SPECS: A non-parameteric method to identify tissue-specific molecular features for unbalanced sample groups

Celine Everaert[1,2*], Pieter-Jan Volders[1,2,3], Annelien Morlion[1,2], Olivier Thas[4,5°] and Pieter Mestdagh[1,2°]

[1]Center for Medical Genetics, Department of Biomolecular Medicine, Ghent University, Ghent, Belgium; [2]Cancer Research Institute Ghent, Ghent, Belgium; [3] Flemish Institute for Biotechnology, Ghent, Belgium [4]I-Biostat, Hasselt University, Hasselt, Belgium; [5]National Institute for Applied Statistics Australia (NIASRA), University of Wollongong, Wollongong, Australia

Abstract

To understand biology and differences among various tissues or cell types, one typically searches for molecular features that display characteristic abundance patterns. Several specificity metrics have been introduced to identify tissue-specific molecular features, but these either require an equal number of replicates per tissue or they can't handle replicates at all. We describe a non-parametric specificity score that is compatible with unequal sample group sizes. To demonstrate its usefulness, the specificity score was calculated on all GTEx samples, detecting known and novel tissue-specific genes. A webtool was developed to browse these results for genes or tissues of interest. An example python implementation of SPECS is available at https://github.ugent.be/ceeverae/SPECs. The precalculated SPECS results on the GTEx data are available through a user-friendly browser at specs.cmgg.be.

1 Introduction

To understand biology and differences among various tissues or cell types, one typically searches for molecular features (i.e. RNA, protein, metabolites) that display characteristic abundance patterns. In the most extreme case, these features display tissue- or cell-type restricted abundance profiles. Such specific features can provide insights in functional, development or disease mechanisms (Leucci *et al.*, 2016) or serve as biomarkers (Stutterheim *et al.*, 2008; Prensner *et al.*, 2013). Various consortium-based efforts have generated vast amounts of molecular data that can be exploited for this purpose. The Genotype-Tissue Expression (GTEx) project (https://gtexportal.org) and The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga) are examples of such rich resources containing RNA-sequencing based molecular features for thousands of samples derived from various individuals and tissue types (Lonsdale *et al.*, 2013). To identify tissue-specific molecular features, several specificity metrics have been introduced, but these can suffer from data loss introduced by the requirement to collapse data from biological replicates. Moreover, those metrics that can handle biological replicates require equal sample sizes. In this application note, we describe a novel non-parametric specificity score that is compatible with unequal sample group sizes and enables the detection of features that are specifically present or absent in one or more tissue types.

2 Methods

Let the index *d* = 1,...,m_d refer to a particular sample state. Depending on the application and whether the user wants to give weight to a certain state, π_d is the prevalence of state *d* in the target population or π_d is equilibrated. Suppose there are m_g candidate features, i.e. g = 1,...,m_g . Let Y_{gd} denote the outcome of feature *g* in state *d* with n_{gd} observations, so that the individual outcomes are denoted by Y_{gdi}, i = 1, …,n_{gd}. The Y_{g-d} notation denotes the outcome of feature *f* in all groups but the state *d*. The index *g* will be dropped in further notations. A feature is a characteristic for a given state if its outcome distribution for the given state shows no overlap with the outcome distributions of the other states. This means a larger AUC, given by:

$$p_d = P\{Y_{-d} < Y_d\} = \sum_{k \neq d} P\{Y_k < Y_d\}\pi_k \qquad (1)$$

If p_d is close to zero or one, the distributions are well separated. The probabilities $P\{Y_k<Y_d\}$ are computationally fast to calculate. The probability $P_{kd} = P\{Y_k<Y_d\}$ is then estimated as:

$$\hat{P}_{kd} = \frac{1}{n_k n_d} \sum_{i=1}^{n_k} \sum_{j=1}^{n_d} I_{ki;dj}$$

with $I_{ki;dj}$ a 0/1 indicator for the event $Y_{ki}<Y_{dj}$.
Hence, an estimator of p_d is given by:

$$\hat{p}_d = \sum_{k \neq d} \hat{P}_{kd}\pi_k$$

Further selection of features can be performed based on the distributions of $\hat{p}_d$ as explained in Supplemental Methods 1. As this is a computationally intensive step for large data matrices, one can opt to select features based on a threshold. In our use case, we defined state-specific features as those where the score ($\hat{p}_d$) for one state was above 0.95 and features that were specifically absent in one state as those with a score ($\hat{p}_d$) lower than 0.05. If the score of 0.95 or 0.05 was reached in multiple states, the feature was defined as specific (present or absent) for all these states. The python implementation of the method is available at https://github.ugent.be/ceeverae/SPECs.

3 Results

The Genotype-Tissue Expression (GTEx v7) project (Lonsdale *et al.*, 2013) consists of RNA sequencing data from 12 766 samples belonging to 31 different tissues (7 to 1854 samples per tissue). We calculated the SPECS specificity score on normalized counts for all Ensembl (GRCh38.v85) genes (n=56 202) using all samples. For 30 of the 31 tissues, 2 (esophagus) to 7948 (testis) specifically expressed genes were identified. Most of these genes are protein coding (n=10 959), followed by lincRNAs (n=3080), antisense genes (n=2022) and pseudogenes (n=1976) (Figure 1A and Supplemental Figure 1). In addition, the method has the ability to identify genes that are highly specific for two (or more) tissues, with specificity scores that are slightly lower. As expected, the tissues with the highest number of common specific genes are biologically related such as spleen and blood, or brain and pituitary or muscle and hart.

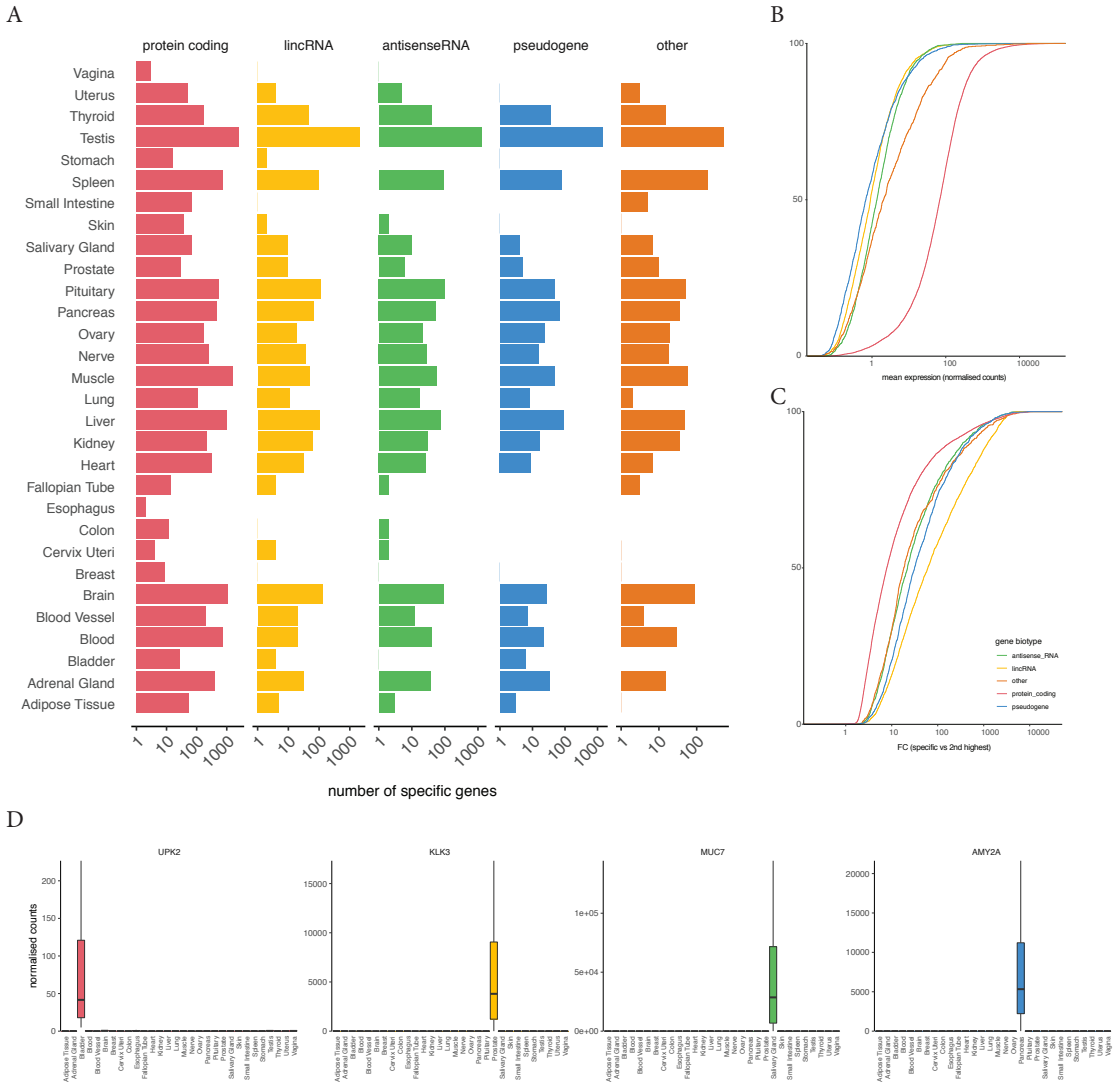

Figure 1 **Known and novel genes are detected as specific for various biotypes.**
A) The number of specific genes for each GTEx tissue and biotype shows that most specific genes are protein-coding. B) Cumulative distribution of the mean expression of specific genes, shows that specific protein-coding genes are higher expressed compared to the other biotypes. C) Cumulative distribution of the fold changes of specific genes and the 2nd tissue shows larger differences for lincRNA genes compared to other biotypes. D) Examples of well-known specific genes; UPK2 for bladder, KLK3 for prostate, MUC7 for adrenal gland and AMY2A for pancreas.

Besides genes that are specifically abundant in a tissue, our method also enables the identification of genes that are specifically repressed in a given tissue. These so-called disallowance genes (Thorrez *et al.*, 2010) were found for 17 tissues ranging from 2 (salivary gland) to 1989 (blood) genes. Most of these are protein coding genes (Supplemental Figure 2).

For all specifically abundant genes we calculated fold changes between the specific tissue(s) and all other tissues. The fold changes for lincRNAs were typically higher than for other biotypes, in line with previous studies in which lincRNAs were shown to be more specific compared to protein coding genes (Cabili *et al.*, 2011) (Figure 1B and Figure 1C).

From our analyses, known specific genes are readily confirmed, such as kallikrein related peptidase 2 (KLK2) and 3 (KLK3, also known as PSA) for prostate, uroplakin 2 (UPK2) for

bladder, mucin 7 (MUC7) for the salivary gland and amylase alpha 2A (AMY2A) for pancreas (Figure 1D). For each tissue in GTEx, rank percentiles for the specific genes are pre-calculated and distilled into a web tool (specs.cmgg.be) where a user can select either their gene of interest to evaluate its specificity or a tissue of interest to identify the most specific genes.

Acknowledgements

Funding

This work has been supported by the Fund for Scientific Research Flanders (FWO), Stichting Tegen Kanker and Vocatio.

Conflict of Interest: none declared.

References

Cabili,M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.

Leucci,E. *et al.* (2016) Melanoma addiction to the long non-coding RNA SAMMSON. *Nature*, **531**, 518–522.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Prensner,J.R. *et al.* (2013) The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat. Genet.*, **45**, 1392–1398.

Stutterheim,J. *et al.* (2008) PHOX2B is a novel and specific marker for minimal residual disease testing in neuroblastoma. *J. Clin. Oncol.*, **26**, 5443–9.

Thorrez,L. *et al.* (2010) Tissue-specific disallowance of housekeeping genes: The other face of cell differentiation. *Genome Res.*, **21**, 95–105.