Additional Files

Additional file 1 — Supplementary Methods

Model Selection for Marginal Distribution f: ICL-BIC

The Bayesian Information Criterion (BIC) [1] is defined as $BIC = 2\mathcal{L}_X(\hat{\theta}) - |\Theta| \log N$, with $\mathcal{L}_X(\hat{\theta})$ being the log-likelihood of the estimated parameters given the observed data and $|\Theta|$ being the size of the parameter space. The Integrated Completed Likelihood (ICL) - BIC is defined as $ICL - BIC = 2\mathcal{L}_{X,\hat{Y}}(\hat{\theta}) - |\Theta| \log N$, with $\hat{Y}$ being the maximum a posteriori (MAP) estimate of the value of the hidden data. All other things being equal, the model with the higher (often "less negative") $ICL - BIC$ is preferred. See [2] for an application of this criterion to models of gene expression, and [3] for a comparison to other model selection criteria, where $ICL - BIC$ outperforms AIC (Akaike's information criterion), BIC, and other criteria in selecting the correct mixture model.

Selection of Weights w^{trn}

The modeling of the observed data is the same as in the unsupervised case. By default, labeled samples are given the same weight as unlabeled samples in the parameter estimations. However, if we have a small training sample, we may choose to assign a higher weight w^{trn} to labeled samples. For further (M-step) calculations involving the posterior probabilities calculated in Equation (7) from the main text, we make the transformation $w_{n,y} \leftarrow w^{trn} w_{n,y}$ for each n such that $t_n \leq K$, while leaving $w_{n,y}$ as-is for each n such that $t_n = K^{trn}$. The effective result of this is to add "copies" of the labeled samples to the data set, thus increasing their influence on parameter estimation. For example, if we choose $w^{trn} = 2$, we are effectively doubling the size of the training data.

Although values of $w^{trn} > 1$ often lead to better parameter estimates and therefore to better model predictions, overfitting can occur if w^{trn} grows too large. We use Monte Carlo cross-validation [4] to choose the best value from a list of candidate values, currently $w^{trn} \in \{1, 5, 10, 20, 50, 100\}$. For a specified number of replications, currently 30, we sample without replacement half the training data, leaving the other half to serve as testing data for the current replication. We then train the model with the first half of the data at each candidate weight, and calculate the receiver operating characteristic (ROC) area under the curve (AUC) for the trained models at each candidate weight. The weight with the highest mean ROC-AUC across all replications is chosen as the final value of w^{trn}.

EM Algorithm for Hierarchical Mixture Model

Details for the estimation of parameters and conditional probabilities for the hidden variables are provided in this section. The unconditional status probability is $p_{0,y_0} = P(Y_0 = y_0)$, where Y_0 generates the distribution for the Y_z's, and the component probability given the status is $q_{z,y_0,y_z} = P(Y_z = y_z | Y_0 = y_0)$. Given observed data $\vec{X} = (\vec{X}_1, \ldots, \vec{X}_z)$ where $\vec{X}_z = (\vec{x}_{z,1}, \ldots, \vec{x}_{z,n})$, and parameters $\theta^{(i-1)}$, denote the conditional probabilities for the hidden variables by

$$\begin{aligned} u_{n,y_0} &= P(y_{0,n} = y_0 | \vec{x}_{\cdot,n}, \theta^{(i-1)}), \\ v_{z,y_0,n,y_z} &= P(y_{0,n} = y_0, y_{z,n} = y_z | \vec{x}_{\cdot,n}, \theta^{(i-1)}) \text{ or} \\ w_{z,n,y_z} &= P(y_{z,n} = y_z | \vec{x}_{\cdot,n}, \theta^{(i-1)}). \end{aligned} \tag{1}$$

Let $I(\mathcal{P})$ denote the indicator function. Then for $\vec{X}$ as above, and hidden data $\vec{Y} = (\vec{y}_1, \ldots, \vec{y}_Z)$ where $\vec{y}_z = (y_{z,1}, \ldots, y_{z,n})$ with $\vec{y}_0 = (y_{0,1}, \ldots, y_{0,n})$, the complete data log-likelihood is

$$\begin{aligned} \mathcal{L}_{X,Y,y_0}(\theta) \quad = \quad & \textstyle\sum_{n,k_0} I(y_{0,n} = k_0) \log p_{0,k_0} \\ & + \textstyle\sum_{n,z,(k),k_z} (I) \log q_{z,(k),k_z} \\ & + \textstyle\sum_{n,z,k_z} I(y_{z,n} = k_z) \log f_{k_z}(x_{z,n}|\theta). \end{aligned} \tag{2}$$

where (k) denotes k_0 and (I) denotes $I(y_{0,n} = k_0, y_{z,n} = k_z)$. The Q-function is thus

$$\begin{aligned} Q(\theta|\theta^{(i-1)}) \quad = \quad & \textstyle\sum_{n,k_0} u_{n,k_0} \log p_{0,k_0} \\ & + \textstyle\sum_{n,z,(k),k_z} v_{z,(k),n,k_z} \log q_{z,(k),k_z} \\ & + \textstyle\sum_{n,z,k_z} w_{z,n,k_z} \log f_{k_z}(\vec{x}_{z,n}|\theta^{(i-1)}). \end{aligned} \tag{3}$$

The first step in joint model fitting is to fit a single mixture model to each data source, as described in the Methods section of the main text, to choose the number of components K_z and marginal distribution that will be used for that data. Then the initialization, execution, and output of the EM algorithm as adapted for the model topologies are as follows:

1 Initialize the parameters for the hierarchical model based on the selected individual mixture models. Note that the individual model fits are used for initialization only, and do not imply any hard categorization of the observed data before fitting the hierarchical model.

2 E-step: for the ith iteration, using the previous iteration's parameter estimates $\theta^{(i-1)}$, estimate the conditional probabilities defined in Equation (1), which are

$$\begin{aligned} u_{n,y_0} \quad &= \quad \frac{p_{y_0}^{(i-1)} \prod_z \sum_{k_z} q_{z,y_0,k_z}^{(i-1)} g_{z,n,k_z}}{\sum_{k_0} p_{k_0}^{(i-1)} \prod_z \sum_{k_z} q_{z,k_0,k_z}^{(i-1)} g_{z,n,k_z}}, \\ v_{z,y_0,n,y_z} \quad &= \quad u_{n,y_0} \frac{q_{y_0,y_z}^{(i-1)} g_{z,n,y_z}}{\sum_{k_z} q_{z,y_0,k_z}^{(i-1)} g_{z,n,k_z}}, \\ w_{z,n,y_z} \quad &= \quad \textstyle\sum_{k_0} v_{z,k_0,n,y_z} \end{aligned} \tag{4}$$

where $g_{z,n,y_z} = f_{y_z}(\vec{x}_{z,n}|\theta^{(i-1)})$.

3 M-step: estimate the current iteration's parameters, $\theta^{(i)} = \arg\max_\theta Q(\theta|\theta^{(i-1)})$. This is a straightforward maximum likelihood estimation for the p's and q's, and a weighted MLE for the parameters relating to the observed variables, using weights $\vec{w}_{z,\cdot,y_z}$ and data $\vec{X}_z$.

4 Repeat steps 2 and 3 until convergence.

5 Report the final estimated parameters $\hat{\theta}$ and posterior status probabilities $\vec{\hat{U}}$, the $N \times K_0$ matrix of which the (n, y_0)th element is $\hat{u}_{n,y_0} = P(y_{0,n} = y_0|\vec{x}_n, \hat{\theta})$. Specifically, $\hat{u}_{n,1}$ is the estimated probability, given the data and the final estimated parameters, that the nth gene is a gene of interest.

EM Algorithm for Semi-Supervised Hierarchical Mixture Model
The conditional probabilities for the hidden variables in the hierarchical semi-supervised models [5] are

$$\begin{aligned} u_{n,y_0} &= P(y_{0,n} = y_0 | t_n, \vec{x}_{\cdot,n}, \theta^{(i-1)}), \\ v_{z,y_0,n,y_z} &= P(y_{0,n} = y_0, y_{z,n} = y_z | t_n, \vec{x}_{\cdot,n}, \theta^{(i-1)}) \text{ layered}, \\ w_{z,n,y_z} &= P(y_{z,n} = y_z | t_n, \vec{x}_{\cdot,n}, \theta^{(i-1)}). \end{aligned} \tag{5}$$

The Q-function is

$$\begin{aligned} Q(\theta|\theta^{(i-1)}) &= \textstyle\sum_{n,k^{trn}} t'_{n,k^{trn}} \log p^{trn}_{k^{trn}} + \sum_{n,k^{trn},(k)} t'_{n,k^{trn}} u_{n,(k)} \log r^{trn}_{k^{trn},(k)} \\ &\quad + \textstyle\sum_{z,n,(k),k_z} v_{z,(k),n,k_z} \log q_{z,(k),k_z} + \sum_{z,n,k_z} w_{z,n,k_z} \log f_{k_z}(\vec{x}_{z,n}) \end{aligned} \tag{6}$$

where $t'_{n,t}$ is as introduced in Equation (6) in the main text, and (k) denotes k_0. The initial calculation in the E-step for the layered semi-supervised model is

$$u_{n,y_0} = \frac{r^{trn(i-1)}_{t_n,y_0} \prod_z \sum_{k_z} q^{(i-1)}_{z,y_0,k_z} g_{z,n,k_z}}{\sum_{k_0} r^{trn(i-1)}_{t_n,k_0} \prod_z \sum_{k_z} q^{(i-1)}_{z,k_0,k_z} g_{z,n,k_z}}. \tag{7}$$

With the value of u_{n,y_0} in hand, the remaining calculations in the E-step proceed exactly as in the unsupervised model described in [5]. Cross-validation as described above is used to choose a value for w^{trn}.

Author details

References

1. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
2. Ji, Y., Wu, C., Liu, P., Wang, J., Coombes, K.R.: Applications of beta-mixture models in bioinformatics. Bioinformatics **21**(9), 2118–2122 (2005). doi:10.1093/bioinformatics/bti318
3. Viroli, C.: Dimensionally reduced model-based clustering through mixtures of factor mixture analyzers. Journal of Classification **27**, 363–388 (2010)
4. Shao, J.: Linear model selection by cross-validation. Journal of the American Statistical Association **88**(422), 486–494 (1993)
5. Dvorkin, D., Biehs, B., Kechris, K.: A graphical model method for integrating multiple sources of genome-scale data. Statistical Applications in Genetics and Molecular Biology **12**(4), 469–487 (2013)