BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 11 2013, pages 1463–1464
doi:10.1093/bioinformatics/btt155

Sequence analysis

Advance Access publication April 4, 2013

A tool for RNA sequencing sample identity check

Jinyan Huang[1,2,3,4], Jun Chen[3,4], Mark Lathrop[2] and Liming Liang[3,4,*]

[1]School of life science, Tongji University, 200092 Shanghai, China, [2]Fondation Jean Dausset-CEPH, 75010 Paris, France and [3]Department of Epidemiology and [4]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: RNA sequencing data are becoming a major method of choice to study transcriptomes, including the mapping of gene expression quantitative trait loci (eQTLs). RNA sample contamination or swapping is a serious problem for downstream analysis and may result in false discovery and lose power to detect the true biological relationships. When genetic data are available, for example, in eQTL studies or samples have been previously genotyped or DNA sequenced, it is possible to combine genetic data and RNA-seq data to detect sample contamination and resolve sample swapping problems. In this article, we introduce a tool (IDCheck) that allows easy assessment of concordance between genotype (from SNP arrays or DNA sequencing) and gene expression (RNA-seq) samples. IDCheck compares the identity of RNA-seq reads and SNP genotypes using a likelihood-based method. Based on maximum likelihood estimates of relevant parameters, we can detect sample contamination and identify correct sample pairs when swapping occurs. Our tool provides an efficient and convenient way to evaluate and resolve these problems.

Availability: A complete description of the software is included on the application home page. The software is freely available in the public domain at http://eqtl.rc.fas.harvard.edu/idcheck/.

Contact: lliang@hsph.harvard.edu

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

Received on January 8, 2013; revised on March 3, 2013; accepted on March 26, 2013

1 INTRODUCTION

Comparing with array-based technology, RNA-seq experiments allow us to more accurately measure gene expression level and provide new insights into the regulatory mechanisms underlying expression quantitative trait loci (eQTLs) (Majewski and Pastinen, 2010). eQTL mapping studies integrate genetic and gene expression data to identify genetic variants associated with expression of genes and have helped interpret findings from genome-wide association studies for many complex diseases and traits (Cookson *et al.*, 2009).

As the scale of the study becomes larger and larger, sample contamination or label swapping is more likely to occur. Sample mislabeling during the manual preparation of DNA samples and RNA samples is more prone to occur because of many steps involved in collecting the data. For example, if some sample labels are mixed and tagged to the wrong sequencing lane or DNA sample, the genotype data and gene expression data are not correctly paired. The mislabeled or contaminated samples could add bias and/or noise to the datasets, resulting in false discovery and loss of statistical power to detect true associations (Jun *et al.*, 2012). Therefore, it is necessary to check the sample identity and potential contamination before any downstream data analysis. A fast and efficient tool to identify these potential problems and offer possible solutions is needed.

Here, we present a tool, IDCheck, to do sample identity checking and detect potential sample contamination. It is particularly useful for eQTL mapping studies using RNA-seq data, but also applicable to any RNA-seq studies with genotype data available. It will output a text format file to report for how each sample is paired. IDCheck also provides a quick overview of RNA-seq data quality using summary graphs. These results are exported to PDF files.

2 METHODS

For each SNP site with RNA-seq information, we remove read bases with base quality <25 and calculate the read coverage using Samtools (Li *et al.*, 2009). Only SNP sites with read coverage ≥20 are included in further analysis. For each RNA-seq and SNP sample pair, we build a likelihood model of the observed RNA-seq reads given the observed genotypes. Denote ε as an aggregate error rate because of sequencing error, contamination, SNP genotyping error and so forth. And μ_A as the allelic expression rate for allele A for a heterozygous site. Depending on the genotype of the SNP in consideration, we model the probability of observed read base given the genotype using:

$$P(\text{read} = A|AA) = 1 - \varepsilon, \quad P(\text{read} \neq A|AA) = \varepsilon \tag{1}$$

$$P(\text{read} = A|AB) = \mu_A(1 - \varepsilon), \quad P(\text{read} \neq A \text{ or } B|AB) = \varepsilon \tag{2}$$

$$P(\text{read} = A|AB) = \mu_A(1 - \varepsilon), \quad P(\text{read} \neq A \text{ or } B|AB) = \varepsilon \tag{3}$$

To further model the heterogeneity of allelic expression, we assume that μ_A follows a β distribution $\beta(\mu, \varphi)$, where μ is the mean of allelic expression rate, which is taken to be 0.5, and φ is the over dispersion parameter with small value indicating large heterogeneity. The likelihood of the reads given the genotypes can then be written as:

$$Likelihood = \prod_{m=1}^{M} \prod_{r=1}^{R_m} P(read_{r,m}|genotype_m) \tag{4}$$

$$= (1-\varepsilon)^C \varepsilon^D \prod_{m \in \{\text{genotype}_m = AB\}} \frac{\Gamma(\varphi)\Gamma(n_m^A + 0.5\varphi)\Gamma(n_m^B + 0.5\varphi)}{\Gamma^2(0.5\varphi)\Gamma(n_m^A + n_m^B + \varphi)}, \tag{5}$$

where M is the number of SNPs, R_m is the number of reads at SNP_m, C is the number of reads consistent with the genotype, D is the number of

*To whom correspondence should be addressed.

reads discordant to the genotype and n_m^A, n_m^B are the number of reads with base A and B at locus m, respectively. Therefore, we include in the model two parameters: the aggregate error rate (ε) and overdispersion of the allelic expression (φ). If the SNP data and RNA-seq come from two different individuals or the RNA is contaminated, we expect a large ε and a small φ. The large overdispersion (small φ) in allelic expression for discordant samples results from the fact that the RNA-seq reads can come from a homozygous SNP. In such a situation, the SNP site will show extremely high or low allelic expression, and the sequencing error model alone cannot capture this part of information. The maximum likelihood estimate (MLE) of ε is $C/(C+D)$. However, no closed-form solution exists for the MLE of φ. Numerical method, such as Newton–Raphson, can be used to estimate the MLE of φ. Finally, to scale the overdispersion parameter between 0 and 1, we redefine overdispersion $d = 1/(1+\varphi)$, with large value indicating more evidence of discordance of the two samples.

3 APPLICATION AND DISCUSSION

We have used IDCheck on a subset of 70 siblings from 35 families from our family-based studies (Liang *et al.*, 2013). Each sample has >5 million (range from 5 to 120 M) paired-end 101 bp reads generated by the Illumina GA II Sequencing platform. SNPs were characterized with the Illumina HumanHap300 BeadChip, which is now available through EGA, accession number: EGAS00000000137. To evaluate the error rate and overdispersion for different situations, we simulate RNA–DNA mismatched samples by first swapping samples within families (i.e. swapping the genotype sample labels between siblings) and then between families (swapping the genotype sample labels between families). We also simulated RNA-contaminated samples by randomly replacing 10 or 50% RNA-seq reads from one sample with reads from another sample. The error rate and overdispersion for correctly matched samples, within-family and between-family swapped samples and RNA-contaminated samples are shown in Figure 1 (Supplementary Fig. S1).

The ranges of error rate for samples correctly matched samples, and those permuted within families and between families are 0.003–0.196, 0.176–0.46 and 0.346–0.582, and the overdispersions are 0.045–0.166, 0.377–0.662 and 0.525–0.816, respectively. These values will provide a good reference for other users to determine potential swapping and contamination. In general, for deeper RNA-seq data, we would expect higher precision for error rate and overdispersion estimates, and, therefore, improves discrimination. Our results suggest that 5 M reads will be largely sufficient for performing an IDCheck analysis. We found that even with 2 M reads, the sample swapping can still be clearly identified (Supplementary Fig. S2). In some cases, it might be desirable to perform ID checking using low-coverage sequencing before undertaking additional large-scale sequencing.

It is worth to note that the combination of error rate and overdispersion parameters clearly outperform the discrimination relying on only either parameters. The correctly matched samples can be clearly separated from contaminated or swapped samples by using a straight line in the 2D plot in Figure 1, whereas there is no such cut-point in either dimension that can clearly distinguish the five groups of samples from each other.

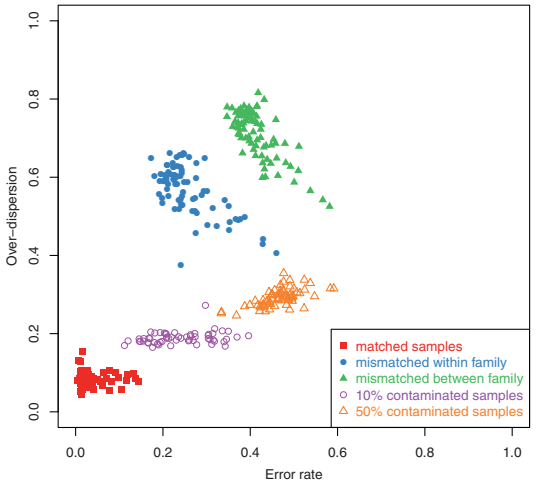

Fig. 1. Distribution of error rate and overdispersion for matched and unmatched genotype/RNA-seq sample pairs. Squares are for the matched pairs, i.e. the genotype and RNA-seq data assigned to the correct individuals. Solid dots show pairs of samples that are randomly swapped within families. Solid triangles show pairs of samples that are randomly swapped between families. Open squares and open triangles are samples with 10 and 50% RNA contamination. The *y*-axis is the transformed overdispersion d

IDCheck is designed for RNA-seq data identify checking. The current approach explicitly models sequencing, genotyping error and potential allele-specific expression. We note that this package can also be applied to DNA sequencing data identify checking when both DNA sequencing and array-based genotype data are available. In this application, mismatching of heterozygous genotypes will be modeled as overdispersion or sequencing error. IDCheck is efficient and can be applied to large-scale data. We estimated that the time needed for each 30 M reads is ~8 min on a standard desktop computer when starting with a mapped BAM file.

Funding: ANR Labex project 'Medical Genomics' to (M.L.); National Natural Science Foundation of China (30900838 to J.H. in part).

Conflict of Interest: none declared.

REFERENCES

Cookson,W. *et al.* (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.

Jun,G. *et al.* (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.*, **91**, 839–848.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liang,L. *et al.* (2013) A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.*, **23**, 716–726.

Majewski,J. and Pastinen,T. (2010) The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.*, **27**, 72–79.