Generation of expression calls for RNA-seq data

Marta Rosikiewicz[1,2], **Marc Robinson-Rechavi**[1,2]✉

[1]Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland
[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

Motivation and Objectives

The Bgee database (database for Gene Expression Evolution; Bastian *et al.,* 2008) provides information about genes that are expressed in different organs and tissues. In order to introduce RNA-seq results into Bgee we had to develop methodology for deriving expressed/unexpressed calls for genes. Such detection calls can be used for characterization of the tissue gene expression profile. Additionally detection calls are widely used in transcriptomic studies for filtering the genes used for differential expression analysis, clustering samples or building more reliable classifiers (Archer and Reese, 2010). The goal of our work is to find an automatic way to define the cut-off value on a transcription level that allows discrimination between expressed and non-expressed genomic features for each library individually.

Methods

RNA-seq data preprocessing

Reads from RNA-seq libraries from experiment GSE30352 (Brawand *et al.,* 2011) were mapped to gene models from Ensembl database and to selected intergenic regions of the reference genome. The mapping of the reads was performed using TopHat2 (Trapnell *et al.,* 2009). The maximum number of mapping sites allowed for a read was set to 1. The intergenic regions are chosen in such a way that the distribution of their lengths matches the distribution of lengths of the transcriptome. Reads that map to the features are summed up using the HTSeq-count software (http://www-huber.embl.de/users/anders/HTSeq/). The RPK (read per kilobase) value for every feature is obtained by dividing the number of reads that match a given feature by its length.

The present/absent calls

Our approach to define present/absent calls is based on Hebenstreit *et al.,* 2011. For each RNA-seq library independently, we define a RPK cut-off, k, for determining "present/absent" calls, set to be equal to the minimal value for which the ratio of relative abundance of intergenic regions and genes, with RPK values above k, is equal or lower than α (in Bgee, α = 0.05). In other words, a RPK threshold is defined for each sample independently, such that a randomly chosen feature, from the set of genes and intergenic regions, with a RPK value above the threshold, has at least 95% probability of being a gene.

Cut-off determination procedure

1) For every value of x define the ratio r: $r = n_{ix}N_g / n_{gx}N_i$ where:

n_{ix}: number of intergenic regions with RPK values higher than x
n_{gx}: number of genes with RPK values higher than x
N_i: number of all intergenic regions
N_g: number of all genes

2) The cutoff value k is the minimal value of x for which r is equal or lower than α.

Results and Discussion

The procedure described for expression calls generation was applied to all samples from the analyzed dataset. In general we decided to use selected random intergenic fragments to estimate transcription level coming from experimental noise or background activity of the transcription machinery. Despite up to 4 times differences in the number of aligned reads between libraries the proportion of genes called expressed by our algorithm remained consistent among different samples reaching in case of mouse data 39.1% ± 1.49 SD and human 34.4% ± 3.9 SD (56.4% ± 2.22 and 71.59% ± 5.28 for protein coding genes respectively). In contrast only 3.4% ± 0.12 in case of mouse intergenic regions and 4% ± 0.55 of human intergenic regions were above the cut-off (example distributions of expression values for different types of genomic features is shown on Figure 1). Less than 15% and 20% of intergenic regions for mouse (n=17) and human (n=16) data accordingly were ever called "expressed".

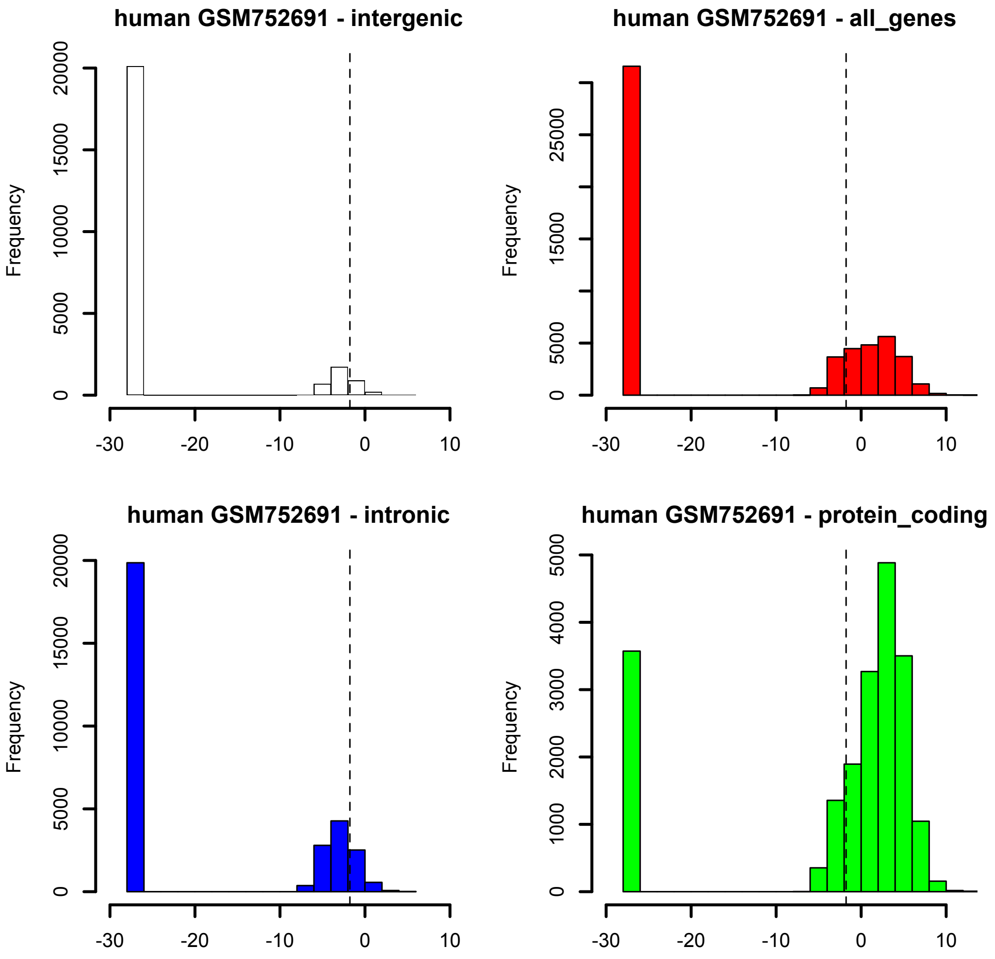

Figure 1. Distribution of log2 (RPK + 1e-08) values for different feature categories, dashed line specify cutoff.

In contrary more than 80% of mouse and 90% of human protein coding genes were at least once called "expressed". Moreover according to our results, among protein coding genes more than 50% in case of human data and 40% in case of mouse data are expressed ubiquitously in all analyzed samples. If we took, as criterion of transcription, the presence of at least one uniquely mapped read then many intergenic regions would have to be classified as expressed, which we believe would be less informative. Moreover, thanks to our methodology it is possible to avoid applying a single arbitrary cut-off for all libraries.

Acknowledgements

This work was supported by the Swiss Institute of Bioinformatics, by the Swiss National Science Foundation [grant number 31003A 133011/1], and by Etat de Vaud.

References

Archer, K.J., and Reese, S.E. (2010). Detection call algorithms for high-throughput gene expression microarray data. *Briefings in bioinformatics* **11**, 244-252.

Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., and Robinson-Rechavi, M. (2008). Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In Data Integration in the Life Sciences, A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux, eds. (Springer Berlin Heidelberg), pp. 124-131.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348.

Hebenstreit, D., Fang, M., Gu, M., Charoensawan, V., van Oudenaarden, A., and Teichmann, S.A. (2011). RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular systems biology* **7**, 497.

HTSeq: Analysing high-throughput sequencing data with Python. [http://www-huber.embl.de/users/anders/HTSeq/]

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111.