BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1183–1184
doi:10.1093/bioinformatics/btu125

Gene expression

Advance Access publication March 27, 2014

CAGExploreR: an R package for the analysis and visualization of promoter dynamics across multiple experiments

Emmanuel Dimont[1], Oliver Hofmann[1], Shannan J. Ho Sui[1], Alistair R. R. Forrest[2,3], Hideya Kawaji[2,3,4], the FANTOM Consortium and Winston Hide[1,*]

[1]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115, USA, [2]RIKEN Omics Science Center, Yokohama, Kanagawa 230-0045 Japan, [3]Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa 230-0045, Japan and [4]RIKEN Preventive Medicine and Diagnosis Innovation Program, Wako, Saitama 351-0198, Japan

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Summary: Alternate promoter usage is an important molecular mechanism for generating RNA and protein diversity. Cap Analysis Gene Expression (CAGE) is a powerful approach for revealing the multiplicity of transcription start site (TSS) events across experiments and conditions. An understanding of the dynamics of TSS choice across these conditions requires both sensitive quantification and comparative visualization. We have developed CAGExploreR, an R package to detect and visualize changes in the use of specific TSS in wider promoter regions in the context of changes in overall gene expression when comparing different CAGE samples. These changes provide insight into the modification of transcript isoform generation and regulatory network alterations associated with cell types and conditions. CAGExploreR is based on the FANTOM5 and MPromDb promoter set definitions but can also work with user-supplied regions. The package compares multiple CAGE libraries simultaneously. Supplementary Materials describe methods in detail, and a vignette demonstrates a workflow with a real data example.

Availability and implementation: The package is freely available under the MIT license from CRAN (http://cran.r-project.org/web/packages/CAGExploreR).

Contact: edimont@mail.harvard.edu

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

Received on October 25, 2013; revised on February 5, 2014; accepted on February 27, 2014

1 INTRODUCTION

It has been predicted that the majority of human genes have multiple promoters. The differential use of transcription start sites (TSSs) in alternative promoters is a complementary mechanism to alternate splicing for the generation of RNA diversity that is now becoming better understood (Pal *et al.*, 2011). Tissue-specific TSS usage has been identified in mammalian genomes (Carninci *et al.*, 2006), and alternate TSS usage has been identified in cancers when compared with normal cells (Thorsen *et al.*, 2011), implying that promoter-specific transcription coupled with gene expression exists as hallmarks of cell state. The profound impact of switches in transcript isoform production is well recognized for its role in regulation (Trapnell *et al.*, 2010).

Cap Analysis Gene Expression (CAGE) captures, sequences and maps capped 5′ RNA tags. In addition to being a platform for measuring gene expression, it has more importantly provided molecular biologists with enhanced resolution of gene regulation by revealing the precise locations of transcription initiation events (Plessy *et al.*, 2010). CAGE data have recently become more plentiful, thanks to the recent ENCODE (2012) and FANTOM5 (Forrest A.R.R. *et al.*, 2014) publications.

Analysis of TSS choice provides insight into the variation of transcription factor binding, epigenetic modifications and regulatory network activation between different cell types. Although CAGE allows for the identification of individual TSS, it is more convenient to group clusters of TSSs detected in close vicinity into ‘promoter’ regions. This makes CAGE an attractive platform for *de novo* promoter identification.

The relative transcription occurring at TSSs among alternative promoters of a gene is termed promoter composition (PC). We describe CAGExploreR, an R package that conveniently summarizes, visualizes and ranks changes in PC (also called promoter ‘switching’) genome-wide across different samples. The dynamics of differential PC is especially intriguing when this phenomenon leads to changes in the abundance of different transcript isoforms or protein products within the cell population under study. Figure 1 highlights the conceptual difference between PC and differential gene expression. Four samples, A–D, are evaluated at four color-coded promoter regions located near or within a gene. Total gene expression measured as mapped tags per million sequenced following optional library normalization with *edgeR* (Robinson *et al.*, 2010) is obtained by summing the number of tags that map to the union of the gene region, all four promoters including the regions between them and dividing by effective library size. PC is measured as a proportion vector. A and B have no change in PC, but the gene is differentially expressed. A and C have no differential gene expression but there is differential PC. Finally, A and D demonstrate both differential gene expression and differential PC.

Existing bioinformatics tools that analyze count data from sequencing technologies treat genes as elementary indivisible units and usually measure differences in expression via a contrast between two groups of samples. Examples include *edgeR*

*To whom correspondence should be addressed.

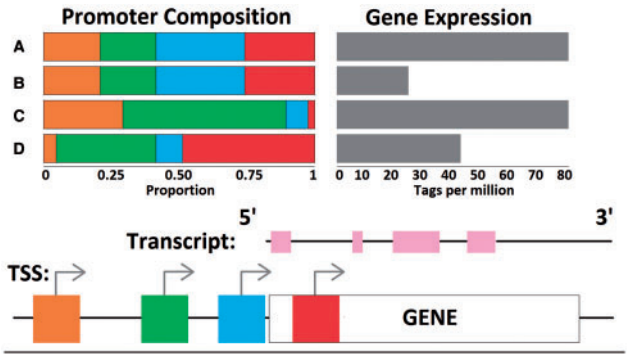

Fig. 1. Differential PC and differential gene expression. A, B, C and D are four arbitrary samples being compared. Gene displayed has four promoters

(Robinson *et al.*, 2010) and *Cuffdiff* (Trapnell *et al.*, 2010) for differential gene expression and transcription analysis, respectively. Unlike these tools, CAGExploreR treats genes as multiunit blocks composed of promoter subunits and compares their relative expression within the gene across samples. It is not restricted to the analysis of contrasts between pairs of experiments but rather is designed to scale to any number of experiments for simultaneous comparison.

2 MODEL AND METHODS

Typical CAGE output consists of a BAM library file that maps sequenced tags to the genome, which CAGExploreR converts to a table of tag counts that correspond to promoter regions using either the built-in FANTOM5, MPromDb (Gupta *et al.*, 2011) or a set of user-specified promoter definitions such as *de novo* identified regions. Internally, a table of counts is generated for each mapped gene, with rows corresponding to the different samples or libraries and columns corresponding to promoters, and displayed accordingly as in Figure 1. The promoter counts are normalized to proportions within each gene and sample. The simplest case would be a 2×2 table when comparing the PC across two samples for a gene with two promoters.

Genes are assigned a proportional entropy reduction score (Theil, 1970), which ranges from 0, when PC stays constant across every sample, to 1, when every sample transcribes exclusively from a unique promoter, for overall promoter switching. For each gene, the test of the null hypothesis of no differential PC corresponds to the test that the entropy reduction score is 0. *P*-values for this test are obtained using a Monte Carlo approach (Supplementary Methods). The switching effect size for promoter pairs is reported using the odds ratios for every nested 2×2 table within a gene. In addition, entropy-based measures are used to quantify the level of heterogeneity in gene expression across samples. All *P*-values are adjusted for multiple comparisons using the Benjamini–Hochberg method to control the false discovery rate or some other appropriate user-specified method.

CAGExploreR generates text-based and visual HTML reports and figures similar to Figure 1 for further analysis. When several conditions are being compared, they can be grouped together via hierarchical clustering on a gene-by-gene basis, demonstrating which conditions have similar PC profiles. This profiling helps demonstrate whether the PC across replicates clusters together within experimental conditions. Consequently, the user can assess replicate agreement at the gene level and so can gain a sense of biological variability.

3 DISCUSSION

We present CAGExploreR, the R package that addresses the important task of detecting changes in PC in CAGE experiments. The method is scalable to any number of conditions and/or promoter regions for simultaneous comparison. The method is flexible and can be applied to any experiment that produces tag counts grouped by classification factors in which the detection of switching or changes in composition is of interest, e.g. gene expression switching within gene sets, pathway activity switching within regulatory and molecular networks, isoform and exon switching using RNA-Seq. To use this software for any of the aforementioned applications, the user need only to change the genomic region definitions from promoter regions to other regions of interest. This work is part of the FANTOM5 project. Data download, genomic tools and copublished manuscripts have been summarized at http://fantom.gsc.riken.jp/5/top/.

ACKNOWLEDGEMENTS

The authors would like to thank all members of the FANTOM5 consortium for contributing to the generation of samples and analysis of the dataset, and GeNAS for data production. The authors would also like to thank Gabriel Altschuler and Christine Wells who have contributed invaluable suggestions.

Riken Omics Science Center ceased to exist as of April 1, 2013 due to RIKEN reorganization.

Funding: RIKEN Omics Science Center from the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) (to Yoshihide Hayashizaki); Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan (to Yoshihide Hayashizaki); MEXT to RIKEN CLST and RIKEN PMI.

Conflict of Interest: none declared.

REFERENCES

Carninci,P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.

ENCODE Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Forrest,A.R.R. *et al.* (2014) A promoter level mammalian expression atlas. *Nature*, http://dx.doi.org/10.1038/nature13182.

Gupta,R. *et al.* (2011) MPromDb update 2010: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-seq experimental data. *Nucleic Acids Res.*, **39**, D92–D97.

Pal,S. *et al.* (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.*, **21**, 1260–1272.

Plessy,C. *et al.* (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat. Methods*, **7**, 528–534.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Theil,H. (1970) On the estimation of relationships involving qualitative variables. *Am. J. Sociol.*, **76**, 103–154.

Thorsen,K. *et al.* (2011) Tumor-specific usage of alternative transcription start sites in colorectal cancer identified by genome-wide exon array analysis. *BMC Genomics*, **12**, 505.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.