Supplementary material
for the manuscript

Identification of transcript regulatory patterns in cell differentiation
by Arief Gusnanto, John Paul Gosling, and Christopher Pope
Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom

Contents
1 Theoretical justification for posterior sample size 3
2 What our prior is worth 4
3 Posterior correlation between cell types: additional figures 9
4 Sensitivity analysis: Additional figures 11
  4.1 Symmetric prior $\Psi$ .......................... 12
  4.2 Diagonal prior $\Psi$ .......................... 15
5 Additional figures in simulation 18
6 Analysis using $t$-test 20
7 Identifying relevant genes in more than one cell type 21
8 Analysis on a second dataset (E-GEOD-24759) 22
  8.1 Experiment .................................. 22
  8.2 Prior and constraints ......................... 24
  8.3 Specific genes ............................... 24
9 References 26
From this point onward in this supplementary material, we use the same notation as in the main manuscript. Let $x_{ijk}$ be the log expression of gene $i$, in person $j$, and cell type $k$, with $i = 1, 2, \ldots, n_g = 46713$, $j = 1, 2, \ldots, n_p = 7$, and $k = 1, 2, \ldots, n_t = 6$. Since the analysis is performed independently for each gene, we drop the index $i$ from the notation without a danger of confusion. We denote $x_j \equiv x_{ij}$ as an $n_t$-vector of log expression of gene $i$ in person $j$, across the different cell types, i.e. $x_j = (x_{j1} \ x_{j2} \ldots \ x_{jn_t})^T$. Furthermore, we also denote $X$ as the matrix of log expressions for each gene, where the columns correspond to cell types $k$ and the rows correspond to person $j$, i.e. $X \equiv (x_1 \ x_2 \ldots \ x_{n_p})^T$ or

$$
X = \begin{pmatrix}
  x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\
  x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} \\
  x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \\
  x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} \\
  x_{51} & x_{52} & x_{53} & x_{54} & x_{55} & x_{56} \\
  x_{61} & x_{62} & x_{63} & x_{64} & x_{65} & x_{66} \\
  x_{71} & x_{72} & x_{73} & x_{74} & x_{75} & x_{76}
\end{pmatrix}
$$

We also denote $\bar{x} \equiv (\bar{x}_1 \ \bar{x}_2 \ \ldots \ \bar{x}_{n_t})^T$, where $\bar{x}_j = \frac{1}{n_p} \sum_{k=1}^{n_p} x_{jk}$. One thing that differs from the main text in terms of notation is that we use $p$ in place of $n_t$ in the notation since $p$ can be interpreted as degrees of freedom in different places. We also denote $N(\cdot)$ and $\text{IW}(\cdot)$ as the normal and inverse-Wishart distribution functions, respectively.
1 Theoretical justification for posterior sample size

As with any Bayesian analysis, i.i.d samples from the posterior distribution are generated and inference is based off these realisations of the posterior distribution. There is, of course, the question of how many samples from the posterior should be drawn. One procedure to decide the number of posterior samples is by using a ‘test run’. This entails generating a small number of samples - say 100 samples and record the amount of time it takes to get the total sample. In our case, there is the added complexity that some genes are much harder to sample from than the others, due to the different characteristics of their gene expression data.

As it will be seen in Eq. (4) below, the expectation of the posterior variance-covariance matrix depends mainly on the data. Therefore, the correlation (between cell types) constraints that are imposed can result in large number of proposals being rejected (see Section 2.3 in the main manuscript). Based on the ‘test run’, and balancing this with an acceptable level of the precision for the estimates, we decided to take 1000 accepted posterior samples for each gene. With 1000 posterior samples for each gene, it was estimated the analysis would take about one week in a standard desktop computer, and the mean square error of the estimates are at most 0.00025 as described below.

As Monte Carlo simulation was used to generate the posterior samples, each draw can be analysed and a posterior probability can be approximated by observing the results of the full sample. For example, with each draw, it can be observed whether a cell type has higher (or lower) mean expression than the other cell types, which can either be true or false, and define this multidimensional space to be \( A \). Then if a sample lies in this space, then it can be counted and given the number 1, otherwise it is assigned the number 0 and not counted. If the \( z \)-th posterior sample from each gene is denoted as \( Y^{(z)} \), then an indicator function is defined as

\[
\mathbb{1}_A(Y^{(z)}) = \begin{cases} 
1 & \text{if } Y^{(z)} \in A \\
0 & \text{if } Y^{(z)} \notin A
\end{cases}
\]

By generating a large number of posterior samples and using the indicator function, an estimate for \( P(\mathbf{Y} \in A) \) can be found (given that the \( Y^{(z)} \) are i.i.d samples of \( \mathbf{Y} \)). In a more formal context, a Monte Carlo estimate of \( P(\mathbf{Y} \in A) \) can be defined as

\[
Z^{MC} = \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} \mathbb{1}_A(Y^{(z)}).
\]

The above Equation 2 is an approximation due to the strong law of large numbers as

\[
P(\mathbf{Y} \in A) = \lim_{n_{\text{post}} \to \infty} \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} \mathbb{1}_A(Y^{(z)}) \approx \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} \mathbb{1}_A(Y^{(z)}).
\]

As it is not possible to generate infinite samples, only an approximate of \( P(\mathbf{Y} \in A) \) can be obtained. Intuitively, as more samples are generated, the approximation is getting better and this can be shown mathematically for the analysis. Firstly, it can be shown that the Monte Carlo estimate \( Z^{MC} \) is unbiased:

\[
E(Z^{MC}) = E \left( \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} \mathbb{1}_A(Y^{(z)}) \right)
= \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} E \left( \mathbb{1}_A(Y^{(z)}) \right)
= \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} E(\mathbb{1}_A(Y))
= E(\mathbb{1}_A(Y))
\]
\[ E(Z^{MC}) = E(1_A(Y)) \]  \hspace{1cm} (3)

Secondly, once we obtain that the estimates are unbiased, our next question would be how accurate the estimates are. A common measure used is the mean squared error (MSE), which for a general estimator \( \theta^* \) is defined as:

\[ \text{MSE}(\theta^*) = \text{Var}(\theta^*) + \text{Bias}(\theta^*). \]

The MSE of the estimates can therefore be written as

\[
\text{MSE}(Z^{MC}) = \text{Var}(Z^{MC}) \\
= \text{Var}\left(\frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} 1_A(Y^{(z)})\right) \\
= \frac{1}{n_{\text{post}}^2} \sum_{z=1}^{n_{\text{post}}} \text{Var}(1_A(Y^{(z)}) \\
= \frac{1}{n_{\text{post}}^2} \sum_{z=1}^{n_{\text{post}}} \left[ E(1_A(Y^{(z)})^2) - E(1_A(Y^{(z)}))^2 \right] \\
= \frac{1}{n_{\text{post}}^2} \sum_{z=1}^{n_{\text{post}}} \left[ E(1_A(Y)) - E(1_A(Y))^2 \right] \\
= \frac{1}{n_{\text{post}}} E(1_A(Y)) \left[ 1 - E(1_A(Y)) \right].
\]

As \( 1_A(Y^{(z)}) \) can only take the values 0 and 1 \( \implies 1_A(Y^{(z)}) = 1_A(Y^{(z)})^2 \)

\[
\text{MSE}(Z^{MC}) = \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} \left[ E(1_A(Y)) - E(1_A(Y))^2 \right] \\
= \frac{1}{n_{\text{post}}} \left[ E(1_A(Y)) - E(1_A(Y))^2 \right] \\
= \frac{1}{n_{\text{post}}} E(1_A(Y)) \left[ 1 - E(1_A(Y)) \right].
\]

In the above equation, it can easily be seen that the function

\[ f(Y) = E(1_A(Y)) \left[ 1 - E(1_A(Y)) \right] \]

is bounded. The maximum of this function is achieved when \( E(1_A(Y)) = 0.5 \), which gives \( \text{Max}(f(Y)) = 0.25 \). By choosing \( n_{\text{post}} = 1000 \), it can therefore be calculated that

\[ \text{MSE}(Z^{MC}) \leq \frac{0.25}{1000} = 0.00025. \]

2 What our prior is worth

An important consideration to contemplate when using an informative prior, is how much influence the prior information is having on the posterior inference. Specifically, we consider how much effect the prior information in \( \Sigma \) has on the posterior inference of \( \Sigma|x_j \). To do this, we can identify the expression for \( E(\Sigma|x_j) \) relative to \( E(\Sigma) = \frac{\Psi}{\nu-\nu-1} \) (see the main text).

Considering the derivation of the posterior probability described in the above section, we can write

\[
E(\Sigma|x_j) = \frac{\Psi + \frac{c_{np}}{n_p+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x}) (x_j - \bar{x})^T}{\nu + p - n_p - 1} \\
= \frac{E(\Sigma)(\nu - p - 1) + \frac{c_{np}}{n_p+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x}) (x_j - \bar{x})^T}{\nu + n_p - p - 1}.
\]

As described in the main text, \( \nu \) is defined to be 10 in all priors, \( c \) is defined to be 1 for all priors, \( n_p \) is known for all genes to be 7 and \( p \) (or \( n_t \)) is known for all genes to be 6. This equation can be
simplified and the influence of the prior expectation of $\Sigma$ on the posterior inference of the expectation of $\Sigma|x_j$ can be written as

$$E(\Sigma|x_j) = \frac{3E(\Sigma) + \frac{7}{8} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_j (x_j - \bar{x}) (x_j - \bar{x})^T}{10}.$$  \hspace{1cm} (4)

Now, the prior worth of the variance in $\Sigma$ (the diagonal elements denoted as $\Sigma^{kk}$) can be calculated. From the main text, we have $\text{Var}(\Sigma^{kk}) = \frac{\frac{2}{\nu} \Psi_{kk}^2}{(\nu - p - 1)^2 (\nu - p - 2)}$. The posterior parameters that were derived above can be used to find the posterior variance of $\Sigma^{kk}$ and the influence of the prior can be derived. From the derivation near the end of this section, it is calculated that

$$\text{Var}(\Sigma^{kk}|x_j) = \frac{2 \left( \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right)_{kk} + \left[ \sum_j (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk}}{(\nu + n_p - p - 1)^2 (\nu + n_p - p - 3)}.$$  \hspace{1cm} (5)

The above equation can be simplified further as

$$\text{Var}(\Sigma^{kk}|x_j) = \text{Var}(\Sigma^{kk}) \left[ \frac{9}{800} + \frac{2 \left( \frac{7}{8} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right)_{kk} + \left[ \sum_j (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk}}{800 \Psi_{kk}} \right] \frac{\frac{2}{\nu} \Psi_{kk}^2}{(\nu - p - 1)^2 (\nu - p - 2)}.$$  \hspace{1cm} (6)

Now, the effect of the prior on the posterior inference of $\mu | \Sigma$ can also be investigated; it is again worth noting that $\Sigma$ is assumed to be known. From the specification of prior distribution and derivation of the posterior distribution above, the expectation of the prior distribution is given by $E(\mu|\Sigma) = \mu^*$. We can see the effect of the prior distribution on the posterior expectation as

$$E(\mu|\Sigma, x_j) = \frac{c \mu^* + n_p \bar{x}}{n_p + c} = \frac{c E(\mu|\Sigma)}{n_p + c} + \frac{n_p \bar{x}}{n_p + c}.$$  \hspace{1cm} (7)

From the main text, $c$ is defined to be 1 in all priors and $n_p$ is known for all genes to be 7; therefore, this equation can be simplified as

$$E(\mu|\Sigma, x_j) = \frac{E(\mu|\Sigma)}{8} + \frac{7 \bar{x}}{8}.$$  \hspace{1cm} (8)

for all genes.

The effect that the prior has on the posterior variance of $\mu | \Sigma$ can also be calculated. Remembering that the prior variance is $\text{Var}(\mu|\Sigma) = \frac{1}{c} \Sigma$, then the posterior variance is calculated as

$$\text{Var}(\mu|\Sigma, x_j) = \frac{1}{n_p + c} \Sigma = \frac{c}{c + n_p} \text{Var}(\mu|\Sigma).$$  \hspace{1cm} (9)
Again, in the main text $c$ is defined to be 1 in all priors and $n_p$ is known for all genes to be 7; this equation can therefore be simplified as

$$\text{Var} (\mu | \Sigma, x_j) = \frac{\text{Var} (\mu | \Sigma)}{8}$$  \hspace{1cm} (7)$$

for all genes. In Eqs. 6 and 7, the natural interpretation for $c$ can be easily seen. This is especially noticeable in equation 6, where it can be explicitly seen that our prior expectation is scaled by $\frac{1}{8}$ and the information from the data are scaled by $\frac{7}{8}$.

The derivation of the posterior of $\text{Var} (\Sigma | x_j)$ in the Eq. (5) above can be described as follows. We have

$$E (\Sigma | x_j) = \frac{\Psi + \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}{\nu + n_p - p - 1}$$

$$= \frac{\nu + n_p - p - 1}{\nu} + \frac{\frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}{\nu + n_p - p - 1}$$

The first term can be rearranged to give:

$$\frac{\nu}{\nu + n_p - p - 1} = \frac{(\nu + n_p - p - 1) + n_p}{(\nu + n_p - p - 1)}$$

Putting this form back into the the original equation gives:

$$E (\Sigma | x_j) = \frac{E (\Sigma)(\nu - p - 1)}{\nu + n_p - p - 1} + \frac{\frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}{\nu + n_p - p - 1}$$

$$= \frac{E (\Sigma)(\nu - p - 1) + \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}{\nu + n_p - p - 1}$$

Therefore, as $\nu$ is defined to be 10 in all priors, $c$ is defined to be 1 for all priors, $n_p$ is known for all genes to be 7, $p$ is known for all genes to be 6; this equation can be simplified and the influence of the prior expectation of $\Sigma$ on the posterior inference of the expectation of $\Sigma | x_j$ can be seen in Eq. (4).

$$E (\Sigma | x_j) = \frac{3E (\Sigma)}{10} + \frac{7}{8} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T$$  \hspace{1cm} (8)$$

Now, the prior worth of the variance of the diagonals of the covariance matrix (which can be denoted as $\Sigma^{(kk)}$) can be calculated. From the last part of Eq. (7) in the main text, the variance of the diagonals is $\text{Var}(\Sigma^{(kk)}) = \frac{2\Psi^2}{(\nu - p - 1)(\nu - p - 3)}$. Bearing this in mind, the posterior parameters that were described in Section 2.6 of the main text can be used to find the posterior variance of $\Sigma^{(kk)}$ and the influence of the prior can be derived.

Specifically,

$$\text{Var} (\Sigma^{(kk)} | x_j) = \frac{2 \left( \Psi_{kk} + \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right)_{kk} + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}_{(\nu + n_p - p - 1)^2(\nu + n_p - p - 3)}$$

$$= \frac{2\Psi_{kk} \left( \Psi_{kk} + \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right)_{kk} + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}_{(\nu + n_p - p - 1)^2(\nu + n_p - p - 3)}$$

$$+ \frac{2 \left( \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right)_{kk} + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}_{(\nu + n_p - p - 1)^2(\nu + n_p - p - 3)}$$

$$+ \frac{2 \left( \frac{c n_p}{n_p + c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right)_{kk} + \sum_{j=1}^{n_p} (x_j - \bar{x})(x_j - \bar{x})^T}_{(\nu + n_p - p - 1)^2(\nu + n_p - p - 3)}$$
The denominator of this large equation can be written in a form that will help to identify the effect of the prior:

\[(\nu + n_p - p - 1)^2(\nu + n_p - p - 3)\]
\[= (\nu - p - 1)^2 + 2\nu n_p - 2n_p^2 - 2n_p p (\nu + n_p - p - 3)\]
\[= (\nu - p - 1)^2(\nu + n_p - p - 3) + (\nu - p - 1)^2 n_p + (\nu + n_p - p - 3) (2\nu n_p - 2n_p - n_p^2 - 2n_p p)\]
\[= (\nu - p - 1)^2(\nu - p - 3) \left[ 1 + \frac{(\nu - p - 1)^2 n_p + (\nu + n_p - p - 3) (2\nu n_p - 2n_p - n_p^2 - 2n_p p)}{(\nu - p - 1)^2(\nu - p - 3)} \right]\]

The first term of this denominator can be used to identify the influence of the prior variance. Now, returning to the \(V ar(\Sigma^{(kk)}| \mathbf{x}_j)\) equation, the first term can be rearranged to give:

\[
2\Psi_{kk} \left( \Psi_{kk} + \left[ \frac{cn_p}{np+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right]_{kk} + \left[ \sum_j^n p (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk} \right)
\]
\[
= \frac{2\Psi_{kk}^2 + 2\Psi_{kk} \left( \left[ \frac{cn_p}{np+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right]_{kk} + \left[ \sum_j^n p (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk} \right)}{(\nu - p - 1)^2(\nu - p - 3) \left[ 1 + \frac{n_p((\nu - p - 1)^2 + (\nu + n_p - p - 3) (2\nu - 2 + n_p - 2p))}{(\nu - p - 1)^2(\nu - p - 3)} \right]}
\]
\[
= \frac{\text{Var}(\Sigma^{(kk)})}{1 + \frac{n_p((\nu - p - 1)^2 + (\nu + n_p - p - 3) (2\nu - 2 + n_p - 2p))}{(\nu - p - 1)^2(\nu - p - 3)}}
\]
\[
+ \frac{2\Psi_{kk} \left( \left[ \frac{cn_p}{np+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right]_{kk} + \left[ \sum_j^n p (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk} \right)}{(\nu - p - 1)^2(\nu - p - 3) \left[ 1 + \frac{n_p((\nu - p - 1)^2 + (\nu + n_p - p - 3) (2\nu - 2 + n_p - 2p))}{(\nu - p - 1)^2(\nu - p - 3)} \right]}
\]

Now, the last part of this term can be added to the second term of the \(V ar(\Sigma^{(kk)}| \mathbf{x}_j)\) equation, giving:

\[
\frac{4\Psi_{kk} \left( \left[ \frac{cn_p}{np+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right]_{kk} + \left[ \sum_j^n p (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk} \right)}{(\nu - p - 1)^2(\nu - p - 3) \left[ 1 + \frac{n_p((\nu - p - 1)^2 + (\nu + n_p - p - 3) (2\nu - 2 + n_p - 2p))}{(\nu - p - 1)^2(\nu - p - 3)} \right]}
\]
\[
= \frac{4\Psi_{kk} \left( \left[ \frac{cn_p}{np+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right]_{kk} + \left[ \sum_j^n p (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk} \right)}{(\nu - p - 1)^2(\nu - p - 3) \left[ 1 + \frac{n_p((\nu - p - 1)^2 + (\nu + n_p - p - 3) (2\nu - 2 + n_p - 2p))}{(\nu - p - 1)^2(\nu - p - 3)} \right]}
\]
\[
= \frac{2\text{Var}(\Sigma^{(kk)}) \left( \left[ \frac{cn_p}{np+c} (\mu^* - \bar{x}) (\mu^* - \bar{x})^T \right]_{kk} + \left[ \sum_j^n p (x_j - \bar{x}) (x_j - \bar{x})^T \right]_{kk} \right)}{(\nu - p - 1)^2(\nu - p - 3) \left[ 1 + \frac{n_p((\nu - p - 1)^2 + (\nu + n_p - p - 3) (2\nu - 2 + n_p - 2p))}{(\nu - p - 1)^2(\nu - p - 3)} \right]}
\]
The terms can now all be combined to give a posterior variance equation:

\[
\text{Var}(\Sigma^{(kk)}|x_j) = \frac{\text{Var}(\Sigma^{(kk)})}{1 + \frac{n_p(\nu-p-1)^2+(\nu+n_p-p-3)(2\nu-2+n_p-2p)}{(\nu-p-1)^2(\nu-p-3)} + 2\frac{\text{Var}(\Sigma^{(kk)})}{\Psi_{kk} 1 + \frac{n_p(\nu-p-1)^2+(\nu+n_p-p-3)(2\nu-2+n_p-2p)}{(\nu-p-1)^2(\nu-p-3)}} \left( \frac{\sum_j n_p (x_j - \bar{x})(x_j - \bar{x})^T}{\nu + n_p - p - 1} \right)}
\]

Now, as \(\nu\) is defined to be 10 in all priors, \(c\) is defined to be 1 for all priors, \(n_p\) is known for all genes to be 7, \(p\) is known for all genes to be 6; this equation can be simplified and the influence of the prior variance of the diagonal elements of \(\Sigma\) on the posterior inference of the variance of the diagonals of \(\Sigma|x_j\) can be seen in equation 9.

\[
\text{Var}(\Sigma^{(kk)}|x_j) = \frac{9}{800} + \frac{2}{\Psi_{kk}} \left( \frac{\sum_j n_p (x_j - \bar{x})(x_j - \bar{x})^T}{\nu + n_p - p - 1} \right)^2.
\]
3 Posterior correlation between cell types: additional figures

In this section, we present some additional figures for the posterior correlation between cell types in genes SLC46A2 and CYFIP2 illustrated in Figure 4 in the main text. These figures are to describe the relationship between posterior correlation between cell types previously presented in Figure 3 in the main text.

Figure S1: Comparison of posterior correlation $\rho_{ab}$ between some cell types for gene SLC46A2. The indices $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{51}$ means the correlation between CD66b and CD14.
Figure S2: Comparison of posterior correlation $\rho_{ab}$ between some cell types for gene CYFIP2. The indices $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{51}$ means the correlation between CD66b and CD14.
4 Sensitivity analysis: Additional figures

In the main analysis, the definition of $\Psi$ is given in Eq. (8) in the main text as

$$\Psi = \begin{pmatrix}
16 & 11 & 10 & 12 & 15 & 10 \\
11 & 16 & 13 & 13 & 11 & 13 \\
10 & 13 & 16 & 12 & 10 & 15 \\
12 & 13 & 12 & 16 & 12 & 12 \\
15 & 11 & 10 & 12 & 16 & 10 \\
10 & 13 & 15 & 12 & 10 & 16 \\
\end{pmatrix}$$

to reflect the path of cell differentiation in Figure 1. The posterior mean $\mu|\Sigma, x_j$ based on this prior are presented in Figure 3 in the main text.

In the sensitivity analysis, we are interested to know the impact of specifying different prior $\Psi$ to the posterior $\Sigma|x_j$ and $\mu|\Sigma, x_j$. We consider two different structure of $\Psi$. The first one is a symmetric matrix

$$\Psi = \begin{pmatrix}
16 & 14 & 14 & 14 & 14 & 14 \\
14 & 16 & 14 & 14 & 14 & 14 \\
14 & 14 & 16 & 14 & 14 & 14 \\
14 & 14 & 14 & 16 & 14 & 14 \\
14 & 14 & 14 & 14 & 16 & 14 \\
14 & 14 & 14 & 14 & 14 & 16 \\
\end{pmatrix}$$,

which means that there are dependencies between cell types (in the context of prior), but not in the structure as the haematopoiesis cell differentiation in Figure 1 of the main text. The second one is a diagonal matrix

$$\Psi = \begin{pmatrix}
16 & 0 & 0 & 0 & 0 & 0 \\
0 & 16 & 0 & 0 & 0 & 0 \\
0 & 0 & 16 & 0 & 0 & 0 \\
0 & 0 & 0 & 16 & 0 & 0 \\
0 & 0 & 0 & 0 & 16 & 0 \\
0 & 0 & 0 & 0 & 0 & 16 \\
\end{pmatrix},$$

which means that cell types are assumed independent a priori.
4.1 Symmetric prior $\Psi$

Figure S3: Posterior samples of mean $\mu|\Sigma, x_j$ for gene SLC46A2 (top panel) and gene CYFIP2 (third panel), and posterior correlation between cell types $\rho_{ab}|x_j$ for those genes in second and bottom panels under a symmetric variance-covariance prior $\Psi$. The index $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{61}|x_j$ means the posterior correlation between CD8 and CD14.
Figure S4: Comparison of posterior correlation $\rho_{ab}$ between some cell types for gene SLC46A2 under a symmetric variance-covariance prior $\Psi$. The indices $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{51}$ means the correlation between CD66b and CD14.
Figure S5: Comparison of posterior correlation $\rho_{ab}$ between some cell types for gene CYFIP2 under a symmetric variance-covariance prior $\Psi$. The indices $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{51}$ means the correlation between CD66b and CD14.
4.2 Diagonal prior $\Psi$

Figure S6: Posterior samples of mean $\mu \mid \Sigma, x_j$ for gene SLC46A2 (top panel) and gene CYFIP2 (third panel), and posterior correlation between cell types $\rho_{ab} \mid x_j$ for those genes in second and bottom panels under a diagonal variance-covariance prior $\Psi$. The index $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{61} \mid x_j$ means the posterior correlation between CD8 and CD14.
Figure S7: Comparison of posterior correlation $\rho_{ab}$ between some cell types for gene SLC46A2 under a diagonal variance-covariance prior $\Psi$. The indices $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{51}$ means the correlation between CD66b and CD14.
Figure S8: Comparison of posterior correlation $\rho_{ab}$ between some cell types for gene CYFIP2 under a diagonal variance-covariance prior $\Psi$. The indices $a$ and $b$ in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, $\rho_{51}$ means the correlation between CD66b and CD14.
5 Additional figures in simulation

Figure S9: Operating characteristics of the proposed method (solid line) and t-test (dashed line) in the simulation study for low signal where the simulated gene expression are correlated across cell types according to Figure 1 in the main manuscript (simulation A), under equal correlation across cell types (simulation B), and under independence (simulation C). The description of the simulation can be found in the main manuscript.
Figure S10: Operating characteristics of the proposed method (solid line) and $t$-test (dashed line) in the simulation study for high signal where the simulated gene expression are correlated across cell types according to Figure 1 in the main manuscript (simulation A), under equal correlation across cell types (simulation B), and under independence (simulation C). The description of the simulation can be found in the main manuscript.
6 Analysis using \textit{t}-test

The main problem of the \textit{t}-statistics (under some sort of independence assumption between cell types) in identifying specific genes in cell differentiation lies in the multiplicity correction that we have to perform. We have outlined this in the introduction of the main manuscript. With six cell types, 15 comparison between cell types, and 46,713 probes, the multiplicity burden stands at 700,695. With this level of multiplicity burden, we are basically hindered from discovering relevant genes. We have done some analysis using the \textit{t}-statistics and the results in Table S1 and Table S2 reflect that. We could only identify limited number of genes that are specific in CD19 and CD66b, but not in the other cell types. Given this, the \textit{t}-test is inappropriate to be considered and is not a suitable alternative analysis. We include the \textit{t}-test in the simulation study only to show that the proposed method is preferable even in the situation where \textit{t}-test is expected to be advantageous. In the simulation study, we deliberately set the number of genes to be 1,000 so that the \textit{t}-test can gain sensitivity due to the low multiplicity burden. In reality, with tens of thousands of probes tested using the \textit{t}-test, the sensitivity is extremely low as indicated in Table S1 and Table S2.

<table>
<thead>
<tr>
<th>\textit{p}-value</th>
<th>CD14</th>
<th>CD19</th>
<th>CD4</th>
<th>CD56</th>
<th>CD66b</th>
<th>CD8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>0</td>
<td>26</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>0.01</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table S1: \textit{Number of probes identified to be specific in each cell type (higher mean expression compared to the other cell types) by their Bonferroni-adjusted \textit{p}-values from \textit{t}-tests. To be considered as specific in a cell type, a probe must be significant in all comparisons between the cell type in consideration and the other cell types.}

<table>
<thead>
<tr>
<th>\textit{p}-value</th>
<th>CD14</th>
<th>CD19</th>
<th>CD4</th>
<th>CD56</th>
<th>CD66b</th>
<th>CD8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>0.001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table S2: \textit{Number of probes identified to be specific in each cell type (lower mean expression compared to the other cell types) by their Bonferroni-adjusted \textit{p}-values from \textit{t}-tests. To be considered as specific in a cell type, a probe must be significant in all comparisons between the cell type in consideration and the other cell types.}

From Table S1, there are 26 probes that are specific in CD19 (higher expression) at the adjusted \textit{p}-value 0.05. They correspond to 25 genes, out of which 19 of them are known genes identified by DAVID (Huang \textit{et al.}, 2009). The 19 known genes are presented in Table S3. From Table S2, there are also 24 and 18 probes that are specific in CD66b (higher and lower expression, respectively). They correspond to 23 and 16 genes (respectively), out of which 21 and 15 genes are known genes identified by DAVID (Huang \textit{et al.}, 2009). The lists of the known genes are presented in Table S4 and Table S5. All of the above probes are a subset of the list of probes we identified in the main analysis with posterior probability more than 99% in their respective cell types.

In the main manuscript, we present the result of over-representation test on the genes that are specific in each cell type. Since the genes we identified using the \textit{t}-test are a subset of that list in the main manuscript, the genes are still meaningful biologically. We list the functional categories and KEGG pathways of the genes identified by the \textit{t}-test, in term of relevant keywords as output from DAVID (Huang \textit{et al.}, 2009), in several tab-delimited files available for download from the manuscript webpage.
<table>
<thead>
<tr>
<th>ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AFF3 AF4/FMR2 family member 3 (AFF3)</td>
</tr>
<tr>
<td>2</td>
<td>ABCB4 ATP binding cassette subfamily B member 4 (ABCB4)</td>
</tr>
<tr>
<td>3</td>
<td>BCL11A B-cell CLL/lymphoma 11A (BCL11A)</td>
</tr>
<tr>
<td>4</td>
<td>CD22 CD22 molecule (CD22)</td>
</tr>
<tr>
<td>5</td>
<td>CD79B CD79b molecule (CD79B)</td>
</tr>
<tr>
<td>6</td>
<td>E2F5 E2F transcription factor 5 (E2F5)</td>
</tr>
<tr>
<td>7</td>
<td>FCRL1 Fc receptor like 1 (FCRL1)</td>
</tr>
<tr>
<td>8</td>
<td>FCRL2 Fc receptor like 2 (FCRL2)</td>
</tr>
<tr>
<td>9</td>
<td>POU2AF1 POU class 2 associating factor 1 (POU2AF1)</td>
</tr>
<tr>
<td>10</td>
<td>SPIB Spi-B transcription factor (SPIB)</td>
</tr>
<tr>
<td>11</td>
<td>COBLL1 cordon-bleu WH2 repeat protein like 1 (COBLL1)</td>
</tr>
<tr>
<td>12</td>
<td>EBF1 early B-cell factor 1 (EBF1)</td>
</tr>
<tr>
<td>13</td>
<td>FAM129C family with sequence similarity 129 member C (FAM129C)</td>
</tr>
<tr>
<td>14</td>
<td>KLHL14 kelch like family member 14 (KLHL14)</td>
</tr>
<tr>
<td>15</td>
<td>HLA-DOB major histocompatibility complex, class II, DO beta (HLA-DOB)</td>
</tr>
<tr>
<td>16</td>
<td>OSBPL10 oxysterol binding protein like 10 (OSBPL10)</td>
</tr>
<tr>
<td>17</td>
<td>KCNG1 potassium voltage-gated channel modifier subfamily G member 1 (KCNG1)</td>
</tr>
<tr>
<td>18</td>
<td>PNOC prepronociceptin (PNOC)</td>
</tr>
<tr>
<td>19</td>
<td>PTPRK protein tyrosine phosphatase, receptor type K (PTPRK)</td>
</tr>
</tbody>
</table>

Table S3: List of genes identified to be specific in CD19 (higher expression) using the t-test at the adjusted 5% p-value. Out of 26 probes from Table 1, we identified 25 genes and 19 of them are identified by keyword in DAVID tools (Huang et al., 2009).

7 Identifying relevant genes in more than one cell type

Some genes may be involved in driving cell differentiation in two cell types, for example. This involvement may be in two forms: first, a gene has higher (or lower) expression in two cell types relative to the other cell types and, second, a gene has higher expression in one cell type but lower expression in a second cell type relative to the other cell types. The proposed method is able to identify those genes by investigating the posterior mean of expression across the different cell types.

Specifically, for the first form, we are interested in identifying genes that are involved in the differentiation of cell types \( k \) and \( k^* \), \( k = 1, \ldots, n_t, k^* = 1, \ldots, n_t \), and \( k \neq k^* \). For this purpose, Equations (11) and (12) in the main manuscript can be set as follows. The probabilities of a gene to be involved in differentiation of cell types \( k \) and \( k^* \) are given by

\[
p_{k,k^*}^+ = \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} I \left( \mu_k^{(z)} > \mu_{k'}^{(z)} \land \mu_{k^*}^{(z)} > \mu_{k'}^{(z)} \right),
\]

\[
p_{k,k^*}^- = \frac{1}{n_{\text{post}}} \sum_{z=1}^{n_{\text{post}}} I \left( \mu_k^{(z)} < \mu_{k'}^{(z)} \land \mu_{k^*}^{(z)} > \mu_{k'}^{(z)} \right),
\]

for \( k' \in \{1, \ldots, n_t\}, k \neq k' \) and \( k^* \neq k' \). The genes are then identified as those with large values of \( p_{k,k^*}^+ \) or \( p_{k,k^*}^- \).

For the second form, we have identified those as ‘non-specific’ genes as described in Section 3.2 of the main manuscript. An illustration is also given in Figure 5 of the main manuscript, in which gene SP140 is inferred to be involved in the differentiation of CD14 and CD19, while gene RP9 in CD19 and CD66b. Genes that are involved in two cell types, but in opposite direction, can be identified from large values of both \( p_{k}^+ \) and \( p_{k}^- \).
<table>
<thead>
<tr>
<th>ID</th>
<th>Gene Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CMTM2 CKLF like MARVEL transmembrane domain containing 2(CMTM2)</td>
</tr>
<tr>
<td>2</td>
<td>FCGR3B Fc fragment of IgG receptor IIIb(FCGR3B)</td>
</tr>
<tr>
<td>3</td>
<td>MXD1 MAX dimerization protein 1(MXD1)</td>
</tr>
<tr>
<td>4</td>
<td>PINK1 PTEN induced putative kinase 1(PINK1)</td>
</tr>
<tr>
<td>5</td>
<td>B3GNT8 UDP-GlcNAc:betaGal beta-1,3-N-acetylglicosaminyltransferase 8 (B3GNT8)</td>
</tr>
<tr>
<td>6</td>
<td>ALPL alkaline phosphatase, liver/bone/kidney(ALPL)</td>
</tr>
<tr>
<td>7</td>
<td>ASPRV1 aspartic peptidase, retroviral-like 1(ASPRV1)</td>
</tr>
<tr>
<td>8</td>
<td>CA4 carbonic anhydrase 4(CA4)</td>
</tr>
<tr>
<td>9</td>
<td>CEACAM3 carcinoembryonic antigen related cell adhesion molecule 3 (CEACAM3)</td>
</tr>
<tr>
<td>10</td>
<td>DGAT2 diacylglycerol O-acyltransferase 2(DGAT2)</td>
</tr>
<tr>
<td>11</td>
<td>XPO6 exportin 6(XPO6)</td>
</tr>
<tr>
<td>12</td>
<td>FFAR2 free fatty acid receptor 2(FFAR2)</td>
</tr>
<tr>
<td>13</td>
<td>GMFG glia maturation factor gamma(GMFG)</td>
</tr>
<tr>
<td>14</td>
<td>IL1R2 interleukin 1 receptor type 2(IL1R2)</td>
</tr>
<tr>
<td>15</td>
<td>LRG1 leucine rich alpha-2-glycoprotein 1(LRG1)</td>
</tr>
<tr>
<td>16</td>
<td>MMP25 matrix metallopeptidase 25(MMP25)</td>
</tr>
<tr>
<td>17</td>
<td>NOV nephroblastoma overexpressed(NOV)</td>
</tr>
<tr>
<td>18</td>
<td>PANX2 pannexin 2(PANX2)</td>
</tr>
<tr>
<td>19</td>
<td>PI3 peptidase inhibitor 3(PI3)</td>
</tr>
<tr>
<td>20</td>
<td>STX3 syntaxin 3(STX3)</td>
</tr>
<tr>
<td>21</td>
<td>ZDHHC18 zinc finger DHHC-type containing 18(ZDHHC18)</td>
</tr>
</tbody>
</table>

Table S4: List of genes identified to be specific in CD66b (higher expression) using the t-test at the adjusted 5% p-value. Out of 24 probes from Table 2, we identified 23 genes and 19 of them are identified by keyword in DAVID tools (Huang et al., 2009).

In this study, we focus on genes that are specific for each cell type because this is the priority in a study of cell differentiation and is much more simpler to interpret first. When this primary objective has been met, then a further investigation might be to search for genes that are involved in the differentiation of more than one cell type. The above formulation shows the flexibility of the proposed method.

8 Analysis on a second dataset (E-GEOD-24759)

8.1 Experiment

Our second analysis is on the study of haematopoiesis by Novershtern et al. (2011), in which gene expressions from 38 different blood cell types were measured 4-7 individuals. To avoid singularity of the covariance matrix of the data, we consider five blood cell types from six individuals in our analysis: basophil (BAS), erythroblast (ERY), megakaryocyte (MK), naive CD8+ Tc-lymphocyte, and naive CD4 Th-lymphocyte. The gene expressions were obtained from Affymetrix HGU133AAofAv2 microarrays, which contain 22,944 probes. Further details of the experiment are described in Novershtern et al. (2011).

Figure 8.1 presents the haematopoiesis paths where the five cell types are indicated.
<table>
<thead>
<tr>
<th>ID</th>
<th>Gene Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ATIC</td>
<td>5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase (ATIC)</td>
</tr>
<tr>
<td>2</td>
<td>ABCE1</td>
<td>ATP binding cassette subfamily E member 1 (ABCE1)</td>
</tr>
<tr>
<td>3</td>
<td>DDX10</td>
<td>DEAD-box helicase 10 (DDX10)</td>
</tr>
<tr>
<td>4</td>
<td>DPH2</td>
<td>DPH2 homolog (DPH2)</td>
</tr>
<tr>
<td>5</td>
<td>AKR1B1</td>
<td>aldo-keto reductase family 1 member B (AKR1B1)</td>
</tr>
<tr>
<td>6</td>
<td>C12orf45</td>
<td>chromosome 12 open reading frame 45 (C12orf45)</td>
</tr>
<tr>
<td>7</td>
<td>CCDC6</td>
<td>coiled-coil domain containing 6 (CCDC6)</td>
</tr>
<tr>
<td>8</td>
<td>EEF1B2</td>
<td>eukaryotic translation elongation factor 1 beta 2 (EEF1B2)</td>
</tr>
<tr>
<td>9</td>
<td>PEPD</td>
<td>peptidase D (PEPD)</td>
</tr>
<tr>
<td>10</td>
<td>PRDX4</td>
<td>peroxiredoxin 4 (PRDX4)</td>
</tr>
<tr>
<td>11</td>
<td>RPL15</td>
<td>ribosomal protein L15 (RPL15)</td>
</tr>
<tr>
<td>12</td>
<td>SARS2</td>
<td>seryl-tRNA synthetase 2, mitochondrial (SARS2)</td>
</tr>
<tr>
<td>13</td>
<td>TTC27</td>
<td>tetra-tricopeptide repeat domain 27 (TTC27)</td>
</tr>
<tr>
<td>14</td>
<td>TULP4</td>
<td>tubby like protein 4 (TULP4)</td>
</tr>
<tr>
<td>15</td>
<td>VPS45</td>
<td>vacuolar protein sorting 45 homolog (VPS45)</td>
</tr>
</tbody>
</table>

**Table S5:** List of genes identified to be specific in CD66b (lower expression) using the t-test at the adjusted 5% p-value. Out of 18 probes from Table 2, we identified 16 genes and 15 of them are identified by keyword in DAVID tools (Huang et al., 2009).

**Figure S11:** Diagram of the development of different blood cells from haematopoietic stem cell to mature cells. Some cell types that are not involved in the second analysis are omitted from the figure.
8.2 Prior and constraints

From Figure 8.1, the dissimilarity matrix of gene expression between cell types can be considered as

\[
\begin{pmatrix}
BAS & ERY & MK & CD8 & CD4 \\
BAS & 0 & 2 & 2 & 6 & 6 \\
ERY & 2 & 0 & 1 & 5 & 5 \\
MK & 2 & 1 & 0 & 5 & 5 \\
CD8 & 6 & 5 & 5 & 0 & 1 \\
CD4 & 6 & 5 & 5 & 1 & 0
\end{pmatrix}
\]  

(12)

The above matrix suggests the hyper parameter variance covariance matrix \( \Psi \) is defined as

\[
\Psi = \begin{pmatrix}
16 & 14 & 14 & 10 & 10 \\
14 & 16 & 15 & 11 & 11 \\
14 & 15 & 16 & 11 & 11 \\
10 & 11 & 11 & 16 & 15 \\
10 & 11 & 11 & 15 & 16
\end{pmatrix}
\]  

(13)

where the columns (and corresponding rows) are correspond to BAS, ERY, MK, CD8, and CD4.

The constraints on the correlation between cell types are given by

\[0 \leq \{ \rho_{41}, \rho_{51} \} \leq \{ \rho_{42}, \rho_{43}, \rho_{52}, \rho_{53} \} \leq \{ \rho_{21}, \rho_{31} \} \leq \{ \rho_{32}, \rho_{54} \}\]

8.3 Specific genes

In this analysis, we obtain posterior samples which allow us to estimate the probability of a gene to have consistent higher \( (p^+_k) \) or lower \( (p^-_k) \) posterior mean in one cell type compared to the others. The number of genes whose probabilities match and pass different levels of threshold are presented in Table S6 and Table S7, for \( p^+_k \) and \( p^-_k \) respectively. Table S6 indicates that the number of genes that have at least 99% probability to have higher posterior mean in erythroblast is 324, while in megakaryocyte it is 122. Table S7 also indicates that there are 117 genes with at least 99% probability to have lower posterior mean in erythroblast compared to the other cell types.

<table>
<thead>
<tr>
<th>( p^+_k )</th>
<th>BAS</th>
<th>ERY</th>
<th>MK</th>
<th>CD8</th>
<th>CD4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \geq 0.00 )</td>
<td>22944</td>
<td>22944</td>
<td>22944</td>
<td>22944</td>
<td>22944</td>
</tr>
<tr>
<td>( \geq 0.50 )</td>
<td>4011</td>
<td>3373</td>
<td>1476</td>
<td>2981</td>
<td>2426</td>
</tr>
<tr>
<td>( \geq 0.80 )</td>
<td>969</td>
<td>1435</td>
<td>549</td>
<td>637</td>
<td>507</td>
</tr>
<tr>
<td>( \geq 0.90 )</td>
<td>371</td>
<td>937</td>
<td>388</td>
<td>253</td>
<td>207</td>
</tr>
<tr>
<td>( \geq 0.95 )</td>
<td>200</td>
<td>639</td>
<td>270</td>
<td>115</td>
<td>124</td>
</tr>
<tr>
<td>( \geq 0.98 )</td>
<td>128</td>
<td>428</td>
<td>184</td>
<td>53</td>
<td>54</td>
</tr>
<tr>
<td>( \geq 0.99 )</td>
<td>101</td>
<td>324</td>
<td>122</td>
<td>28</td>
<td>32</td>
</tr>
<tr>
<td>( \geq 1.00 )</td>
<td>37</td>
<td>111</td>
<td>28</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Table S6: Number of probes whose \( p^+_k \)’s match and pass different thresholds. \( p^+_k \) is defined as the probability of a gene to have a higher posterior mean in each cell type than the other cell types. The probability for each probe is presented in an Excel file downloadable from the manuscript webpage (see the main text).

The full list of the probability of each probe to be specific in each cell type (higher or lower mean expression) is available in the manuscript webpage (see the main text).
Table S7: Number of probes whose $p^-_k$’s match and pass different thresholds. $p^-_k$ is defined as the probability of a gene to have a higher posterior mean in each cell type than the other cell types. The probability for each probe is presented in an Excel file downloadable from the manuscript webpage (see the main text).

<table>
<thead>
<tr>
<th>$p^-_k$</th>
<th>BAS</th>
<th>ERY</th>
<th>MK</th>
<th>CD8</th>
<th>CD4</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥0.00</td>
<td>22944</td>
<td>22944</td>
<td>22944</td>
<td>22944</td>
<td>22944</td>
</tr>
<tr>
<td>≥0.50</td>
<td>4095</td>
<td>2274</td>
<td>2921</td>
<td>1808</td>
<td>961</td>
</tr>
<tr>
<td>≥0.80</td>
<td>1128</td>
<td>689</td>
<td>693</td>
<td>135</td>
<td>169</td>
</tr>
<tr>
<td>≥0.90</td>
<td>506</td>
<td>430</td>
<td>246</td>
<td>43</td>
<td>75</td>
</tr>
<tr>
<td>≥0.95</td>
<td>220</td>
<td>280</td>
<td>100</td>
<td>17</td>
<td>44</td>
</tr>
<tr>
<td>≥0.98</td>
<td>90</td>
<td>159</td>
<td>35</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>≥0.99</td>
<td>52</td>
<td>117</td>
<td>18</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>≥1.00</td>
<td>4</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
9 References
