Bioinformatics, 31(19), 2015, 3163–3171
doi: 10.1093/bioinformatics/btv344
Advance Access Publication Date: 4 June 2015
Original Paper

OXFORD

Gene expression

Sparse multi-view matrix factorization: a multivariate approach to multiple tissue comparisons

Zi Wang[1], Wei Yuan[2] and Giovanni Montana[1,3,*]

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, [2]Department of Twin Research and Genetic Epidemiology, King's College London, St Thomas' Hospital, London SE1 7EH and [3]Department of Biomedical Engineering, King's College London, St Thomas' Hospital, London SE1 7EH, UK

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on February 26, 2015; revised on May 12, 2015; accepted on May 29, 2015

Abstract

Motivation: Within any given tissue, gene expression levels can vary extensively among individuals. Such heterogeneity can be caused by genetic and epigenetic variability and may contribute to disease. The abundance of experimental data now enables the identification of features of gene expression profiles that are shared across tissues and those that are tissue-specific. While most current research is concerned with characterizing differential expression by comparing mean expression profiles across tissues, it is believed that a significant difference in a gene expression's variance across tissues may also be associated with molecular mechanisms that are important for tissue development and function.

Results: We propose a sparse multi-view matrix factorization (sMVMF) algorithm to jointly analyse gene expression measurements in multiple tissues, where each tissue provides a different 'view' of the underlying organism. The proposed methodology can be interpreted as an extension of principal component analysis in that it provides the means to decompose the total sample variance in each tissue into the sum of two components: one capturing the variance that is shared across tissues and one isolating the tissue-specific variances. sMVMF has been used to jointly model mRNA expression profiles in three tissues obtained from a large and well-phenotyped twins cohort, TwinsUK. Using sMVMF, we are able to prioritize genes based on whether their variation patterns are specific to each tissue. Furthermore, using DNA methylation profiles available, we provide supporting evidence that adipose-specific gene expression patterns may be driven by epigenetic effects.

Availability and implementation: Python code is available at http://wwwf.imperial.ac.uk/~gmontana/.

Contact: giovanni.montana@kcl.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA abundance, as the results of active gene expression, affects cell differentiation and tissue development (Coulon *et al.*, 2013). As such, it provides a snapshot of the undergoing biological process within certain cells or a tissue. Except for house-keeping genes, the expressions of a large number of genes vary from tissue to tissue, and some may only be expressed in a particular tissue or a certain cell type (Xia *et al.*, 2007). The regulation of tissue-specific expression is a complex process in which a gene's enhancer plays a key role regulating gene expressions via DNA methylation (Ong and

Corces, 2011). Genes displaying tissue-specific expressions are widely associated with cell type diversity and tissue development (Reik, 2007), and aberrant tissue-specific expressions have been associated with diseases that originated in the underlying tissue (Lage et al., 2008; van't Veer et al., 2002). Distinguishing tissue-specific expressions from expression patterns prevalent in all tissues holds the promise to enhance fundamental understanding of the universality and specialization of molecular biological mechanisms and potentially suggest candidate genes that may regulate traits of interest (Xia *et al.*, 2007). As collecting genome-wide transcriptomic profiles from many different tissues of a given individual is becoming more affordable, large population-based studies are being carried out to compare gene expression patterns across human tissues (Liu *et al.*, 2008; Yang *et al.*, 2011).

A common approach to detecting tissue-specific expressions consists of comparing the *mean* expression levels of individual genes across tissues. This can be accomplished using standard univariate test statistics. For instance, Wu *et al.* (2014) used the two-sample *Z*-test to compare non-coding RNA expressions in three embryonic mouse tissues: they reported approximately 80% of validated *in vivo* enhancers exhibited tissue-specific RNA expression that correlated with tissue-specific enhancer activity. Yang *et al.* (2011) applied a modified version of Tukey's (1949) range test, a test statistic based on the standardized mean difference between two groups, to compare expression levels of 127 human tissues, and results of this study are publicly available in the VeryGene database. A related database, TiGER (Liu *et al.*, 2008), has also been created by comparing expression sequence tags in 30 human tissues using a binomial test on expression sequence tag counts. Both VeryGene and TiGER contain up-to-date annotated lists of tissue-specific gene expressions, which generated hypotheses for studies in the area of pathogenic mechanism, diagnosis and therapeutic research (Wu *et al.*, 2009).

More recent studies have gone beyond the single-gene comparison and aimed at extracting multivariate patterns of differential gene expression across tissues. Xiao *et al.* (2014) applied the higher-order generalized singular value decomposition method proposed by Ponnapalli *et al.* (2011) and compared co-expression networks from multiple tissues. This technique is able to highlight co-expression patterns that are equally significant in all tissues or exclusively significant in a particular tissue. The rationale for a multivariate approach is that when a gene regulator is switched on, it can raise the expression level of all its downstream genes in specific tissues. Hence, a multi-gene analysis may be a more powerful approach.

While most studies explore the differences in the mean of expression, the sample variance is another interesting feature to consider. Traditionally, comparison of expression variances has been carried out in case-control studies (Mar *et al.*, 2011). Using an *F*-test, significantly high or low gene expression variance has been observed in many disease populations including lung adenocarcinoma and colorectal cancer, whereas the difference in mean expression levels was not found significant between cases and controls (Ho *et al.*, 2008). In a tissue-related study, Cheung *et al.* (2003) carried out a genome-wide assessment of gene expressions in human lymphoblastoid cells. Using an *F*-test, the authors showed that high-variance genes were mostly associated with functions such as cytoskeleton, protein modification and transport, whereas low-variance genes were mostly associated with signal transduction and cell death/proliferation.

In this work, we introduce a novel multivariate methodology that can detect patterns of differential variance across tissues. We regard the gene expression profiles in each tissue as providing a different 'view' of the underlying organism and propose an approach to carry out such a multi-view analysis. Our objective is to identify genes that jointly explain the same amount of sample variance in all tissues—the 'shared' variance—and genes that explain substantially higher variances in each specific tissue separately—the 'tissue-specific' variances—while the shared variance has been accounted for. During this process, we impose a constraint that the factors driving shared and tissue-specific variability must be uncorrelated, so that the total sample variance can be decomposed into the two corresponding components. The proposed methodology, called sparse multi-view matrix factorization (sMVMF), can be interpreted as an extension of principal component analysis (PCA), which is traditionally used to identify a handful of latent factors (LFs) explaining a large portion of sample variance separately in each tissue.

The rest of this article is organized as follows. The sMVMF methodology is presented in Section 2, where we also discuss connections with a traditional PCA and derive the parameter estimation algorithm. In Section 3, we demonstrate the main feature of the proposed method on simulated data and report on comparison with alternative univariate and multivariate approaches. In Section 4, we apply the sMVMF to compare mRNA expressions in three tissues obtained from a large twin population, the TwinsUK cohort. We conclude in Section 5 with a discussion.

2 Methods

2.1 Sparse multi-view matrix factorization

We assume to have collected p gene expression measurements for M different tissues. Ideally, the data for all tissues should be derived from the same underlying random sample (as in our application, Section 4) to remove sources of biological variability that can potentially induce differences in gene expression profiles across tissues. In practice, however, cross-tissue experiments rarely collect samples from the same set of subjects or may fail quality control. In our setting, therefore, we assume M different random samples, each one contributing a different tissue dataset. The mth dataset consists of n_m subjects, and the expression profiles are arranged in an $n_m \times p$ matrix. All matrices are collected in $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(M)}\}$, where the superscripts refer to tissue indices. For each $X^{(m)}$, we subtract the column mean from each column, such that each diagonal entry of the scaled gram matrix, $\frac{1}{n_m}(X^{(m)})^T X^{(m)}$, is proportional to the sample variance of the corresponding variable and the trace is the total sample variance. We aim to identify genes that jointly explain a large amount of sample expression variances in all tissues and genes that explain substantially higher variances in a specific tissue. Our strategy involves approximating each $\frac{1}{\sqrt{n_m}}X^{(m)}$ by the sum of a shared variance component and a tissue-specific component:

$$\frac{1}{\sqrt{n_m}}X^{(m)} \approx \underbrace{S^{(m)}}_{\text{shared variance component}} + \underbrace{T^{(m)}}_{\text{tissue-specific variance component}} \quad (1)$$

for $m = 1, 2, \ldots, M$, where $1/\sqrt{n_m}$ is a scaling factor such that the trace of the gram matrix of the left-hand-side equals the sample variance. These components are defined so as to yield the following properties:

a. The rank of $S^{(m)}$ and $T^{(m)}$ are both much smaller than $\min(n_m, p)$, so that the two components provide insights into the intrinsic structure of the data while discarding redundant information.
b. The variation patterns captured by shared component are uncorrelated to the variation patterns captured by tissue-specific

component. As a consequence of this, the total variance explained by $S^{(m)}$ and $T^{(m)}$ altogether equals the sum of the variance explained by each individual component.

c. The shared component explains the same amount of variance of each gene expression in all tissues. As such, the difference in expression variance between tissues is exclusively captured in tissue-specific variance component.

We start by proposing a factorization of both $S^{(m)}$ and $T^{(m)}$ which, by imposing certain constraints, will satisfy the above properties. Suppose rank $(S^{(m)}) = d$ and rank $(T^{(m)}) = r$, where $d, r \ll \min(n_m, p)$ following property (a). For a given r, $T^{(m)}$ can be expressed as the product of an $n_m \times r$ full rank matrix $W^{(m)}$ and the transpose of a $p \times r$ full rank matrix $V^{(m)}$, that is:

$$T^{(m)} = W^{(m)}(V^{(m)})^T = \sum_{j=1}^{r} W_j^{(m)}(V_j^{(m)})^T = \sum_{j=1}^{r} T_{[j]}^{(m)} \tag{2}$$

where the superscript ^{T} denotes matrix transpose, and the subscript j denotes the jth column of the corresponding matrix. Each $T_{[j]}^{(m)} := W_j^{(m)}(V_j^{(m)})^T$ has the same dimension as $T^{(m)}$ and is composed of a tissue-specific LF. An LF is an unobservable variable assumed to control the patterns of observed variables and hence may provide insights into the intrinsic mechanism that drives the difference of expression variability between tissues. The matrix factorization in (2) is not unique, since for any $r \times r$ non-singular square matrix R, $T^{(m)} = W^{(m)}(V^{(m)})^T = (W^{(m)}R)(R^{-1}(V^{(m)})^T) = \tilde{W}^{(m)}(\tilde{V}^{(m)})^T$. We introduce an orthogonal constraint $(W^{(m)})^T W^{(m)} = I_r$ so that the matrix factorization is unique subject to an isometric transformation. Similarly, we can factorize the shared component as:

$$S^{(m)} = U^{(m)}(V^*)^T = \sum_{k=1}^{d} U_k^{(m)}(V_k^*)^T = \sum_{k=1}^{d} S_{[k]}^{(m)} \tag{3}$$

where $U^{(m)}$ is orthogonal and V^* is tissue independent which we shall explain. Each $S_{[k]}^{(m)}$ has the same dimension as $S^{(m)}$ and is composed of one shared variability LF. The resulting multi-view matrix factorization (MVMF) then is:

$$\frac{1}{\sqrt{n_m}} X^{(m)} \approx U^{(m)}(V^*)^T + W^{(m)}(V^{(m)})^T \tag{4}$$

The matrix factorizations (2) and (3) are intimately related to the singular value decomposition (SVD) of $S^{(m)}$ and $T^{(m)}$. Specifically, $U^{(m)}$ and $W^{(m)}$ are analogous to the matrix of left singular vectors and also the principal components (PCs) in a standard PCA. They represent gene expression patterns in a low-dimensional space where each dimension is derived from the original gene expression measurements such that the maximal amount of variance is explained. We shall refer the columns of $U^{(m)}$ and $W^{(m)}$ as the principal projections (PPJ). $(V^*)^T$ and $(V^{(m)})^T$ are analogous to the product of the diagonal matrix of eigenvalues and the matrix of right singular vectors. Since the singular values determine the amount of variance explained and the right singular vectors correspond to the loadings in the PCA which quantifies the importance of the genes to the expression variance explained, using the same matrix V^* for all tissues in the shared component results in the same amount of shared variability explained for each gene expression probe, such that property (c) is satisfied. We shall refer to matrices V^* and $V^{(m)}$ as transformation matrices.

A sufficient condition to satisfy property (b) is:

$$(U^{(m)})^T W^{(m)} = 0_{d \times r} \tag{5}$$

This constraint, in addition to the orthogonality of $U^{(m)}$ and $W^{(m)}$, results in the $(d + r)$ PPJs represented by $[U^{(m)}, W^{(m)}]$ being pairwise orthogonal, which is analogous to the standard PCA where the PCs are orthogonal. Intuitively, this means for each tissue, the LFs driving shared and tissue-specific variability are uncorrelated. The amount of variance explained in tissue m, $\hat{\sigma}_m^s$, can be computed as (subject to a constant factor):

$$\hat{\sigma}_m^s = \mathrm{Tr}\{(S^{(m)})^T S^{(m)} + (T^{(m)})^T T^{(m)} + 2(S^{(m)})^T T^{(m)}\} \tag{6}$$

where Tr denotes the matrix trace. Recalling that $S^{(m)} = U^{(m)}(V^*)^T$ and $(U^{(m)})^T U^{(m)} = I_d$, the amount of shared variance explained is:

$$\sigma_* = \mathrm{Tr}\{(S^{(m)})^T S^{(m)}\} = \mathrm{Tr}\{V^*(V^*)^T\} \tag{7}$$

Likewise, recalling that $T^{(m)} = W^{(m)}(V^{(m)})^T$ and $(W^{(m)})^T W^{(m)} = I_r$, the amount of tissue-specific variance explained is:

$$\sigma_m = \mathrm{Tr}\{(T^{(m)})^T T^{(m)}\} = \mathrm{Tr}\{V^{(m)}(V^{(m)})^T\} \tag{8}$$

Making the same substitutions into (6), we obtain:

$$\hat{\sigma}_m^s = \mathrm{Tr}\{V^*(V^*)^T + V^{(m)}(V^{(m)})^T + 2V^*(U^{(m)})^T W^{(m)}(V^{(m)})^T\}$$

Substituting (5) into the above equation, we reach:

$$\hat{\sigma}_m^s = \mathrm{Tr}\{V^*(V^*)^T + V^{(m)}(V^{(m)})^T\} = \sigma_* + \sigma_m \tag{9}$$

which satisfies (b).

2.2 Sparsity constraints and estimation

The factorization (4) is obtained by minimizing the squared error. This amounts to minimizing the loss function:

$$\ell = \sum_{m=1}^{M} \| \frac{1}{\sqrt{n_m}} X^{(m)} - U^{(m)}(V^*)^T - W^{(m)}(V^{(m)})^T \|_{\mathcal{F}}^2 \tag{10}$$

where $\|.\|_{\mathcal{F}}$ refers to the Frobenius norm, subject to the following orthogonality constraints:

$$(U^{(m)})^T U^{(m)} = I, \ (W^{(m)})^T W^{(m)} = I, \ (U^{(m)})^T W^{(m)} = 0. \tag{11}$$

For fixed $U^{(m)}(V^*)^T$, the optimal $T^{(m)} = W^{(m)}(V^{(m)})^T$ is a low-rank approximation of $\| \frac{1}{\sqrt{n_m}} X^{(m)} - S^{(m)} \|_{\mathcal{F}}^2$, where each rank sequentially captures the maximal variance remained in each data matrix after removing the shared variability. Likewise, for fixed $W^{(m)}(V^{(m)})^T$, each rank of the optimal $S^{(m)} = U^{(m)}(V^*)^T$ sequentially captures the maximal variance remained across all tissues after removing the tissue-specific variance.

In transcriptomics studies, it is widely believed that the differences in gene expressions between cell and tissue types are largely determined by transcripts derived from a small number of tissue-specific genes (Jongeneel *et al.*, 2005). Therefore, it seems reasonable that in our application of multi-tissue comparison of gene expressions, for each PPJ, the corresponding column in the transformation matrix should feature a limited number of non-zero entries. In such a scenario, a sparse representation will not only generate more reliable statistical models by excluding noise features but also offer more biological insight into the underlying cellular mechanism (Ma and Huang, 2008).

In the context of MVMF, we induce sparse estimates of V^* and $V^{(m)}$ by adding penalty terms to the loss function $\ell\ (U, W, V^*, V)$ as in (10). Specifically, we minimize:

$$\ell\ (U, W, V^*, V) + 2 \cdot M \cdot \ \|V^*\Lambda^*\|_1 + 2\sum_{m=1}^{M} \ \|V^{(m)}\Lambda^{(m)}\|_1 \tag{12}$$

where $\|\|_1$ denotes the ℓ_1 norm. Λ^* and $\Lambda^{(m)}$ are $d \times d$ and $r \times r$ diagonal matrices, respectively. In both matrices, the kth diagonal entry

is a non-negative regularization parameter for the kth column of the corresponding transformation matrix and the kth column tends to have more zero entries as the kth diagonal entry increases. In practice, a parsimonious parametrization may be employed where $\Lambda^* = \lambda_1 I_d$ and $\Lambda^{(m)} = \lambda_2 I_r$ for $m = 1, \ldots, M$ so that the number of parameters to be specified is greatly reduced. Alternatively, Λ^* and $\Lambda^{(m)}$ may be set such that a specified number of variables are selected in each column of $\hat{V}^*$ and $\hat{V}^{(m)}$.

The optimization problem (12) with constraints (11) is not jointly convex in $U^{(m)}, W^{(m)}, V^{(m)}$ and V^* for $m = 1, 2, ..., M$ (for instance the orthogonality constraints are non-convex in nature), hence gradient descent algorithms will suffer from multiple local minima (Gorski *et al.*, 2007). We propose to solve the optimization problem by alternately minimizing with respect to one parameter in $U^{(m)}$, $W^{(m)}$, V^*, $V^{(m)}$ while fixing all remaining parameters and repeating this procedure until the algorithm converges numerically. The minimization problem with respect to V^* or $V^{(m)}$ alone is strictly convex, hence in these steps, a coordinate descent algorithm (CDA) is guaranteed to converge to the global minimum (Friedman *et al.*, 2007). CDA iteratively update the parameter vector by cyclically updating one component of the vector at a time, until convergence. On the other hand, the minimization problem with respect to $W^{(m)}$ or $U^{(m)}$ is not convex. For fixed V^* and $V^{(m)}$, the estimates of $W^{(m)}$ and $U^{(m)}$ that minimize (12) can be jointly computed via a closed form solution. Assuming we have obtained initial estimates of V^* and $V^{(m)}$, we cyclically update the parameters in the following order:

$$(U^{(m)}, W^{(m)}) \rightarrow V^{(m)} \rightarrow V^*$$

Here $U^{(m)}$ and $W^{(m)}$ are jointly estimated in the first step, and in the subsequent steps, $V^{(m)}$ and V^* are updated separately, while keeping the previous estimates fixed. A detailed explanation of how each update is performed is in order.

First we reformulate the estimation problem as follows: we bind the columns of $U^{(m)}$ and $W^{(m)}$ and define the $n_m \times (d + r)$ augmented matrix: $\tilde{U}^{(m)} = [U^{(m)} \ , \ W^{(m)}]$; we then bind the columns of V^* and $V^{(m)}$ and define the $p \times (d + r)$ matrix: $\tilde{V}^{(m)} = [V^* \ , \ V^{(m)}]$. As such:

$$\ell\ (U, W, V^*, V^{(m)}) = \sum_{m=1}^{M} \ \| \frac{1}{\sqrt{n_m}} X^{(m)} - \tilde{U}^{(m)} (\tilde{V}^{(m)})^T \|_{\mathcal{F}}^2$$

and the constraints in (11) can be combined into:

$$(\tilde{U}^{(m)})^T \tilde{U}^{(m)} = I_{d+r}$$

Fixing $\tilde{V}^{(m)}$, the estimate of $\tilde{U}^{(m)}$ can be obtained by the reduced-rank Procrustes rotation procedure which seeks the optimum rotation of $X^{(m)}$ such that the error $\| \frac{1}{\sqrt{n_m}} X^{(m)} - \tilde{U}^{(m)} (\tilde{V}^{(m)})^T \ \|_{\mathcal{F}}^2$ is minimal. For a proof of this, see Zou *et al.* (2006). We obtain the SVD of $\frac{1}{\sqrt{n_m}} X^{(m)} \tilde{V}^{(m)}$ as PQR^T and compute the estimate of $\tilde{U}^{(m)}$ by: $\hat{\tilde{U}}^{(m)} = PR^T$.

Next, we fix $U^{(m)}$, $W^{(m)}$ and V^* while minimizing (12) with respect to $V^{(m)}$. For each fixed m, varying $V^{(m)}$ only changes the objective function via the summand indexed (m). Hence, it is sufficient to minimize:

$$\| \frac{1}{\sqrt{n_m}} X^{(m)} - U^{(m)} (V^*)^T - W^{(m)} (V^{(m)})^T \|_{\mathcal{F}}^2 + 2 \| V^{(m)} \Lambda^{(m)} \|_1. \quad (13)$$

This function is strictly convex in $V^{(m)}$, and the CDA is guaranteed to converge to the global minimum. We drop the superscript (m) in the following derivation for convenience and denote the jth column of the matrix V by V_j. In each iteration, the estimate of V_j is found by equating the first derivative of (13) with respect to V_j to zero. Hence:

$$-2 \left(\frac{1}{\sqrt{n_m}} X - U V^* - W V^T \right)^T W_j + 2 \Lambda_j \cdot \nabla (|V_j|) = 0,$$

where ∇ is the gradient operator. Substitute (11) and rearrange to give:

$$V_j = \frac{1}{\sqrt{n_m}} X^T W_j - \Lambda_j \cdot \nabla (|V_j|)$$

We define the sign function $\sigma(y)$ which equals 1 if $y > 0$, –1 if $y < 0$ and 0 if $y = 0$. First, note the derivative of the function $|y|$ is $\sigma(y)$ if $y \neq 0$ and a real number in the interval $(-1, 1)$ otherwise. Rearrange the previous equation to obtain the updated estimate in each iteration:

$$\hat{V}_j^{(m)} = S_{\Lambda_j^{(m)}} \left(\left(\frac{1}{\sqrt{n_m}} X^{(m)} \right)^T W_j^{(m)} \right) \quad (14)$$

where $S_\lambda(y)$ is a soft-thresholding function on vector y with non-negative parameter λ, such that $S_\lambda(y) = \sigma(y) \cdot \max\{|y| - \lambda, 0\}$ and $\Lambda_j^{(m)}$ is the jth diagonal entry of $\Lambda^{(m)}$.

In the third step, we fix the estimates of $U^{(m)}$, $W^{(m)}$ and $V^{(m)}$ and minimize (12) with respect to V^*. The objective function becomes:

$$\ell + 2 \cdot M \cdot \ \| V^* \Lambda^* \ \|_1 \quad (15)$$

where ℓ is defined in (10). As in the second step, we use a CDA in each iteration and the updated estimate of V_i^* is found by equating the first derivative of (15) to zero. Specifically:

$$-2 \sum_{m=1}^{M} \left\{ \left[\frac{1}{\sqrt{n_m}} X^{(m)} - U^{(m)} V^* - W^{(m)} \left(V^{(m)} \right)^T \right]^T U_i \right\} + 2 \cdot M \cdot \Lambda_i^* \cdot \nabla (|V_i^*|) = 0,$$

where Λ_i^* is the ith diagonal entry of Λ^*. Applying (11), this can be re-arranged into:

$$M \cdot V_i^* = \sum_{m=1}^{M} \left(\frac{1}{\sqrt{n_m}} X^{(m)} \right)^T U_i^{(m)} - M \cdot \Lambda_i^* \cdot \nabla (|V_i^*|),$$

Using the soft-thresholding and the sign functions, the updated estimate in each iteration can be re-written as:

$$\hat{V}_i^* = S_{\Lambda_i^*} \left(\frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{\sqrt{n_m}} X^{(m)} \right)^T U_i^{(m)} \right) \quad (16)$$

The cyclic CDA requires initial estimates of V^* and $V^{(m)}$, which are obtained as follows. First, we set an initial value to V^*, which explains as much variance in all datasets in $\mathcal{X}$ as possible. This amounts to a PCA on the $(\sum_{m=1}^{M} n_m) \times p$ matrix $\check{X}$ obtained by binding the rows of all data matrices $\frac{1}{\sqrt{n_m}} X^{(m)}$, $m = 1, ..., M$. We compute the truncated SVD of $\check{X}$ and obtain $\check{X} = \check{U} D B^T$ where D contains the d largest eigenvalues of $\check{X}^T \check{X}$. The initial estimate of V^* is then defined as:

$$(\hat{V}^*)^T = \frac{1}{M} D B^T, \quad (17)$$

and $\hat{U}^{(m)}$ is defined by the corresponding rows of $\check{U}$ in the SVD. For the tissue-specific transformation matrices $V^{(m)}$, we compute the

SVD of the residuals after removing the shared variance component from $\frac{1}{\sqrt{n_m}}X^{(m)}$, which gives: $\frac{1}{\sqrt{n_m}}X^{(m)} - \hat{U}^{(m)}\hat{V}^{*} = W^{(m)}R^{(m)}(Q^{(m)})^{T}$. The initial estimate of $V^{(m)}$ is defined as:

$$(\hat{V}^{(m)})^{T} = R^{(m)}(Q^{(m)})^{T}. \quad (18)$$

A summary of the estimation procedure is given in Algorithm 1.

Algorithm 1. sMVMF estimation algorithm

Input: data $\mathcal{X}$; parameters d, r, $\Lambda^{(m)}$, Λ^{*} for $m = 1, 2, \ldots, M$.
Output: $U^{(m)}$, $W^{(m)}$, $V^{(m)}$, for $m = 1, 2, \ldots, M$ and V^{*}.
1: Get initial estimates of $V^{(m)}$ for $m = 1, 2, \ldots, M$ and V^{*} as in (18) and (17).
2: **while** not convergent **do:**
3: Apply SVD: $\frac{1}{\sqrt{n_m}}X^{(m)}\tilde{V}^{(m)} = PQR^{T}$ and set $\hat{U}^{(m)} = PR^{T}$.
4: Use CDA to estimate $V^{(m)}$ according to (14).
5: Use CDA to estimate V^{*} according to corollary (16).

The sMVMF contains two sets of parameters: the tissue-specific sparsity parameters $\Lambda^{(m)}$, Λ^{*} and the (d, r) pair. Both d and r balance model complexity and the amount of variance explained. We select the smallest possible values of d and r such that a prescribed proportion of variance is explained. For a fixed (d, r) pair, we propose to optimize the model with respect to the choice of sparsity parameters using a variable selection procedure called 'stability selection', which is particularly effective in improving variable selection accuracy and reducing the number of false positives in high-dimensional settings (Meinshausen and Bühlmann, 2010). Stability selection consists of fitting the sparse model to a large number of randomly generated subsamples, each of which typically contains half of the subjects. Variable selection results across all subsamples are collected to compute empirical selection probabilities. A cutoff probability value is then chosen and the variables whose selection probability (SP) is larger than this threshold are selected by this procedure. One of the appealing features of this approach is that the ranking of variables, especially the high-ranking variables, is generally insensitive to the choice of regularization parameters. An overview of the stability selection procedure is given in Supplementary Material, Section A.

3 Illustration with simulated data

In this section, we present simulation studies to characterize how the sMVMF method is able to distinguish between shared and tissue-specific variance. We simulate shared and tissue-specific variance patterns as illustrated by the middle and right panels in Figure 1. We then test whether sMVMF correctly decomposes the total sample variance (left panel) while detecting variables contributing to the non-random variability within each variance component. We also compare sMVMF with two alternative methods: standard PCA and Levene's test (Gastwirth *et al.*, 2009) of the equality of variance between population groups.

3.1 Simulation setting

Our simulation study consists of 1000 independent experiments. In each experiment, we simulate three data matrices or datasets (tissues) of dimension $n = 100$ (samples) and $p = 500$ (genes). Each simulated data matrix $X^{(m)}$ is obtained via:

$$X^{(m)} = Y^{(m)} + Z^{(m)} + E^{(m)},$$

where $Y^{(m)}$ is a component designed to control the shared variance, $Z^{(m)}$ is introduced to control the tissue-specific variance and $E^{(m)}$ is a random error. They are all $n \times p$ random matrices. Since we ultimately wish to test whether our method is able to distinguish between signal and noise variables, we assume that only the first 30 variables carry the signal, whereas the remaining 470 only introduce noise.

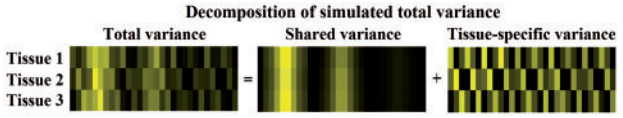

Fig. 1. Simulated patterns of sample variance: the total, non-random, sample variance of 30 signal-carrying random variables is generated so that it can be decomposed into the sum of shared and tissue-specific components. Rows correspond to tissues (datasets) and columns correspond to 30 variables. Brighter colours represent large variance and darker colours represent low variance. Although by construction the underlying shared and tissue-specific variances have very different patterns, sMVMF is able to discriminate between them

We suppose that the shared variability is controlled by the activation of 3 LFs, each regulating the variance of a different block of variables. To this end, we further group the 30 signal variables into three blocks of 10 normally distributed random variables each (numbered 1–10, 11–20 and 21–30), as illustrated in Figure 2A. We design the simulations so that each of the first 30 variables in Y has the same variance in different datasets; moreover, the variance decreases while moving from the first to the third block. Further details and simulation parameters are available in the Supplementary Material, Section B. This procedure generates shared variance patterns that look like those reported in the middle panel of Figure 1.

The variables in Z are also assumed to be normally distributed. They are generated such that exactly 10 of them have the largest variance across datasets. The resulting 'mosaic' structure of the simulated variance patterns is illustrated in right panel of Figure 1. The data matrices $Y^{(m)}$ and $Z^{(m)}$ are generated such that the total non-random sample variance of each variable in a tissue equals the sum of its shared and tissue-specific variances, which is also illustrated in Figure 1. The random error term $E^{(m)}$ is generated from independent and identical standard normal distributions for all variables in all datasets. As a result of this simulation design, we are able to characterize the true underlying architecture that explains the total sample variance.

3.2 Simulation results

The data generated in each experiment were analysed by fitting the sMVMF algorithm. To focus on the ability of the model to disentangle the true sources of variability, we take $d = 3$ and $r = 1$, which equal the true number of shared and tissue-specific LFs used to generate the data. The regularization parameters Λ^{*} and $\Lambda^{(m)}$ are tuned such that each PPJ consists of 10 variables, the true number of signal variables.

For comparison, we propose two additional approaches that are able to identify variables featuring dataset-specific sample variances, although they do not attempt to model the shared variance. The first method consists of carrying out a separate PCA on each dataset; for each PCA/dataset, we then select the 10 variables having the largest loadings in the first PC. The second method consists of applying a standard Levene's test of equality of population variances independently for each variable, which is then followed by a Bonferroni adjustment to control the family-wise error rate; if a test rejects the null hypothesis at the 5% significance level, we select the variable having the largest sample variance among the three datasets.

By averaging across 1000 experiments, we are able to estimate the probability that each one of the 30 signal variables is selected by

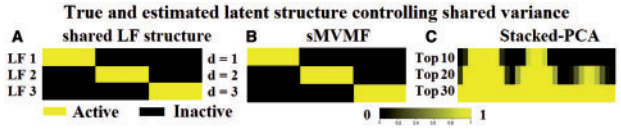

Fig. 2. Each LF is only active in a block of 10 signal-carrying variables and controls the amount of variance of those variables that is shared among datasets. (**A**) The true latent structure used to generate the data. (**B**) and (**C**) The estimated probabilities that each variable has been selected as signal-carrier using sMVMF and a stacked-PCA approach, respectively. sMVMF accurately captures the true shared LF structure, whereas stacked-PCA tends to identify variables with large variance but fails to identify the LF structure

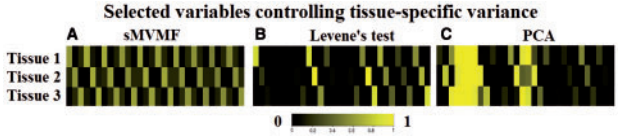

Fig. 3. Three different methods—sMVMF, Levene's test and PCA—are used to detect random variables whose variance pattern is dataset-specific. Each heatmap represents the selection probabilities estimated by each method. (**A**) sMVMF produces patterns that closely match the true tissue-specific variances shown in the right panel of Figure 1. (**B**) Levene's test performs well for variables those variance is mostly driven by tissue-specific factors but fails to detect those variables having a strong shared-variance component. (**C**) The PCA-based method cannot distinguish between shared and tissue-specific variability, and fails to recover the true pattern

each one of the three competing methods. The heatmaps (A–C) in Figure 3 visually represent these selection probabilities. Here sMVMF perfectly identifies the variables that introduce dataset-specific variability. The results obtained using Levene's tests are somewhat similar, except for some variables in the first block (indexed 3–8) and second block (indexed 14–17). By reference to the middle panel of Figure 1, it can be noted that these variables are precisely those featuring large shared variability by construction. On the other hand, the PCA-based approach performs poorly because it can only select variables that contribute to explaining the total sample variance but is unable to capture dataset-specific patterns. This example is meant to illustrate the limitations of both univariate and multivariate approaches that do not explicitly account for factors driving shared and dataset-specific effects. sMVMF has been designed to address exactly these limitations.

Both Levene's test and the individual-PCA approach are not designed to capture shared variance patterns. As a way of direct comparison with sMVMF, we therefore propose an alternative PCA-based approach that has the potential to identify variables associated to the direction of largest variance across all three datasets. This method consists of performing a single PCA on a 'stacked' matrix of dimension $(Mn) \times p$ containing measurements collected from all three datasets and obtained by coalescing the rows of the three individual data matrices. By varying the cutoff value for thresholding the loadings of the first PC, we are able to select the top 10, 20 and 30 variables. We shall refer to this approach as stacked-PCA.

Results produced by sMVMF and stacked-PCA are summarized by the heatmaps (B) and (C) in Figure 2 and can be directly compared with the true simulated patterns in (A). As expected, stacked-PCA tends to select variables having large total sample variances, whereas sMVMF can identify variables affected by each shared LF which jointly explain a large amount of variance. This example shows that sMVMF is able to identify the variables associated to the LFs controlling the shared variance.

We also carried out a simulation, based upon the same setting, with smaller signal-to-noise ratio, i.e. by sampling the random error terms in $E^{(m)}$ from independent normal distributions having larger variance. The results were very similar to the previous setting, except that Levene's test was hardly able to identify any tissue-specific genes. The heatmaps summarizing model performances are given in Supplementary Material, Section C.

4 Application to the TwinsUK cohort

4.1 Data preparation

TwinsUK is one of the most deeply phenotyped and well-characterized adult twin cohort in the world (Moayyeri *et al.*, 2013). It has been widely used in studying the genetic basis of aging procession and complex diseases (Codd *et al.*, 2013). More importantly, it contains a broad range of 'omics' data including genomic, epigenomic and transcriptomic profiles among others (Bell *et al.*, 2012). In this study, we focus on comparing the variance of mRNA expressions in adipose (subcutaneous fat), lymphoblastoid cell lines (LCLs) and skin tissues. The microarray data used in this study were obtained from the Multiple Tissue Human Expression Resource (Nica *et al.*, 2011), with participants being recruited from the TwinsUK registry. Peripheral blood samples were artificially transformed from mature blood cells by infecting them with the Epstein–Barr virus (Glass *et al.*, 2013). All tissue samples were collected from 856 female Caucasian twins (154 monozygotic twin pairs, 232 dizygotic twin pairs and 84 singletons) aged between 39 and 85 years (mean 62 years). Genome-wide expression profiling was performed using Illumina Human HT-12 V3 BeadChips, which included 48 804 probes. Log2-transformed expression signals were normalized per tissue using quantile normalization of the replicates of each individual followed by quantile normalization across all individuals, as described in Nica *et al.* (2011). In addition, we also had access to 450 K methylation data of the same adipose biopsies profiled using Infinium HumanMethylation 450 K BeadChip Kit (Wolber *et al.*, 2014). We only retained probes whose expression levels were measured in all three tissues and removed subjects comprising unmeasured expressions in any tissue. Using the same notation introduced before, this resulted in three data matrices each of dimension $n = 618$ and $p = 26\,017$. For each probe in each tissue, a linear regression model was fitted to regress out the effects of age and experimental batch, following the same procedure as in Grundberg *et al.* (2012). Residuals in adipose, LCL and skin tissues were arranged in $n \times p$ matrices $X^{(1)}, X^{(2)}, X^{(3)}$, respectively, for further analysis using the proposed MVMF method.

4.2 Experimental results

Non-sparse MVMF was initially fitted for all combination of parameter pairs (d, r) in a grid. For each model fit, we computed the percentage of variance explained in each tissue. These are shown in the 3D bar charts presented in the Supplementary Figure S5. The percentages of variance explained varied between 25.2% ($d = r = 1$, LCL) and 87.3% ($d = r = 160$, skin). The following analyses are based on the $d = r = 3$ setting, which explains at least 40% of expression variance across tissues. Given that there are more than 26 000 probes, and this is much larger than the sample size, this choice of parameters offers a good balance between dimensionality reduction and retaining a large portion of total variance. Although two other combinations of (d, r), i.e. (2, 4) and (4, 2), also explain a similar amount of total variance, we have found that the gene ranking results are not extremely sensitive to these values. For

more details on this sensitivity analysis, see Supplementary Material, Section D.

The sparse version of our model, sMVMF, was then fitted to each subsample in stability selection procedure to rank gene expressions explaining a large amount of shared and tissue-specific variances respectively. A detailed description of the procedure is presented in Supplementary Material, Section A. In summary, 1000 random subsamples were generated each consisting of 309 subjects randomly and independently sampled without replacement from a total of 618. No twin pair was included in any subsample to remove possible correlations due to zygosity. sMVMF was fitted to each subsample, where the sparsity parameters were fixed such that each column of the transformation matrices comprised exactly 100 non-zero entries. There were 3274 mRNA expression probes that were selected at least once from any of the transformation matrices.

Probes that explain a large amount of expression variance exclusively in one tissue are of particular interest. To make such probes visually discernible, we propose a new visualization tool, the SPOW (Selection PrObability Wheel) plot. The plot in Figure 4 consists of 3274 fan slices corresponding to probes that are selected at least once in all subsamples, re-ordered by their selection probabilities in $\hat{V}^*$. The wheel is further divided into four rings, representing shared, adipose-, LCL- and skin tissue, respectively. Each ring is assigned a unique colour spectrum to illustrate selection probabilities of the probes: brighter colours denote a higher probability and darker colours denote a lower probability. Probes featuring exclusively shared or tissue-specific variability can be found along the radii where only one part is painted in a bright colour and the other three parts are coloured in black. The SPOW plots for the top 200 probes that explain shared and tissue-specific variability, respectively, are presented in Supplementary Figures S6–S9, where such probes can be more easily captured.

Four groups of mRNA expressions were selected for further investigation, corresponding to shared-exclusive, adipose-, LCL- and skin-exclusive expressions. Each group consisted of probes whose selection probabilities were larger than 0.5 in the corresponding transformation matrix and less than 0.005 in the other transformation matrices. These thresholds were set to give a manageable number of featured gene probes while tolerating occasional selection in the other groups. This procedure selected 294 genes for further study, including 114 adipose-exclusive, 83 LCL-exclusive, 64 skin-exclusive and 33 shared-exclusive genes. We summarize the results in Table 1. A Venn-diagram representation of the results is given in Supplementary Material, Section E.

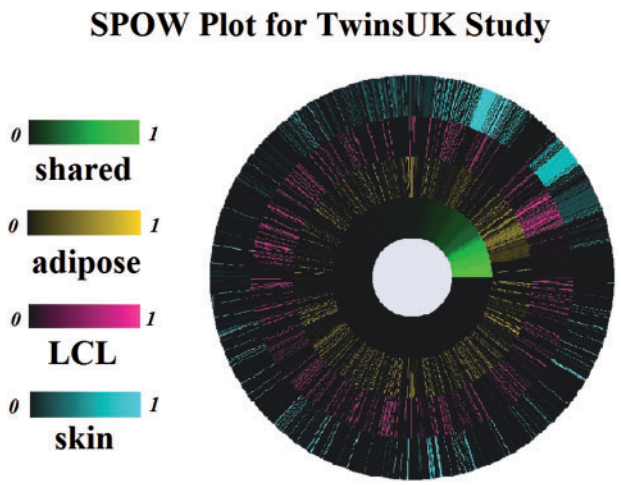

Fig. 4. TwinsUK study: resulting SPOW plot. The wheel comprises four rings, which correspond to shared, adipose-, LCL- and skin-specific variability from the inner ring. It is also evenly divided into 3274 fan slices, corresponding to 3274 mRNA expression probes that are selected at least once in all subsamples. Probes are re-ordered by their selection probabilities in the transformation matrix in the shared component. Brighter colour denotes higher probability, whereas darker colour denotes lower probability. We are particularly interested in probes with high SP exclusively in one ring

For each tissue, we performed an enrichment test by overlapping genes in our list with genes contained in the TiGER and VeryGene databases to examine the extent of agreement. In addition, a Gene Ontology biological process pathway enrichment test (Ashburner *et al.*, 2000) and a Cytoscape pathway (CP) analysis (Saito *et al.*, 2012) were carried out to reveal the function of the pathways which the 261 tissue-exclusive genes belonged to and FDR-corrected *P* values were reported (see Supplementary Tables T1 and T2 for full results). Below we present test results for each group of genes separately for each tissue. We also report the SP for some selected probes.

4.2.1 Skin-exclusive genes

Fifteen of the 64 genes from our skin-exclusive list are contained in the combined TiGER/VeryGene list, giving rise to significant enrichment of our list with Fisher exact test *P* value $P < 10^{-16}$. The overlapping genes include serine protease family genes KLK5 (SP: 1.000) and KLK7 (SP: 1.000), which are highly expressed in the epidermis and related to various skin conditions, such as cell shedding (desquamation) (Brattsand and Egelrud, 1999). Another member ALOX12B (SP: 1.000) controls producing 12 R-LOX, which adds an oxygen molecule to a fatty acid to produce the 12 R-hydroperoxyeicosatetraenoic acid that has major function in the skin cell proliferation and differentiation (de Juanes *et al.*, 2009). The skin-exclusive genes have also been found significantly enriched in two biological processes, namely epidermis development and cell–cell adhesion ($P < 0.001$ and $P = 0.03$, respectively).

4.2.2 LCL-exclusive genes

LCLs are not natural human cells: they are laboratory-induced immortal cells that have abnormal telomerase activity and tumorigenic property (Sie *et al.*, 2009). Since neither TiGER nor VeryGene assessed transcriptomic profile in LCL cells, we obtained LCLs data from Li *et al.* (2010), in which the authors compared LCLs expression profile in four human populations and reported 282 LCL-specific expression genes. Nine of those genes are contained in our LCL-exclusive gene list, giving a Fisher exact test $P < 10^{-16}$. These include CDK5R1 (SP: 0.961) and HEY1 (SP: 1.000), which are key genes in the transformation of B lymphocytes to LCLs (Zhao *et al.*, 2006). Pathway analysis of the LCL-exclusive genes reveals several aging and cell-death-related pathways such as regulation of telomerase (CP enrichment test, $P = 0.014$), small cell lung cancer

Table 1. TwinsUK study: summary of results

Tissues	% of variance explained by tissue-specific component	% of variance explained by shared component	Number of tissue-exclusive probes	Number of tissue-exclusive genes
Adipose	27.0	14.7	132	114
LCL	30.8	12.1	91	83
Skin	32.6	11.5	74	64

There are additionally 33 shared-exclusive genes.

(CP enrichment test, $P = 0.019$) and cell cycle checkpoints (CP enrichment test, $P = 0.021$). These results show that our tissue-exclusive genes represent tissue unique molecular functions and biological pathways, which may be used to validate known pathways or discover new biological mechanisms.

4.2.3 Adipose-exclusive genes

ApoB (SP: 1.000) is the only member in our adipose-exclusive list which is also contained in the list of known adipose-specific expression genes (Fisher exact test, $P = 0.05$). ApoB is one of the primary apolipoproteins that transport cholesterol to peripheral tissues (Knott *et al.*, 1986) and it has been widely linked to fat formation (Riches *et al.*, 1999). In adipose, the selected genes are found significantly enriched in triglyceride catabolic process pathway ($P = 0.022$), which is in line with the fact that adipose tissue is the major storage site for fat in the form of triglycerides. Pathway analysis reveals that genes in the adipose-exclusive list are significantly enriched in triglyceride catabolic process pathway ($P = 0.022$), which agrees with the fact that adipose tissue is the major storage site for fat in the form of triglycerides. In addition, these genes are enriched in inflammation pathways, such as lymphocyte chemotaxis ($P = 0.016$) and neutrophil chemotaxis ($P = 0.027$). This coincides with previous findings of the complex and strong link between metabolism and immune system in adipose tissue (Tilg and Moschen, 2006).

For this tissue, we were also able to further investigate the causes for the observed adipose-exclusive gene expression variability. One possible explanation could be that environmental factors influenced an individual's epigenetic status, which subsequently regulated gene expression (Razin and Cedar, 1991). As a mediator of gene regulatory mechanisms, DNA methylation is crucial to genomic functions such as transcription, chromosomal stability, imprinting and X-chromosome inactivation (Lokk *et al.*, 2014), which consequently influence an individual's tissue development (Ziller *et al.*, 2013). It thus seemed reasonable to hypothesize that the expression of tissue-exclusive genes could be modified by their methylation status in the same tissue.

We sought to identify genes featuring a statistically significant linear relationship between the gene's methylation profile and its expression value from the same tissue. In adipose biopsies, where both transcriptome and methylation data are available, we found that 68.4% (78 out of 114 genes) of the genes had expression levels significantly associated with their methylation status using a linear fit (Bonferroni correction, $P < 0.05$) (see Supplementary Table T3, for full lists). We then wanted to assess whether a similar number of linear associations could be found by chance only by randomly selecting any genes, not only those that feature adipose-exclusive variability and testing for association between gene expression and methylation levels. This was done by randomly extracting the same, fixed number (132) of expression probes and corresponding methylation levels from adipose tissue and fitting a linear model as before. By repeating this experiment 1000 times, we obtained the empirical distribution reported in Supplementary Figure S10. This distribution suggested that all the proportions were below 0.2, compared with our observed proportion of 0.684, which provided overwhelming evidence that DNA methylation was an important factor affecting the expression of the tissue-exclusive genes. It was notable that the adipose-exclusive variability of ApoB was regulated by methylation at 50 bp upstream of the Transcriptional Starting Site (linear fit, $P = 2.1 \times 10^{-5}$), which agreed with the findings that the promoter of ApoB has tissue-specific and species-specific methylation property (Apostel *et al.*, 2002). Apart from ApoB, we also found that methylation in Syk was associated with Syk expression level, which was potentially involved in B cell development and cell apoptosis (Ma *et al.*, 2010).

5 Conclusion

The proposed sMVMF method facilitates the comparison of gene expression variances across multiple tissues. The primary challenge of this task arises from the interference between substantial co-variability of gene expressions across all tissues and substantial variability of gene expressions featured only in specific tissues. Characterizing tissue-specific variability can shed light on the biological processes involved with tissue differentiation. Analysing shared variability can potentially reveal genes that are involved in complex or basic biological processes and may as well enhance the estimation of tissue-specific variability.

sMVMF has been used here to compare gene expression variances in three human tissues from the TwinsUK cohort. Two hundred sixty-one genes having substantial expression variability exclusively featured in one tissue have been identified. Enrichment tests showed significant overlaps between our lists of tissue-exclusive genes and those reported in the TiGER and VeryGene databases, which were established by comparing mean expression levels. This confirms the link between tissue-specific expression variance and the biological functions associated with particular tissues. In future work, it would be interesting to explore the functions of the tissue-exclusive genes from our list that have not been reported in existing databases. We further showed adipose-exclusive expression variability was driven by an epigenetic effect. Using these results as a guiding principle, we expect our methods and results could improve efficiencies in mapping functional genes by reducing the multiple testing and enhancing the knowledge of gene function in tissue development and disease phenotypes. Future works would consist of investigating the outcome of tissue-exclusive expression variability, for which we can perform association studies between expressions of tissue-exclusive genes and disease phenotypes related to adipose and skin tissues.

Funding

The Biological Research Council has supported Z.W. (DCIM-P31665) and the TwinsUK study. We also thank the European Community's Seventh Framework Programme (FP7/2007-2013) and the National Institute for Health Research (NIHR) for their support in the TwinsUK study.

Conflict of Interest: none declared.

References

Apostel,F. *et al.* (2002) Reduced expression and increased cpg dinucleotide methylation of the rat apobec-1 promoter in transgenic rabbits. *Biochim. Biophys. Acta*, **1577**, 384–394.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bell,J. *et al.* (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. *PLoS Genet.*, **8**, e1002629.

Brattsand,M. and Egelrud,T. (1999) Purification, molecular cloning, and expression of a human stratum corneum trypsin-like serine protease with possible function in desquamation. *J. Biol. Chem.*, **274**, 30033–30040.

Cheung,V. *et al.* (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.

Codd,V. *et al*. (2013) Identification of seven loci affecting mean telomere length and their association with disease. *Nat. Genet.*, **45**, 422–427.

Coulon,A. *et al*. (2013) Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat. Rev. Genet.*, **14**, 572–584.

de Juanes,S. *et al*. (2009) Development of an ichthyosiform phenotype in alox12b-deficient mouse skin transplants. *J. Invest. Dermatol.*, **129**, 1429–1436.

Friedman,J. *et al*. (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.*, **2**, 302–332.

Gastwirth,J. *et al*. (2009) The impact of Levene's test of equality of variances on statistical theory and practice. *Stat. Sci.*, **24**, 343–360.

Glass,D. *et al*. (2013) Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol.*, **14**, R75.

Gorski,J. *et al*. (2007) Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math Methods Oper. Res.*, **66**, 373–401.

Grundberg,E. *et al*. (2012) Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.*, **44**, 1084–1089.

Ho,J. *et al*. (2008) Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics*, **24**, 390–398.

Jongeneel,C. *et al*. (2005) An atlas of human gene expression from massively parallel signature sequencing (mpss). *Genome Res.*, **15**, 1007–1014.

Knott,T. *et al*. (1986) Complete protein sequence and identification of structural domains of human apolipoprotein b. *Nature*, **323**, 134–138.

Lage,K. *et al*. (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl. Acad. Sci. USA.*, **105**, 20870–20875.

Li,J. *et al*. (2010) Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comput. Biol.*, **6**, e1000910.

Liu,X. *et al*. (2008) Tiger: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Lokk,K. *et al*. (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.*, **15**, r54.

Ma,L. *et al*. (2010) The relationship between methylation of the syk gene in the promoter region and the genesis of lung cancer. *Clin. Lab.*, **56**, 407–416.

Ma,S. and Huang,J. (2008) Penalized feature selection and classification in bioinformatics. *Brief. Bioinform.*, **9**, 392–403.

Mar,J. *et al*. (2011) Variance of gene expression identifies altered network constraints in neurological diseases. *PLoS Genet.*, **7**, e1002207.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. B*, **72**, 417–473.

Moayyeri,A. *et al*. (2013) Cohort profile: Twinsuk and healthy ageing twin study. *Int. J. Epidemiol.*, **42**, 76–85.

Nica,A. *et al*. (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.*, **7**, e1002003.

Ong,C.-T. and Corces,V. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.

Ponnapalli,S.P. *et al*. (2011) A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. *PLoS One*, **6**, 1–11.

Razin,A. and Cedar,H. (1991) DNA methylation and gene expression. *Microbiol. Mol. Biol. Rev.*, **55**, 451–458.

Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.

Riches,F. *et al*. (1999) Reduction in visceral adipose tissue is associated with improvement in apolipoprotein b-100 metabolism in obese men. *J. Clin. Endocrinol. Metab.*, **84**, 2854–2861.

Saito,R. *et al*. (2012) A travel guide to cytoscape plugins. *Nat. Methods*, **9**, 1069–1076.

Sie,L. *et al*. (2009) Utility of lymphoblastoid cell lines. *J. Neurosci. Res.*, **87**, 1953–1959.

Tilg,H. and Moschen,A. (2006) Adipocytokines: mediators linking adipose tissue, inflammation and immunity. *Nat. Rev. Immunol.*, **6**, 772–783.

Tukey,J. (1949) Comparing individual means in the analysis of variance. *Biometrics*, **5**, 99–114.

van't Veer,L. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wolber,L. *et al*. (2014) Epigenome-wide DNA methylation in hearing ability: new mechanisms for an old problem. *PLoS One*, **9**, e105729.

Wu,C. *et al*. (2009) Combinatorial control of suicide gene expression by tissue-specific promoter and microRNA regulation for cancer therapy. *Mol. Ther.*, **17**, 2058–2066.

Wu,H. *et al*. (2014) Tissue-specific RNA expression marks distant-acting developmental enhancers. *PLoS Genet.*, **10**, e1004610.

Xia,Q. *et al*. (2007) Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm *Bombyx mori*. *Genome Biol.*, **8**, R162.

Xiao,X. *et al*. (2014) Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules. *PLoS Genet.*, **10**, e1004006.

Yang,X. *et al*. (2011) Verygene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery. *Physiol. Genomics*, **43**, 457–460.

Zhao,B. *et al*. (2006) RNAs induced by Epstein-Barr virus nuclear antigen 2 in lymphoblastoid cell lines. *Proc. Natl. Acad. Sci. USA.*, **103**, 1900–1905.

Ziller,M. *et al*. (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.

Zou,H. *et al*. (2006) Sparse principal component analysis. *J. Comput. Graph. Stat.*, **15**, 265–286.