BIOINFORMATICS

Vol. 00 no. 00 2016
Pages 1–3

HDTD: Analyzing multi-tissue gene expression data

Anestis Touloumis [1,2*], John C. Marioni [1,3] and Simon Tavaré [1]

[1] CRUK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, United Kingdom.
[2] Computing, Engineering and Mathematics, University of Brighton, Brighton, BN2 4GJ, United Kingdom.
[3] EMBL-European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.

Received on October 29, 2015; revised on March 30, 2016; accepted on April 15, 2016

Associate Editor: Inanc Birol

ABSTRACT

Motivation: By collecting multiple samples per subject, researchers can characterise intra-subject variation using physiologically relevant measurements such as gene expression profiling. This can yield important insights into fundamental biological questions ranging from cell type identity to tumour development. For each subject, the data measurements can be written as a matrix with the different subsamples (e.g., multiple tissues) indexing the columns and the genes indexing the rows. In this context, neither the genes nor the tissues are expected to be independent and straightforward application of traditional statistical methods that ignore this two-way dependence might lead to erroneous conclusions. Herein, we present a suite of tools embedded within the R/Bioconductor package *HDTD* for robustly estimating and performing hypothesis tests about the mean relationship and the covariance structure within the rows and columns. We illustrate the utility of *HDTD* by applying it to analyze data generated by the Genotype-Tissue Expression consortium.

Availability: The R package *HDTD* is part of Bioconductor. The source code and a comprehensive user's guide are available at http://bioconductor.org/packages/release/bioc/html/HDTD.html.

Contact: A.Touloumis@brighton.ac.uk

Supplementary information: Supplementary materials, including R code, data and results from the data analysis, are available at *Bioinformatics* online.

1 INTRODUCTION

The term "transposable data" refers to data that are naturally written in a matrix whose dimensions correspond to two distinct features of interest, while the term "high-dimensional" reflects the fact that the dimension of the subject-specific data matrix is larger than the number of subjects. High-dimensional transposable data can be found in genetics, e.g., when, for each subject, gene expression levels are measured in multiple tissues (Piccirillo *et al.*, 2015), in different fragments of the same tumour (Sottoriva *et al.*, 2013) or in a well-defined spatial order (Petretto *et al.*, 2010), in yeast expression studies (Smith and Kruglyak, 2008), in protein-signaling networks (Sachs *et al.*, 2005), in eQTL analysis (Bhadra and Mallick, 2013) and in other studies with EEG, fMRI and time-series data (cf. Touloumis *et al.*, 2014). To analyze robustly such datasets, we developed the R package *HDTD* (High-Dimensional Transposable Data).

In multiple-tissue gene expression studies, the rows correspond to genes and the columns to tissues, and genes and tissues might to be correlated with each other. Ignoring a potential tissue-wise correlation could be misleading in determining the strength of the gene-wise correlation (Touloumis *et al.*, 2014) and it may hinder the discovery of differentially expressed genes, since traditional ANOVA-type tests suffer from extremely low power and/or false positive findings (Touloumis *et al.*, 2015). The unique feature of *HDTD* is the implementation of sound statistical methods that account for and estimate both the tissue- and gene-wise correlation, thus facilitating reliable inference about the form of the mean gene expression levels and the functional relationship among genes and/or tissues.

2 STATISTICAL BACKGROUND

To introduce the notation, suppose that the gene expression levels for subject i are recorded in an $r \times c$ matrix $\boldsymbol{X}_i$ with rows the same set of r genes and columns the same set of c tissues. We assume that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ are independently and identically distributed. Inference about the mean relationship of the genes across the tissues and about the dependence structure relies on estimating and/or testing hypotheses about the mean matrix $\boldsymbol{M}$, the gene covariance matrix $\boldsymbol{\Sigma}_{\mathrm{R}}$, and the tissue covariance matrix $\boldsymbol{\Sigma}_{\mathrm{C}}$. In particular, the (a, b) element of $\boldsymbol{M}$ determines the mean expression level for gene a in tissue b, the (c, d) element of $\boldsymbol{\Sigma}_{\mathrm{R}}$ the covariance of genes c and d, and the (e, f) element of $\boldsymbol{\Sigma}_{\mathrm{C}}$ the covariance of tissues e and f. The covariance structure between two elements of a typical $\boldsymbol{X}$ has a Kronecker product form: $\mathrm{Cov}(X_{ij}, X_{lm}) = \boldsymbol{\Sigma}_{\mathrm{R}il}\, \boldsymbol{\Sigma}_{\mathrm{C}jm}$.

In practice it is often of interest to identify differentially expressed genes. For example, it is important to assess whether the overall mean pattern of gene expression levels remains constant across all or pre-specified tissue groups. To do this, *HDTD* implements the testing methods proposed by Touloumis *et al.* (2015).

To estimate $\boldsymbol{\Sigma}_{\mathrm{R}}$ and $\boldsymbol{\Sigma}_{\mathrm{C}}$, shrinkage approaches are employed. These have been found to be extremely useful in constructing

*to whom correspondence should be addressed

reliable gene networks (see Schäfer and Strimmer, 2005). The novel shrinkage covariance estimators derived in the Supplementary Material are statistically efficient and practical because they are invertible and easy to calculate regardless the number of genes and tissues. In addition, *HDTD* allows users to study correlation patterns of the genes or tissues by testing against known covariance structures (Touloumis *et al.*, 2014). The non-parametric nature of our analysis provides some robustness against non-normality.

3 MULTIPLE TISSUE EXAMPLE

Melé *et al.* (2015) investigated variability in the human transcriptome across multiple tissues by analyzing RNA sequencing (RNAseq) data from the Genotype-Tissue Expression project. This project identified, among other things, genes whose expression signature characterized particular tissues. To accomplish this, Melé *et al.* (2015) used essentially all available tissue-samples from each of the 175 individuals by aggregating gene expression levels across the tissue tested and the remaining tissues (see §3.5 in Supplementary Material in Melé *et al.*, 2015). This approach does not acknowledge the tissue-wise correlation and consequently, this can affect the discovery of tissue-specific gene lists (Touloumis *et al.*, 2015). Since *HDTD* requires measurements from the same set of tissues across subjects, we considered a subset of this dataset including only the subjects ($N = 11$) with available RNAseq samples across all the most frequently collected tissues (skin, nerve, adipose, artery, lung, skeletal muscle, heart, blood and thyroid). A $44,781 \times 9$ data matrix was created for each subject, with rows corresponding to genes, columns corresponding to the samples from the nine tissues and entries corresponding to the RPKM values. We use RPKM values for consistency with the original publication but we excluded genes where the sum of the RPKM values across the tissues was less than 0.1. To illustrate benefits when utilizing *HDTD*, we focused on two important inferential aspects: i) study of the dependence structure among the nine tissues and ii) corroboration of the gene signatures when the dependence between tissues is accounted for.

To study the tissue-specific variability, we estimated the corresponding covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{C}}$ (Table 1 in the Supplementary Material). Blood was by far the most variable tissue ($SE = 870.4$), with SE at least four times that of the other tissues. To study the tissue-wise correlation, we calculated the correlation matrix from $\widehat{\boldsymbol{\Sigma}}_{\mathrm{C}}$ (Table 2 in the Supplementary Material). We observed that lung, skeletal muscle, heart and thyroid were mildly correlated with each other (correlations ≥ 0.1), while the remaining tissues showed weaker strength of correlation. To investigate the statistical significance of our observation, we employed the sphericity test (Touloumis *et al.*, 2014) to all possible tissue pairs so as to identify correlated pairs of tissues. After applying an FDR correction, we failed to reject the sphericity hypothesis for the tissue pairs listed in Table 3 in the Supplementary Material. To summarize these results, there seems to exist a weak but statistically significant tissue-wise correlation pattern that needs to be considered when analyzing the gene expression pattern across tissues.

Melé *et al.* (2015) generated lists of genes that showed tissue-specific expression (Table S5 in Melé *et al.*, 2015). For a given tissue, we tested the hypotheses of conservation of the overall mean gene-expression levels of the corresponding genes-list between this tissue and any of the other eight, leading to a total of eight $p-$values, to which we applied an FDR correction. Failure to reject all hypotheses means that we do not have enough evidence that these genes are tissue-specific in their expression. After performing this analysis, we confirmed the validity of the tissue-specific gene-lists for skin, nerve, lung, skeletal muscle, heart and blood tissue. However, we failed to confirm that the overall mean gene-expression levels of the thyroid-specific gene-list is different in skeletal muscle (p-value $= 0.782$); that of the adipose-specific gene-list different in the skin (p-value $= 0.105$), and that of the artery-specific gene-list different in skin (p-value $= 0.668$), in adipose (p-value $= 0.716$), and in blood (p-value $= 0.145$). We also failed to reject the hypothesis that the mean gene-expression pattern for the artery-specific genes is simultaneously preserved across artery, skin, adipose and blood tissues (p-value$= 0.412$), which is in accordance with the pairwise tissue analysis. The difference in our conclusions compared to those in Melé *et al.* (2015) presumably arises because the methods in *HDTD* account for the presence of the tissue-wise correlation, regardless of its strength, a key inferential property that is not discussed by Melé *et al.* (2015).

4 SUMMARY

Although *HDTD* was motivated by and illustrated using multi-tissue gene expression data, we emphasize that *HDTD* is suitable for analyzing other types of high-dimensional transposable data including single-cell transcriptomics data (see Lee *et al.*, 2014; Lovatt *et al.*, 2014) sampled from different tissues. In these studies, *HDTD* should lead to more robust inference since it accounts for both the gene- and tissue-wise correlation and is reliable for large numbers of cells without a dramatic increase in the computational cost.

5 ACKNOWLEDGMENT

We acknowledge the support of The University of Cambridge, Cancer Research UK (C14303/A17197) and Hutchison Whampoa Limited.

REFERENCES

Bhadra, A. and Mallick, B.K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, **69**, 447–457.

Lee, J. H., *et al.* (2014). Highly multiplexed subcellular RNA sequencing in situ. *Science*, **343**, 1360–1363.

Lovatt, D., *et al.* (2014). Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue. *Nature Methods*, **11**, 190-196.

Melé, M., *et al.* (2015). The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Piccirillo, S.G.M., *et al.* (2015). Contributions to drug resistance in glioblastoma derived from malignant cells in the sub-ependymal zone. *Cancer Research*, **75**, 194–202.

Petretto, E., *et al.* (2010). New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology*, **6**, e1000737.

Sachs, K., *et al.* (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, **308**, 523–529.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, Art. 32.

Sottoriva, A., *et al.* (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, **110**, 4009–4014.

Smith, E.N. and Kruglyak, L. (2008). Gene environment interaction in yeast gene expression. *PLoS Biology*, **6**, e83.

Touloumis, A., Marioni, J.C. and Tavaré, S. (2014). Hypothesis testing for the covariance matrix in high-dimensional transposable data with Kronecker product dependence structure. *Submitted. arXiv:1404.7684v2*.

Touloumis, A., Tavaré, S. and Marioni, J.C. (2015). Testing the mean matrix in high-dimensional transposable data. *Biometrics*, **71**, 157–166.