Bioinformatics, 35(19), 2019, 3842–3845
doi: 10.1093/bioinformatics/btz138
Advance Access Publication Date: 1 March 2019
Applications Note

OXFORD

Gene expression

deTS: tissue-specific enrichment analysis to decode tissue specificity

Guangsheng Pei[1], Yulin Dai [1], Zhongming Zhao [1,2,3,*] and Peilin Jia [1,*]

[1]School of Biomedical Informatics, Center for Precision Health, [2]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on August 23, 2018; revised on January 27, 2019; editorial decision on February 20, 2019; accepted on February 26, 2019

Abstract

Motivation: Diseases and traits are under dynamic tissue-specific regulation. However, heterogeneous tissues are often collected in biomedical studies, which reduce the power in the identification of disease-associated variants and gene expression profiles.

Results: We present *deTS*, an R package, to conduct tissue-specific enrichment analysis with two built-in reference panels. Statistical methods are developed and implemented for detecting tissue-specific genes and for enrichment test of different forms of query data. Our applications using multi-trait genome-wide association studies data and cancer expression data showed that *deTS* could effectively identify the most relevant tissues for each query trait or sample, providing insights for future studies.

Availability and implementation: https://github.com/bsml320/deTS and CRAN https://cran.r-project.org/web/packages/deTS/

Contact: peilin.jia@uth.tmc.edu or zhongming.zhao@uth.tmc.edu

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

1 Introduction

Genome-wide association studies (GWASs) and next-generation sequencing technologies have identified hundreds of thousands of disease-associated variants and genes. Interpretation of these variants, however, remains an open challenge. Tissue-specific regulation, which is affected by many genetic variants, is a critical factor leading to diseases or traits. So far, many diseases or traits have not been reported their causal tissues or cell types, or in particular with its tissue-specific regulation. Recent success of the Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2013) enables us to systematically investigate tissue-specific gene (TSG) expression and regulation. Leveraging these non-disease reference data could open new avenues to infer causal tissues for diseases and to unveil underlying biological mechanisms.

Tissue transcriptome data are often heterogeneous. It includes the genes that are ubiquitously expressed (e.g. housekeeping genes) and other genes that are expressed in specific tissues. Several methods have been developed to identify TSGs from expression profiles. For example, SpeCond fits a normal mixture model to each gene using microarray data for 32 human tissues (Cavalli *et al.*, 2011). Zhao *et al.* (2015) used the Tukey test to identify TSGs from RNA-sequencing (RNA-seq) data. Here, we present a convenient R package, *deTS*, to identify the most relevant tissues for candidate genes or gene expression profiles by tissue-specific enrichment analysis. *deTS* builds on two pre-processed reference panels. We developed a statistical method to identifying TSGs while controlling potential confounding factors. We implemented different statistic tests for different forms of query data. We validated the reference

panels by comparing each other and demonstrated *deTS* using multi-trait GWAS data and cancer RNA-seq data.

2 Materials and methods

2.1 Data collection

We prepared two reference panels: GTEx and the Encyclopedia of DNA Elements project (ENCODE). The GTEx RNA-seq data included 14 725 protein-coding, non-housekeeping genes in 47 tissues (Eisenberg and Levanon, 2013) (details in Supplementary Text S1 and Table S1). The ENCODE panel included 14 031 protein-coding, non-housekeeping genes for 44 tissues (accessed August 2018, Supplementary Table S2). We downloaded and processed GWAS summary statistics for 26 traits (Supplementary Table S3). We defined trait-associated genes by using the gene-based *P*-values (Supplementary Text S1). In addition, we downloaded RNA-seq data for 635 normal samples matched to 14 cancers from The Cancer Genome Atlas (TCGA, Supplementary Table S4) (Weinstein *et al.*, 2013).

2.2 Measurement of tissue specificity

For GTEx data, we implemented a previous method (Finucane *et al.*, 2018) by fitting an ordinary regression model for each gene and computed *t*-statistics to measure the tissue specificity. Notably, several tissues in GTEx dataset were biologically related, such as some brain sub-regions. Treating these naturally related tissues as independent and including all of them in one regression model would underestimate the tissue specificity. Finucane *et al.* (2018) defined a new variable called 'tissue group' (Supplementary Table S1) and used it as the explanatory variable. We followed their work and fitted the regression model for each tissue as: $Y \sim X + \text{age} + \text{sex}$, where Y was the log2-transformed gene expression, the nominal variable X was the tissue group status, and age and sex were sample covariates. We fitted the above model for each tissue, instead of fitting one model including all tissues. Specifically, for a tissue in examination, we defined $X = \{x_i\}$, $i = 1,\ldots, N$, where N is the total number of samples, $x_i = 1$ if the sample belonged to the tissue in examination and $x_i = 0$ if the sample belonged to any tissues not in the same group. Samples from other tissues of the same tissue group as the examined tissue would not be included. Accordingly, N varies by the tissues in examination. After fitting the model, we selected the *t*-statistic for the explanatory variable X for the gene.

Considering that sample size per tissue is small in ENCODE dataset, we employed *z*-score to measure tissue specificity. For each gene, a *z*-score is calculated as $z_i = (e_i - \text{mean}(E))/sd(E)$, where e_i is the average expression of the gene in the ith tissue, E represents the collection of its average expression in all tissues, and *sd* indicates the standard deviation of E.

A higher *t*-statistic or *z*-score indicates that the gene is more specifically expressed in the corresponding tissue.

2.3 Tissue-specific enrichment analysis

For each tissue, TSGs are defined by high *t*-statistics or *z*-scores. We allow the user to define the cutoff values, e.g. the top 5% genes as TSGs. Depending on the query data, two tests are implemented.

Test 1: if the query is a list of genes, we implement Fisher's Exact Test to identify TSGs enriched in the tissue(s).

Test 2: if the query is an expression matrix, we use *t*-test to identify the most relevant tissue(s). Two methods are used to normalize the query expression data so that the query data will be scaled appropriately with the reference data. (i) The *z*-score strategy that normalizes the query data using the tissue parameters as below: $e_n = (e_q - u_s)/sd_s$, where e_q and e_n are the query and normalized expression, and u_s and sd_s are the mean and *sd* of a reference tissue s. (ii) The abundance correction approach (Skene *et al.*, 2018) that normalizes the query data by $e_n = \log_2(e_q+1)/(\log_2(u_s+1)+1)$. Finally, two-sample *t*-test is used to examine the difference between e_n of TSGs and e_n of non-TSGs in each reference tissue (Supplementary Fig. S1).

3 Applications

We provided two reference panels in *deTS*: the GTEx panel (47 tissues, *t*-statistic) and the ENCODE panel (44 tissues, *z*-score). Validation across the two panels showed high concordance (91.5–93.2%, Supplementary Text S1). We further compared the two panels with a previously reported database of TSGs (TiGER) (Liu *et al.*, 2008). We found that the majority of the tissues on our panels shared the highest TSGs with their matched tissues in TiGER (Supplementary Text S1 and Fig. S7). *deTS* can be applied in many scenarios, such as inferring the causal tissues for diseases based on disease-associated genes, assessing bulk RNA-seq data (e.g. finding the most related tissues, outliers, sample contamination/purity), and cross-validation of other genomics data (e.g. tissue specificity of microRNAs or transcription factors by their targets), among others. Below we demonstrate the utility of *deTS* with two applications.

3.1 Application 1: multi-trait GWAS data

We tested trait-associated genes (gene-based $P < 5 \times 10^{-3}$) identified from GWAS for 26 traits (Table 1) using *deTS* Test 1 for candidate genes. In most traits, trait-associated genes were found enriched in the trait-related tissues whereas instances of variation implied novel insights into the disease origin(s). For example, anthropometric trait genes were mainly enriched in artery tissues, metabolic traits in liver, immune-related traits in blood and spleen, and neurodegenerative/neuropsychiatric disease in brain (Fig. 1A) . However, autism spectrum disorder, waist–hip ratio and fasting insulin failed to be linked with any tissues, possibly due to weak GWAS signals or the causal tissues not included in our panels.

3.2 Application 2: TCGA normal samples

The RNA-seq data for TCGA normal samples were organized as an expression matrix and were analyzed using *deTS* Test 2. For demonstration purpose, we pre-defined the biologically matched tissue of each cancer type (Supplementary Text S1 and Table S4). We normalized the original RNA-seq data using the abundance correction strategy based on the GTEx panel. As a result, in nine cancer types (denoted by triangle in Fig. 1B), the matched tissues were most enriched in the samples. In four cancer types (circles in Fig. 1B), the biological tissues were ranked within top 3 and the related tissues could be implied. For example, stomach was the most enriched tissue for esophagus cancer, providing insights into cancer origination or tissue/organ relatedness. In breast cancer, only 50% samples were most enriched in breast while others in minor salivary gland and uterus. One possible reason is sample purity.

4 Conclusion

We present an R package, *deTS*, for tissue-specific enrichment analysis. *deTS* runs fast—it took only 15 s for a gene list matrix with

Table 1. Overview of the 26 GWAS traits

Abbreviations	Traits	References
Neurodegenerative/neuropsychiatric disease		
ALZ	Alzheimer's disease	Lambert *et al.* (2013)
ADHD	Attention deficit-hyperactivity disorder	Martin *et al.* (2018)
ASD	Autism spectrum disorder	Wood *et al.* (2014)
BD	Bipolar disorder	Ruderfer *et al.* (2018)
MDD	Major depressive disorder	Wray *et al.* (2018)
SCZ	Schizophrenia	Ripke *et al.* (2014)
Anthropometric and social trait		
BMI	Body mass index	Locke *et al.* (2015)
FN-BMD	Bone mineral density (femoral neck)	Zheng et al. (2015)
LS-BMD	Bone mineral density (lumbar spine)	Zheng et al. (2015)
EDU	Educational attainment	Okbay *et al.* (2016)
HEIGHT	Height	Wood *et al.* (2014)
WHR	Waist–hip ratio	Shungin *et al.* (2015)
Immune-related trait		
CD	Crohn's disease	Wood *et al.* (2014)
IBD	Inflammatory bowel disease	Wood *et al.* (2014)
RA	Rheumatoid arthritis	Shungin *et al.* (2015)
UC	Ulcerative colitis	Wood *et al.* (2014)
Metabolic phenotype		
AAM	Age at menarche	Day *et al.* (2015)
CAD	Coronary artery disease	Schunkert *et al.* (2011)
FG	Fasting glucose	Dupuis *et al.* (2010)
FI	Fasting insulin	Dupuis *et al.* (2010)
HDL	High-density lipoproteins	Teslovich *et al.* (2010)
LDL	Low-density lipoproteins	Teslovich *et al.* (2010)
TC	Total cholesterol	Teslovich *et al.* (2010)
TG	Triglycerides	Teslovich *et al.* (2010)
T1D	Type 1 diabetes	Bradfield *et al.* (2011)
T2D	Type 2 diabetes	Morris *et al.* (2012)

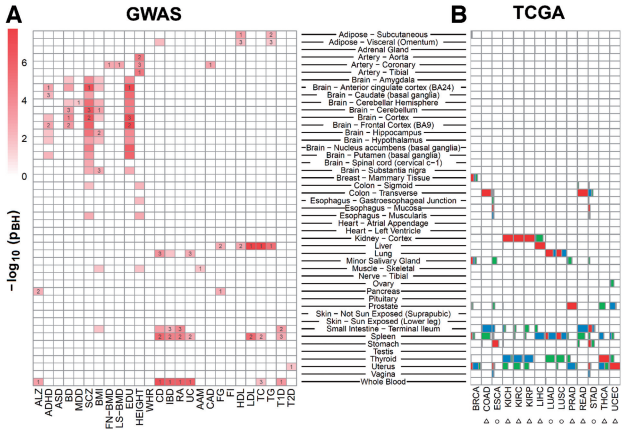

Fig. 1. Tissue-specific enrichment analysis of GWAS and TCGA data. Full names of traits and cancer types and more details are available in Table 1, Supplementary Tables S3 and S4. The *P*-value (adjusted by the Benjamini and Hochberg method) was from Fisher's exact test **(A)** and *t*-test **(B)**, respectively. In (A), the tissues labeled with '1', '2' and '3' indicate those top 3 ranked tissues from the enrichment test. In (B), boxes in red (top 1), blue (top 2) and green (top 3) indicate the proportion of samples in a given cancer type enriched in a tissue

26 GWAS traits, and 26 s for a RNA-seq matrix with 635 samples on an i7-7700HQ desktop. *deTS* can not only identify novel relationships between gene expression and phenotypes, but also is helpful to determine the tissue purity and composition in a gene expression dataset. As expression data from cell types and single cells has been rapidly generated recently, we will expand *deTS* to detect cell types and cell origins from such data. *deTS* is useful to study tissue features and underlying mechanisms for diseases or traits.

Funding

This work was partially supported by National Institutes of Health grant [R01LM012806] and Cancer Prevention & Research Institute of Texas (CPRIT) grant [RP180734].

Conflict of Interest: none declared.

References

Bradfield,J.P. *et al.* (2011) A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.*, **7**, 29.

Cavalli,F.M. *et al.* (2011) SpeCond: a method to detect condition-specific gene expression. *Genome Biol.*, **12**, R101.

Day,F.R. *et al.* (2015) Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.*, **47**, 1294–1303.

Dupuis,J. *et al.* (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, **42**, 105–116.

Eisenberg,E. and Levanon,E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.

Finucane,H.K. *et al.* (2018) Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.*, **50**, 621–629.

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Lambert,J.C. *et al.* (2013) Meta-analysis of 74, 046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, **45**, 1452–1458.

Liu,X. *et al*. (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 1471–2105.

Locke,A.E. *et al*. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.

Martin,J. *et al*. (2018) A genetic investigation of sex bias in the prevalence of attention-deficit/hyperactivity disorder. *Biol. Psychiatry*, **83**, 1044–1053.

Morris,A.P. *et al*. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet*., **44**, 981–990.

Okbay,A. *et al*. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, **533**, 539–542.

Ripke,S. *et al*. (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**, 421–427.

Ruderfer,D.M. *et al*. (2018) Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, **173**, 1705–1715.

Schunkert,H. *et al*. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet*., **43**, 333–338.

Shungin,D. *et al*. (2015) New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, **518**, 187–196.

Skene,N.G. *et al*. (2018) Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet*., **50**, 825–833.

Teslovich,T.M. *et al*. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

Weinstein,J.N. *et al*. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet*., **45**, 1113–1120.

Wood,A.R. *et al*. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet*., **46**, 1173–1186.

Wray,N.R. *et al*. (2018) Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet*., **50**, 668–681.

Zhao,G. *et al*. (2015) An effective analytic method for detecting tissue-specific genes in RNA-seq experiments. *Pharmacogenomics*, **16**, 1769–1779.

Zheng,H.F. *et al*. (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature*, **526**, 112–117.