BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 8 2013, pages 1083–1085
doi:10.1093/bioinformatics/btt090

Gene expression

Advance Access publication February 21, 2013

DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data

Ting Gong* and Joseph D. Szustakowski

Biomarker Development, Translational Medicine, Novartis Institutes for BioMedical Research, Cambridge, MA 02139, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: For heterogeneous tissues, measurements of gene expression through mRNA-Seq data are confounded by relative proportions of cell types involved. In this note, we introduce an efficient pipeline: DeconRNASeq, an R package for deconvolution of heterogeneous tissues based on mRNA-Seq data. It adopts a globally optimized non-negative decomposition algorithm through quadratic programming for estimating the mixing proportions of distinctive tissue types in next-generation sequencing data. We demonstrated the feasibility and validity of DeconRNASeq across a range of mixing levels and sources using mRNA-Seq data mixed *in silico* at known concentrations. We validated our computational approach for various benchmark data, with high correlation between our predicted cell proportions and the real fractions of tissues. Our study provides a rigorous, quantitative and high-resolution tool as a prerequisite to use mRNA-Seq data. The modularity of package design allows an easy deployment of custom analytical pipelines for data from other high-throughput platforms.

Availability: DeconRNASeq is written in R, and is freely available at http://bioconductor.org/packages.

Contact: tinggong@gmail.com

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

Received on July 30, 2012; revised on January 22, 2013; accepted on February 18, 2013

1 INTRODUCTION

High-throughput next-generation sequencing (NGS) technologies are becoming quickly established in biomedical research and discovery (Huang *et al.*, 2011). NGS-based sequencing of mRNA (referred to as mRNA-Seq) offers several potential advantages over microarray-based methods (Haas and Zody, 2010; Tang *et al.*, 2009). Heterogeneous tissues are frequently collected (e.g. blood, tumour, etc.) from humans or model animals (Zhao and Simon, 2010). Therefore, mRNA-Seq samples are often heterogeneous with regard to those cell types, making it difficult to distinguish gene expression variability that reflects shifts in cell populations from variability that reflects change of cell-type-specific expression (Kuhn *et al.*, 2011).

Recently, we have described the deconvolution of microarray data from complex samples such as blood samples collected in a clinical trial (Gong *et al.*, 2011). Quon and Morris (2009) described a probabilistic framework to address the similar problem using high-throughput sequencing. Their approach requires strong assumptions about the distributions of gene abundance and is not able to generate per-sample estimates of tissue or cell abundance. mRNA-Seq data are inherently different than microarray data, which are able to give the more refined information necessary to tackle the deconvolution problem; we therefore expected that these methods need to be updated and re-validated. However, transferring the approach of microarray data into a robust analysis pipeline is an inherently study-specific task and poses ongoing challenges. Although mRNA-Seq is promising, it is widely accepted that many factors could introduce variation and bias to mRNA-Seq data. Therefore, proper pre-processing methods are needed to de-noise and fine-tune the mRNA-Seq data before they are used in downstream deconvolution. In light of these considerations, we have developed DeconRNASeq, an R/Bioconductor-based pipeline for constituent fraction estimation and quality assessment of deconvolution on mRNA-seq datasets.

Here, we describe the DeconRNASeq package that encapsulates the method in a convenient-to-use format. DeconRNASeq is based on a linear model of a sum of pure tissue- or cell-type-specific reads of all cell types, weighted by the respective cell-type proportions. To estimate the proportions of known tissue types in a sample, DeconRNASeq solves a non-negative least-squares constraint problem with quadratic programming to obtain the globally optimal solution for estimated fractions. Results from mRNA-Seq data described here demonstrate that our method is able to accurately predict mixing fractions for multiple species of tissues or cells. Using the mixing samples containing relatively rare cell types ($\leq$2%), we show that correlation coefficient between the estimate proportions and the true proportions can reach 0.9754. In principle, DeconRNASeq is also applicable to other types of profiling data from heterogeneous samples. For example, the same approach could be used to track species in metagenomics data (Knights *et al.*, 2011) or DNA contamination in tumour samples (Cibulskis *et al.*, 2011).

2 METHODS

Figure 1a depicts the workflow of our transcriptome quantification and deconvolution pipeline. Given the information of the short-read alignment, followed by transcript quantification, we apply a statistical approach to model expression from a mixed cell population as the weighted average of expression from different cell types. We solve these equations using quadratic programming, which efficiently identifies the

*To whom correspondence should be addressed.

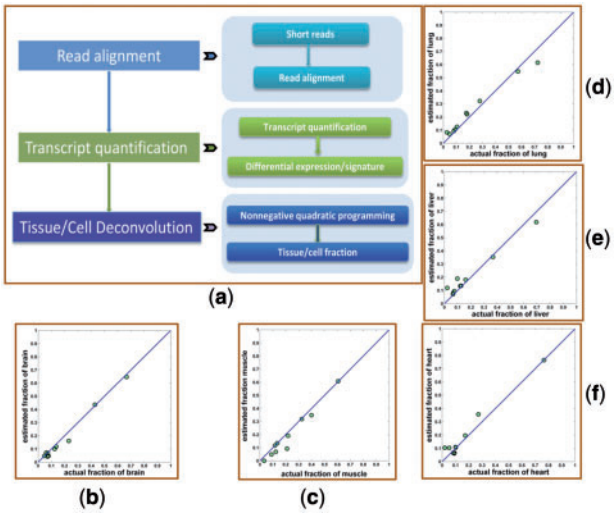

Fig. 1. (**a**) The workflow of our transcriptome deconvolution analysis pipeline for mRNA-Seq data. The preprocessing steps include short-read alignment onto a reference genome, followed by transcript quantification. A nonnegative least-squares constraint problem is solved by quadratic programming to obtain the globally optimal solution for estimated fractions of known tissue types in a sample. (**b–f**) Statistical deconvolution of mRNA-Seq yields accurate estimates of pure tissue fractions for multiple tissues. Plotting of estimated proportions (y axis) vs. actual proportions (x axis) shows strong congruity for brain (b), muscle (c), lung (d), liver (e) and heart (f) tissues. Each green dot represents a sample

globally optimal solution while preserving non-negativity of the fraction of the cells. Consequently, we can accurately estimate the fractions of various cell populations. The details are described as follows.

DeconRNASeq requires input in the form of two R data frames that contain normalized transcriptional measurements from pure tissues (**S**) and the heterogeneous samples (**X**) to be analysed respectively. The gene was used as the expression unit in our example, although the methodology can be extended to transcripts or exons, provided the quantification of their expression is properly obtained.

More specifically, the expression level x_{jk} of gene j in a sample k is the average of expected expression levels across the cell types s_{ij}, weighted by the respective cell-type proportions a_{ki} ($i=1\ldots N$, N: the total number of cell types):

$$x_{jk} = \sum_{i=1}^{N} a_{ki} s_{ij} \tag{1}$$

Or more generally, we represent the above equation in matrix form as:

$$\mathbf{X}=\mathbf{AS}. \tag{2}$$

Here, **X** denotes the observed mixture gene expression matrix (genes by samples). **S**, the signature matrix derived from the training data (homogeneous samples), gives the tissue-type-specific gene expression profiles (genes by tissue types), and **A**, the proportion matrix, is the quantity to estimate tissue proportions over samples (tissue types by samples). We solve this weighted non-negative least squares problem for each gene by

$$\min_{\mathrm{A}}(\| \mathbf{AS}-\mathbf{X} \|^2), \quad s.t. \begin{cases} \sum_i a_{ki} = 1 \\ a_{ki} \geq 0, \forall i \end{cases} \tag{3}$$

where the coefficient a_{ki} is a scalar parameter between 0 and 1 to represent the fraction of tissue type. The formulation can then be efficiently solved by quadratic programming (Bertsekas, 1999; Mackey *et al.*, 1996).

3 RESULTS

To systematically examine the accuracy of our deconvolution algorithm, we designed several *in silico* mixing experiments that made use of mRNA-Seq data from Illumina's Human BodyMap 2.0 (generated on a HiSeq 2000) [GSE30611] to generate tissue-specific transcriptional profiles (Training Data). *In silico* mixed data were simulated using Pan *et al.'s* (2008) data, with disparate proportions drawn from random numbers. The mixing proportions used by each type of tissue are shown in Supplementary Table S4. It should also be noted that we investigated the influence of extremely low fractions of contaminating cell types (<2%). More experiments including the deconvolution of biologically related cell population can be found in Supplementary Material.

We estimated the number of expression signatures (genes) through the condition number of the signature matrix (see Supplementary Sections 3.2 and 3.3 for its detailed derivation). Consequently, we seleted first 1570 genes that consist of the signatures for the five tissues and deconvoluted the data (refer to the Supplementary Section 3.3 for the validation of the 'optimal' number of expression signatures in terms of condition number). The results are shown in Figure 1(b–f), with correlation coefficient between the estimation and the real frations = 0.9754. Reassuringly, our algorithm was able to accurately predict the level of mixing across a wide range of proportions including multiple pure sources for mRNA-Seq data.

4 CONCLUSION

We have introduced a statistical workflow—DeconRNASeq—to estimate proportions of tissue or cell types by incorporating tissue-type-specific genes. Simulations show that our deconvolution analysis accurately detects expression heterogeneity and assesses proportions of multiple tissue types in newly advent NGS platform. The main objective of our work is to develop comprehensive and flexible deconvolution software effectively applicable to expression profiling experiments. Thereby, it is worth noting that the algorithmic items are shared with our previous work (Gong *et al.*, 2011) and can also be applied to microarray data. However, microarray data require appropriate normalization and other pre-processing methods as described in Gong *et al.* (2011). The independent profile-generating module in DeconRNASeq grants great freedom to users, who can combine with other R or Bioconductor packages to perform upstream and downstream analysis of NGS data.

ACKNOWLEDGEMENTS

The authors thank N.R. Nirmala for critical review of the manuscript. The authors especially thank Robin Ge for running the pre-processing programs of mRNA-Seq and Claudia Hon for assistance with the software release. We also thank the bioinformatics team in the Biomarker Development department (BMD) for fruitful discussions. We are grateful to Dr Isaac S. Kohane (Harvard Medical School, Children's Hospital Medical Center) for his guidance. T.G. thanks the NIBR Education Office for their support via a Presidential Postdoctoral Fellowship.

Conflict of Interest: T.G. was an employee of Novartis when she conducted this research. J.S. is an employee of Novartis and holds stock in the company.

REFERENCES

Bertsekas,D.P. (1999) *Nonlinear Programming*, 2nd edn. Athena Scientific, Belmont, MA.

Cibulskis,K. *et al.* (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics*, **27**, 2601–2602.

Gong,T. *et al.* (2011) Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS One*, **6**, e27156.

Haas,B.J. and Zody,M.C. (2010) Advancing RNA-Seq analysis. *Nat. Biotechnol.*, **28**, 421–423.

Huang,W. *et al.* (2011) Efficiently identifying genome-wide changes with next-generation sequencing data. *Nucleic Acids Res.*, **39**, e130.

Knights,D. *et al.* (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods*, **8**, 761–763.

Kuhn,A. *et al.* (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, **8**, 945–947.

Mackey,M.D. *et al.* (1996) CHEMTAX—a program for estimating class abundances from chemical markers: application to HPLC measurements of phytoplankton. *Mar. Ecol. Prog. Ser.*, **144**, 265–283.

Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet.*, **40**, 1413–1415.

Quon,G. and Morris,Q. (2009) ISOLATE: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, **25**, 2882–2889.

Tang,F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*, **6**, 377–382.

Zhao,Y. and Simon,R. (2010) Gene expression deconvolution in clinical samples. *Genome Med.*, **2**, 93.