Supplemental Information

Bayesian context-specific infinite mixture model for clustering of gene expression profiles across diverse microarray datasets

Xiangdong Liu, Siva Sivaganesan, Ka Yee Yeung, Junhai Guo, Roger Bumgarner, Mario Medvedovic*

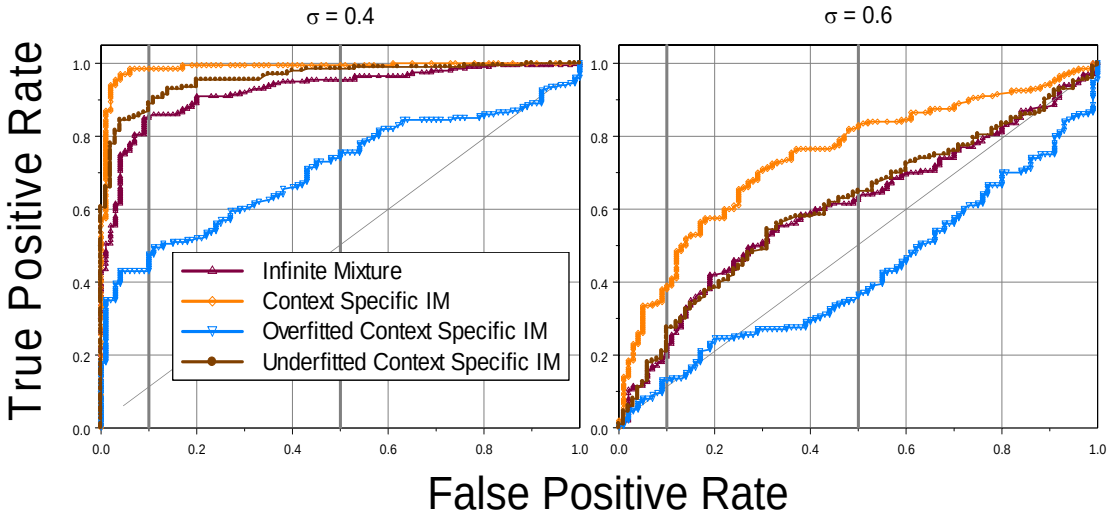
*Mario.Medvedovic@uc.edu

OUTLINE:

1. Additional ROC curves for the simulation study
2. Patterns of gene expression based on the joint analysis of cell cycle and sporulation data.
3. Patterns of gene expression based on the analysis of individual datasets (cell cycle and sporulation) separately.
4. Prior and posterior conditional probability distributions in the context-specific infinite mixture model.
5. Dynamic annealing modification of the Gibbs sampler.
6. Computational complexity and run times

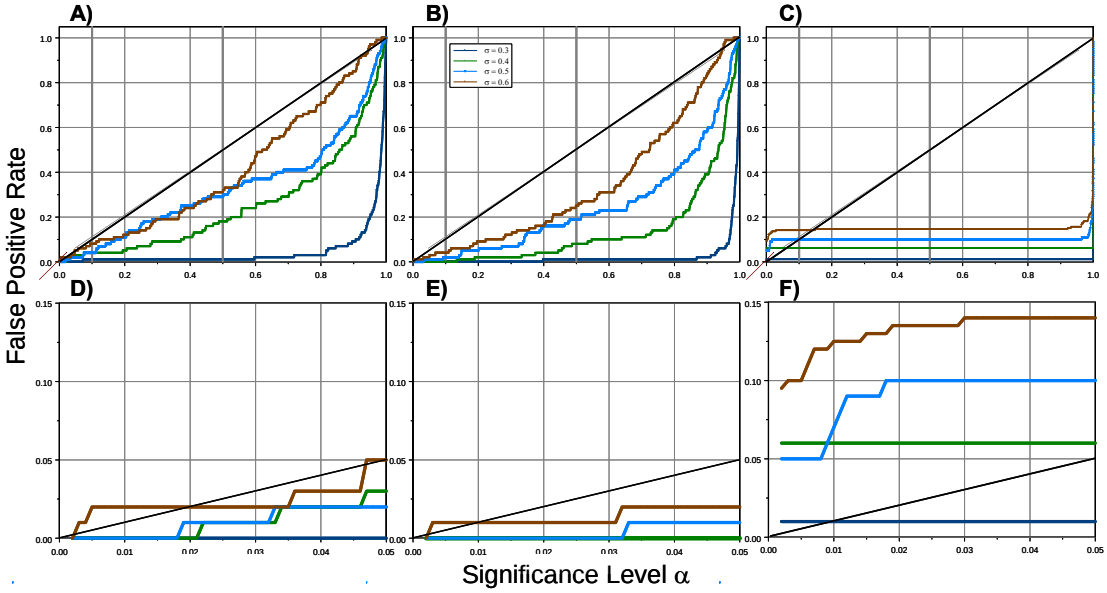
1. Additional ROC curves for the simulation study

Figure S1: ROC curves for different levels of precision in specifying contexts.



Posterior pairwise probabilities are valid measures of statistical significance: In **Figure 5** we plotted observed false positive rates against corresponding statistical significance levels from the finite and infinite mixture analyses (Medvedovic and Guo 2004). Given a significance level α, all gene-pairs whose PPP was lower than α were assumed to belong to different patterns. For a valid statistical procedure the false positive rates following this decision making scheme should be less than α for all α levels. As the noise in the data increases, the finite mixture model based PPP's become progressively worse as estimates of statistical significance, especially in the range relevant for assessing statistical significance of differences, between 0 and 0.1. On the other hand PPPs based on the infinite mixture model remain valid measures of statistical significance at all noise levels.

Figure S2: Posterior probabilities as measures of statistical significance. A) Simple infinite mixtures. B) Context-specific infinite mixtures. C) Finite mixtures. D, E and F are "zoomed-in" versions of A, B and C respectively.



2. Patterns of gene expression based on the joint analysis of cell cycle and sporulation data

Hierarchical clustering based on CSIMM for 135 genes which were co-clustered with at least one other gene after cutting the tree at the average linkage distance of 0.05 is displayed in **Figure S3.** The red-green heatmap depicts the gene expression levels, blue heatmap specifies membership in different KEGG pathways, and the yellow heatmap specifies the binding of a specific transcription factors in "Chip-on-Chip" experiments (3). The hierarchical tree in **Figure S3** was cut in 8 clusters. Six of these clusters had more than two genes and they were tested for over-representation of genes whose promoters are substrates of any one single transcription factor using the Fisher's exact test. Eight transcription factors were significantly associated with at least one of the cluster (Fisher's exact p-value<0.05). It is interesting that two "sporulation clusters" (denoted by the black bar on the right-hand side of the heatmap) both correlate with the

transcription factor SUM1 which seems to be involved in both sporulation and cell-cycle regulation (4). This suggests the functional relatedness of genes in these clusters which are not associated with any KEGG pathway.
In comparison, cutting the tree formed by the Euclidian-distance based hierarchical clustering to obtain 135 genes that were co-clustered with at least one other gene generated diffused patterns without any obvious clustering structure (**Figure S4**).

Figure S3: CSIMM clustering for 135 “closest” genes.

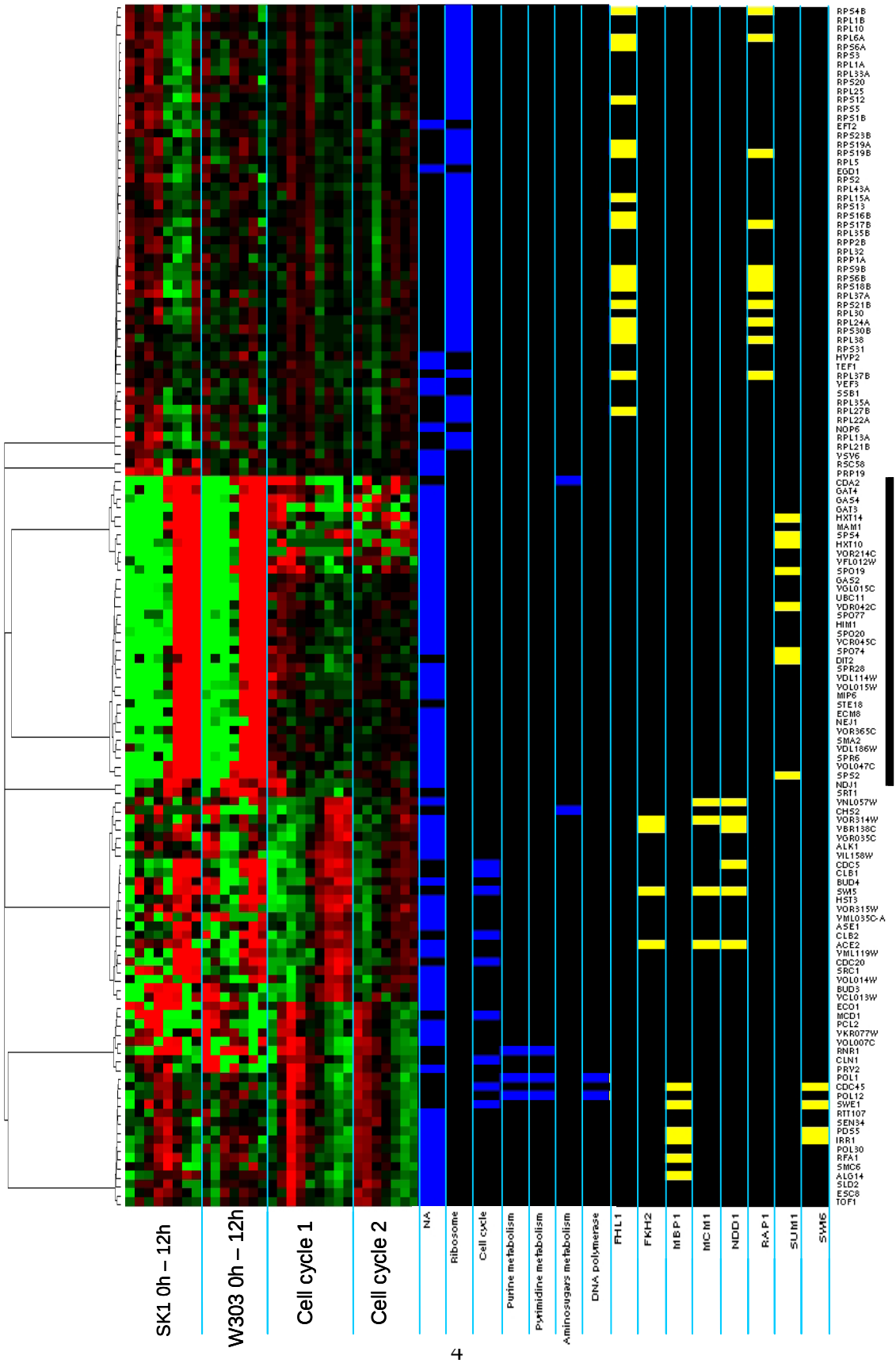
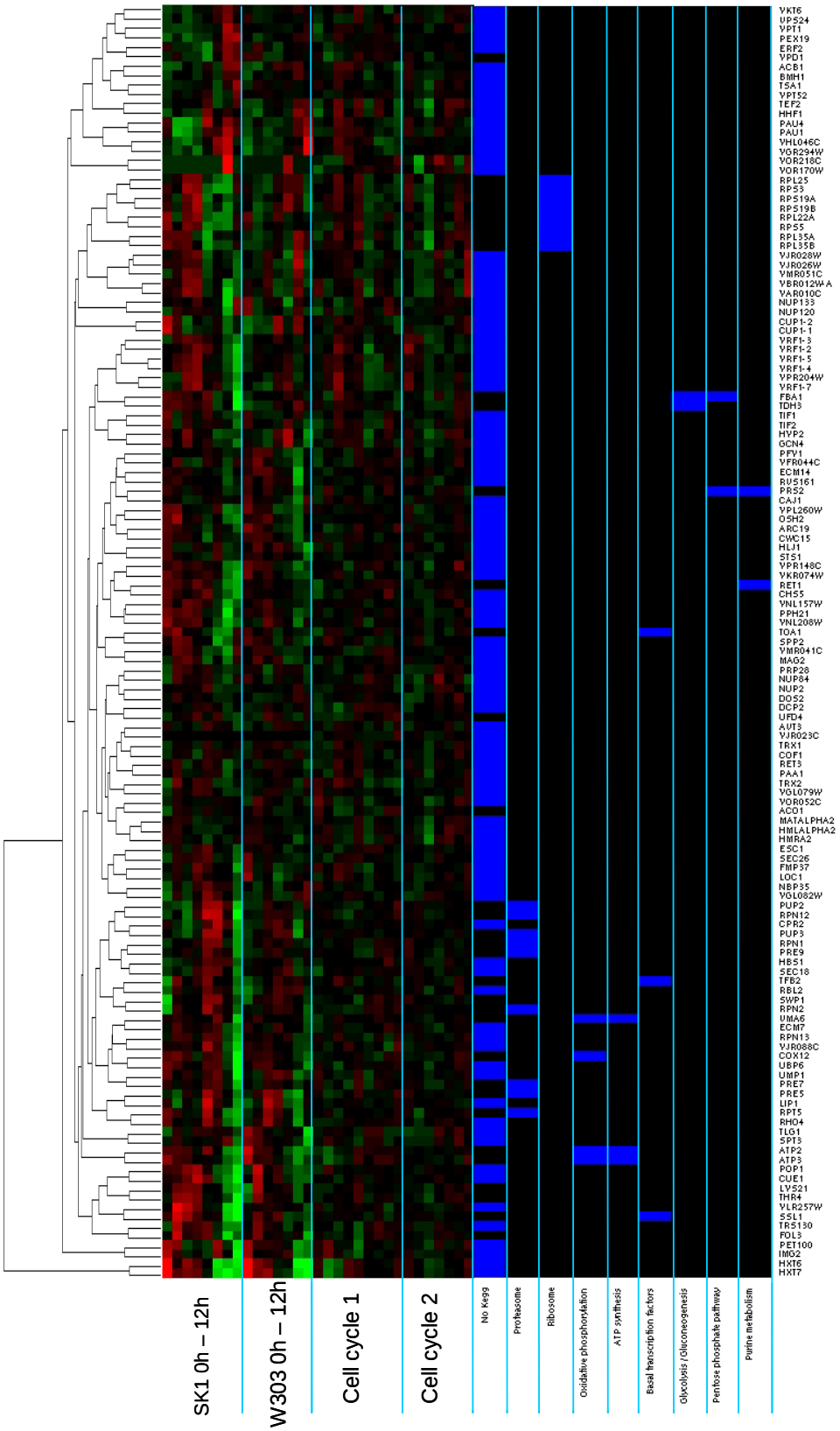


Figure S4: Euclidian distance based clustering for 135 “closest” genes.



3. Patterns of gene expression based on the analysis of individual datasets (cell cycle and sporulation) separately

In the main text of the paper we included ROC curves for individual data sets (sporulation and cell cycle) based on again using CSIMM approach in which data from two strains of yeas (in sporulation experiments) and data from two successive cell-cycles (cell cycle data) were designated as different contexts. This was motivated by the fact that such approaches offered higher precision than the traditional analyses that would not distinguish between such contexts (**Figure S5**) and we wanted to demonstrate the increased precision of combining the sporulation and cell cycle datasets was not an artifact of an inferior analytical approach in the case of individual datasets. Here we examine further the characteristics of gene expression patterns generated based on individual datasets alone (**Figure S6, S7, S8 and S9**). Results are very similar to what we observed in the joint analysis of both datasets. The CSIMM approach results in clearly defined and functionally relevant clusters, while Euclidian-distance based clusterings are diffused without very few clearly defined and meaningful patterns.

Figure S5: ROC curves for cell cycle and sporulation data analyzed separately.

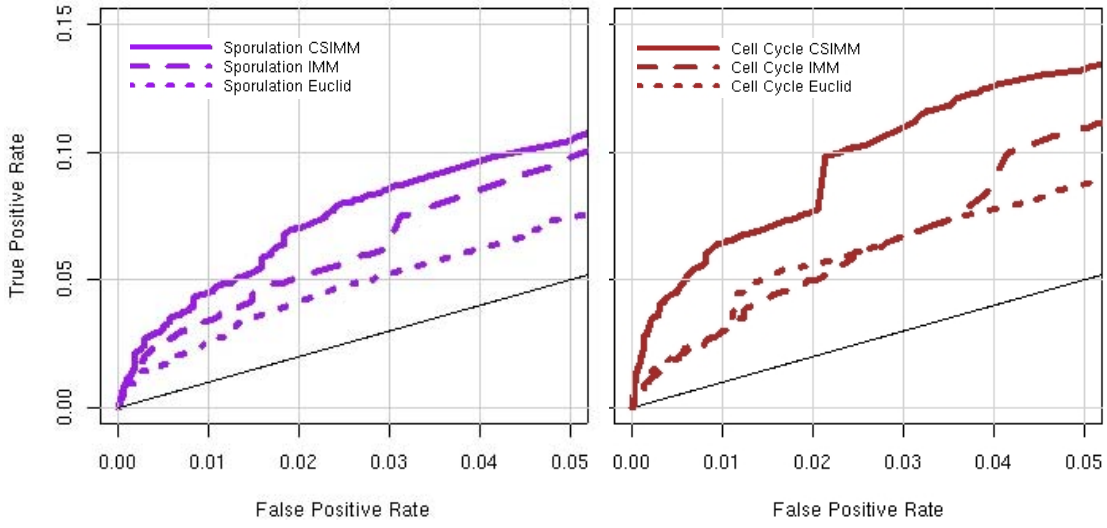


Figure S6: CSIMM clustering for 135 “closest” genes based only on cell cycle data.

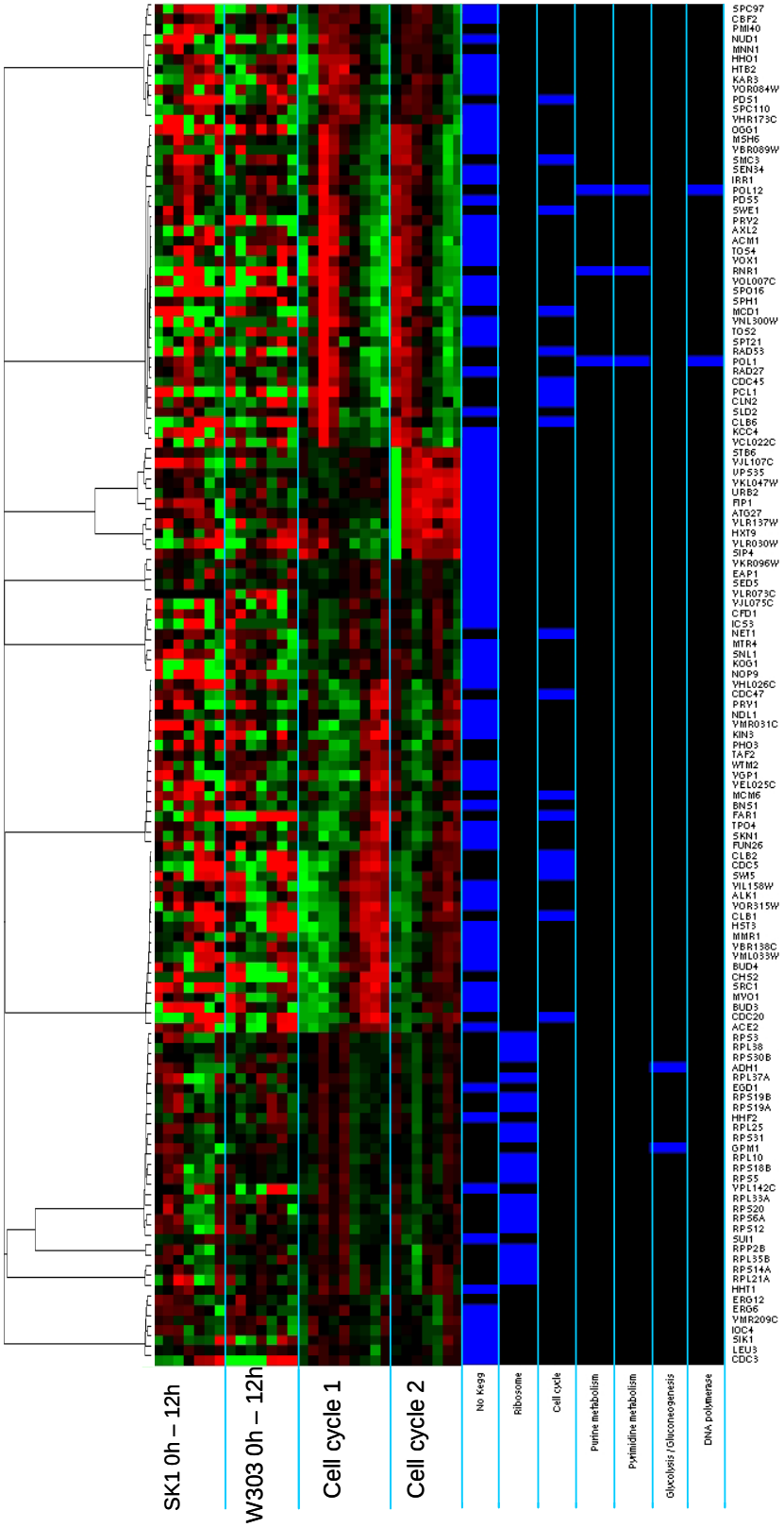


Figure S7 Euclidian distance based clustering for 135 “closest” genes based only on cell cycle data.

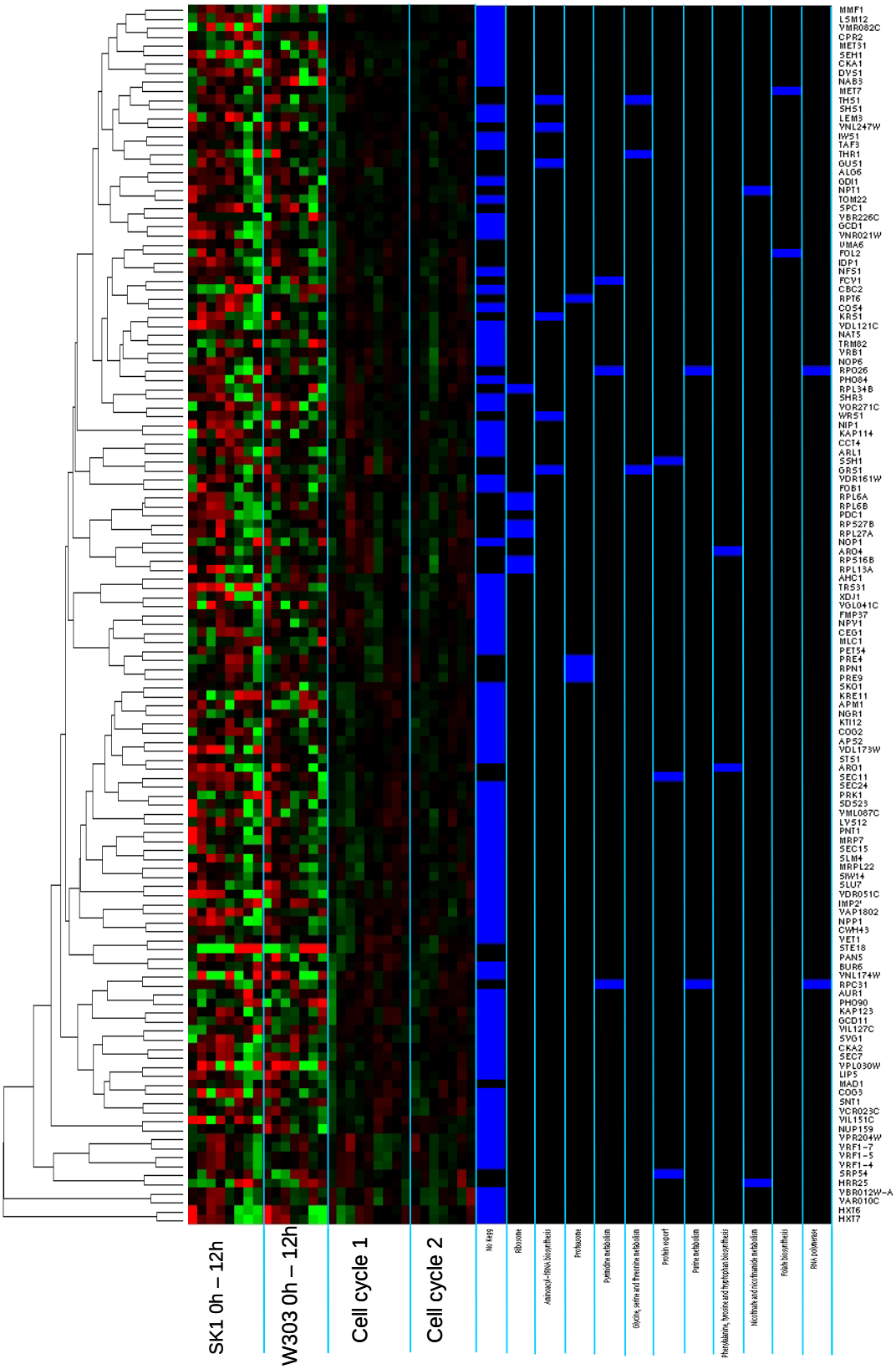


Figure S8 CSIMM clustering for 135 "closest" genes based only on sporulation data.

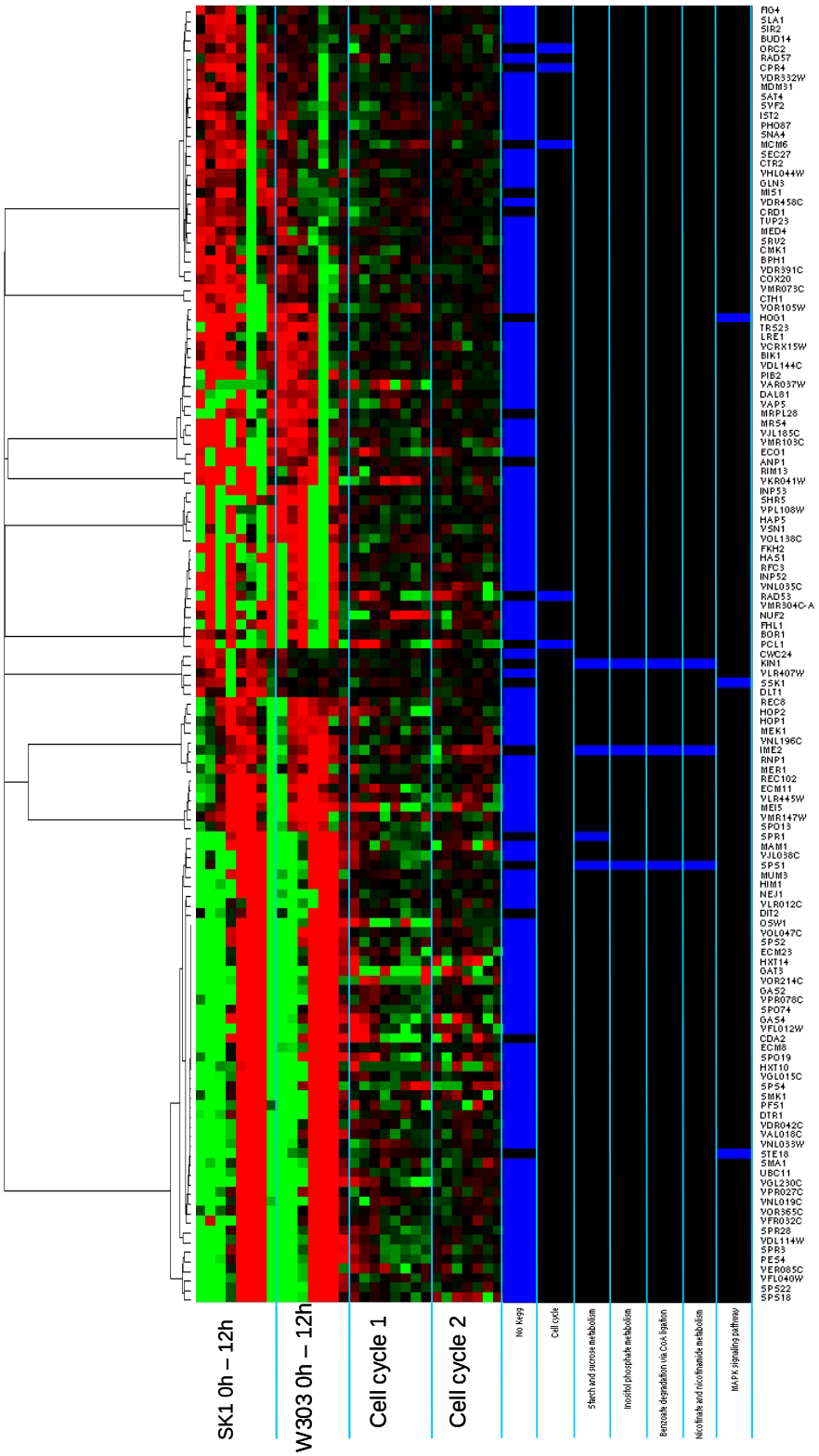
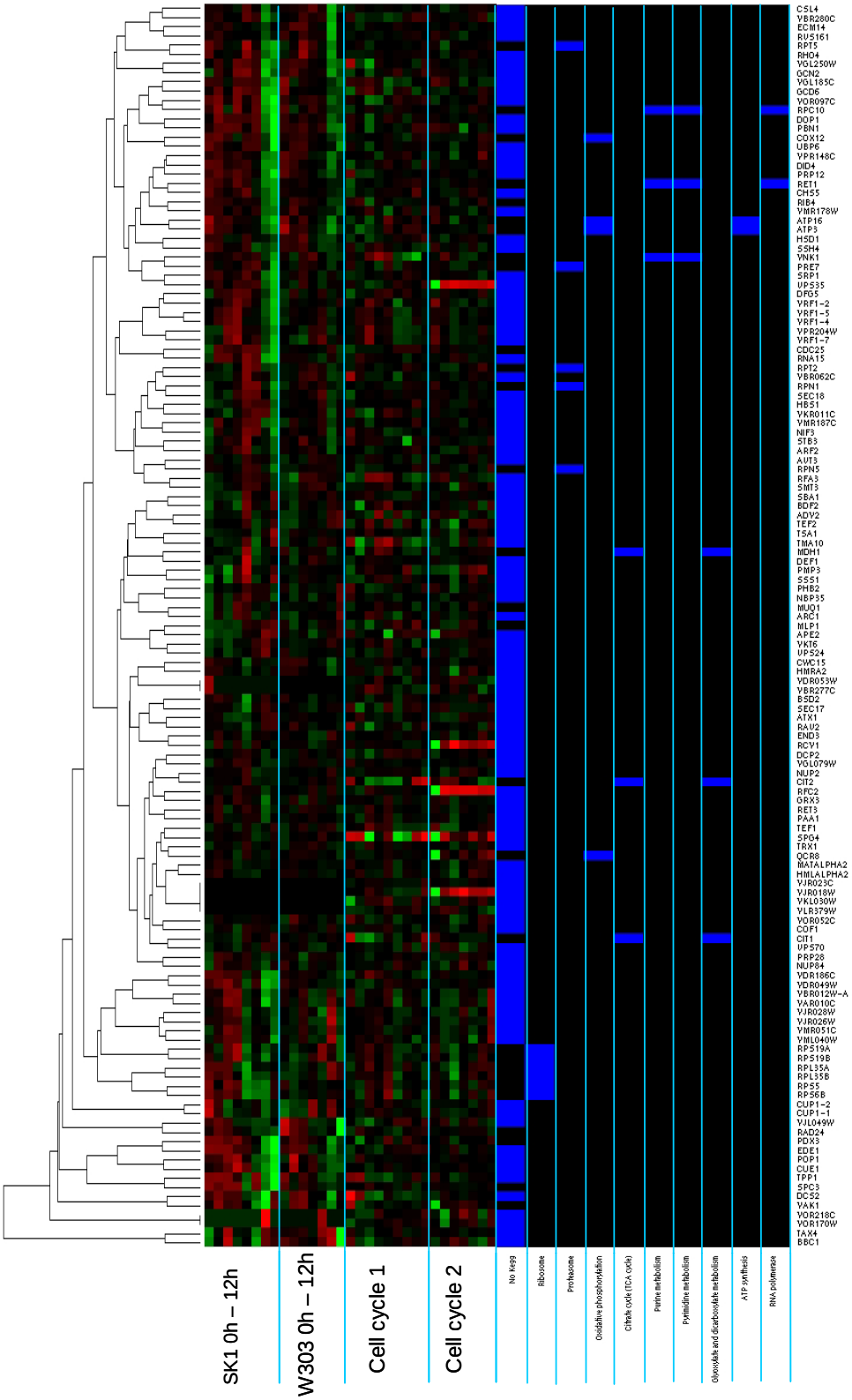


Figure S9 Euclidian distance based clustering for 135 “closest” genes based only on sporulation data.



4. Prior and posterior conditional probability distributions in the context-specific infinite mixture model:

Variables in the model:

$\mathbf{x}_i=(x_{i1}, x_{i2},\ldots, x_{iM})$, i=1,…,T observed gene expression profiles for all T genes

$\boldsymbol{\mu}_q=(\mu_{q1},\ldots, \mu_{qM})$, q=1,…,Q the mean profile for global cluster q

$\mathbf{x}_i^f=(x_{i\,r'_f+1},\ldots,x_{i\,r'_f+r_f})$ where $r'_f=r_1+\ldots+r_{f-1}$, f=1,…,R is the expression profile for gene i within context f, i=1,…,Q, f=1,…,R

$\boldsymbol{\mu}^*_{tf}$, mean expression profile for the local cluster t within context f

$\mathbf{M}=(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_Q)$

$\mathbf{S}=(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_Q)$, where each $\boldsymbol{\Sigma}_q$ is a diagonal matrix with context-specific cluster variances on the diagonal. That is $\boldsymbol{\Sigma}_q=\mathrm{diag}((\sigma^2_{q1},\ldots,\sigma^2_{q1}),(\sigma^2_{q2},\ldots,\sigma^2_{q2}),\ldots,(\sigma^2_{qR},\ldots,\sigma^2_{qR}))$

$\mathbf{M}^*=[(\boldsymbol{\mu}^*_{11},\ldots,\boldsymbol{\mu}^*_{K_1 1}),\ldots,(\boldsymbol{\mu}^*_{1R},\ldots,\boldsymbol{\mu}^*_{K_R R})]$

$\mathbf{S}^*=[\Sigma^*_1,\ldots,\Sigma^*_R]$, where $\Sigma^*_f=(\sigma^*_q)^2\mathbf{I}$ for f=1,…,R

Hyperparameters $\boldsymbol{\tau}$, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are all assumed to be context-specific:
$\boldsymbol{\tau}=(\tau_1,\ldots,\tau_R)$, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_R)$, $\boldsymbol{\phi}=(\phi_1,\ldots,\phi_R)$.

The joint distribution of all variables from the model in Figure 1 of the paper:

$$p(\mathbf{X}, \mathbf{C}, \mathbf{L}, \mathbf{M}, \mathbf{M}^*, \mathbf{S}, \alpha, a, \boldsymbol{\lambda}, \tau, \beta, \phi) = p(\mathbf{X}|\,\mathbf{C}, \mathbf{M}, \mathbf{S})p(\mathbf{C}|\alpha)p(\mathbf{M}|\mathbf{L},\mathbf{M}^*)p(\mathbf{S}|\beta, \phi)$$
$$p(\mathbf{L}|\mathbf{C},a)p(\mathbf{M}^*|\boldsymbol{\lambda}, \tau)p(\alpha)p(a)p(\boldsymbol{\lambda})p(\tau)p(\beta)p(\phi)$$

Conditional distributions given parent nodes:

$$p(\mathbf{x}_i\,|\,c_i=q,\ \mathbf{M},\mathbf{S}) = f_N(\mathbf{x}_i\,|\,\boldsymbol{\mu}_q,\Sigma_q)\,.$$

$$p((\boldsymbol{\mu}_{q1},\boldsymbol{\mu}_{q2},\ldots,\boldsymbol{\mu}_{qR})\,|\,\mathbf{L}\ ,\mathbf{M}^*,\sigma^*) = (f_N(\boldsymbol{\mu}_{q1}\,|\,\boldsymbol{\mu}^*_{L_{q1}1},\boldsymbol{\Sigma}^*_1), f_N(\boldsymbol{\mu}_{q2}\,|\,\boldsymbol{\mu}^*_{L_{q2}2},\boldsymbol{\Sigma}^*_2),\ldots, f_N(\boldsymbol{\mu}_{qR}\,|\,\boldsymbol{\mu}^*_{L_{qR}R},\boldsymbol{\Sigma}^*_R)$$

$p(c_i=q\,|\,\mathbf{C}_{-i},\alpha)\propto\frac{n_{-i,q}}{T-1+\alpha}$, q=1,…,Q, $p(c_i\neq c_{i'},i'\neq i\,|\,\mathbf{C}_{-i},\alpha)\propto\frac{\alpha}{T-1+\alpha}$, i=1,…,T, where $n_{-i,q}$ is the number of profiles placed in global cluster q not counting the profile i

$p(L_{qf}=t|C,a)\propto\frac{n_{-qft}}{Q-1+a}$, t=1,..,Q, $p(L_{kf}\neq L_{kr},\forall k'\neq k|C,a)\propto\frac{a}{Q-1+a}$ where n_{-qft} is the number of global clusters currently placed in local cluster t within context f without counting the q^{th} global cluster

$$p(\boldsymbol{\mu}^*_{tf}\,|\,\boldsymbol{\lambda}_f,\tau_f)=f_N(\boldsymbol{\mu}^*_{tf}\,|\,\boldsymbol{\lambda}_f,\tau_f^{-1}\mathbf{I}) \qquad p(\sigma^{-2}_{tf}\,|\,\beta_f,\varphi_f)=f_G(\sigma^{-2}_{tf}\,|\,\frac{\beta_{tf}}{2},\frac{\beta_f\varphi_f}{2}),\ f=1,\ldots,R$$

$$p(\varphi_f\,|\,\sigma^2_{xf})=f_G(\varphi_f\,|\,\frac{1}{2},\frac{\sigma^{-2}_{xf}}{2}) \qquad p(\beta_f)=f_G(\beta_f\,|\,\frac{1}{2},\frac{1}{2})$$

$$p(\tau_f\,|\,\sigma^2_{\mathbf{xf}})=f_G(\tau_f\,|\,\frac{1}{2},\frac{\sigma^2_{\mathbf{xf}}}{2}) \qquad p(\boldsymbol{\lambda}_f\,|\,\boldsymbol{\mu}_{xf},\sigma^2_{xf})=f_N(\boldsymbol{\lambda}_f\,|\,\boldsymbol{\mu}_{xf},\sigma^2_{xf}\mathbf{I})$$

$$p(\alpha^{-1})=f_G(\alpha^{-1}\,|\,\frac{1}{2},\frac{1}{2})$$

where

$$\boldsymbol{\mu}_{xf} = \frac{\sum_{i=1}^{T} \mathbf{x}_i^f}{T} \qquad \sigma_{xf}^2 = \frac{\sum_{i=1}^{T} (\mathbf{x}_i^f - \boldsymbol{\mu}_{xf})`(\mathbf{x}_i^f - \boldsymbol{\mu}_{xf})}{T\, r_f - 1}$$

Posterior Conditional Distributions:

$$p(\boldsymbol{\mu}_{tf}^* \mid \mathbf{C}, L, \sigma_{tf}^2, \mathbf{X}, \boldsymbol{\lambda}_f, \tau_f) = f_N(\boldsymbol{\mu}_{tf}^* \mid \frac{\tau_f^{-1}\bar{\mathbf{x}}_{tf\bullet} + \frac{\sigma_{tf}^2}{n_{tf}^*}\boldsymbol{\lambda}_f}{\tau_f^{-1} + \frac{\sigma_{tf}^2}{n_{tf}^*}}, \frac{\tau_f^{-1}\frac{\sigma_{tf}^2}{n_{tf}^*}}{\tau_f^{-1} + \frac{\sigma_{tf}^2}{n_{tf}^*}}\mathbf{I}), \text{ where } \bar{\mathbf{x}}_{tf\bullet} = \frac{\sum_{L_{c_i f}=t} \mathbf{x}_i^f}{n_{tf}^*} \text{ and}$$

n_{tf}^* is the total number of expression profiles grouped in global clusters which are place in the local cluster t within the context f. Similarly, the variance for all global clusters place in the local cluster t within the context f is

$$p(\sigma_{tf}^{-2} \mid \mathbf{X}, \mathbf{M}, \beta_f, \varphi_f) = f_G(\sigma_{tf}^{-2} \mid \frac{r_f n_{tf}^* + \beta_f}{2}, \frac{s_{tf}^2 + \beta_f \varphi_f}{2}), \text{ where } s_{tf}^2 = \sum_{L_{c_i f}=t} (\mathbf{x}_i^f - \boldsymbol{\mu}_{tf}^*)`(\mathbf{x}_i^f - \boldsymbol{\mu}_{tf}^*)$$

$$f(\boldsymbol{\lambda}_f \mid \boldsymbol{\mu}_{1f}^*, ..., \boldsymbol{\mu}_{K_f f}^*, \tau_f) = f_N(\boldsymbol{\lambda}_f \mid \frac{\sigma_{xf}^2 \frac{\sum_{t=1}^{K_f} \boldsymbol{\mu}_{tf}^*}{K_f} + \frac{\tau_f^{-1}}{K_f}\boldsymbol{\mu}_{xf}}{\sigma_x^2 + \frac{\tau^{-1}}{Q}}, \frac{\sigma_{xf}^2 \frac{\tau_f^{-1}}{K_f}}{\sigma_{xf}^2 + \frac{\tau_f^{-1}}{K_f}}\mathbf{I})$$

$$f(\tau_f^{-1} \mid \boldsymbol{\mu}_{1f}^*, ..., \boldsymbol{\mu}_{K_f f}^*, \boldsymbol{\lambda}_f) = f_G(\tau_f^{-1} \mid \frac{r_f K_f + 1}{2}, \frac{\sum_{t=1}^{K_f} (\boldsymbol{\mu}_{tf}^* - \boldsymbol{\lambda}_f)`(\boldsymbol{\mu}_{tf}^* - \boldsymbol{\lambda}_f) + \sigma_{xf}^2}{2})$$

$$f(\varphi_f \mid \sigma_{1f}^{-2}, ..., \sigma_{K_f f}^{-2}, \beta_f) = f_G(\varphi_f \mid \frac{K_f \beta_f + 1}{2}, \frac{\beta_f \sum_{t=1}^{K_f} \sigma_{tf}^{-2} + \sigma_{xf}^{-2}}{2})$$

$$f(\beta_t \mid \sigma_{tf}^{-2}, ..., \sigma_{tf}^{-2}, \varphi_f) \propto \Gamma(\frac{\beta_f}{2})(\frac{\beta_f}{2})^{\frac{(K_f \beta_f - 3)}{2}} \exp\left(-\frac{\beta_f^{-1}}{2}\right) \prod_{t=1}^{K_f} \left[(\varphi_f \sigma_{tf}^{-2})^{\frac{\beta_f}{2}} \exp\left(-\frac{\sigma_{tf}^{-2} \varphi_f \beta_f)}{2}\right)\right]$$

$$f(\alpha \mid Q, T) \propto \frac{\alpha^{Q-\frac{3}{2}} \exp(-\frac{1}{2\alpha})\Gamma(\alpha)}{\Gamma(T + \alpha)}$$

The posterior distributions for **C** and **L** are now:

$$p(c_i = q \mid \mathbf{C}_{-i}, \mathbf{L}, \mathbf{x}_i, \mathbf{M}, \boldsymbol{\Sigma}) \propto \frac{n_{-i,q}}{T - 1 + \alpha} f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$$

$$p(c_i \neq c_{i'}, i' \neq i \mid \mathbf{C}_{-i}, \mathbf{x}_i, \boldsymbol{\mu}_x, \sigma_x^2) \propto \frac{\alpha}{T - 1 + \alpha} \left(\int f_N(\mathbf{x}_i \mid \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)\, p(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q \mid \lambda, \tau)\, d\boldsymbol{\mu}_q d\boldsymbol{\Sigma}_q \right)$$

$$p(L_{qf} = t \mid \mathbf{X}, a) \propto \frac{n_{-qft}}{Q - 1 + a} f_N(\bar{\mathbf{x}}_f^q \mid \boldsymbol{\mu}_{tf}^*, \frac{\sigma_{tf}^2}{n_q}\mathbf{I})$$

$$p(L_{kf} \neq L_{k'f}, \forall k' \neq k \mid \mathbf{X}, a) \propto \frac{a}{Q - 1 + a} \int f_N(\bar{\mathbf{x}}_f^q \mid \boldsymbol{\mu}_{tf}^*, \frac{\sigma_{tf}^2}{n_q}\mathbf{I})\, p(\boldsymbol{\mu}_{ft}^*, \frac{\sigma_{tf}^2}{n_q}) d(\boldsymbol{\mu}_{tf}^*, \sigma_{tf}^2),$$

where $n_{-i,q}$ is the number of profiles in global cluster q without counting profile i, and n_{-qft} is the number of global clusters grouped into local cluster t within context f not counting q^{th} global cluster, and $\bar{\mathbf{x}}_f^q = \frac{\sum_{c_i = q} \mathbf{x}_i^f}{n_q}$.

5. Dynamic annealing modification of the Gibbs sampler.

Two aspects of the Gibbs sampler convergence that generally need to be assessed are the appropriateness of the "burn-in" period, after which a Gibbs sampler has attained its stationary distribution, and the mixing of the sampler, which describes how well a finite sample obtained by Gibbs sampler approximates the target distribution. It has generally been well documented that the simple Gibbs sampler often has poor mixing properties in when fitting finite or infinite mixture models (1, 2). In such situations, the sampler will be unable to describe the whole posterior distribution in a computationally feasible number of steps. This is often due to the sampler getting trapped in a sub-optimal mode of the posterior distribution resulting in sub-optimal clustering results and inappropriate confidence estimates. Previously, we described a heuristic algorithm for "heating up" the Markov chain described by the Gibbs sampler by using "reverse annealing." The optimal annealing schedule was chosen based on running a significant number of independent chains with different maximum annealing constants. However, it turned out that in some situations choosing the appropriate parameters in such a way was virtually impossible. Therefore we developed a heuristic algorithm that adjusts the annealing exponent dynamically. Consequently, only a single run is needed to estimate the posterior distribution.

If $\pi(.)$ is the target posterior distribution, "reverse annealing" refers to "flattening" of the posterior distribution using the transformation $\pi^{(\xi)}(x) = \frac{\pi^{\xi}(x)}{K(\xi)}$, $\xi < 1$, where $K(\xi)$ is the normalizing constant. Based on this general idea, if $p(c_i=j|C_{-i}, \Theta)$ is the conditional posterior probability of placing the i^{th} profile into the j^{th} cluster then "flattened probabilities" are defined as

$$p(c_i = j \mid C_{-i}, \Theta)^{(\xi)} = \frac{p(c_i = j \mid C_{-i}, \Theta)^{\xi}}{K(\xi)}, \quad \xi < 1.$$

Since the mixing problem with the Gibbs sampler for the IM model can be particularly pronounced in its inability to generate new clusters, we keep track of the posterior probability of placing a profile in a new cluster. If this probability p_{new} is below the given threshold p_{min}, we decrease ξ by the value ξ_{step}. If p_{new} is above p_{min}, we increase ξ by ξ_{step}. Possible values of ξ are further constrained by the requirement that $0<\xi_{min}<\xi<\xi_{max}\leq 1$. Our modified Gibbs sampler now proceeds by generating n_{cold} samples from the unmodified conditional posteriors (cold cycles). It then generates a single sample using "heated" classification probabilities (heated cycle). The p_{new} from the heated cycle is used to increase or decrease the value of ξ by ξ_{step}. However, only the samples from "cold" cycles are used in the estimation of the posterior distribution of clusterings. We established a set of appropriate parameters by extensive testing on both simulated and real world data ($n_{cold}=1$, $\xi_{min}=0.01$, $\xi_{max}=1$, $p_{new}=0.01$ and $p_{step}=0.01$). All analyses presented here for both CSIMM and the simple IMM model used this set of dynamic annealing parameters. These values are also set as default values in the GIMM software package that can be downloaded from our web page and user is not expected to supply different values for different datasets.

6. Computational complexity and run times.

The Gibbs sampling algorithm described here is fairly computationally complex. The number of mathematical operations for a single iteration is proportional to $(M*T*Q+Q*(r_1*L_1+\ldots+ r_R*L_R))$ where M is the dimension of the global patterns, T is the number of expression profiles being clustered, Q is the average number of global clusters, r_f is the dimensionality of the context f and L_f is the number of local clusters within context f.

To achieve the precision of the analysis presented in this paper, it suffice to run 20,000 iterations of the Gibbs sampler. First 10,000 are discarded as "burn-in" and second 10,000 are used to calculate posterior pairwise probabilities of co-expression. Computing times will obviously depend on the capacity of the computing platform. We timed the execution times for several scenarios using the code compiled with Intel C++ compiler running on the dual 3.6 GHz Xeon machine under Suse Linux 9.2. It took 182 minutes for the Gibbs sampler to generate 20,000 samples on the full datasets (5685 genes across 4 contexts with total of 31 experiments). We also recorded execution times on smaller problems after applying a traditional "variation filter". It took 44 minutes to cluster 2842 most variable genes (top 50%), 28 minutes for 1421 (top 25%) most variable genes, and only 10 minutes to cluster 569 (top 10%) most variable genes. On the other hand, the execution time for all 5685 genes on just sporulation data (2 contexts with total of 15 experiments) took 114 minutes.

For the reasons unclear to us, execution times on the 3GHz Xeon Windows boxes with the code compiled using the MS Visual C++ compiler were significantly longer (24 hours for the full dataset, and 30 minutes for clustering 569 most variable genes). Even after accounting for the fact that Windows code runs on a single and somewhat slower CPU, the run-times are disproportionally long.

Reference List

1. Medvedovic, M., Yeung, K. Y. & Bumgarner, R. E. (2004) *Bioinformatics.* **20**, 1222-1232.

2. Celeux, G., Hurn, M. & Robert, C. P. (2000) *JASA* **95**, 957-970.

3. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. *et al.* (2002) *Science* **298**, 799-804.

4. Xie, J., Pierce, M., Gailus-Durner, V., Wagner, M., Winter, E. & Vershon, A. K. (1999) *EMBO J.* **18**, 6448-6454.