Identification of transcript regulatory patterns in cell differentiation

Arief Gusnanto [1]*, John Paul Gosling [1], and Christopher Pope [1]

[1] Department of Statistics, University of Leeds, Leeds LS2 9JT, United Kingdom

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Studying transcript regulatory patterns in cell differentiation is critical in understanding its complex nature of the formation and function of different cell types. This is done usually by measuring gene expression at different stages of the cell differentiation. However, if the gene expression data available are only from the mature cells, we have some challenges in identifying transcript regulatory patterns that govern the cell differentiation.

Results: We propose to exploit the information of the lineage of cell differentiation in terms of correlation structure between cell types. We assume that two different cell types that are close in the lineage will exhibit many common genes that are co-expressed relative to those that are far in the lineage. Current analysis methods tend to ignore this correlation by testing for diffferential expression assuming some sort of independence between cell types. We employ a Bayesian approach to estimate the posterior distribution of the mean of expression in each cell type, by taking into account the cell formation path in the lineage. This enables us to infer genes that are specific in each cell type, indicating the genes are involved in directing the cell differentiation to that particular cell type. We illustrate the method using gene expression data from a study of haematopoiesis.

Availability: R codes to perform the analysis are available in http://www1.maths.leeds.ac.uk/∼arief/R/CellDiff/

Contact: a.gusnanto@leeds.ac.uk

1 INTRODUCTION

Haematopoiesis is a formation of mature blood cells from their precursor stem cells. In the process, a stem cell will experience changes in gene expression and other complex processes that will direct it to a specific mature cell type. As a stem cell matures, it undergoes changes in gene expression that limit the cell types that it can become and moves it closer to a specific cell type. These changes can often be tracked by monitoring the presence of proteins on the surface of the cell, designated as cluster of differentiation (CD) markers (Zola *et al.*, 2005), which we use in this study to identify the different blood cell types. Each successive change moves the cell closer to the final cell type and further limits its potential to become a different cell type.

Some studies have investigated the role of some speficic genes in haematopoiesis. For example, Tanaka *et al.* (2011) reported that genes *ASH1* and *MLL1* regulate the development of myelomonocytic lineages from haematopoietic stem cells. Mancini *et al.* (2012) showed that *FOG1* and *GATA1* are involved in the differentiation between megakaryocytic and erythroid cells. Furthermore, *RUNX1* has been identified to be highly expressed in megakaryocytic cells and supressed in erythroid cells, hence also characterising the lineage between the two (Kuvardina *et al.*, 2015; Draper *et al.*, 2016). Ungerbäck *et al.* (2015) also indicated that genes *EBF1* and *PAX5* play a significant role in the differentiation of the T- and B-lymphocyte cells. These studies are only to name a few; it is therefore of main interest to identify transcript regulatory patterns in cell differentiation to have a global view and understanding of the complex process.

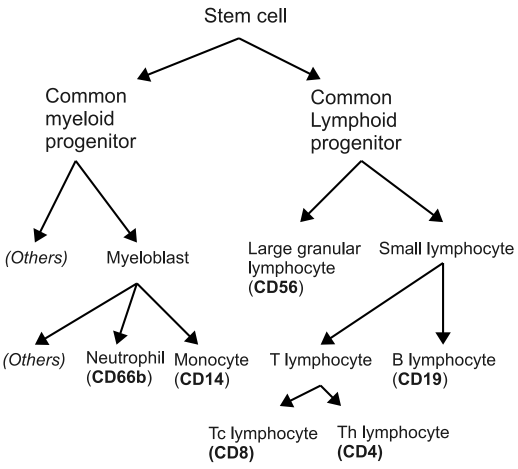

Fig. 1. *Diagram of the development of different blood cells from haematopoietic stem cell to mature cells, identified by their cluster of differentiation (CD) markers (Zola et al., 2005). Some cell types that are not involved in this study are omitted from the figure.*

Figure 1 shows the development of different blood cell types from stem cell to mature cells, omitting other cell types that are not involved in this study. The stem cell differentiates into two progenitor cell types before differentiating further into mature cells. In the process, different genes are involved, either producing more or less mRNA, to direct the cell differentiation. We expect that the

*to whom correspondence should be addressed

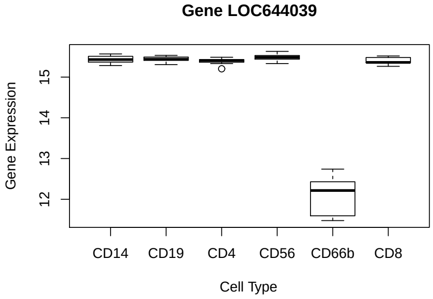

Fig. 2. *Boxplot of expression of gene LOC644039 across different cell types in our study.*

gene expression profiles between two cell types that are close in the lineage will generally be also more correlated than between two cell types that are far in the lineage.

Our objective in this study is to identify genes that are specifically involved in driving the differentiation in each cell type by incorporating the haematopoietic information described in Figure 1. This objective translates into identifying genes that significantly and consistently have higher or lower mean of expressions compared to the other cell types, given the blood-cell formation path. In our analysis, we will assume that the formation path as shown in Figure 1 is fixed and known in advance.

A naïve and inappropriate approach to identifying genes that are specific to each cell types is to perform a pairwise test of mean equality, assuming some sort of independence between cell types, either (1) between a cell type and the averaged expression of the other cell types, or (2) between pairs of cell types in each gene. For example, suppose we are interested in identifying genes that are specific to CD56. In the first approach, we perform a pairwise t-test between CD56 and average expressions of CD14, CD19, CD4, CD66b, and CD8 for each gene. We declare a gene to be specific in CD56 if its multiplicity-adjusted p-value passes a certain threshold. In the second approach, we perform a paired t-test between CD56 and each other cell types, e.g. CD56 vs. CD14, CD56 vs. CD19, CD56 vs. CD4, etc. The genes that are specific to CD56 are identified as those whose multiplicity-adjusted p-values pass a threshold in *all* of the pairwise comparisons.

The above procedures suffer from two problems:

1. In the first procedure, the averaging of gene expression across different cell types can mislead us, as illustrated in Figure 2 for gene LOC644039. When we performed a paired t-test between CD56 and the average of the other cell types, we obtain a significant result (p-value < 0.0001 after false discovery rate correction), suggesting that the gene is specific in CD56. However, Figure 2 clearly does not support this conclusion. The main reason for this significance is because the gene's expression in CD66b cell type brings down the average for non-CD56 cell types. Hence, the gene may appear to be specific to CD 56 when in fact it is not.
2. In the second procedure, the amount of multiplicity involved due to the pairwise comparison will severely restrict us in discovering cell type-specific genes. With six cell types involved in our study, there are 15 pairwise comparisons for each probe. Considering that there are more than 46 thousand probes in the data, the number of hypothesis testing involved is in the order of 700 thousands. In a case where we have 10 cell types to be compared, for example, the total number of hypothesis involved is in the order of 2.1 millions. With this level of hypothesis testing burden, the power to detect specific genes in each cell type will be extremely low, if any at all.

A major drawback that makes the above procedures to be inappropriate is that they ignore the correlation structure between the cell types as indicated in Figure 1. Since statistical tests generally assume independence between observations, the correlation is bypassed to arrive at independent observations. When there are only two cell types to compare, then this is not a problem. However, when there are multiple cell types to compare, we are losing valuable information on the global landscape of gene expression between cell types that direct the cell differentiation.

We believe that the key to solve the problems is to respect and take into account the correlation structure between the cell types. Rather than considering the correlation between the cell types as nuisance, we accommodate it in our proposed model as described in the next section. We consider a Bayesian approach to deal with the problem, which allows us to perform rigorous statistical inference.

2 METHODS

2.1 Samples

Whole blood units from seven healthy donors were obtained at the National Health Service (NHS) Blood and Transplant. Six cell types, CD4 Th lyphocyte, CD8 Tc lymphocyte, CD14 monocyte, CD19 B lymphocyte, CD56 natural killer cells, and CD66 granulocyte, were isolated from each donor. Total RNA were isolated, checked for quality, and amplified. The biotinylated cRNA was applied to Illumina Human WG-6 v2 Expression BeadChips and hybridized overnight. Further details of the samples' preparation are described in Watkins *et al.* (2009). The results of analysis of this dataset are presented mainly in this manuscript with some supporting information available in the Supplementary Material.

In addition to the above dataset, we also consider a second dataset from a study on haematopoiesis by Novershtern *et al.* (2011). The experiment involved 38 blood cell types from 4-7 individuals. For our analysis, we only consider five cell types: Basophyl, CD4 Th-lymphocyte, CD8 Tc-lymphocyte, Erythrocyte, and Megakaryocyte, from six individuals. 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). The results of analysis of this dataset are presented solely in the Supplementary Material.

2.2 Gene expression data and notation

Before we describe the statistical modelling involved, we first describe the notation that we use in this paper. 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 that we will describe later is performed independently for each gene, the index i can safely be dropped from the notation without a danger of confusion. When there is a danger of confusion, we will put the index back in the notation. We denote $\boldsymbol{x}_j \equiv \boldsymbol{x}_{ij}$ as an n_t-vector of log expression of gene i in person j, across the different cell types, i.e. $\boldsymbol{x}_j = (x_{j1} \;\; x_{j2} \;\; \ldots \;\; x_{jn_t})^T$. Furthermore, we also denote $\boldsymbol{X}$ 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. $\boldsymbol{X} \equiv (\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \ldots \;\; \boldsymbol{x}_{n_p})^T$ or

$$\boldsymbol{X} = \begin{array}{c} \\ \text{Person 1} \\ \text{Person 2} \\ \text{Person 3} \\ \text{Person 4} \\ \text{Person 5} \\ \text{Person 6} \\ \text{Person 7} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{CD14} & \text{CD19} & \text{CD4} & \text{CD56} & \text{CD66b} & \text{CD8} \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} \\ x_{51} & x_{52} & x_{53} & x_{54} & x_{55} & x_{56} \\ x_{61} & x_{62} & x_{63} & x_{64} & x_{65} & x_{66} \\ x_{71} & x_{72} & x_{73} & x_{74} & x_{75} & x_{76} \end{pmatrix} \end{array} . \qquad (1)$$

The gene expression data involved in this study basically consist of $n_g =$ 46,713 matrices of $\boldsymbol{X}$ in (1). The expression data have been properly normalised prior to analysis. Given this, we assume that the expressions between genes (between different $\boldsymbol{X}$'s) to be considered independent following Ploner *et al.* (2005). Since we have tens of thousands of genes in our analysis, the departure from this assumption is too weak to have practical importance. Taking the correlation structure between genes is important in some studies e.g. gene network modelling. However, this is a future research topic and is outside the scope of our current study which aims to identify specific genes in cell diferentiation. We also assume that the individuals are independent (because e.g. they are not related genetically). The dependency structure that we take into account in the modelling is the correlation of gene expression between cell types as illustrated in Figure 1. This is described further in the following sections.

2.3 Dissimilarity and correlation between cell types

An advantage of using a Bayesian approach is the ability to encode uncertainty and prior knowledge within an analysis, which, in our case, is the haematopoietic paths in Figure 1. Based on the figure, a dissimilarity between cell types can be defined as the number of split transformations between two cell types. For example, a small lymphocyte splitting into B lymphocyte and T lymphocyte is considered a split transformation and would count as a dissimilarity. Based on this definition and Figure 1, we obtain the following dissimilarity matrix

$$\begin{array}{c} \\ \text{CD14} \\ \text{CD19} \\ \text{CD4} \\ \text{CD56} \\ \text{CD66b} \\ \text{CD8} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{CD14} & \text{CD19} & \text{CD4} & \text{CD56} & \text{CD66b} & \text{CD8} \end{array} \\ \begin{pmatrix} 0 & & & & & \\ 5 & 0 & & & & \\ 6 & 2 & 0 & & & \\ 4 & 2 & 3 & 0 & & \\ 1 & 5 & 6 & 4 & 0 & \\ 6 & 2 & 1 & 3 & 6 & 0 \end{pmatrix} \end{array} . \qquad (2)$$

From Eq. (2), cell types that have a small dissimilarity are expected to be much more correlated than the cell types that have a large dissimilarity. With this in mind, we can obtain a relationship of correlation between the different cell types which we will take into account in the inference as described in Section 2.7. Denoting ρ_{ab} as the correlation between cell type a and b, the correlation structure that would be expected based on Figure 1 is

$$\begin{aligned} 0 \le \{\rho_{13}, \rho_{16}, \rho_{53}, \rho_{56}\} < \{\rho_{12}, \rho_{52}\} < \{\rho_{14}, \rho_{45}\} < \\ \{\rho_{34}, \rho_{46}\} < \{\rho_{23}, \rho_{26}, \rho_{24}\} < \{\rho_{15}, \rho_{36}\} \end{aligned} \qquad (3)$$

where the indices $a, b \in \{1, \ldots, 6\}$ correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively.

2.4 Bayesian modelling

In general context, Bayesian modelling can be described briefly as follows. We denote the data we observe as $\boldsymbol{x}$, and they are assumed to come from a model with parameter θ. The probability density for the data given θ is denoted as $\pi(\boldsymbol{x}|\theta)$ and is proportional to the likelihood. Our uncertainty or belief held about the parameter θ (before any data are seen) is called the prior probability density and denoted as $\pi(\theta)$. As an inference, our interest is in the posterior probability of the parameter given the data denoted $\pi(\theta|\boldsymbol{x})$. Using Bayes's theorem, this is given by

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\boldsymbol{x}|\theta)\pi(\theta). \qquad (4)$$

To proceed with the Bayesian analysis in our study, we model, for each gene, the n_t-vector of gene expression in person j across different cell types as

$$\boldsymbol{x}_j|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (5)$$

where MVN$(\cdot)$ is a multivariate normal distribution function, with mean $\boldsymbol{\mu}$ (an n_t-vector) and variance-covariance matrix $\boldsymbol{\Sigma}$ of size $n_t \times n_t$, which can be written as (for simplicity, only the diagonal and lower triangular elements are printed)

$$\begin{bmatrix} \sigma_1^2 & & & & & \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & & & & \\ \rho_{31}\sigma_3\sigma_1 & \rho_{32}\sigma_3\sigma_2 & \sigma_3^2 & & & \\ \rho_{41}\sigma_4\sigma_1 & \rho_{42}\sigma_4\sigma_2 & \rho_{43}\sigma_4\sigma_3 & \sigma_4^2 & & \\ \rho_{51}\sigma_5\sigma_1 & \rho_{52}\sigma_5\sigma_2 & \rho_{53}\sigma_5\sigma_3 & \rho_{54}\sigma_5\sigma_4 & \sigma_5^2 & \\ \rho_{61}\sigma_6\sigma_1 & \rho_{62}\sigma_6\sigma_2 & \rho_{63}\sigma_6\sigma_3 & \rho_{64}\sigma_6\sigma_4 & \rho_{65}\sigma_6\sigma_5 & \sigma_6^2 \end{bmatrix} \qquad (6)$$

where ρ_{ab} is the correlation of gene expression between cell type a and b.

Here, it can be seen that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ define the distribution of the observations fully. Our beliefs about these parameters are then encoded into the prior distributions. For more interpretable results, the distribution of $\boldsymbol{\mu}$ was encoded given the covariance matrix $\boldsymbol{\Sigma}$. This is due to the fact that we are encoding our beliefs about all of the parameters, which is the joint distribution, $\pi(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Firstly, we define our belief about the mean $\boldsymbol{\mu}$ given the variance-covariance matrix $\boldsymbol{\Sigma}$ to follow a multivariate normal distribution

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \text{MVN}\left(\boldsymbol{\mu}^*, \frac{1}{c}\boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\mu}^*$ is an n_t-vector of hyper mean parameter, and c is a scalar that will be described next in Section 2.5. Secondly, our belief in the variance-covariance parameter is defined as

$$\boldsymbol{\Sigma} \sim \text{IW}(\boldsymbol{\Psi}, \nu)$$

where IW$(\cdot)$ denotes the Inverse Wishart distribution with hyperparameters $\boldsymbol{\Psi}$ and ν.

The specification of prior distribution can be described as follows. From the above formulation, we can infer that the mean of $\boldsymbol{\mu}$ in each gene is given by $E(\boldsymbol{\mu}) = \boldsymbol{\mu}^*$, and similarly for $\boldsymbol{\Sigma}$

$$\begin{aligned} E(\boldsymbol{\Sigma}) &= \frac{1}{\nu - n_t - 1}\boldsymbol{\Psi}, \\ \text{Var}(\boldsymbol{\Sigma}_{ab}) &= \frac{(\nu - n_t + 1)\Psi_{ab}^2 + (\nu - n_t - 1)\Psi_{aa}\Psi_{bb}}{(\nu - n_t)(\nu - n_t - 1)^2(\nu - n_t - 3)}, \\ \text{Var}(\boldsymbol{\Sigma}_{aa}) &= \frac{2\boldsymbol{\Psi}_{aa}^2}{(\nu - n_t - 1)^2(\nu - n_t - 3)}, \end{aligned} \qquad (7)$$

where $n_t = 6$ is the number of cell types in our data. A strategy for picking a diffuse prior (and to make $\boldsymbol{\Psi}$ positive definite) is to set $\nu = n_t + 4$, and select $\boldsymbol{\Psi} = 3E(\boldsymbol{\Sigma})$. We could also set such that $\boldsymbol{\Sigma}_{aa}^{0.5}$ gives a standard deviation that would cover all possible $\boldsymbol{\mu}$ values. In our study $\boldsymbol{\mu}^*$ is just set to be the mid-point of the distribution of x_{ijk}.

2.5 Consideration for the selection of the prior distributions

In Bayesian analysis, a natural choice for the prior distribution is conjugate as described above. In our case of multivariate normal distribution for $\boldsymbol{x}_j|\boldsymbol{\mu}, \boldsymbol{\Sigma}$, a conjugate prior distribution for the mean $\boldsymbol{\mu}$ is multivariate normal with hyperparameters $\boldsymbol{\mu}^*$ and $\frac{1}{c}\boldsymbol{\Sigma}$. A conjugate prior distribution for $\boldsymbol{\Sigma}$ is Inverse Wishart distribution with hyperparameters $\boldsymbol{\Psi}$ and ν.

The hyperparameters for the prior distributions need to be carefully selected so that the priors have little effect on the inference. The prior distribution of the mean was chosen to be $\boldsymbol{\mu}^* = (9, 9, 9, 9, 9, 9)^T$. This was chosen as the median/mean of the usual range of (log) expressions between 2 and 16. The hyperparamater c represents the number of observations our prior is worth. By choosing $c = 1$, it ensures that our prior belief in $\boldsymbol{\mu}$ is relatively weak and is imposing as little information as possible on the analysis (see also the Supplementary Material on the effect of prior on the posterior).

The choice of $\boldsymbol{\Psi}$ was chosen so that the magnitude of the variances and covariances is large enough to explore the posterior space well. To cover the expression between 2 and 16 from median/mean 9, we need standard deviation of 4, or variance of 16. Furthermore, it needs to reflect the correlation structure that we would expect to see in the data given our knowledge on the haematopiesis in Figure 1. For the covariances, we then subtract the dissimilarities between cell types in Eq. (2) from the variance 16. Hence, cell types that are close in the differentiation will be expected to have higher correlation due to lower dissimilarity.

After all of these factors were taken into consideration, the hyperparamater $\boldsymbol{\Psi}$ is defined as

$$\boldsymbol{\Psi} = \begin{pmatrix} 16 & 11 & 10 & 12 & 15 & 10 \\ 11 & 16 & 14 & 14 & 11 & 14 \\ 10 & 14 & 16 & 13 & 10 & 15 \\ 12 & 14 & 13 & 16 & 12 & 13 \\ 15 & 11 & 10 & 12 & 16 & 10 \\ 10 & 14 & 15 & 13 & 10 & 16 \end{pmatrix}. \tag{8}$$

The covariances between cell types in $\boldsymbol{\Psi}$ are still large enough to be able to cover the posterior sample space of $\boldsymbol{\Sigma}$. Our experience in the analysis suggests that a small change to the covariances in $\boldsymbol{\Psi}$, as long as their magnitudes are reasonably large, does not affect the inference.

The choice of ν in our study was due to the characteristics of the Inverse Wishart distribution as described in the previous section. For the variance to be defined and to ensure that our prior beliefs have as little effect as possible for the posterior and allowing $\boldsymbol{\Sigma}$ as much freedom as possible, $\nu = 10$ was chosen. All of the hyperparameters are set the same for all of the genes.

2.6 Posterior probability

Distributions are often used in their proportional form. As such, the proportional probability distribution functions of the likelihood and their parameters for each gene are given by (Mardia *et al.*, 1980):

$$\begin{aligned} \pi(\boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{\frac{-(\nu+n_t+1)}{2}} \exp\left\{-\frac{1}{2}\text{trace}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right\}, \\ \pi(\boldsymbol{\mu}|\boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}^*)^{\mathrm{T}} c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}^*)\right\}, \\ \pi(\mathbf{x}_j|\boldsymbol{\mu},\boldsymbol{\Sigma}) &\propto |\boldsymbol{\Sigma}|^{-\frac{n_p}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_p}(\mathbf{x}_j-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_j-\boldsymbol{\mu})\right\}. \end{aligned}$$

Using the conjugacy of the prior distributions, a posterior distribution can be obtained which gives our updated beliefs about the parameters, in light of the data. So, after the data are observed the posterior distributions for each gene are given by (O'Hagan and Forster, 2004)

$$\boldsymbol{\mu}|\boldsymbol{\Sigma},\boldsymbol{x}_j \sim \mathrm{N}\left(\mathbf{m}^*, \frac{1}{s^*}\boldsymbol{\Sigma}\right), \tag{9}$$

$$\boldsymbol{\Sigma}|\boldsymbol{x}_j \sim \mathrm{IW}\left(\boldsymbol{\Psi}^*, \nu^*\right), \tag{10}$$

where

$$\begin{aligned} \mathbf{m}^* &= \frac{c\boldsymbol{\mu}^* + n_p\bar{\mathbf{x}}}{n_p+c}, \quad s^* = n_p + c, \quad \nu^* = \nu + n_p, \\ \boldsymbol{\Psi}^* &= \boldsymbol{\Psi} + \frac{cn_p}{n_p+c}\left(\boldsymbol{\mu}^* - \bar{\mathbf{x}}\right)\left(\boldsymbol{\mu}^* - \bar{\mathbf{x}}\right)^{\mathrm{T}} + \sum_{j=1}^{n_p}\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)\left(\mathbf{x}_j - \bar{\mathbf{x}}\right)^{\mathrm{T}}, \\ \bar{\mathbf{x}} &\equiv \left(\bar{x}_1 \ \ \bar{x}_2 \ \ \ldots \ \ \bar{x}_{n_t}\right)^{\mathrm{T}}, \text{ and } \bar{x}_j = \frac{1}{n_p}\sum_{k=1}^{n_p} x_{jk}. \end{aligned}$$

In the above updating, $\bar{x}_j$ is the mean of gene expression of j-th cell type and $\bar{\boldsymbol{x}}$ is an n_t-vector of cell-type means.

2.7 Inference

To obtain the posterior samples, we draw from the posterior distribution $\boldsymbol{\Sigma}|\boldsymbol{x}_j$ and then $\boldsymbol{\mu} \equiv \{\mu_1, \ldots, \mu_{n_t}\}|\boldsymbol{\Sigma},\boldsymbol{x}_j$, for each gene, denoted as $\boldsymbol{\mu}^{(z)} \equiv \{\mu_1^{(z)}, \ldots, \mu_{n_t}^{(z)}\}$ and $\boldsymbol{\Sigma}^{(z)}$ for $z = 1, 2, \ldots, n_{\text{post}}$, where n_{post} is the number of *accepted* posterior samples. Among the samples drawn from the posterior distributions, we accept those that fulfill the condition on the correlation structure in Eq. (3).

To identify whether a gene is specific in directing a cell differentiation, we calculate the probability of the k-th cell type to have higher (or lower) posterior $\mu_k|\boldsymbol{\Sigma},\boldsymbol{x}_j$ than those of the other cell types, i.e.

$$p_k^+ = \frac{1}{n_{\text{post}}}\sum_{z=1}^{n_{\text{post}}} I\left(\mu_k^{(z)} > \mu_{k'}^{(z)}\right), \text{ for } k' \in \{1, \ldots, n_t\} \text{ and } k \neq k' \tag{11}$$

$$p_k^- = \frac{1}{n_{\text{post}}}\sum_{z=1}^{n_{\text{post}}} I\left(\mu_k^{(z)} < \mu_{k'}^{(z)}\right), \text{ for } k' \in \{1, \ldots, n_t\} \text{ and } k \neq k' \tag{12}$$

where the summation is across the accepted posterior samples, and $I(\cdot)$ is an indicator function which is equal to one if the argument inside the brackets is true and zero otherwise.

Having posterior samples $\mu_k^{(z)}$, $k = 1, \ldots, n_t$, also enables us to construct 95% credible interval for each of $\mu_k|\boldsymbol{\Sigma},\boldsymbol{x}_j$. The limits of the interval are defined as the 2.5 and 97.5 percentiles of the accepted posterior samples $\mu_k^{(z)}$ across $z = 1, 2, \ldots, n_{\text{post}}$ for each $k = 1, \ldots, n_t$. In our analysis, the number of accepted posterior samples n_{post} is set to be 1000. The reasoning of the choice of this number is because it achieved an acceptable mean square error on the posterior mean (see also the Supplementary Material).

The above inference has some flexibilities, for example to identify genes that are involved in the differentiation of more than one cell types. In the first situation, we can identify them as those with high p_k^+ in one cell type and high p_k^- in another cell type. They are referred to as non-specific genes in Section 3.2. In the second situation, we can identify them as those that have higher (or lower) posterior means in two cell types compared to the other cell types. This is done by including another inequality for the second cell type in each of the Equations (11) and (12), as described in the Supplementary Material.

2.8 Simulation study

We perform a simulation study to investigate the proposed method's operating characteristics in acknowledging correlation between cell types in the analysis of cell differentiation. We anticipate that respecting lineage in cell differentiation, in terms of correlation structure of gene expression between cell types, would result in higher sensitivity to detect genes that direct cell differentiation.

We generate gene expressions for 1000 genes, under independence between genes. Within each gene, we generate gene expressions for seven individuals and six cell types from the same mean $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}$. Among the genes, 100 of them are set to have different (true) mean for the first cell type (μ_1) to indicate that the 100 genes are directing differentiation of the first cell-type. The mean for the first cell-type is differed by one to three, corresponding to 0.35σ to 1.1σ, to represent low, medium, and high signals. Three different scenarios that we consider are based on the form of correlation between cell types in $\boldsymbol{\Sigma}$:

1. under different correlation structure based on the path of cell differentiation in Figure 1 (scenario A)
2. under equal correlation between cell-types, which means $\boldsymbol{\Sigma}$ is symmetric matrix with the same non-zero off-diagonal elements (scenario B)

3. under independence between cell-types, which means $\boldsymbol{\Sigma}$ is diagonal matrix (scenario C).

We then estimate the operating characteristics, in terms of sensitivity and specificity, of the proposed method based on 100 simulated datasets per setting. As a comparison, we will also consider the t-test. In withdrawing samples from posterior distribution, we do not apply the constraint in the correlation structure in Eq. (3).

3 RESULTS

3.1 Posterior

An illustration of the posterior samples $\boldsymbol{\mu}^{(z)}$'s and the estimated correlation between cell types a and b, ρ_{ab}, for genes SLC46A2 (9q32) and CYFIP2 (5q33.3) are presented in Figure 3. These two genes in our analysis are among genes that are identified as unique in CD14 and involved in directing its differentiation. Gene SLC46A2 and CYFIP2, respectively, have higher and lower posterior mean in CD14 compared to the other cell types. These figures are from 1000 posterior samples in each cell types, and it can be shown that the probabilities p_1^+ for gene SLC46A2 is 1 and p_1^- for gene CYFIP2 is 1. We are confident that these two genes are involved in directing the cell differentiation of CD14.

Figure 3 also presents the accepted posterior correlation based on the cell differentiation diagram in Figure 1. For example, based on Figure 1, we constrain that the correlation between CD4 and CD8 (ρ_{63}) is higher that the correlation between CD4 and CD14. We also constrain the correlation between CD66b and CD14 (ρ_{51}) to be higher than the correlation between CD66b and CD56. The impact of these constraints are not immediately visible in Figure 3; this is more clearly visible if we create a scatterplot between posterior correlations as in Figure 4. The figure shows that the constraints are imposed in the result that we observe previously in Figure 3.

3.2 Specific and non-specific genes

In our analysis, we obtain posterior samples that are illustrated in Figure 3 for each gene. This allows 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 theshold are presented in Table 1 and Table 2, for p_k^+ and p_k^- respectively. Table 1 indicates that the number of genes that have at least 99% probability to have higher posterior mean in CD8 is only seven, while in CD66b it is 1,029. Table 2 also indicates that there are 12 genes with at least 99% probability to have lower posterior mean in CD56 compared to the other cell types.

Table 1 and Table 2 provide a profile of the distribution of p_k^+ and p_k^- in the data, which in turn suggest how the genes are involved in directing haematopoiesis. For example, both tables show that more genes are involved in directing the CD66b cell differentiation compared to the other cell types, either by actively increasing or lowering gene expression. Similarly, there are not many genes specifically involved in CD8 cell differentiation. Both tables also suggest that, with the exception of CD66b cell type, more genes are involved in directing cell differentiation by actively increasing gene expression than lowering them.

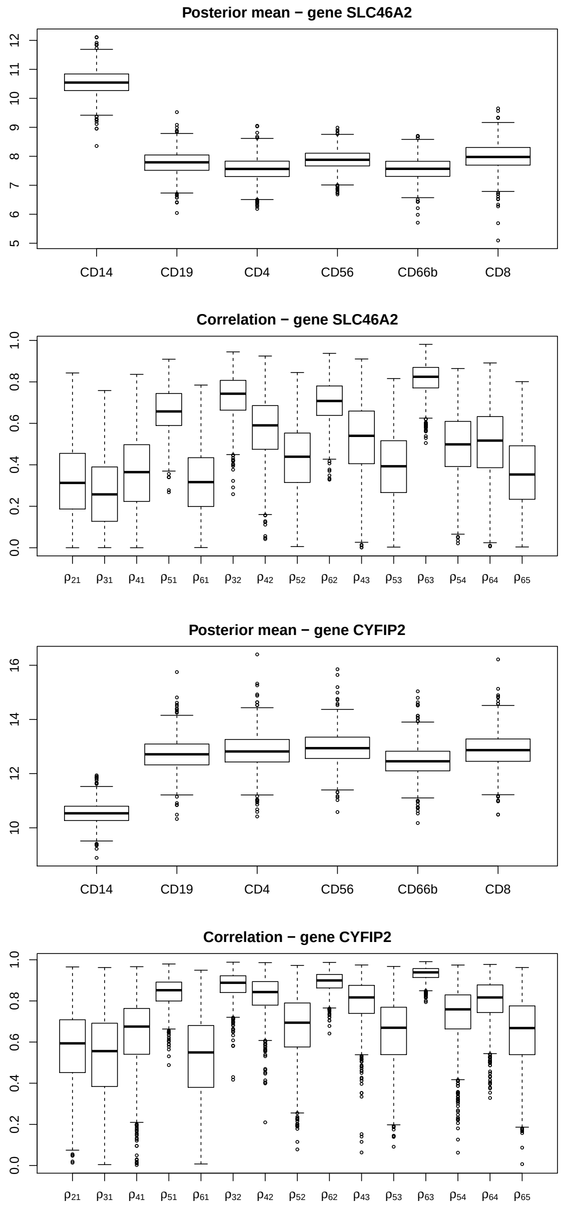

Fig. 3. *Posterior mean samples $\boldsymbol{\mu}^{(z)}$ for gene SLC46A2 (top panel) and gene CYFIP2 (third panel), and correlation between cell types ρ_{ab} for those two genes in second and bottom panels. The indices a and b in the correlation are from 1 to 6, which correspond to CD14, CD19, CD4, CD56, CD66b, and CD8, respectively. For example, ρ_{61} means the correlation between CD8 and CD14.*

Due to the definition of p_k^+ of p_k^- in Eqs. (11) and (12), it is possible to have a consistently higher posterior means in one cell type and at the same time lower posterior means in another cell type, and vice versa. We consider these genes to be non-specific (as opposed to specific to one cell type). However, the terms ’specific’ and ’non-specific’ need to be interpreted in relative sense, and not in absolute sense, due to the definition in Eqs. (11) and (12). The numbers of such genes are shown in Table 3. The table indicates the number of specific and non-specific probes across all cell types at different probability thresholds. The posterior mean samples from

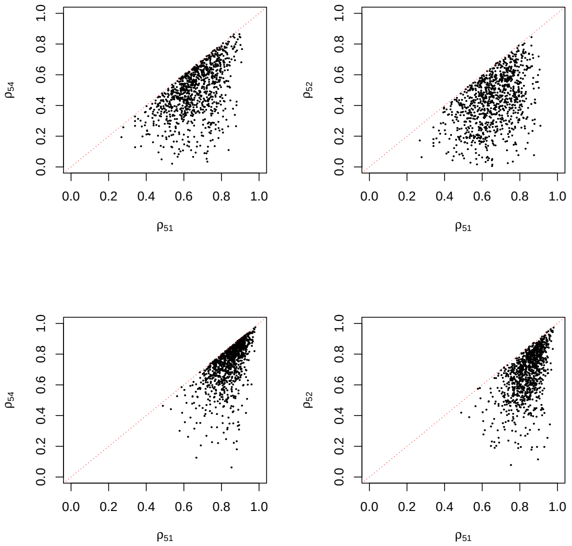

Fig. 4. *Left column: Comparison of correlation posterior samples between cell types CD66b and CD14 (ρ_{51}) and those between CD66b and CD56 (ρ_{54}). Right column: Comparison of correlation posterior samples between cell types CD66b and CD14 (ρ_{51}) and those between CD66b and CD19 (ρ_{52}). The top panels are for gene SLC46A2 and bottom panels are for gene CYFIP2.*

p_k^+	CD14	CD19	CD4	CD56	CD66b	CD8
≥0.00	46713	46713	46713	46713	46713	46713
≥0.50	2030	1713	2242	1216	5816	156
≥0.80	922	881	767	500	3223	41
≥0.90	658	638	445	323	2127	23
≥0.95	505	494	313	235	1675	15
≥0.99	268	320	158	116	1029	7
1.00	83	169	42	38	402	3

Table 1. *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 the Supplementary Material.*

p_k^-	CD14	CD19	CD4	CD56	CD66b	CD8
≥0.00	46713	46713	46713	46713	46713	46713
≥0.50	1590	1066	1055	363	4970	55
≥0.80	494	409	410	85	3102	2
≥0.90	294	268	270	42	2427	1
≥0.95	179	200	191	27	1957	1
≥0.99	79	111	88	12	1221	0
1.00	28	36	13	1	447	0

Table 2. *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 lower posterior mean in each cell type than the other cell types. The probability for each probe is presented in the Supplementary Material.*

p_k	≥0.50	≥0.80	≥0.90	≥0.95	≥0.99	1.00
Specific	9060	7816	6074	4938	3163	1242
Non-specific	6606	1510	721	427	123	10

Table 3. *Number of probes whose p_k^-'s or p_k^+'s match and pass different thresholds that are specific or non-specific to a cell type. Non-specific means that the probes have p_k^+ above the threshold in one cell type and p_k^- above the threshold in another cell-type. The ten non-specific probes at probability one are from the genes RNF149, HS.579530, NUP88, SP140, RP9, RGS2, CPD, HSPA6, TNFRSF1A, and FAM129A. The posterior means for genes SP140 and RP9 are shown in Figure 5.*

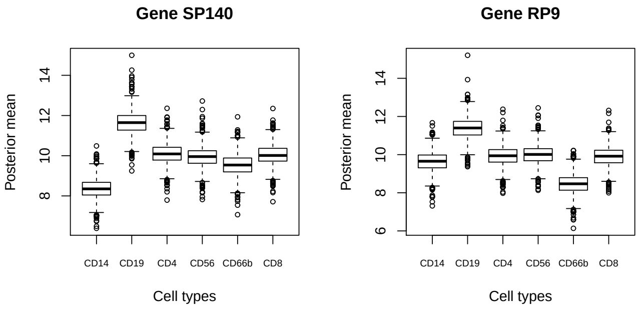

Fig. 5. *Posterior mean samples $\boldsymbol{\mu}^{(z)}$ for genes SP140 and RP9, which are identified as non-specific genes in the haematopoiesis. The genes are identified to be involved in the direction of haematopoiesis in two different cell types: CD14 and CD19 for SP140 and CD19 and CD66b for RP9.*

two non-specific genes (SP140 and RP9) are presented in Figure 5. The figure illustrates the non-specificity of the two genes, in which the posterior mean samples are consistently higher in one cell type and lower in another cell type. More details, including the probability and information for each probe, are available in the Supplementary Material.

3.3 Gene ontology

Table 4 presents some of the gene ontology (GO) biological processes of genes with p_k^+ and p_k^- greater than 0.95 in Table 1 and Table 2, based on the PANTHER classification system (Thomas *et al.*, 2003; Mi *et al.*, 2005). The full list of the GO biological processes is available in the Supplementary Material as Excel files, which indicates the full extent of biological processes of the genes identified by our method.

To highlight few genes, our method identifies *CYP1B1, C9ORF88 (FAM129B)*, and *CEPBA* to be significant. *CYP1B1* is involved in the signalling of haematopoietic stem cells as recently described in (Rentas *et al.*, 2016). *FAM129B* was identified to suppress apoptosis (Chen *et al.*, 2011), and suppression of apoptosis was recognised to allow differentiation and development of a multipotent hemopoietic cell line (Fairbairn *et al.*, 1993). With regard to *CEPBA*, Wölfler *et al.* (2010) showed that *CEPBA/EYFP*(+) cells represent a significant subset of multipotent hematopoietic progenitors, which predominantly give rise to myeloid cells in steady-state haematopoiesis.

GO Biological Process	Observed	Expected	Fold	p-value
leukocyte diff. (GO:0002521)	103	53.73	1.92	9.02E-06
lymphocyte diff. (GO:0030098)	73	38.3	1.91	2.72E-03
reg. of leukocyte diff. (GO:1902105)	80	42.66	1.88	1.45E-03
reg. of haematopoiesis (GO:1903706)	102	55.18	1.85	6.59E-05
imm. syst. process (GO:0002376)	641	360.12	1.78	5.91E-42
apoptotic process (GO:0006915)	288	171.35	1.68	3.51E-13
locomotion (GO:0040011)	274	202.02	1.36	3.34E-03
metabolic process (GO:0008152)	2336	1802.96	1.3	4.11E-64
cellular process (GO:0009987)	3155	2620.86	1.2	5.95E-85
biological regulation (GO:0065007)	2372	2049.82	1.16	1.44E-22
develop. process (GO:0032502)	1079	949.13	1.14	6.78E-03

Table 4. *Some gene ontology (GO) biological processes from the list of genes in both Table 1 and Table 2 with p_k^+ and p_k^- greater than 0.95 based on PANTHER classification system (Thomas et al., 2003; Mi et al., 2005). The complete lists are available in the Supplementary Material as Excel files. The p-value is the result from an over-representation test, which compared the observed count of genes in each category to the expected count based on the GO reference list. Bonferroni multiplicity correction has been applied to the p-value.*

3.4 Sensitivity analysis

In the above analysis, the structure on $\mathbf{\Psi}$ in the prior is defined according to the structure of the cell differentiation in Figure 1 as indicated in Eq. (8). To check whether our analysis does not depend largely on the choice of prior, we also consider other structures of $\mathbf{\Psi}$ (see also the Supplementary Material for mathematical derivation on how much the prior is worth). The first one we consider is that $\mathbf{\Psi}$ is a diagonal matrix, i.e. the off-diagonal entries of $\mathbf{\Psi}$ in Eq. (8) are zero. In this setting, we assume *a priori* that the gene expression between the different cell types are independent. The second one is that $\mathbf{\Psi}$ is a symmetric matrix, by which we assume *a priori* that the genes are equally correlated (i.e. there is a correlation between cell types, but not in the structure in Figure 1). The results are presented in the supplementary material.

The results indicate that the posterior samples of $\boldsymbol{\mu}^{(z)}$ under diagonal $\mathbf{\Psi}$ are relatively consistent to those under general $\mathbf{\Psi}$ in Figure 3. However, the posterior correlation samples between cell types under diagonal prior $\mathbf{\Psi}$ are higher than those in Figure 3 under general $\mathbf{\Psi}$.

3.5 Simulation study

The simulation results are presented in Figure 6. The figure shows the operating characteristics of the proposed method (solid line) in three different scenarios for the medium signal (the figures for the low and high signals are presented in the Supplementary Material). The figure indicates reasonably good characteristics of the proposed method (solid line). The area-under-curve for the proposed method in simulation A is slightly more than that in simulation B, which is also more than that in simulation C. This result is as expected. The setting for simulation A and B is very close; both have correlation structure in the expression data across cell types, while in simulation C, the gene expressions between cell types are independent. The figure also indicates that respecting the correlation structure between cell types gives better operating characteristics than ignoring them, as is the case in using the t-test (dashed lines).

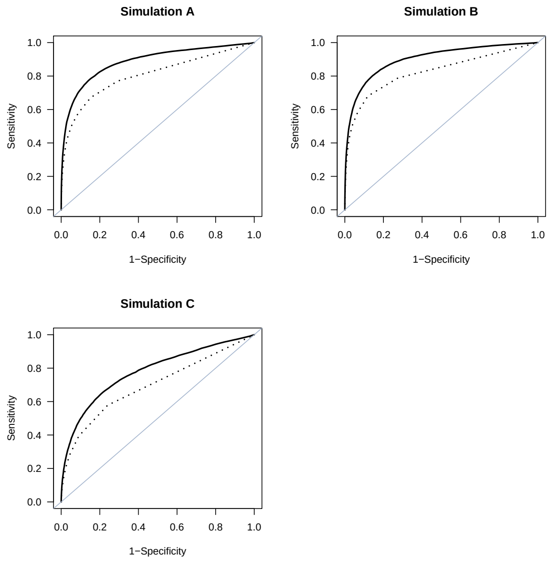

Fig. 6. *Operating characteristics of the proposed method (solid line) and t-test (dashed line) in the simulation study for medium signal where the simulated gene expression are correlated across cell types according to Figure 1 (simulation A), under equal correlation across cell types (simulation B), and under independence (simulation C). The figures for low and high signals are available in the Supplementary Material.*

4 DISCUSSION

Identifying specific genes in cell differentiation is a challenging task, especially when gene expression data available are from mature cells. In the ideal case where gene expressions were obtained from cells at different stages in the cell differentiation, then the identification of specific genes can be performed in a straightforward manner. However, when the gene expression data available are from the final stage in cell differentiation, then the cell differentiation paths need to be taken into account in the inference. Failing to take into account these information means that we only identify genes that are differentially expressed between cell types under some sort of independence assumption. To take into account the cell differentiation paths, we consider Bayesian modelling as a natural and intuitive method, where the cell differentiation paths serve as prior. In this study, we present how this methodology can address the challenge under some distributional assumptions and conjugacy.

The proposed method enabled us to identify genes that are specifically involved in the differentiation in each cell type. The results indicate that the number of such genes in each cell type varies. It turns out many more genes that are responsibe in directing the cells to mature to CD66b (neutrophil) and CD19 (B lymphocytes) than those to the other cell types. Looking into the GO biological processes involved in the significant genes, the results indicate that the haematopoiesis is controlled by a wide transcription regulatory networks. Further downstream analysis also indicates that many genes that are specific in each cell type share common transcription factors (see also the Supplementary Material). This study is an important effort to identify genes that control lineage commitment, albeit from a difficult context, in which the information come from final mature cells in the differentiation.

In the proposed method, there are two steps in the analysis pipeline where the cell differentiation paths are taken into account as a correlation structure between cell types. Firstly, it is in the formulation of the prior, and secondly, in the inference (Section 2.7). The correlation structure, as prior, has little influence on the posterior distribution of the mean and variance, as indicated in Section 3.4 and the Supplementary Material. Our sensitivity analysis on the choice of $\mathbf{\Psi}$ as hyperparameter in the prior distribution of $\mathbf{\Sigma}$ indicates that the mean posterior samples $\boldsymbol{\mu}^{(z)}$'s are relatively consistent; i.e. the mean posterior samples of $\boldsymbol{\mu}^{(z)}$'s are relatively consistent whether the correlation structure between cell types are reflected in the prior distribution or not. However, a difference is visible on the correlation posterior samples between the two cases of the prior. The results indicate that if the correlation structure between cell types are not included in the prior, the correlation posterior samples are generally higher than those when the structure are not included in the prior. In the second step, the correlation between cell types imposes a stronger structure in the posterior sample. As illustrated in Figure 4 and the Supplementary Material, the posterior samples that we accept are those that respect the constraints on the correlation between cell types in Eq. (3).

Simulation results indicate that the proposed method has a reasonably good operating characteristics. Respecting the correlation structure between cell types in the analysis certainly gives an advantage in the inference, even if the data were generated assuming independence between cell types. The results (see also the Supplementary Material) suggest that when the amount of signal is low and medium, this advantage is notable. As the signal increases to high, this advantage is reduced because the gene signal already stands out. This suggests that when the gene expression in a particular cell type is relatively high, the correlation between cell types are somehow less relevant.

5 CONCLUSION

We have some challenges in identifying transcript regulatory patterns that govern cell differentiation when gene expression data available are only from mature cells. To identify specific genes that are involved in directing cell differentiation, we propose to take into account the information of cell differentiation paths in the analysis using Bayesian approach. It is natural and intuitive to incorporate cell differentiation paths as prior information and the method is able to identify the relevant genes in haematopoiesis. The simulation indicates that we obtain the best advantage among low to moderate signal when we take into account the correlation stucture.

REFERENCES

Chen, S *et al.* (2011) FAM129B/MINERVA, a Novel Adherens Junction-associated Protein, Suppresses Apoptosis in HeLa Cells, *The Journal of Biological Chemistry*, **286**, 10201-10209

Draper JE, *et al.* (2016) RUNX1B Expression Is Highly Heterogeneous and Distinguishes Megakaryocytic and Erythroid Lineage Fate in Adult Mouse Hematopoiesis, *PLoS Genetics*, **12**(1), e1005814

Fairbairn, LJ *et al.* (1993) Suppression of apoptosis allows differentiation and development of a multipotent hemopoietic cell line in the absence of added growth factors, *Cell*, **74**(5), 823-832

O'Hagan A, and Forster JJ (2004) Bayesian Inference 2nd Ed. in *Kendall's Advanced Theory of Statistics Volume 2B*, Arnold, London

Kuvardina ON *et al.* (2015) RUNX1 represses the erythroid gene expression program during megakaryocytic differentiation, *Blood*, **125**(23), 3570–3579

Mancini E *et al.* (2012) FOG-1 and GATA-1 act sequentially to specify definitive megakaryocytic and erythroid progenitors, *The EMBO Journal*, **31**(2), 351–365

Mardia KV, Kent JT, and Bibby JM (1980) *Multivariate Analysis*, Academic Press, London

Mi, H *et al.* (2005) The PANTHER database of protein families, subfamilies, functions and pathways, *Nucleic Acids Research*, **33**(Suppl 1), D284-D288

Novershtern N, *et al.* (2011) Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis, *Cell*, **144**(2), 296-309

Ploner A, Miller LD, Hall P, Bergh J, Pawitan Y (2005) Correlation test to assess low-level processing of high-density oligonucleotide microarray data, *BMC Bioinformatics*, **6**, 80

Rentas, S *et al.* (2016) Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells, *Nature*, **532**, 508-511

Tanaka Y, *et al.* (2011) Dual Function of Histone H3 Lysine 36 Methyltransferase ASH1 in Regulation of Hox Gene Expression, *PLoS One*, **6**(11): e28171

Thomas, PD *et al.* (2003) PANTHER: A library of protein families and subfamilies indexed by function, *Genome Research*, **13**, 2129-2141

Ungerbäck J *et al.* (2015) Combined heterozygous loss of Ebf1 and Pax5 allows for T-lineage conversion of B cell progenitors, *The Journal of Experimental Medicine*, **212**(7), 1109–1123

Watkins N *et al.* (2009) HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood*, **113**(19), e1-9

Wölfler *et al.* (2010) Lineage-instructive function of C/EBPα in multipotent hematopoietic cells and early thymic progenitors, *Blood*, **116**, 4116-4125

Zola H, *et al.* (2005) CD molecules 2005: human cell differentiation molecules, *Blood*, **106**, 3123–3126