldsymbol{\Psi}^{-1}\;\boldsymbol{\Lambda}\right|\;,
\end{aligned}$$

where we applied *Sylvester's theorem for determinants*,

$$|\boldsymbol{I}_n \;+\; \boldsymbol{U}\;\boldsymbol{V}^T| = |\boldsymbol{I}_p \;+\; \boldsymbol{V}^T\;\boldsymbol{U}|,$$

which is a special case of the generalization of the *matrix determinant lemma*:[1]

$$\left|\boldsymbol{A} \;+\; \boldsymbol{U}\;\boldsymbol{V}^T\right| \;=\; \left|\boldsymbol{I} \;+\; \boldsymbol{V}^T\;\boldsymbol{A}^{-1}\;\boldsymbol{U}\right|\;|\boldsymbol{A}|\;.$$

The above formula can also be obtained from

$$\begin{aligned}
\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) &= \mathrm{H}(\tilde{\boldsymbol{z}}_j) \;-\; \mathrm{H}(\tilde{\boldsymbol{z}}_j \mid \boldsymbol{x}_j) \\
&= \ln\left(\sqrt{(2\,\pi\,e)^n\;|\boldsymbol{\Xi}_j|}\right) \;-\ln\left(\sqrt{(2\,\pi\,e)^n\;\left|\left(\boldsymbol{\Lambda}^T\;\boldsymbol{\Psi}^{-1}\;\boldsymbol{\Lambda} \;+\; \boldsymbol{\Xi}_j^{-1}\right)^{-1}\right|}\right) \\
&= \frac{1}{2}\;\ln\left|\boldsymbol{\Xi}_j\;\left(\boldsymbol{\Lambda}^T\;\boldsymbol{\Psi}^{-1}\;\boldsymbol{\Lambda} \;+\; \boldsymbol{\Xi}_j^{-1}\right)\right| \\
&= \frac{1}{2}\;\ln\left|\boldsymbol{I}_p \;+\; \boldsymbol{\Xi}_j\;\boldsymbol{\Lambda}^T\;\boldsymbol{\Psi}^{-1}\;\boldsymbol{\Lambda}\right|\;.
\end{aligned}$$

The independence of the factors (up to the second moment) implies that the covariance matrix of $\tilde{\boldsymbol{z}}_j$, i.e. $\boldsymbol{\Xi}_j$, is diagonal. This allows for the expansion

$$\boldsymbol{\Lambda}\;\boldsymbol{\Xi}_j\;\boldsymbol{\Lambda}^T \;=\; \sum_{i=1}^{p} \xi_{ij}\;\boldsymbol{\lambda}_i\;\boldsymbol{\lambda}_i^T\;,$$

[1] `http://en.wikipedia.org/wiki/Matrix_determinant_lemma`

and we obtain

$$\begin{aligned}
\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) &= \frac{1}{2} \ln \left| \boldsymbol{I}_n + \boldsymbol{\Lambda}\ \boldsymbol{\Xi}_j\ \boldsymbol{\Lambda}^T\ \boldsymbol{\Psi}^{-1} \right| \\
&= \frac{1}{2} \ln \left| \boldsymbol{I}_n + \boldsymbol{\Psi}^{-1}\ \boldsymbol{\Lambda}\ \boldsymbol{\Xi}_j\ \boldsymbol{\Lambda}^T \right| \\
&= \frac{1}{2} \ln \left| \boldsymbol{I}_n + \boldsymbol{\Psi}^{-1} \sum_{i=1}^{p} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T \right| \\
&= \frac{1}{2} \ln \left| \boldsymbol{I}_n + \boldsymbol{\Psi}^{-1} \sum_{i=1}^{p-1} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T + \boldsymbol{\Psi}^{-1}\ \xi_{jp}\ \boldsymbol{\lambda}_p\ \boldsymbol{\lambda}_p^T \right| \\
&= \frac{1}{2} \ln \left(\left| \boldsymbol{I}_n + \boldsymbol{\Psi}^{-1} \sum_{i=1}^{p-1} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T \right| \left(1 + \xi_{jp} \boldsymbol{\lambda}_p^T \boldsymbol{\Psi}^{-1} \left(\boldsymbol{I}_n + \boldsymbol{\Psi}^{-1} \sum_{i=1}^{p-1} \xi_{ij} \boldsymbol{\lambda}_i \boldsymbol{\lambda}_i^T \right)^{-1} \boldsymbol{\lambda}_p \right) \right) \\
&= \frac{1}{2} \ln \left(\left| \boldsymbol{I}_n + \boldsymbol{\Psi}^{-1} \sum_{i=1}^{p-1} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T \right| \left(1 + \xi_{jp}\ \boldsymbol{\lambda}_p^T \left(\boldsymbol{\Psi} + \sum_{i=1}^{p-1} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T \right)^{-1} \boldsymbol{\lambda}_p \right) \right),
\end{aligned}$$

where we again used the generalization of the *matrix determinant lemma* with

$$\boldsymbol{A} = \boldsymbol{I}_n + \sum_{i=1}^{p-1} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T\ \boldsymbol{\Psi}^{-1}\,.$$

In information-theoretic terms, this can be interpreted as

$$\begin{aligned}
\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) &= \mathrm{H}(\boldsymbol{x}_j) - \mathrm{H}(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j) \\
&= \mathrm{H}(\boldsymbol{x}_j) - \mathrm{H}(\boldsymbol{x}_j \mid z_{pj}) + \mathrm{H}(\boldsymbol{x}_j \mid z_{pj}) - \mathrm{H}(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j) \\
&= \mathrm{I}(\boldsymbol{x}_j; z_{pj}) + \mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j \mid z_{pj})
\end{aligned}$$

with

$$\mathrm{I}(\boldsymbol{x}_j; z_{pj}) = \frac{1}{2} \ln \left(1 + \xi_{pj}\ \boldsymbol{\lambda}_p^T \left(\boldsymbol{\Psi} + \sum_{i=1}^{p-1} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T \right)^{-1} \boldsymbol{\lambda}_p \right)$$

and

$$\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j \mid z_{pj}) = \frac{1}{2} \ln \left| \boldsymbol{I}_n + \boldsymbol{\Psi}^{-1} \sum_{i=1}^{p} \xi_{ij}\ \boldsymbol{\lambda}_i\ \boldsymbol{\lambda}_i^T \right|.$$

By applying the same decomposition inductively, we obtain

$$\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) = \frac{1}{2} \sum_{i=1}^{p} \ln \left(1 + \xi_{ij}\ \boldsymbol{\lambda}_i^T \left(\boldsymbol{\Psi} + \sum_{t=1}^{i-1} \xi_{tj}\ \boldsymbol{\lambda}_t\ \boldsymbol{\lambda}_t^T \right)^{-1} \boldsymbol{\lambda}_i \right).$$

In information-theoretic terms, this can be interpreted as

$$\begin{aligned}
\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) &= \mathrm{I}(\boldsymbol{x}_j; z_{pj}) + \mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j \mid z_{pj}) \\
&= \mathrm{I}(\boldsymbol{x}_j; z_{pj}) + \mathrm{I}(\boldsymbol{x}_j; z_{p-1j} \mid z_{pj}) + \mathrm{I}(\boldsymbol{x}_j; z_{p-2j} \mid z_{p-1j}, z_{pj}) + \\
&\quad \ldots + \mathrm{I}(\boldsymbol{x}_j; z_{2j} \mid z_{3j}, \ldots, z_{pj}) + \mathrm{I}(\boldsymbol{x}_j; z_{j1} \mid z_{2j}, z_{3j}, \ldots, z_{pj})\,.
\end{aligned}$$

It is obvious that there is no necessity to proceed by isolating the variables in the exact order from p down to 1. Instead, we can use any permutation $(i_1, \ldots, i_p)$. Then the above formula generalizes to the following:

$$\begin{aligned}\mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j) \;&=\; \mathrm{I}(\boldsymbol{x}_j; z_{i_1 j}) \;+\; \mathrm{I}(\boldsymbol{x}_j; \tilde{\boldsymbol{z}}_j \mid z_{i_1 j}) \\ &=\; \mathrm{I}(\boldsymbol{x}_j; z_{i_1 j}) + \mathrm{I}(\boldsymbol{x}_j; z_{i_2 j} \mid z_{i_1 j}) \;+\; \mathrm{I}(\boldsymbol{x}_j; z_{i_3 j} \mid z_{i_2 j}, z_{i_1 j}) + \\ &\quad \ldots \;+\; \mathrm{I}(\boldsymbol{x}_j; z_{i_{p-1} j} \mid z_{i_{p-2} j}, \ldots, z_{i_1 j}) \;+\; \mathrm{I}(\boldsymbol{x}_j; z_{i_p j} \mid z_{i_{p-1} j}, \ldots, z_{i_1 j})\end{aligned}$$

Finally, the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is the sum of mutual information between each $\boldsymbol{x}_j$ and its corresponding $\tilde{\boldsymbol{z}}_j$. This follows from the independence of $\boldsymbol{x}_j$, the independence of $\tilde{\boldsymbol{z}}_j$ across j, and the fact that the entropy is additive for independent variables:

$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{Z}) \;=\; \frac{1}{2} \sum_{j=1}^{l} \ln \left| \boldsymbol{I}_p \;+\; \boldsymbol{\Lambda}^T \, \boldsymbol{\Psi}^{-1} \, \boldsymbol{\Lambda} \, \boldsymbol{\Xi}_j \right| \;.$$

S4.2 Information of One Latent Variable on Observations

Now we can consider the case that one factor z_i is removed from the final model. Thus, the explained covariance

$$\xi_{ij} \, \boldsymbol{\lambda}_i \, \boldsymbol{\lambda}_i^T$$

is lost and must be considered as noise. Hence, we have

$$\boldsymbol{x}_j \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij}) \;\sim\; \mathcal{N}\left(\boldsymbol{\Lambda} \, \tilde{\boldsymbol{z}}_j |_{z_{ij}=0} \;,\; \boldsymbol{\Psi} \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i \, \boldsymbol{\lambda}_i^T \right) \;.$$

Then the information of this factor in the context of the other factors can be expressed by

$$\begin{aligned}\mathrm{I}(\boldsymbol{x}_j; z_{ij} \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij})) \;&=\; \mathrm{H}(\boldsymbol{x}_j \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij})) \;-\; \mathrm{H}(\boldsymbol{x}_j \mid \tilde{\boldsymbol{z}}_j) \\ &=\; \ln \left(\sqrt{(2\,\pi\,e)^n \, \left| \boldsymbol{\Psi} \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i \, \boldsymbol{\lambda}_i^T \right|} \right) \;-\; \ln \left(\sqrt{(2\,\pi\,e)^n \, |\boldsymbol{\Psi}|} \right) \\ &=\; \frac{1}{2} \, \ln \left| \left(\boldsymbol{\Psi} \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i \, \boldsymbol{\lambda}_i^T \right) \, \boldsymbol{\Psi}^{-1} \right| \;=\; \frac{1}{2} \, \ln \left| \boldsymbol{I}_n \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i \, \boldsymbol{\lambda}_i^T \boldsymbol{\Psi}^{-1} \right| \\ &= \frac{1}{2} \, \ln \left(1 \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}_i \right) \;.\end{aligned}$$

The information of the i-th factor in the context of the other factors can also be expressed as

$$\begin{aligned}\mathrm{I}(\boldsymbol{x}_j; z_{ij} \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij})) \;&=\; \mathrm{H}(z_{ij} \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij})) \;-\; \mathrm{H}(z_{ij} \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij}), \boldsymbol{x}_j) \\ &=\; \ln \left(\sqrt{(2\,\pi\,e)\,\xi_{ij}} \right) \;-\; \ln \left(\sqrt{(2\,\pi\,e) \, \left(\xi_{ij}^{-1} \;+\; \boldsymbol{\lambda}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}_i \right)^{-1}} \right) \\ &=\; \frac{1}{2} \, \ln \left(1 \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}_i \right) \;.\end{aligned}$$

Finally, the mutual information between the observed data $\boldsymbol{X}$ and the i-th factor $\boldsymbol{z}_i^T$ is the sum of mutual information between $\boldsymbol{x}_j$ and z_{ij}. This follows from the independence of $\boldsymbol{x}_j$, the independence of z_{ij}, and the fact the entropy is additive for independent variables:

$$\mathrm{I}(\boldsymbol{X}; \boldsymbol{z}_i^T \mid (\boldsymbol{Z} \setminus \boldsymbol{z}_i^T)) \;=\; \frac{1}{2} \sum_{j=1}^{l} \ln \left(1 \;+\; \xi_{ij} \, \boldsymbol{\lambda}_i^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda}_i \right)$$

Normalizing the Information

If the bicluster size is not of interest, the information content can be normalized to the range $[0, 1]$ through

$$\sum_{j=1}^{l} \mathrm{H}(z_{ij} \mid (\tilde{\boldsymbol{z}}_j \setminus z_{ij})) \;=\; \frac{l}{2} \ln(2\,\pi\,e) \;+\; \frac{1}{2} \sum_{j=1}^{l} \ln \xi_{ij}\;.$$

S5 Extracting Members of Biclusters

Since we have

$$\boldsymbol{\Lambda}\,\boldsymbol{Z} \;=\; \boldsymbol{\Lambda} \sqrt{\frac{1}{l}\,\mathrm{diag}\,(\boldsymbol{Z}\,\boldsymbol{Z}^T)}\; \left(\sqrt{\frac{1}{l}\,\mathrm{diag}\,(\boldsymbol{Z}\,\boldsymbol{Z}^T)}\right)^{-1} \boldsymbol{Z}\;,$$

we can define

$$\hat{\boldsymbol{\Lambda}} \;=\; \boldsymbol{\Lambda} \sqrt{\frac{1}{l}\,\mathrm{diag}\,(\boldsymbol{Z}\,\boldsymbol{Z}^T)} \qquad \text{and} \qquad \hat{\boldsymbol{Z}} \;=\; \left(\sqrt{\frac{1}{l}\,\mathrm{diag}\,(\boldsymbol{Z}\,\boldsymbol{Z}^T)}\right)^{-1} \boldsymbol{Z}\;.$$

This scaling normalizes the moments of the factors to 1.

Now a threshold thresZ on the factors can be chosen. We set $\mathrm{thresZ} = 0.5$ in our experiments. For the Laplace distribution with variance 1, we then obtain $\frac{1}{2}\exp(-\sqrt{2}/2) \approx 0.25$, which means 25% of the samples can belong to a bicluster. For the improper distribution $\mathrm{spz} = 1$, the percentage of samples that belong to a bicluster would be smaller, because more z_{ij} are close to zero.

Since biclusters may overlap, the contribution of $\lambda_{ki} z_{ij}$ that are relevant must be estimated. Therefore, we we first estimate the variance of $\boldsymbol{\Lambda}\boldsymbol{Z}$ by

$$\begin{aligned}
\mathrm{vLZ} \;=\; \mathrm{E}\Big(\big(\hat{\lambda}_{ki}\;\hat{z}_{ij}\big)^2\Big) \;&=\; \frac{1}{p\;l\;n} \sum_{(i,j,k)=(1,1,1)}^{(p,l,n)} \big(\hat{\lambda}_{ki}\;\hat{z}_{ij}\big)^2 \\
&=\; \frac{1}{p\;n} \sum_{(i,k)=(1,1)}^{(p,n)} \big(\hat{\lambda}_{ki}\big)^2\; \frac{1}{l} \sum_{j=1}^{l} \big(\hat{z}_{ij}\big)^2 \\
&=\; \frac{1}{p\;n} \sum_{(i,k)=(1,1)}^{(p,n)} \big(\hat{\lambda}_{ki}\big)^2 \\
&=\; \mathrm{var}(\boldsymbol{\Lambda})\;.
\end{aligned}$$

Here we used the fact that the factors are normalized and $\frac{1}{l}\sum_{j=1}^{l}\big(\hat{z}_{ij}\big)^2 = 1$. We set the standard deviation $\mathrm{sdLZ} = \sqrt{\mathrm{vLZ}}$ to the product of both thresholds which is solved for thresL:

$$\mathrm{thresL} \;=\; \frac{\mathrm{sdLZ}}{\mathrm{thresZ}} \;=\; \frac{\mathrm{sdL}}{\mathrm{thresZ}}\;,$$

where $\mathrm{sdL} = \sqrt{\mathrm{var}(\boldsymbol{\Lambda})}$.

Note that the average contribution sdLZ includes elements close to zero that do not belong to any bicluster. Therefore, sdLZ may underestimate the contribution of a bicluster, because the non-bicluster elements should not count. On the other hand, both λ_{ki} and z_{ij} are assumed to stem from sparse distributions which favor large values that might dominate the second moment. In this case, sdLZ overestimates the contribution of a bicluster, because large values dominate. Summarizing, the *choice of* thresL *is a trade-off between underestimation due to sparseness and overestimation due to large values.*

S6 Experiments

S6.1 Evaluating Biclustering Results

Let $\mathcal{A}_1, \ldots, \mathcal{A}_p$ be the true biclusters and $\mathcal{B}_1, \ldots, \mathcal{B}_q$ biclusters extracted by some biclustering method. Here a bicluster is supposed to be a set of index pairs (k, j) to identify matrix elements, i.e. expression values, which are grouped together. Then the similarity index matrix $\boldsymbol{I}$ is given as

$$I_{rs} \;=\; \mathrm{ja}\left(\mathcal{A}_r, \mathcal{B}_s\right) \;,$$

where $r \in \{1, \ldots, p\}$ and $s \in \{1, \ldots, q\}$ and ja is the Jaccard index. These indices measure the similarity of two biclusters — here the similarity between biclusters $\mathcal{A}_r$ and $\mathcal{B}_s$.

We use the Munkres algorithm implemented in the R package `truecluster` (Oehlschlägel, 2006) to compute an optimal assignment of biclusters $\mathcal{B}_1, \ldots, \mathcal{B}_q$ to the true biclusters $\mathcal{A}_1, \ldots, \mathcal{A}_p$. The optimal assignment is given as a set of pairs

$$\left\{(r_1, s_1), \ldots, (r_{\min(p,q)}, s_{\min(p,q)})\right\} \;,$$

where all $r_1, \ldots, r_{\min(p,q)}$ are pairwise different and all $s_1, \ldots, s_{\min(p,q)}$ are pairwise different. The optimal score is given as

$$v \;=\; \sum_{i=1}^{\min(p,q)} \mathrm{ja}\left(\mathcal{A}_{r_i}, \mathcal{B}_{s_i}\right)$$

The final *consensus score* s is computed as

$$s \;=\; \frac{v}{\max(p, q)}$$

in order to ensure that sets with a single bicluster and sets with all possible biclusters do not obtain the maximal score.

S6.2 Compared Methods

S6.2.1 Settings of Compared Methods

We compared FABIA and FABIAS with 11 other biclustering methods. Some methods were tested for more than one setting. We denote these variants as *method_variant* (e.g. plaid_ss). Table S1 provides a complete overview of all methods and the settings with which they were run.

Table S1: Compared biclustering methods and their settings. An "*na*" entry means that the methods has not been tested on simulated data sets.

Method	**General settings and remarks**	**Changes for simulated data (if any)**
FABIA	$\mathrm{thresZ} = 0.5$ (for a Laplace prior, on average 25% of the samples are assumed to belong to a bicluster), $\alpha = 0.1$, $p = 5$	$\alpha = 0.4$, $p = 13$
FABIAS	$\alpha = 0.4$, $p = 5$, otherwise same as for FABIA	$\alpha = 0.6$, $p = 13$
MFSC	see S6.2.2 for details, $\alpha_{L/Z} = 0.4$, $p = 5$	$\alpha_{L/Z} = 0.6$, $p = 13$
plaid_ss	seekss, $l = 5$, layer $a + b$	$l = 13$
plaid_ms	seekms, $l = 5$, layer $a + b$	$l = 13$
plaid_ms_5	seekms, $l = 5$, layer $a + b$, 5 iterations	$l = 13$
plaid_a_ss	seekms, $l = 5$, layer a	*na*
plaid_a_ms	seekms, $l = 5$, layer a	*na*
plaid_a_ms_5	seekms, $l = 5$, layer a, 5 iterations	*na*
ISA_1	$tc = 2.0$, $tg = 2.0$	
ISA_2	$tc = 1.0, tg = 1.0$	
ISA_3	$tc = 1.1$, $tg = 0.7$	
OPSM	passed models = 10	
SAMBA	opt="valsp_3ap", overlap=0.5	
SAMBA_01	opt="valsp_3ap", overlap=0.1	*na*
xMOTIF	preprocessing by discretization, $\alpha = 5$ (minimal number of samples in bicluster), ns=100 (number of i trials), nd=100 (number of j trials), and sd=5 (seed size of samples), alpha=0.05, number=5	number=13
Bimax	preprocessing by binarization, number=5	number=13
CC	α=1.2 according to (Cheng and Church, 2000), $\delta = 0.03$ is computed from the data in (Cheng and Church, 2000) by rescaling the data range, number=5	number=13
plaid_t_ab	cluster="b", background="TRUE", row.release = 0.7, col.release = 0.7, shuffle = 3, back.fit = 0, iter.startup = 5, iter.layer = 10, fit.model = $y \sim m + a + b$, max.layers = 5	max.layers = 13
plaid_t_a	fit.model = $y \sim m + a$, otherwise same as for plaid_t_ab	max.layers = 13
FLOC	$M = 5$ (minimal number of samples in a bicluster), $N = 30$ (minimal number of genes in a bicluster), pGene=0.1, pSample=0.1, k = 5, t = 500	k = 13
spec_1	$\exp$ preprocessing with $\log$ normalization (to neutralize the preprocessing), numberOfEigenvalues=1, withinVar=100	
spec_2	preprocessing as for spec_1, numberOfEigenvalues=3, withinVar=20	

S6.2.2 Sparse Matrix Factorization

Nonnegative matrix factorization (Lee and Seung, 2001) is a popular multiplicative model for gene expression data. Nonnegative matrix factorization is concerned with computing a multiplicative decomposition of a positive data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times l}$ into two positive matrices $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{Z} \in \mathbb{R}^{p \times l}$ in the following way:

$$\boldsymbol{X} \;=\; \boldsymbol{\Lambda}\, \boldsymbol{Z} \;=\; \sum_{i=1}^{p} \boldsymbol{\lambda}_i\, \boldsymbol{z}_i^T \;,$$

The right hand side of this equation expresses the model as the sum of outer products $\boldsymbol{\lambda}_i \in \mathbb{R}^n$ and the $\tilde{\boldsymbol{z}}_i \in \mathbb{R}^l$, where $\boldsymbol{z}_i^T$ is the i-th row of matrix $\boldsymbol{Z}$.

To allow biclustering, in (Caldas and Kaski, 2008), the indicator variables ρ_{ki} and κ_{ij} of the plaid model are used. In such a way they showed the connection of the plaid model to *binary matrix factorization* (Meeds *et al.*, 2007). We follow this idea and also introduce plaid model indicator variables ρ_{ki} and κ_{ij} for *nonnegative matrix factorization*:

$$\boldsymbol{X} \;=\; \sum_{i=1}^{p} \mathrm{diag}\,(\boldsymbol{\rho}_i)\;\, \boldsymbol{\lambda}_i\, \boldsymbol{z}_i^T\, \mathrm{diag}\,(\boldsymbol{\kappa}_i)\;\;,$$

where $\mathrm{diag}(\boldsymbol{\rho}_i)$ is the $n \times n$ diagonal matrix with entries ρ_{ki} and $\mathrm{diag}(\boldsymbol{\kappa}_i)$ is the $l \times l$ diagonal matrix with entries κ_{ij}.

In the final solution the indicator variables are binary which leads to binary matrix factorization. However, during learning the plaid model, the variables ρ_{ki} and κ_{ij} are not binary, which means that the values $\mathrm{diag}(\boldsymbol{\rho}_i)\;\boldsymbol{\lambda}_i$ and $\mathrm{diag}(\boldsymbol{\kappa}_i)\;\boldsymbol{z}_i$ are sparse vectors, i.e. most components are close to zero. If we do not enforce binary ρ_{ki} and κ_{ij} but small values, then this would lead to sparse nonnegative matrix factorization.

If we skip the non-negativity constraints, then the task is *sparse matrix factorization*, for which an algorithm has been suggested by Hoyer (2004). We call this method MFSC and test it in our experiments. Note that sparse matrix factorization is suited for biclustering, but not a generative approach.

S6.3 Simulated Data Sets with Known Biclusters

S6.3.1 Prelic Data

The characteristics of the data sets published in Prelic *et al.* (2006) are given as follows. It is obvious that most of these characteristics deviate substantially from the characteristics of gene expression data:

- Data sets are small: 50 to 100 genes
- Biclusters are equally sized
- Biclusters are constant, i.e. all genes are up-regulated to exactly the same value

- Biclusters overlap only simultaneously on rows and columns (in gene expression, however, more than one pathway can be switched on in a sample even if the pathways do not share genes)
- Low noise
- Data distribution is bimodal (see Fig. S2) in contrast to observed distributions in gene expression data sets (see Figs. S8 and S9).
- Skewness of data (0.64) is higher than observed in gene expression data sets, whereas the excess kurtosis (0.15) is too low. In our experiments, we have on average zero skewness (note that the skewness in gene expression data set is sometimes positive and sometimes negative) and average excess kurtosis of 0.65.

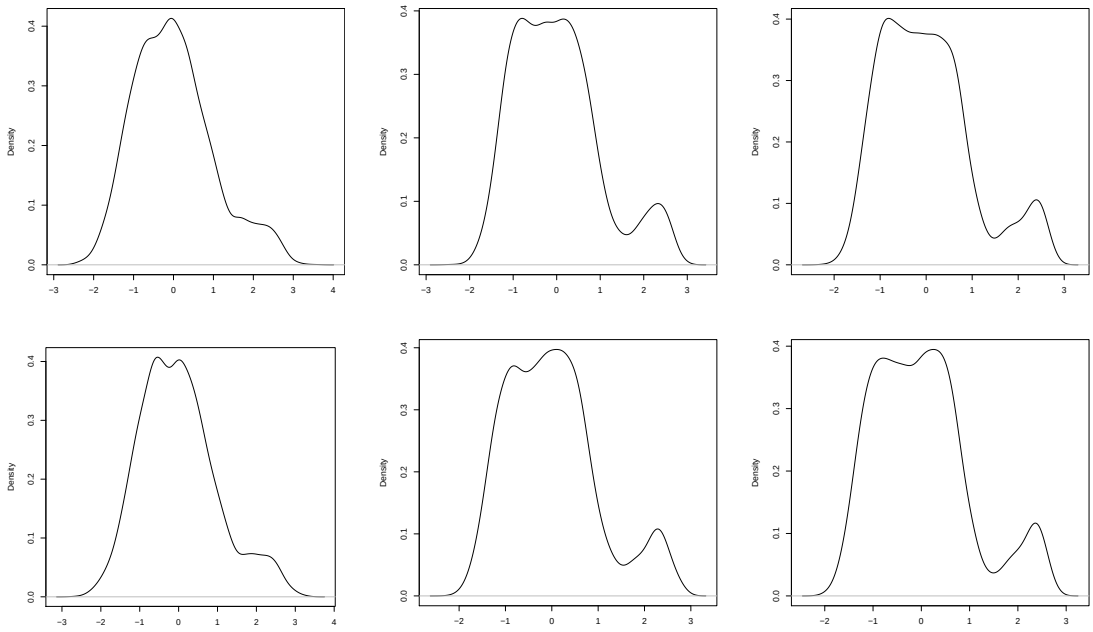

Figure S2: Prelic data sets with noise variance 1, 0.4, 0.2 per row. The skewness is on average 0.64 and the excess kurtosis is on average 0.15.

S6.3.2 Qubic Data

Li *et al.* (2009) have modified the data sets of Prelic *et al.* (2006) to make them more realistic. They do not only contain constant biclusters, but also scaling patterns. This accounts for the fact that there are also down-regulated genes in gene expression. However, the benchmark data set of Li *et al.* (2009) still contains most of the characteristics of the data sets of Prelic *et al.* (2006). Especially the moments do not agree to the moments observed in gene expression data sets.

The data in (Li *et al.*, 2009, Supplementary material) were generated as follows:

> *"In scenario 1, we generated two datasets. We first implant ten non-overlapping biclusters of scaling patterns of size 10(genes)x10(conditions) into a background matrix of size 100x100*

with its σ ranging from 0 to 0.25, and then implant ten non-overlapping biclusters of scaling patterns of size 5(genes)x10(conditions) into the background matrix of size 50x50 with its σ ranging from 0 to 0.25. In scenario 2, the background variation parameter σ was 0, and we set all entries of the first (last) two rows to 1 (-1) so that we can simulate the situation where some transcription factors can be in more than one transcription modules, i.e., all the implanted biclusters shared the first two and the last two genes. We then implant ten biclusters of scaling patterns with size (10+d)x(10+d) into a (100+d)x(100+d) matrix at the interval of 10 genes and 10 conditions, forcing the biclusters to overlap with each other at different levels."

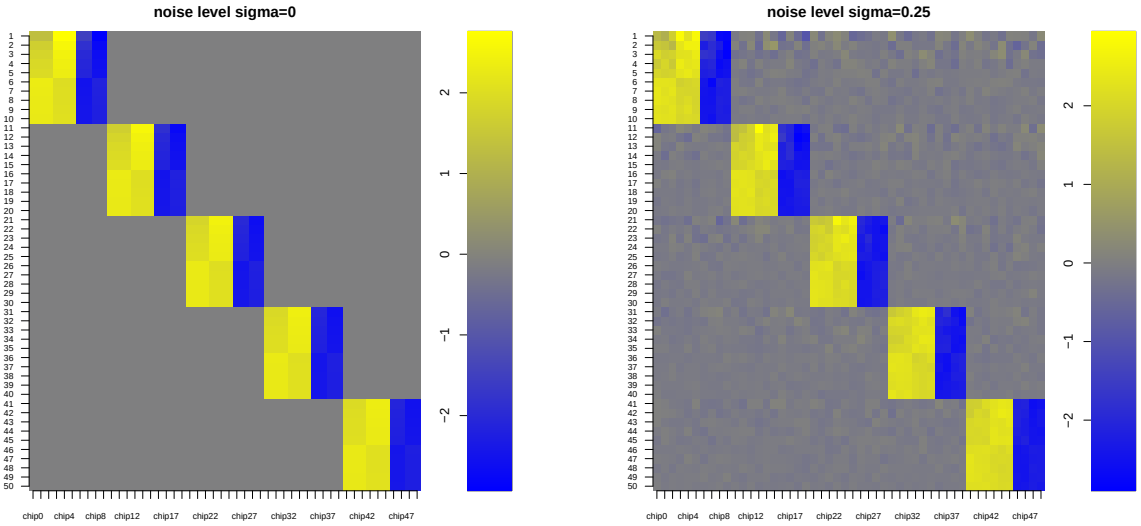

Figure S3: Examples of Qubic data sets. A noise-free one (left) and a noisy one (right).

For larger data sets, only computation times are reported in (Li *et al.*, 2009), but no results about biclustering performance are made available.

Figure S4 shows the result of FABIA on the most complex data set from Li *et al.* (2009). FABIA shows better performance than any other method except Qubic with non-standard parameter settings.

S6.3.3 Our Simulated Data Sets

Figure S5 shows the first three of our simulated data sets, with and without noise added. Figure S6 shows another three data sets created according to the same data generation procedure as all other simulated data sets, but this time with biclusters arranged as blocks. Figure S7 shows distributions of three of our simulated data sets.

For the simulated data set, we computed the information content of the biclusters. Additionally we computed the similarity of the biclusters to the true biclusters assigned by the Munkres algorithm.

In order to validate our method for extracting the information content of biclusters, we sorted biclusters according to the similarity to true biclusters and computed the information content for

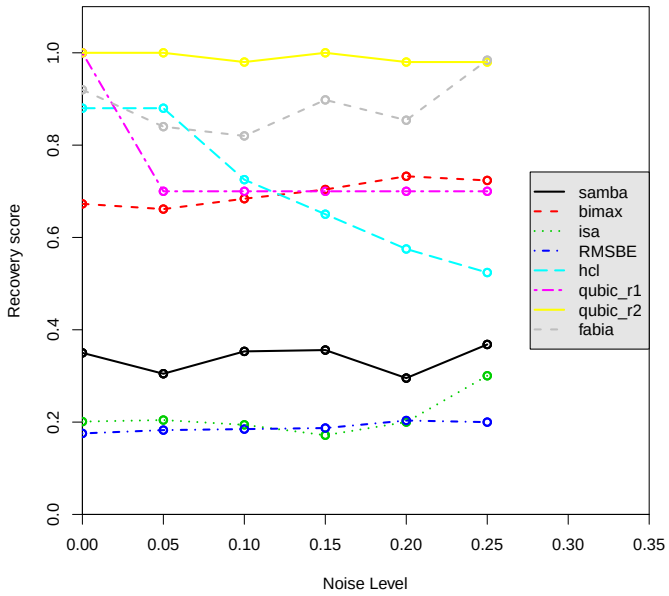

Figure S4: Results on the most complex data set from Li *et al.* (2009) using the quality measure in Li *et al.* (2009).

every bicluster. The left-hand side of Table S2 shows that the biclusters with highest similarity to true biclusters have highest information content. Thus, the information content is useful for selecting and ranking the biclusters. The right-hand side of Table S2 shows the average similarity rank after the biclusters are sorted according to the information content. Biclusters with high information content also have high similarity to true biclusters. Finally, we applied a two-sided Spearman rank correlation test to evaluate the to which extent information content and similarity to true biclusters are monotonically correlated. The resulting p-values are presented in the main paper.

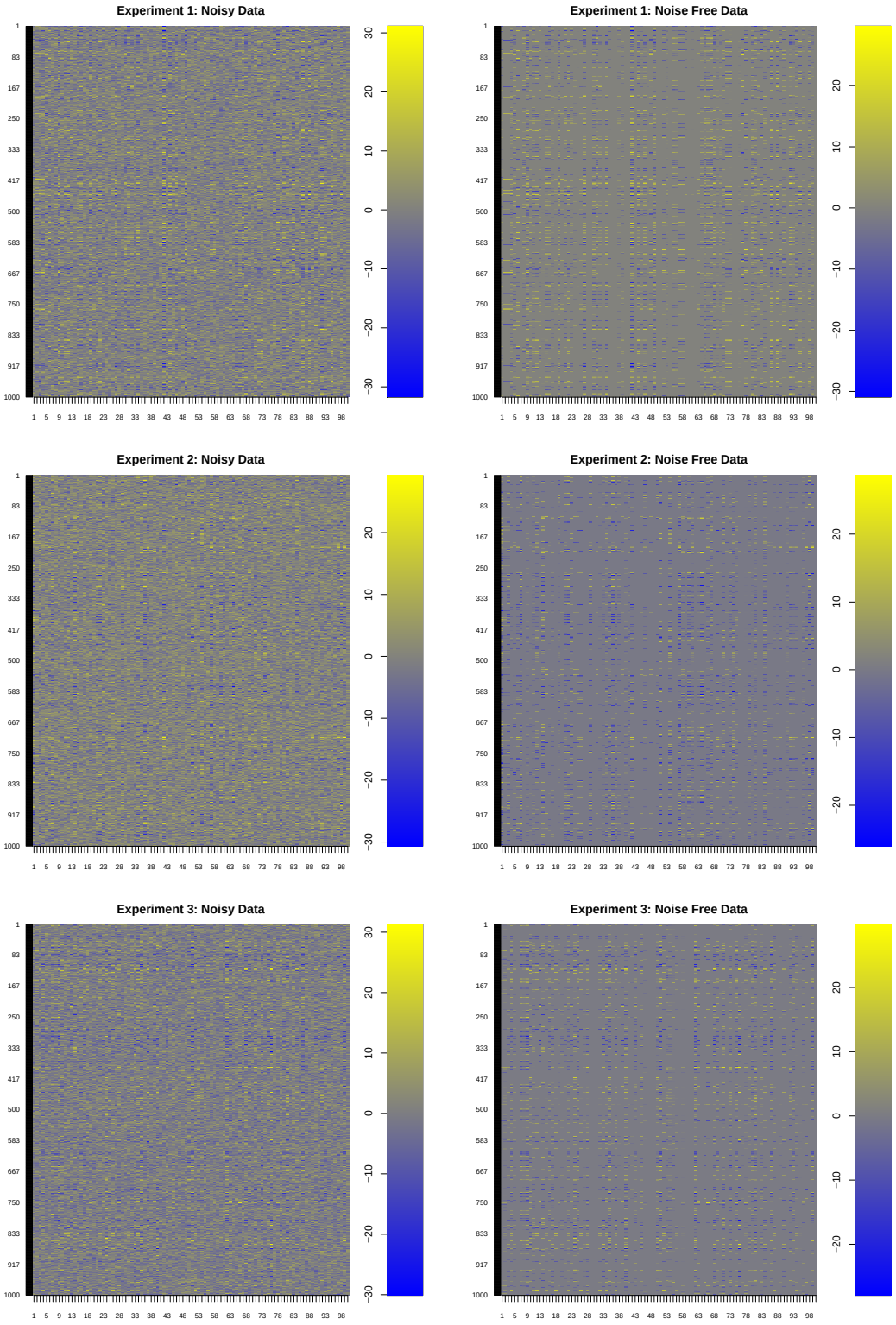

Figure S5: The first three data sets of our experiments. Left: noisy data. Right: noise free data.

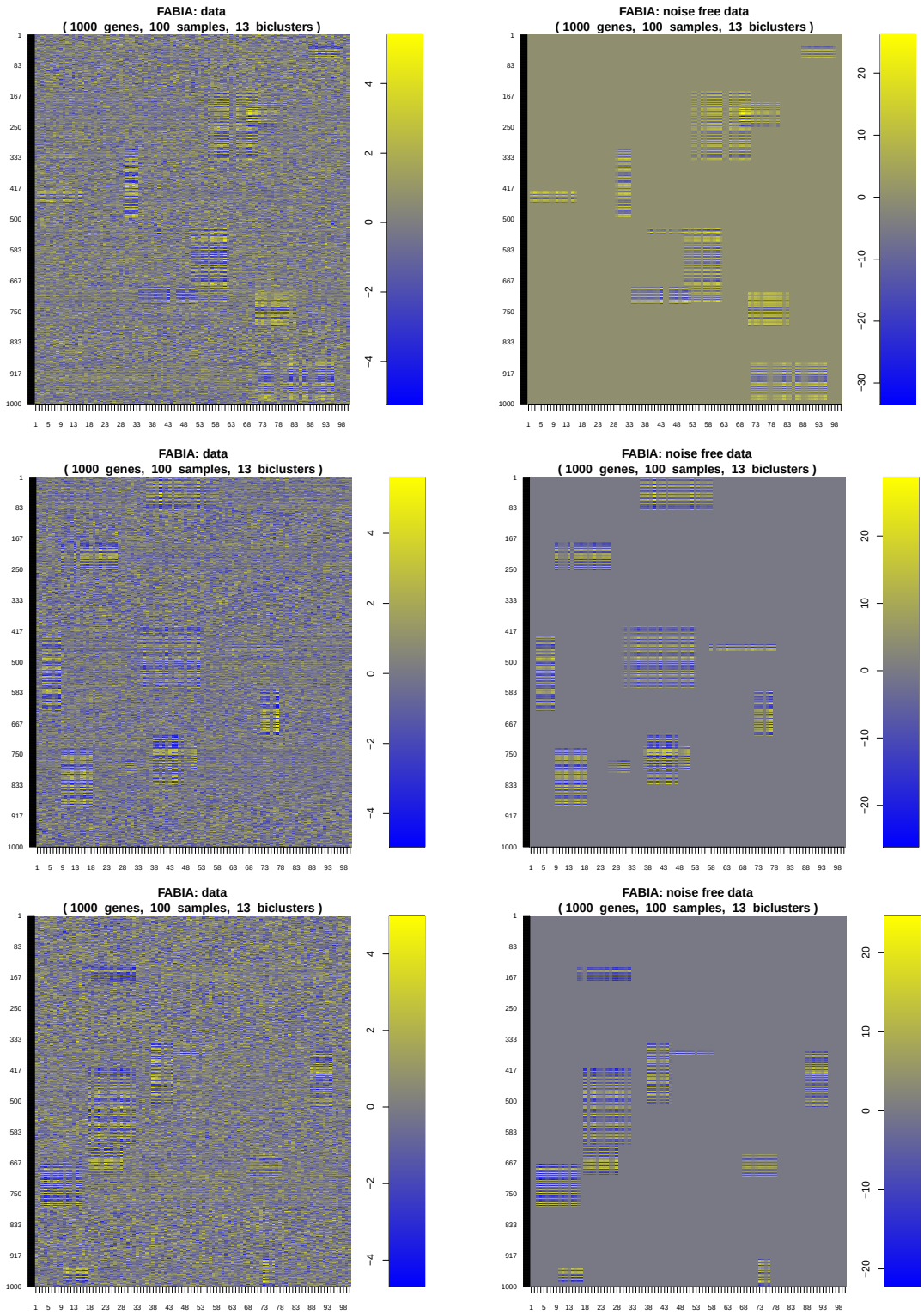

Figure S6: Three data sets with the same parameters (noise, size, overlap, etc.) as in our experiments, but with biclusters arranged in blocks for better visualization.

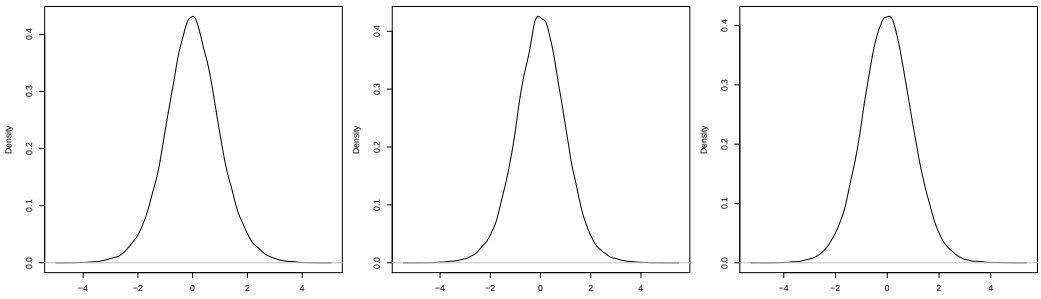

Figure S7: Our experimental data sets. The skewness is on average 0.0 and the excess kurtosis is on average 0.65.

Table S2: Left: average of information content sorted according to the bicluster similarity (Jaccard index). Biclusters with highest similarity to true biclusters have highest information content which is therefore useful for selecting and ranking the biclusters. Right: The average similarity rank after the biclusters are sorted according to the information content. In both tables, the values in parentheses are the standard deviations.

	FABIA	FABIAS
1	330 (4.8)	373 (9.2)
2	324 (6.4)	366 (2.5)
3	318 (5.5)	365 (4.3)
4	287 (2.6)	351 (11)
5	263 (0.5)	331 (4.3)
6	232 (21)	313 (1.9)
7	199 (4.2)	302 (0.5)
8	154 (9.6)	242 (7.7)
9	96 (12)	154 (1.9)
10	17 (20)	57 (23)
11	0 (0)	5 (28)
12	0 (0)	0 (0)
13	0 (0)	0 (0)

	FABIA	FABIAS
1	2.9 (0.3)	3.0 (0.1)
2	3.5 (0.2)	3.6 (0.1)
3	3.6 (0.3)	4.0 (0.2)
4	4.3 (0.1)	4.5 (0.1)
5	4.7 (0.2)	4.7 (0.3)
6	4.9 (0.1)	4.9 (0.2)
7	5.6 (0.1)	5.9 (0.01)
8	6.1 (0.1)	6.7 (0.6)
9	7.3 (0.03)	8.0 (0.1)
10	8.2 (0.1)	8.9 (0.1)
11	8.4 (0.1)	9.6 (0.04)
12	8.4 (0.1)	9.7 (0.03)
13	8.4 (0.1)	9.7 (0.03)

S6.3.4 Data with Additive Biclusters

We also generated data according to an additive model structure in order to analyze how well FABIA and FABIAS perform on data that do not satisfy the multiplicative model assumptions. We generated the biclusters as in previous Subsection S6.3.3:

Genes: Randomly choose the number N_i^{λ} of genes in bicluster i from $\{10, \ldots, 210\}$, choosing N_i^{λ} genes randomly from $\{1, \ldots, 1000\}$.

Samples: Randomly choose the number $N_i^{\boldsymbol{z}}$ of samples in bicluster i from $\{5, \ldots, 25\}$, choosing $N_i^{\boldsymbol{z}}$ samples randomly from $\{1, \ldots, 100\}$.

Noise: Finally, we draw the $\boldsymbol{\Upsilon}$ entries (additive noise on all entries) according to $\mathcal{N}(0, 3^2)$. Using this procedure, noisy biclusters of random size between 10×5 and 210 (genes$\times$samples) are generated.

In contrast to the previous experiments, we now use a general additive model for each bicluster

$$\theta_{kij} \;=\; \mu_i \;+\; \alpha_{ki} \;+\; \beta_{ij},$$

where i is the index of the bicluster under consideration, k is the k-th row and j is the j-th column belonging to the i-th bicluster.

We used three different models that differ in their signal-to-noise ratios. We realized this by choosing μ_i from three different ranges:

M1 (low signal): For each bicluster i, μ_i is chosen from $\mathcal{N}(0, 2^2)$.

M2 (moderate signal): For each bicluster i, μ_i is chosen from $\mathcal{N}(\pm 2, 0.5^2)$, and the sign is randomly chosen.

M3 (high signal): For each bicluster i, μ_i is chosen from $\mathcal{N}(\pm 4, 0.5^2)$, and the sign is randomly chosen.

The values α_{ki} are chosen from $\mathcal{N}(0.5, 0.2^2)$ and β_{ij} are chosen from $\mathcal{N}(1, 0.5^2)$. Apart from that, we use the same experimental setting as with the multiplicative data for all methods.

The Tables S3, S4, and S5 present the results for low, moderate, and high signal, respectively. They show the average consensus scores with the true biclusters as defined in the main paper (standard deviation in parentheses). In all experiments, FABIAS gives the best results followed by FABIA. The next best methods are either plaid_ms_5 or ISA_2.

We explain the superiority of FABIA and FABIAS on data sets where the model does not match the data generation model as follows:

1. Biclusters are constructed simultaneously, thereby, overlaps are taken into account at all time. Therefore, large values need not be explained by separate biclusters, but can be explained as an overlap of biclusters.

2. The decorrelation of factors avoids redundant biclusters. Note, that the biclusters can still overlap.

3. The simplicity of the model ensures low parameter interdependencies, which facilitates model selection. FABIA and FABIAS are quadratic in their parameters. Plaid models are cubic in their parameters, where the indicator variables are multiplied to the general additive model that is defined on the whole data matrix. Parameter interdependencies lead to large entries outside the main diagonal of the Fisher information matrix and may lead to small eigenvalues. In turn, these small eigenvalues lead to a high Cramer-Rao bound on the variance of the estimator.

Table S3: Results on the 100 simulated **additive** data sets with model **M1 (low signal)**. The numbers denote average consensus scores with the true biclusters as defined in the main paper (standard deviations in parentheses) The best results are printed bold and the second best in italics. FABIAS has the highest score followed by FABIA and plaid_ms_5.

method	score	method	score
FABIA	*0.109* (6e-2)	SAMBA	0.002 (6e-4)
FABIAS	**0.150** (7e-2)	xMOTIF	0.002 (4e-4)
MFSC	0.000 (0)	Bimax	0.009 (8e-3)
plaid_ss	0.039 (2e-2)	CC	4e-4 (3e-4)
plaid_ms	0.064 (3e-2)	plaid_t_ab	0.021 (2e-2)
plaid_ms_5	0.098 (4e-2)	plaid_t_a	0.039 (3e-2)
ISA_1	0.039 (4e-2)	FLOC	0.005 (9e-4)
ISA_2	0.081 (5e-2)	spec_1	0.000 (0)
ISA_3	0.040 (4e-2)	spec_2	0.000 (0)
OPSM	0.007 (2e-3)		

Table S4: Results on the 100 simulated **additive** data sets with model **M2 (moderate signal)**. The numbers denote average consensus scores with the true biclusters as defined in the main paper (standard deviations in parentheses) The best results are printed bold and the second best in italics. FABIAS has the highest score followed by FABIA and then plaid_ms_5 as well as ISA_2.

method	score	method	score
FABIA	*0.196* (8e-2)	SAMBA	0.002 (5e-4)
FABIAS	**0.268** (7e-2)	xMOTIF	0.002 (4e-4)
MFSC	0.000 (0)	Bimax	0.010 (9e-3)
plaid_ss	0.041 (1e-2)	CC	3e-4 (2e-4)
plaid_ms	0.072 (2e-2)	plaid_t_ab	0.005 (6e-3)
plaid_ms_5	0.143 (4e-2)	plaid_t_a	0.010 (9e-3)
ISA_1	0.033 (2e-2)	FLOC	0.005 (1e-3)
ISA_2	0.143 (4e-2)	spec_1	0.000 (0)
ISA_3	0.037 (2e-2)	spec_2	0.000 (0)
OPSM	0.007 (2e-3)		

Table S5: Results on the 100 simulated **additive** data sets with model **M3 (high signal)**. The numbers denote average consensus scores with the true biclusters as defined in the main paper (standard deviations in parentheses) The best results are printed bold and the second best in italics. FABIAS has the highest score followed by FABIA and ISA_2.

method	score	method	score
FABIA	*0.475* (1e-1)	SAMBA	0.003 (8e-4)
FABIAS	**0.546** (1e-1)	xMOTIF	0.001 (4e-4)
MFSC	0.000 (0)	Bimax	0.014 (1e-2)
plaid_ss	0.074 (3e-2)	CC	1e-4 (1e-4)
plaid_ms	0.112 (3e-2)	plaid_t_ab	0.022 (2e-2)
plaid_ms_5	0.221 (5e-2)	plaid_t_a	0.051 (4e-2)
ISA_1	0.140 (7e-2)	FLOC	0.003 (9e-4)
ISA_2	0.229 (5e-2)	spec_1	0.000 (0)
ISA_3	0.139 (7e-2)	spec_2	0.000 (0)
OPSM	0.008 (2e-3)		

S6.4 Gene Expression Data Sets

S6.4.1 Statistics of the Expression Data Sets

Figures S8 and S9 show plots of the data distributions of the three gene expression data sets discussed in Section 6.4 of the main paper.

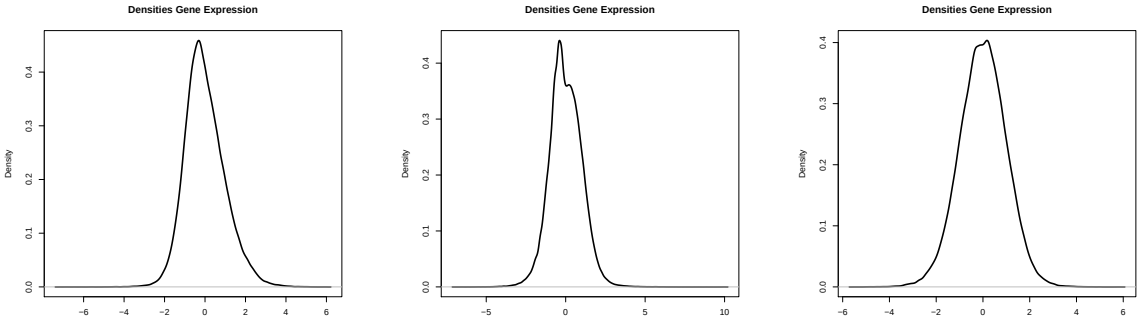

Figure S8: Data densities of the gene expression data sets. Left: breast cancer data set, skewness is 0.45 and excess kurtosis is 0.93; Middle: multiple tissues data set, skewness is 0.15 and excess kurtosis is 1.3; Right: DLBCL data set, skewness is -0.05 and excess kurtosis is 0.35.

S6.4.2 Biological Interpretation

One of the goals of biclustering is to extract biological knowledge from gene expression data sets. In this section, we want to look into the biclusters generated by FABIA.

We performed gene set enrichment analysis and created protein interaction networks. To this end, we applied the following methods which have different levels of specificity:

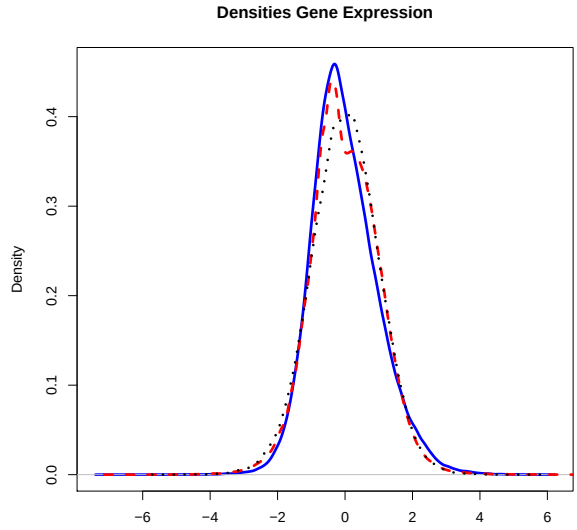

Figure S9: Density of gene expression data sets plotted together. Breast cancer: solid blue; multiple tissue types: dashed red; DLBCL: dashed black.

Low specificity, on biological process level: we performed a *GO (gene ontology) analysis based on the biological process (BP)*. We compute the p-values and plot the hierarchy of GO classes.

Higher specificity, on pathway level: we performed a *KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis* (Kanehisa *et al.*, 2010; `http://www.genome.jp/kegg/`) to see whether genes interact in a pathway. KEGG has 24,329 genes in its database, but not all Affymetrix gene identifiers could be matched to KEGG (though the large majority matches).

High specificity, on protein interaction level: we applied the "STRING 8" software (Jensen *et al.*, 2009; `http://string-db.org/`) to the genes in the biclusters and generated protein interaction networks. This software displays a protein network for each cluster along with its interconnections. Connections are labeled by colors in the following way:

Green: genes occur in close neighborhood on the chromosome;

Blue: same protein-protein interactions found across species;

Red: gene fusion has been observed for the two genes;

Purple: experimentally verified protein-protein interaction;

Black: genes that are co-expressed in the same or in other species (transferred by homology);

Light blue: other significant protein interaction from curated databases;

Light green: relationship identified by text mining in abstracts of scientific literature.

Additionally, for the breast cancer data set, we compared our results to another data set (Finak *et al.*, 2008), where gene signatures for "bad outcome", "mixed outcome", and "good outcome" are provided.

Breast cancer data set

For the breast cancer data set, FABIA found three biclusters, the sizes of which are given as follows:

Bicluster	1	2	3
Genes	98	127	50
Samples	29	38	27

Bicluster 1:

- Table S6 and Figure S10 provide the GO analysis results. The bicluster has p-values that range from $2.5 \cdot 10^{-15}$ to $1.2 \cdot 10^{-8}$ for terms concerning the M phase of the cell cycle and mitosis.
- The bicluster has a KEGG overlap of 93 genes. The KEGG result can be found in Table S7, where the p-values range from $4.7 \cdot 10^{-10}$ to $4.1 \cdot 10^{-4}$.
- The protein network graph for the first bicluster is displayed in Figure S11. It shows a cluster with the central protein CDC2 (cell division control) which interacts with other CDCs. This inner cluster is connected to ZBTB16 (a transcription factor) which, in turn, is connected to RUNX1T which binds to histon deacetylases and transcription factors. Another cluster is the KIF-related cluster that interacts with mitosis.

The most significant pathways are related to the cell cycle, where both the GO and the KEGG analysis are in agreement. Proteins which drive this cluster are the cell division control protein CDC2 and the mitosis related KIF proteins.

Table S6: GO analysis for biological process (BP) on FABIA bicluster 1 obtained for the breast cancer data set.

CL	GO-BP-ID	p-value	Odds ratio	Count	Size	Term
1	GO:0000279	2.5e-15	20	21	37	M phase
2	GO:0022403	5.3e-13	11	23	55	cell cycle phase
3	GO:0022402	2.6e-11	8	24	70	cell cycle process
4	GO:0051301	3.1e-11	19	15	26	cell division
5	GO:0000087	3.1e-11	16	16	30	M phase of mitotic cell cycle
6	GO:0000280	3.1e-11	16	16	30	nuclear division
7	GO:0007067	3.1e-11	16	16	30	mitosis
8	GO:0048285	3.1e-11	16	16	30	organelle fission
9	GO:0007049	2.8e-09	6	26	99	cell cycle
10	GO:0000278	1.2e-08	7	19	58	mitotic cell cycle

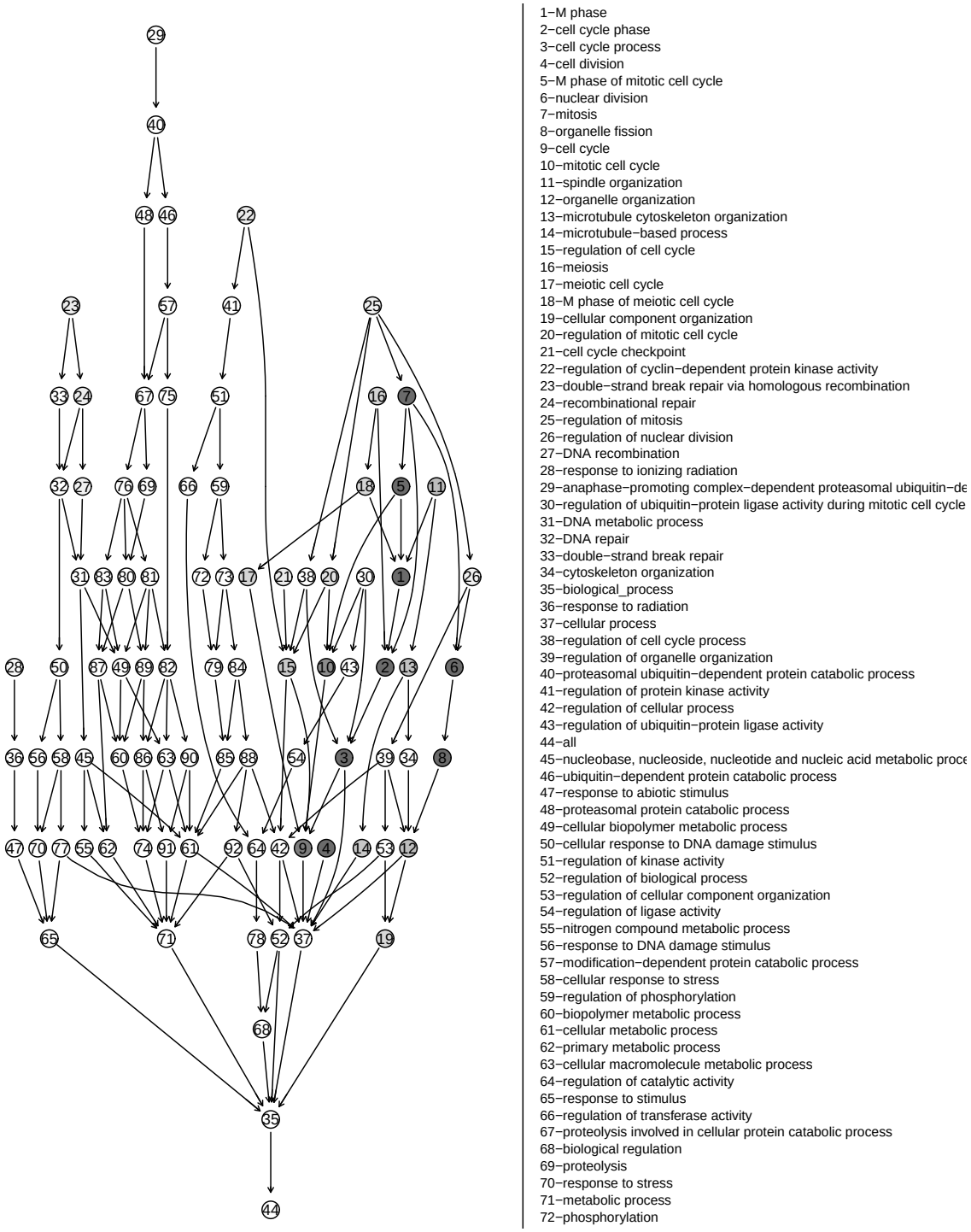

Figure S10: GO analysis for biological process (BP) on FABIA bicluster 1 obtained for the breast cancer data set. The GO hierarchy is shown. The darker the circles, the higher is the significance (the lower the p-value).

Table S7: KEGG analysis of FABIA bicluster 1 obtained for the breast cancer data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:04110	4.7e-10	9.7e-08	9	112	Cell cycle
2	path:04610	2.6e-07	2.7e-05	6	69	Complement and coagulation cascades
3	path:00641	2.4e-05	1.6e-03	3	15	3-Chloroacrylic acid degradation
4	path:00010	1.0e-04	6.9e-03	4	63	Glycolysis / Gluconeogenesis
5	path:00624	1.2e-04	8.0e-03	3	25	1- and 2-Methylnaphthalene degradation
6	path:04640	3.7e-04	2.5e-02	4	88	Hematopoietic cell lineage
7	path:00120	4.1e-04	2.8e-02	3	38	Bile acid biosynthesis

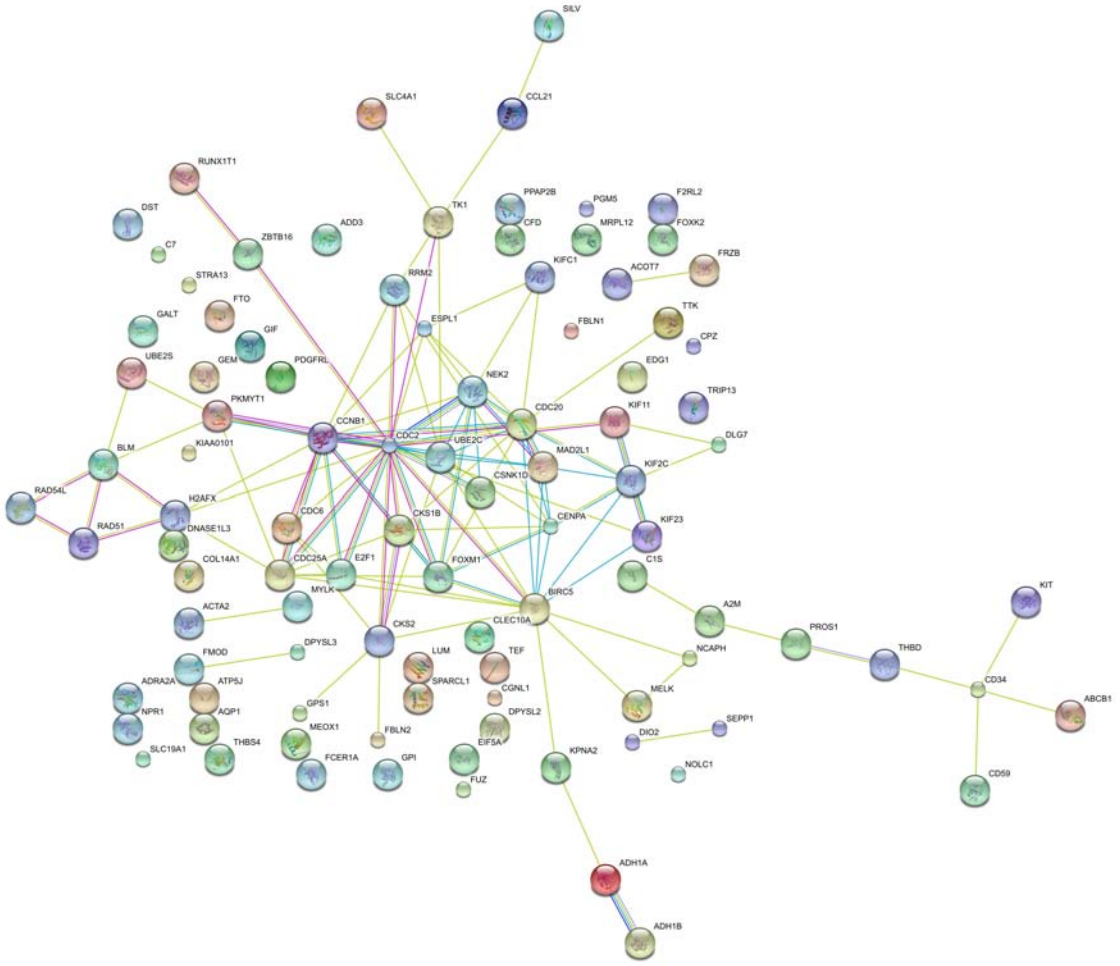

Figure S11: STRING protein network derived from genes contained in FABIA bicluster 1 of the breast cancer data set. Connections are labeled as described on p. 29.

Bicluster 2:

- Table S8 and Figure S12 present the GO analysis results. The bicluster has p-values between $1.4 \cdot 10^{-26}$ and $2.3 \cdot 10^{-5}$ for terms concerning the immune response and chemotaxis.
- The bicluster has a KEGG overlap of 118 genes. The results are shown in Table S9, where the p-values range from 0 (too small to be displayed by the program) to $2.6 \cdot 10^{-4}$.
- Figure S13 visualizes the protein interaction network for the second bicluster. It contains proteins like CCR5 (C-C chemokine receptor type 5), STAT1 (Signal transducer and activator of transcription 1-alpha/beta), CCL4 (Small inducible cytokine A4 precursor), CSF2RB (Cytokine receptor common beta chain precursor), IL2RB (Interleukin-2 receptor subunit beta precursor), CD86 (T-lymphocyte activation antigen CD86 precursor), which belong to a family of signaling and regulative proteins related to the immune system. Cytokine and interleukin are the main regulatory signal carriers in this protein network.

Again we have a perfect agreement between the GO and the KEGG analysis. The most significant pathways are related to cytokine-cytokine receptor interaction, but also other immune system responses are relevant. Note that cytokines are important regulators and mobilizers of the immune response.

Table S8: GO analysis for biological process (BP) on FABIA bicluster 2 obtained for the breast cancer data set.

CL	GO-BP-ID	p-value	Odds ratio	Count	Size	Term
1	GO:0006955	1.4e-26	13	51	113	immune response
2	GO:0002376	1.8e-23	9	58	167	immune system process
3	GO:0050896	1.3e-13	5	70	350	response to stimulus
4	GO:0006952	3.2e-09	5	30	102	defense response
5	GO:0009615	1.6e-06	10	11	22	response to virus
6	GO:0006935	1.7e-06	6	15	40	chemotaxis
7	GO:0042330	1.7e-06	6	15	40	taxis
8	GO:0051707	1.3e-05	5	15	46	response to other organism
9	GO:0006954	1.6e-05	4	19	70	inflammatory response
10	GO:0007626	2.3e-05	5	15	48	locomotory behavior

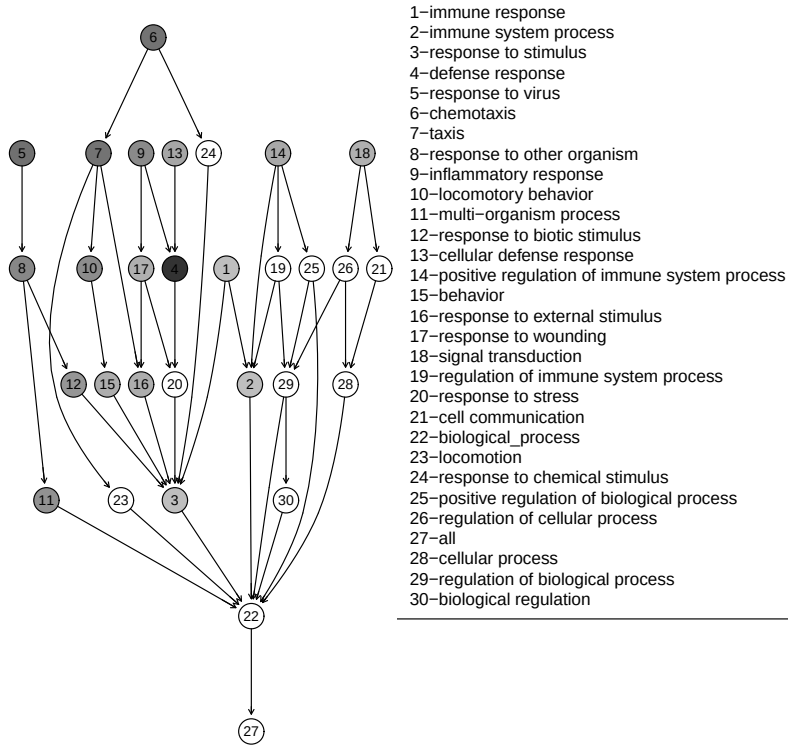

Figure S12: GO analysis for biological process (BP) on FABIA bicluster 2 obtained for the breast cancer data set. The GO hierarchy is shown, where darker circles indicate higher significance (lower p-values).

Table S9: KEGG analysis of FABIA bicluster 2 obtained for the breast cancer data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:04060	0.0e+00	0.0e+00	20	259	Cytokine-cytokine receptor interaction
2	path:04620	3.6e-08	3.1e-06	8	102	Toll-like receptor signaling pathway
3	path:05332	5.2e-08	3.6e-06	6	42	Graft-versus-host disease
4	path:04640	2.4e-07	1.7e-05	7	88	Hematopoietic cell lineage
5	path:04612	2.6e-07	1.8e-05	7	89	Antigen processing and presentation
6	path:05330	1.1e-06	7.5e-05	5	38	Allograft rejection
7	path:04940	2.3e-06	1.6e-04	5	44	Type I diabetes mellitus
8	path:04650	3.8e-06	2.6e-04	7	132	Natural killer cell mediated cytotoxicity
9	path:04514	4.0e-06	2.7e-04	7	133	Cell adhesion molecules (CAMs)
10	path:05320	5.9e-06	4.1e-04	5	53	Autoimmune thyroid disease
11	path:04630	1.0e-05	6.9e-04	7	153	Jak-STAT signaling pathway
12	path:04670	2.6e-04	1.7e-02	5	116	Leukocyte transendothelial migration

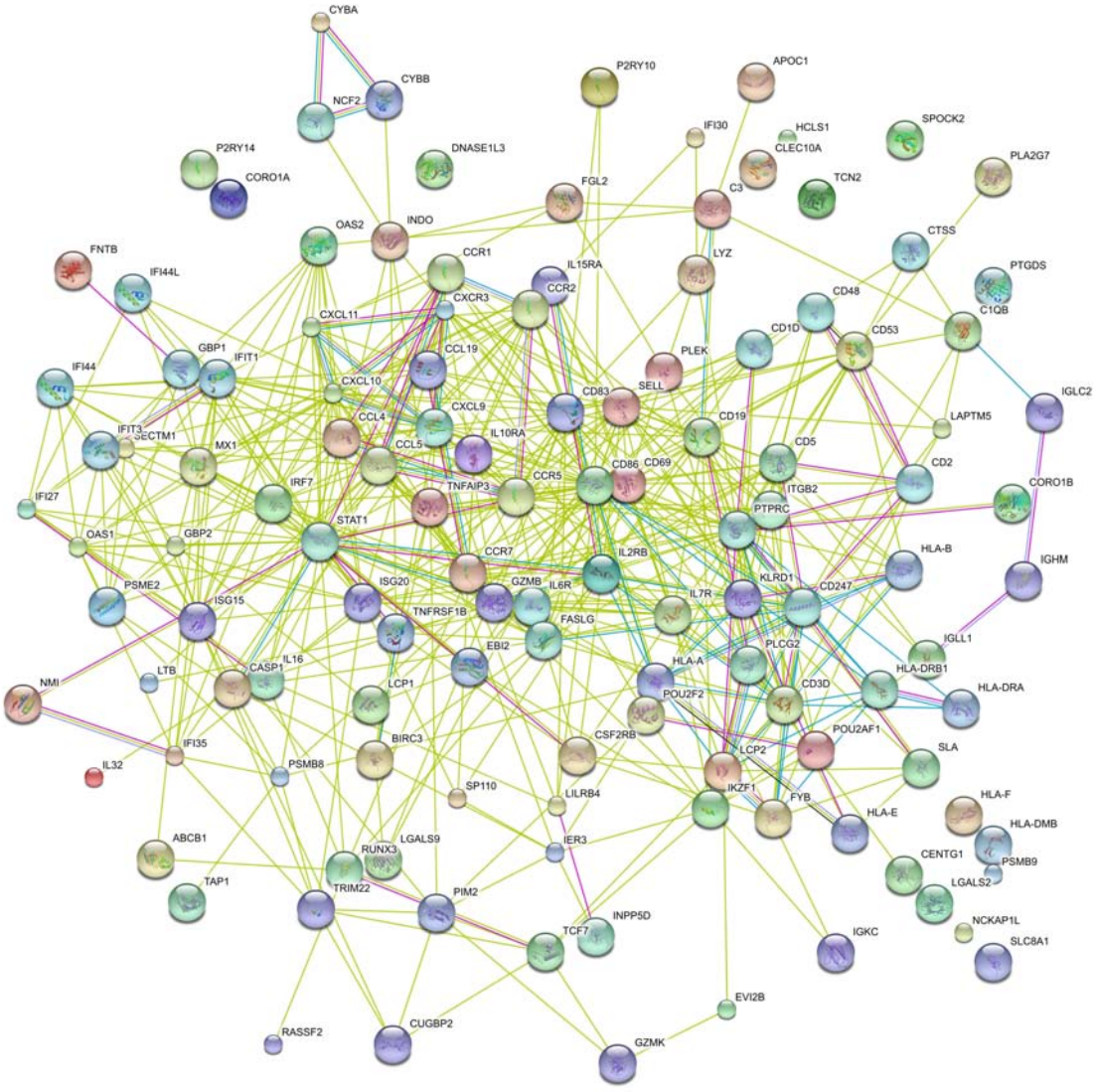

Figure S13: STRING protein network derived from genes found by FABIA in cluster 2 of the breast cancer data set. Connections are labeled as described on p. 29.

Bicluster 3:

- Table S10 shows the GO analysis results. The bicluster has p-values between $1.8 \cdot 10^{-3}$ and $5.3 \cdot 10^{-3}$. However, these values are obtained by hits on small clusters and we think they should not be interpreted.
- The bicluster has a KEGG overlap of 51 genes and is the smallest cluster. The results are shown in Table S11, where the only p-value is $4.5 \cdot 10^{-10}$.
- The protein network graph for the third bicluster is shown in Figure S14.

This bicluster is too small to allow for reliable biological interpretations.

Table S10: GO analysis for biological process (BP) on FABIA bicluster 3 obtained for the breast cancer data set.

CL	GO-BP-ID	p-value	Odds ratio	Count	Size	Term
1	GO:0015914	1.8e-03	∞	2	2	phospholipid transport
2	GO:0033700	1.8e-03	∞	2	2	phospholipid efflux
3	GO:0009612	2.3e-03	18	3	7	response to mechanical stimulus
4	GO:0009628	4.4e-03	5	6	38	response to abiotic stimulus
5	GO:0007605	5.2e-03	12	3	9	sensory perception of sound
6	GO:0050954	5.2e-03	12	3	9	sensory perception of mechanical stimulus
7	GO:0051606	5.2e-03	12	3	9	detection of stimulus
8	GO:0010872	5.3e-03	46	2	3	regulation of cholesterol esterification
9	GO:0015810	5.3e-03	46	2	3	aspartate transport
10	GO:0034375	5.3e-03	46	2	3	high-density lipoprotein particle remodeling

Table S11: KEGG analysis of FABIA bicluster 3 obtained for the breast cancer data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:01430	4.5e-10	9.5e-08	8	138	Cell junctions

Finally, we compared the FABIA results with the results in (Finak *et al.*, 2008) where gene signatures for "bad outcome", "mixed outcome", and "good outcome" are provided. The overlap of genes in the data set of Finak *et al.* (2008) and our breast cancer data set is 1111 genes. From the 22 gene signature for "good outcome" of Finak *et al.* (2008), only 12 genes overlap with our data set. Ten out of these 12 genes can be found in bicluster 2 of the FABIA result. A Fisher test led to a p-value of $5.4 \cdot 10^{-9}$. From the 24 gene signature for "mixed outcome" of Finak *et al.* (2008), only 8 genes overlap with our data set. Three out of these 8 genes can be found in bicluster 1 of the FABIA result. A Fisher test led to a p-value of 0.02. This shows that other studies have identified groups of genes similar to the ones identified by FABIA.

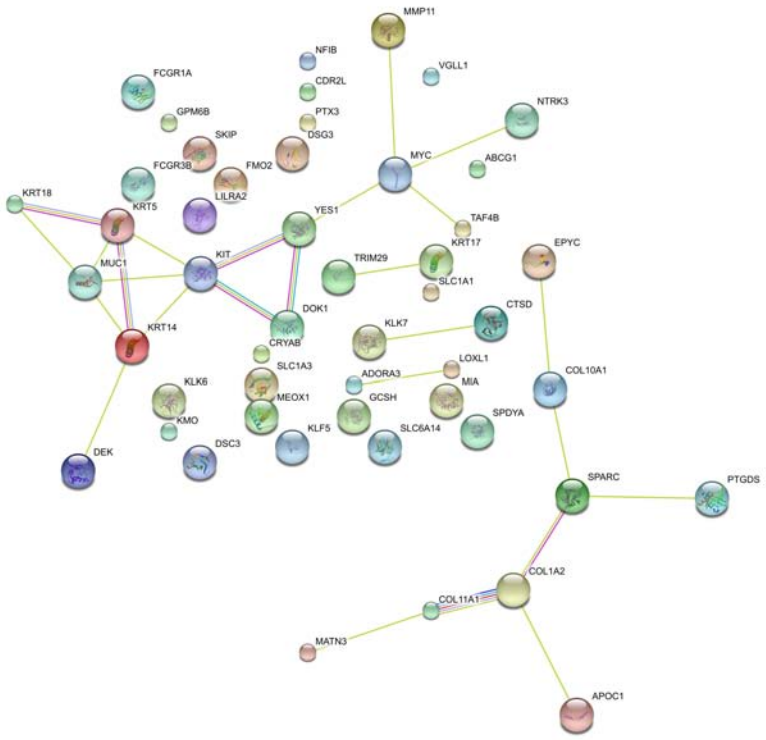

Figure S14: STRING protein network derived from genes found by FABIA in cluster 3 of the breast cancer data set. The connectivity is here much smaller than at the other clusters which indicates that no key pathway has been found. Connections are labeled as described on p. 29.

DLBCL data set

For the DLBCL data set, FABIA found two biclusters, the sizes of which are given as follows:

Bicluster	1	2
Genes	48	70
Samples	63	61

Bicluster 1:

- The GO analysis results are shown in Table S12 and in Figure S15. The bicluster has p-values between $2.2 \cdot 10^{-6}$ and $3.3 \cdot 10^{-2}$. The GO terms point to transcriptomic effects.
- The bicluster has a KEGG overlap of 43 genes. The KEGG result can be found in Table S13, where the p-values range from $1.3 \cdot 10^{-8}$ to $3.4 \cdot 10^{-4}$.
- The protein network graph for bicluster 1 is shown in Figure S16. A strong cluster can be seen that is formed by ribosomal proteins RBS and RBL. The second strong cluster is governed by BLNK (B-cell linker protein), LYN (Tyrosine-protein kinase Lyn), PRKCB1

(Protein kinase C beta type). The latter is considered as a novel component of the NF-kappa-B signaling axis responsible for the survival and activation of B-cells after BCR cross-linking Thus, this cluster is related to B-cell/Tyrosine-kinase. Again this would make sense with respect to the origin of the samples.

Also here the GO and KEGG analysis agree on translational processes — more specifically on ribosome-related effects. However, KEGG found additional pathways that are related to cell receptor signaling, which corresponds well to the origin of this data set — diffuse large-B-cell lymphoma.

Table S12: GO analysis for biological process (BP) on FABIA bicluster 1 obtained for the DLBCL data set.

CL	GO-BP-ID	p-value	Odds ratio	Count	Size	Term
1	GO:0006414	2.2e-06	17	8	15	translational elongation
2	GO:0006412	2.9e-04	7	8	26	translation
3	GO:0044260	8.2e-03	2	27	257	cellular macromolecule metabolic process
4	GO:0034960	1.3e-02	2	26	250	cellular biopolymer metabolic process
5	GO:0043283	1.6e-02	2	27	267	biopolymer metabolic process
6	GO:0010467	1.8e-02	2	17	142	gene expression
7	GO:0044237	1.8e-02	2	30	312	cellular metabolic process
8	GO:0043170	2.5e-02	2	27	275	macromolecule metabolic process
9	GO:0006270	3.0e-02	13	2	4	DNA replication initiation
10	GO:0034645	3.3e-02	2	16	139	cellular macromolecule biosynthetic process

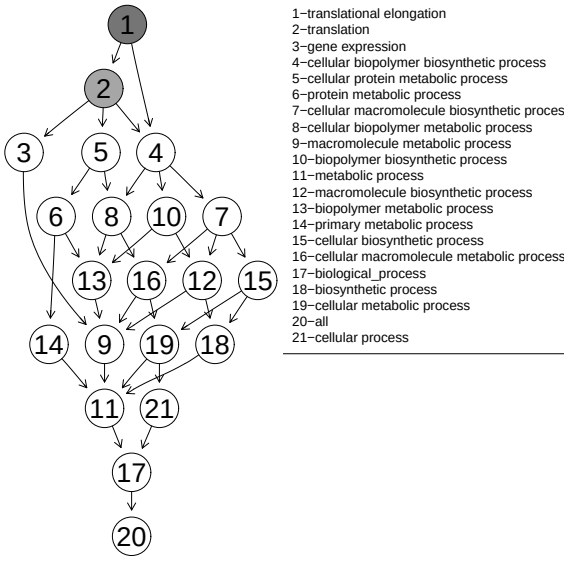

Figure S15: GO analysis for biological process (BP) on FABIA bicluster 1 obtained for the DLBCL data set. The GO hierarchy is shown, where darker circles indicate higher significance (lower p-values).

Table S13: KEGG analysis of FABIA bicluster 1 obtained for the DLBCL data set.

Name	p-value	q-value	Count	Size	Term
path:03010	1.3e-08	2.8e-06	6	92	Ribosome
path:04662	9.6e-08	9.9e-06	5	64	B cell receptor signaling pathway
path:04110	4.6e-05	3.2e-03	4	112	Cell cycle
path:04520	3.2e-04	2.2e-02	3	75	Adherens junction
path:04664	3.4e-04	2.3e-02	3	77	Fc epsilon RI signaling pathway

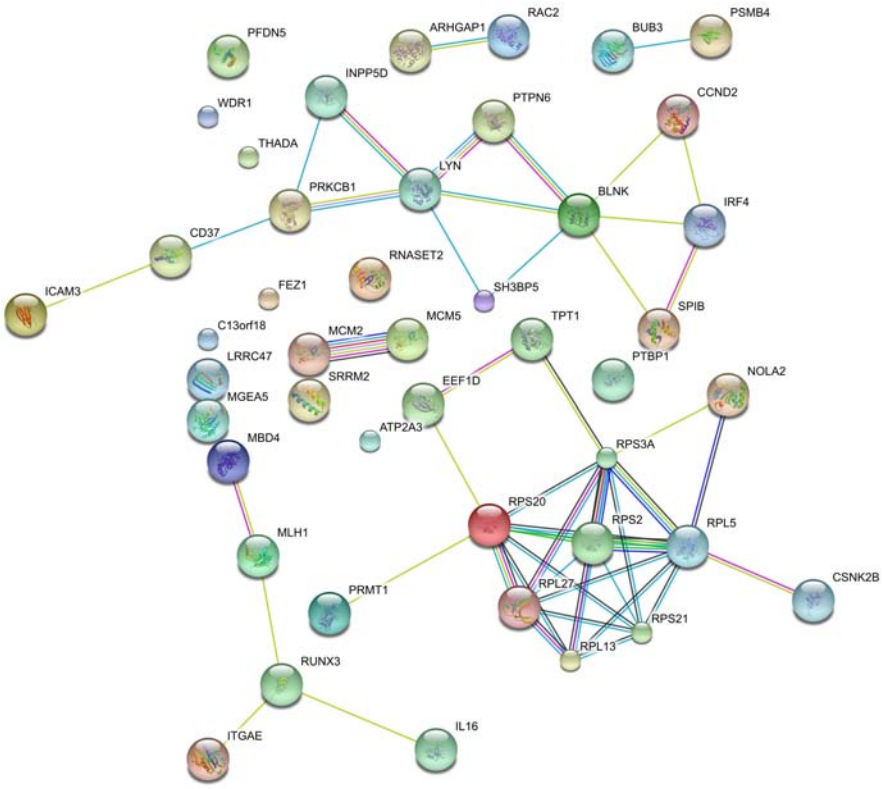

Figure S16: STRING protein network derived from genes found by FABIA in bicluster 1 of the DLBCL data set. Connections are labeled as described on p. 29.

Bicluster 2:

- The GO analysis results are shown in Table S14 and in Figure S17. The bicluster has p-values between $3.2 \cdot 10^{-6}$ and $5.1 \cdot 10^{-2}$ and are all related to the immune system.
- The bicluster has a KEGG overlap of 64 genes. The KEGG result is shown in Table S15; the p-values range from $5.7 \cdot 10^{-8}$ to $7.4 \cdot 10^{-6}$.
- The protein network graph for bicluster 2 is shown in Figure S18. The strongest cluster is related to histocompatibility complex class I molecules like B2M (Beta-2-microglobulin precursor) which is the beta-chain of major histocompatibility complex class I molecules, CD3D (T-cell surface glycoprotein CD3 delta chain precursor), HLA proteins (HLA class I histocompatibility antigen). Another cluster is visible around ICAM1 (Intercellular adhesion molecule 1 precursor), a protein which is a ligand for the leukocyte adhesion protein LFA-1. This cluster also includes ITGAL (Integrin alpha-L precursor) which serves as a receptor for ICAM1. Moreover, we see a cluster around IRF1 (Interferon regulatory factor 1) which binds to the upstream regulatory region of type I IFN and IFN-inducible MHC class I genes. The last strong cluster contains CCL (small inducible cytokine precursor) proteins. All these clusters may be related to each other as text mining indicates.

Again the GO and KEGG analysis agree on genes that are related to the immune system. The most significant pathways are related to cytotoxicity, cell signaling, and cytokine-cytokine receptor interactions. Many proteins are related to the histocompatibility complex (which alerts the immune system) or T-cell surface glycoproteins.

Table S14: GO analysis for biological process (BP) on FABIA bicluster 2 obtained for the DLBCL data set.

CL	GO-BP-ID	p-value	Odds ratio	Count	Size	Term
1	GO:0002376	3.2e-06	4	32	142	immune system process
2	GO:0006955	1.7e-05	4	25	102	immune response
3	GO:0006952	2.9e-05	4	19	67	defense response
4	GO:0050896	1.6e-04	3	36	199	response to stimulus
5	GO:0009607	3.2e-04	5	11	32	response to biotic stimulus
6	GO:0051707	1.1e-03	5	9	26	response to other organism
7	GO:0042742	1.9e-03	17	4	6	defense response to bacterium
8	GO:0009617	2.6e-03	8	5	10	response to bacterium
9	GO:0006954	4.2e-03	3	11	42	inflammatory response
10	GO:0002474	5.1e-03	25	3	4	proc. and pres. of peptide antigen via MHC class I

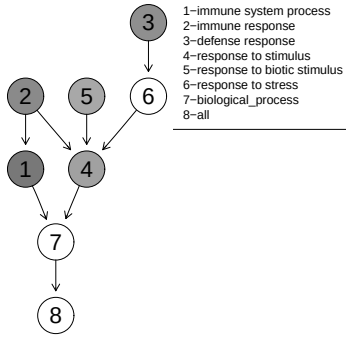

Figure S17: GO analysis for biological process (BP) on FABIA bicluster 2 obtained for the DLBCL data set. The GO hierarchy is shown, where darker circles indicate higher significance (lower p-values).

Table S15: KEGG analysis of FABIA bicluster 2 obtained for the DLBCL data set.

Name	p-value	q-value	Count	Size	Term
path:04650	5.7e-08	6.2e-06	7	132	Natural killer cell mediated cytotoxicity
path:04514	6.0e-08	6.2e-06	7	133	Cell adhesion molecules (CAMs)
path:04060	5.3e-06	3.7e-04	7	259	Cytokine-cytokine receptor interaction
path:04620	7.4e-06	5.1e-04	5	102	Toll-like receptor signaling pathway

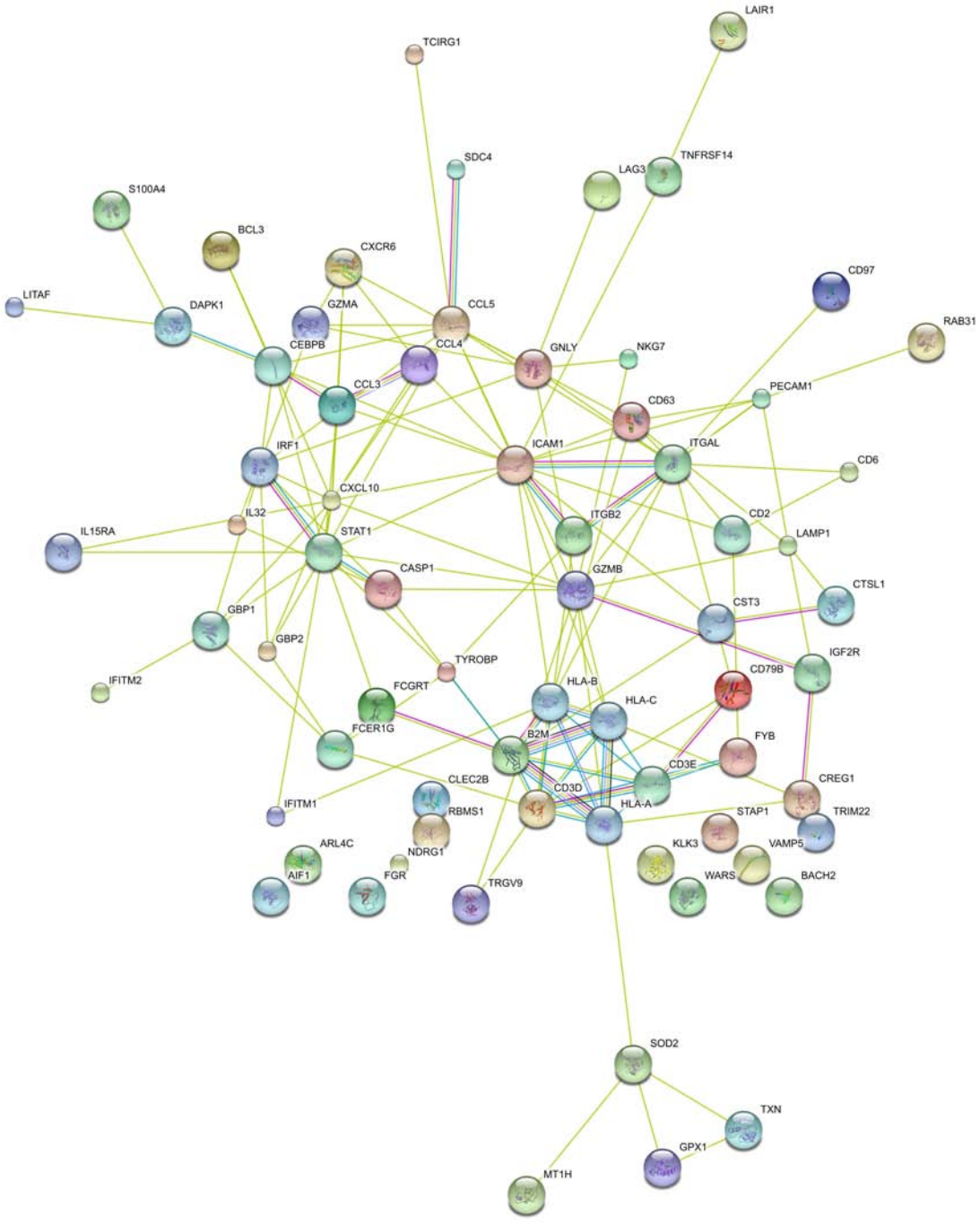

Figure S18: STRING protein network derived from genes found by FABIA in bicluster 2 of the DLBCL data set. Connections are labeled as described on p. 29.

Multiple tissues data set

The last data set we look at is the multiple tissue data set for which FABIA found five biclusters, the sizes of which are given as follows:

Bicluster	1	2	3	4	5
Genes	171	116	114	679	700
Samples	37	27	23	29	28

Here we restrict ourselves to a KEGG analysis, because other methods performed better than FABIA. However, we still can demonstrate that FABIA found biologically interesting groups.

Bicluster 1 has a KEGG overlap of 147 genes. The results can be found in Table S16; the p-values range from $4.8 \cdot 10^{-9}$ to $7.8 \cdot 10^{-4}$. The most significant pathways are related to the ribosome.

Table S16: KEGG analysis of FABIA bicluster 1 obtained for the multiple tissue data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:03010	4.8e-09	9.6e-07	9	92	Ribosome
2	path:04512	1.9e-04	1.9e-02	5	87	ECM-receptor interaction
3	path:00220	7.8e-04	4.8e-02	3	30	Urea cycle and metabolism of amino groups

Bicluster 2 has a KEGG overlap of 95 genes. The results are shown in Table S17. The p-values range from $6.3 \cdot 10^{-6}$ to $6.7 \cdot 10^{-4}$. There are only few genes members of the pathways, therefore, the biological interpretation is questionable.

Table S17: KEGG analysis of FABIA bicluster 2 obtained for the multiple tissue data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:05040	6.3e-06	7.8e-04	4	31	Huntington's disease
2	path:04720	7.5e-06	7.8e-04	5	69	Long-term potentiation
3	path:04020	6.5e-05	4.5e-03	6	175	Calcium signaling pathway
4	path:05214	1.2e-04	8.0e-03	4	64	Glioma
5	path:04740	2.4e-04	1.6e-02	5	31	Olfactory transduction
6	path:04912	5.7e-04	3.8e-02	4	97	GnRH signaling pathway
7	path:04916	6.7e-04	4.4e-02	4	101	Melanogenesis

Bicluster 3 has a KEGG overlap of 109 genes. The results are shown in Table S18; p-values range from $1.9 \cdot 10^{-4}$ to $4 \cdot 10^{-4}$. The p-values are not very small, though they survive FDR correction.

Table S18: KEGG analysis of FABIA bicluster 3 obtained for the multiple tissue data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:00410	1.9e-04	3.7e-02	3	25	beta-Alanine metabolism
2	path:01430	4.0e-04	3.9e-02	5	138	Cell junctions

Bicluster 4 has a KEGG overlap of 613 genes. The KEGG results are shown in Table S19. The p-values range from 0 (too small to be computed precisely) to $1.3 \cdot 10^{-2}$. The small p-values can be explained by the large cluster size, where genes are grouped together even if they show a weak signal.

Table S19: KEGG analysis of FABIA bicluster 4 obtained for the multiple tissue data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:04110	0.0e+00	0.0e+00	28	112	Cell cycle
2	path:03030	5.7e-10	3.8e-08	11	35	DNA replication
3	path:03050	1.4e-09	6.3e-08	9	22	Proteasome
4	path:00010	4.9e-08	1.6e-06	12	63	Glycolysis / Gluconeogenesis
5	path:00230	1.9e-07	5.1e-06	17	147	Purine metabolism
6	path:00240	3.4e-06	7.5e-05	12	92	Pyrimidine metabolism
7	path:04115	7.5e-06	1.4e-04	10	68	p53 signaling pathway
8	path:04010	1.2e-05	2.0e-04	20	262	MAPK signaling pathway
9	path:04810	3.5e-05	5.1e-04	17	215	Regulation of actin cytoskeleton
10	path:03430	1.5e-04	1.8e-03	5	21	Mismatch repair
11	path:04020	1.5e-04	1.8e-03	14	175	Calcium signaling pathway
12	path:00710	2.3e-04	2.5e-03	5	23	Carbon fixation in photosynthetic organisms
13	path:05210	2.6e-04	2.7e-03	9	84	Colorectal cancer
14	path:04510	5.5e-04	5.0e-03	14	199	Focal adhesion
15	path:00620	6.1e-04	5.3e-03	6	42	Pyruvate metabolism
16	path:04912	7.7e-04	6.4e-03	9	97	GnRH signaling pathway
17	path:00071	1.1e-03	8.5e-03	6	47	Fatty acid metabolism
18	path:00640	1.5e-03	1.1e-02	5	34	Propanoate metabolism
19	path:04512	1.6e-03	1.1e-02	8	87	ECM-receptor interaction
20	path:00590	2.3e-03	1.5e-02	6	54	Arachidonic acid metabolism
21	path:05212	2.4e-03	1.5e-02	7	73	Pancreatic cancer
22	path:01430	2.7e-03	1.6e-02	10	138	Cell junctions
23	path:00051	3.9e-03	2.1e-02	5	42	Fructose and mannose metabolism
24	path:03420	4.4e-03	2.3e-02	5	43	Nucleotide excision repair
25	path:00280	4.8e-03	2.4e-02	5	44	Valine, leucine and isoleucine degradation
26	path:04120	5.7e-03	2.7e-02	9	130	Ubiquitin mediated proteolysis
27	path:00260	5.8e-03	2.7e-02	5	46	Glycine, serine and threonine metabolism
28	path:05222	6.4e-03	2.9e-02	7	87	Small cell lung cancer
29	path:00670	7.0e-03	3.0e-02	3	16	One carbon pool by folate
30	path:00565	7.3e-03	3.1e-02	4	31	Ether lipid metabolism
31	path:00251	7.3e-03	3.1e-02	4	31	Glutamate metabolism
32	path:00900	8.9e-03	3.6e-02	2	6	Terpenoid biosynthesis
33	path:03410	9.1e-03	3.7e-02	4	33	Base excision repair
34	path:00330	1.1e-02	4.3e-02	4	35	Arginine and proline metabolism
35	path:05220	1.2e-02	4.5e-02	6	76	Chronic myeloid leukemia
36	path:04664	1.3e-02	4.7e-02	6	77	Fc epsilon RI signaling pathway

Bicluster 5 has a KEGG overlap of 617 genes. The KEGG result is shown in Table S20; p-values range from $6.4 \cdot 10^{-10}$ to $9.4 \cdot 10^{-3}$. Again the small p-values are a result of the large cluster size.

Table S20: KEGG analysis of FABIA bicluster 5 obtained for the multiple tissue data set.

CL	Name	p-value	q-value	Count	Size	Term
1	path:04640	6.4e-10	9.9e-08	16	88	Hematopoietic cell lineage
2	path:04080	8.7e-09	6.7e-07	25	254	Neuroactive ligand-receptor interaction
3	path:04020	1.1e-07	5.8e-06	19	175	Calcium signaling pathway
4	path:04060	3.2e-06	1.2e-04	21	259	Cytokine-cytokine receptor interaction
5	path:04514	7.5e-06	2.3e-04	14	133	Cell adhesion molecules (CAMs)
6	path:00232	1.3e-05	3.4e-04	4	7	Caffeine metabolism
7	path:04630	3.7e-05	7.6e-04	14	153	Jak-STAT signaling pathway
8	path:00983	5.3e-05	9.7e-04	8	53	Drug metabolism - other enzymes
9	path:05222	7.1e-05	1.1e-03	10	87	Small cell lung cancer
10	path:00982	9.4e-05	1.3e-03	9	73	Drug metabolism - cytochrome P450
11	path:03420	9.5e-05	1.3e-03	7	43	Nucleotide excision repair
12	path:04510	1.8e-04	2.3e-03	15	199	Focal adhesion
13	path:04810	4.2e-04	4.6e-03	15	215	Regulation of actin cytoskeleton
14	path:05215	4.3e-04	4.7e-03	9	89	Prostate cancer
15	path:04110	5.7e-04	5.8e-03	10	112	Cell cycle
16	path:04620	1.2e-03	1.1e-02	9	102	Toll-like receptor signaling pathway
17	path:04012	1.6e-03	1.5e-02	8	87	ErbB signaling pathway
18	path:05223	2.4e-03	2.0e-02	6	54	Non-small cell lung cancer
19	path:04670	2.8e-03	2.3e-02	9	116	Leukocyte transendothelial migration
20	path:05060	4.8e-03	3.3e-02	3	14	Prion disease
21	path:05210	5.5e-03	3.6e-02	7	84	Colorectal cancer
22	path:05214	5.6e-03	3.7e-02	6	64	Glioma
23	path:04530	7.6e-03	4.4e-02	9	135	Tight junction
24	path:04614	8.5e-03	4.7e-02	3	17	Renin-angiotensin system
25	path:04910	8.7e-03	4.8e-02	9	138	Insulin signaling pathway
26	path:05218	9.2e-03	4.9e-02	6	71	Melanoma
27	path:04660	9.4e-03	5.0e-02	7	93	T cell receptor signaling pathway

S6.5 Drug Design

Figure S19 shows a plot of the data distribution of the drug design gene expression data set discussed in Section 6.5 of the main paper.

References

Bithas, P. S., Sagias, N. C., Tsiftsis, T. A., and Karagiannidis, G. K. (2007). Distributions involving correlated generalized gamma variables. In *Proc. Int. Conf. on Applied Stochastic Models and Data Analysis*, volume 12, Chania.

Caldas, J. and Kaski, S. (2008). Bayesian biclustering with the plaid model. In *Proc. IEEE Int. Workshop on Machine Learning for Signal Processing*, volume XVIII, pages 291–296.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Proc. Int. Conf. on Intelligent Systems for Molecular Biology*, volume 8, pages 93–103.

Dharmadhikari, S. W. and Jogdeo, K. (1976). Multivariate unimodality. *Ann. Stat.*, **4**(3), 607–613.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., Hallett, M., and Park, M. (2008). Stromal gene expression predicts clinical outcome in breast cancer. *Nat. Med.*, **14**(5), 518–527.

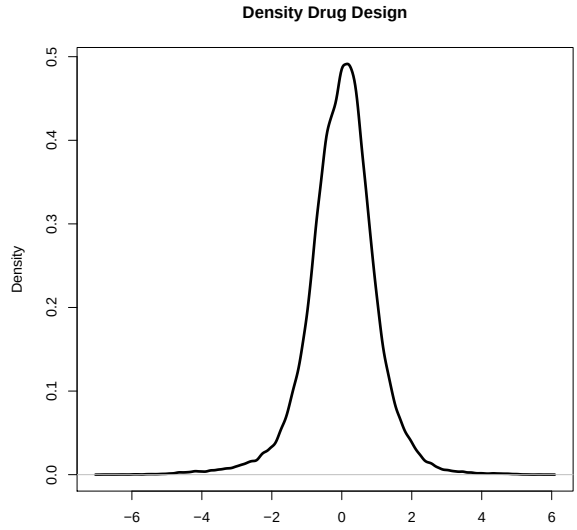

Figure S19: Density of the drug design data set. The skewness is -0.39 and the excess kurtosis is larger than 3.0 (heavier tails than Laplace).

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Comput.*, **13**(11), 2517–2532.

Haldane, J. B. S. (1942). The mode and median of a nearly normal distribution with given cumulants. *Biometrika*, **32**(3-4), 294–299.

Hall, P. (1980). On the limiting behaviour of the mode and median of a sum of independent random variables. *Ann. Probab.*, **8**(3), 419–430.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**, 1457–1469.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–416.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562.

Li, G., Ma, Q., Tang, H., Paterson, A. H., and Xu, Y. (2009). QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res.*, **37**(15), e101.

Meeds, E., Ghahramani, Z., Neal, R. M., and Roweis, S. T. (2007). Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems 19*, pages 977–984.

Oehlschlägel, J. (2006). Truecluster: robust scalable clustering with model selection. arXiv:cs/0601001.

Palmer, J., Wipf, D., Kreutz-Delgado, K., and Rao, B. (2006). Variational EM algorithms for non-Gaussian latent variable models. In *Advances in Neural Information Processing Systems 18*, pages 1059–1066.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**(9), 1122–1129.

Talloen, W., Clevert, D.-A., Hochreiter, S., Amaratunga, D., Bijnens, L., Kass, S., and Göhlmann, H. W. H. (2007). I/NI-calls for the exclusion of non-informative genes: a highly effective feature filtering tool for microarray data. *Bioinformatics*, **23**(21), 2897–2902.