Available online at www.sciencedirect.com

ScienceDirect

www.sciencedirect.com/science/journal/16720229

ELSEVIER

GENOMICS
PROTEOMICS &
BIOINFORMATICS

Application Note

Gene2DGE: A Perl Package for Gene Model Renewal with Digital Gene Expression Data

Xiaoli Tang[1#], Libin Deng[1,2,3#], Dake Zhang[3], Jiari Lin[2], Yi Wei[2], Qinqin Zhou[2], Xiang Li[1], Guilin Li[1], and Shangdong Liang[1*]

[1]*Faculty of Basic Medical Science, Nanchang University, Nanchang 330006, China;*
[2]*Institute of Translational Medicine, Nanchang University, Nanchang 330006, China;*
[3]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China.*

Genomics Proteomics Bioinformatics 2012 Feb; 10(1): 51-54 DOI: 10.1016/S1672-0229(11)60033-8

Received: Sep 20, 2011; Accepted: Jan 31, 2012

Abstract

For transcriptome analysis, it is critical to precisely define all the transcripts across the whole genome. More and more digital gene expression (DGE) scannings have indicated the presence of huge amount of novel transcripts in addition to the known gene models. However, almost all these studies still depend crucially on existing annotation. Here, we present Gene2DGE, a Perl software package for gene model renewal with DGE data. We applied Gene2DGE to the mouse blastomere transcriptome, and defined 98,532 read-enriched regions (RERs) by read clustering supported by more than four reads for each base pair. Taking advantage of this *ab initio* method, we refined 2,104 exonic regions (4% of a total of 48,501 annotated transcribed regions) with remarkable extension into un-annotated regions (>50 bp). For 5% of uniquely mapped reads falling within intron regions, we identified 13,291 additional possible exons. As a result, we renewed 4,788 gene models, which account for 39% of a total of 12,277 transcribed genes. Furthermore, we identified 12,613 intergenic RERs, suggesting the possible presence of novel genes outside the existing gene models. In this study, therefore, we have developed a suitable tool for renewal of known gene models by *ab initio* prediction in transcriptome dissection. The Gene2DGE package is freely available at http://bighapmap.big.ac.cn/.

Key words: transcriptome, annotation, *ab initio* prediction

Introduction

Digital gene expression sequencing, namely DGE-seq, refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about RNA content of a sample (*1*). It can provide researchers with a powerful tool to obtain unbiased and unparalleled information about gene transcripts (*2, 3*). Currently, computational methods are being developed to identify and annotate these transcripts with alternative splice forms (*4, 5*). Although most DGE-seq studies have identified expression outside of known loci (in intronic or intergenic regions) (*6-10*), few attempts have been made to *ab initio* define the read-enriched regions (RERs) in detail and

[#] Equal contribution.
*Corresponding author.
E-mail: liangsd@hotmail.com

compare them with known gene models.

Here, we present Gene2DGE, a free Perl software package for RER detection and gene model update. This novel method consists of RER definition based on read clustering followed by annotation comparison with known gene models. The input of Gene2DGE is the file of mapped reads from RNA-seq data and a gene annotation file of the corresponding genome. In addition, a cmap file needs to be prepared for application to different species to correct the chromosome numbers. The output of Gene2DGE includes a text file containing a set of RERs and a series of text files containing annotated information of the eligible RERs. The Gene2DGE package is freely available at http://bighapmap.big.ac.cn/.

Implementation

We developed Gene2DGE as an *ab initio* tool to annotate the transcriptome using mapped reads from the SOLiD platform (Applied Biosystems) and annotation information downloaded from Ensembl. Gene2DGE consists of three steps. First, we filter the "uniquely mapped" reads from the aligned results of the SOLiD Whole Transcriptome Pipeline. A uniquely mapped read is defined as one with a max scoring alignment to the genome scoring at least 24 and at least four higher than any of the other alignments of that read to the genome (*11*). Considering the restrictions of computer memory, this process will be performed for each chromosome in order, so it is relatively convenient for use on any personal computer.

Second, based on uniquely mapped reads, we construct the RERs by grouping overlapped reads with a number greater than a threshold (at least four reads) (*12*). We list each RER including start position, end position and the number of mapped reads. In addition, we set a parameter for the "maximal distance between RERs", defined by the start positions of RERs minus the start positions of the first one upstream. It can be customized according to the specific requirement of the experiment (the default value is 50 bp).

Finally, we compare the RERs to existing gene models, and generate a catalogue of candidate genes with new annotation information, including exon extension, possible additional exons, and novel genes. The file of existing gene models in *gtf* format can be downloaded from the Ensembl website for the candidate species. We picked out eligible RERs and then checked their overlap with known gene models. As a result, a series of annotated files will be output and then can be used in further analysis.

Application

We applied Gene2DGE to the RNA-seq data from the mouse blastomere dataset obtained from a single-cell whole transcriptome (*13*). The mRNA-Seq short reads were analyzed using whole-transcriptome software tools (Applied Biosystems, http://www.solidsoftwaretools.com/). The reads generated were mapped to the mouse genome (mm9, NCBI build 37). We got more than 6.6 million reads that could be uniquely aligned to the mouse genomic reference ("uniquely mapped reads"). Based on Ensembl annotation (NCBI M37.61), 89% reads (5.9 million) were mapped to annotated regions in exons including coding sequences (CDS) and untranslated regions (UTR), which is significantly higher than those mapped to intronic (0.3 million, 5%) and intergenic (0.4 million, 6%) regions (**Figure 1A**).

Across the mouse genome, 98,532 RERs were identified and each contained more than 4 reads. A total of 62.3% of RER boundaries were within 10 bp of the ends for the corresponding exons (**Figure 1B**). Meanwhile, we identified 2,217 exon ends with remarkable extension into un-annotated regions (>50 bp), suggesting that the mouse transcriptome was more complex than we expected. The transcript levels for RERs overlapping with known exons (exonic RERs) were significantly higher than those of novel RERs (Mann-Whitney U test, $P<10^{-35}$).

We detected 12,277 expressed transcripts (with at least 1 RER across the genic region), in which 11,261 (92%) transcribed genes contained at least one exonic RER. For the 72,628 (74% of all 98,532 RERs) exonic RERs, we found that the known gene models were well defined by the *ab initio* method. For example, read distribution on chr 7 (56900000-57200000) revealed sharp boundaries of RER regions (**Figure 2**).

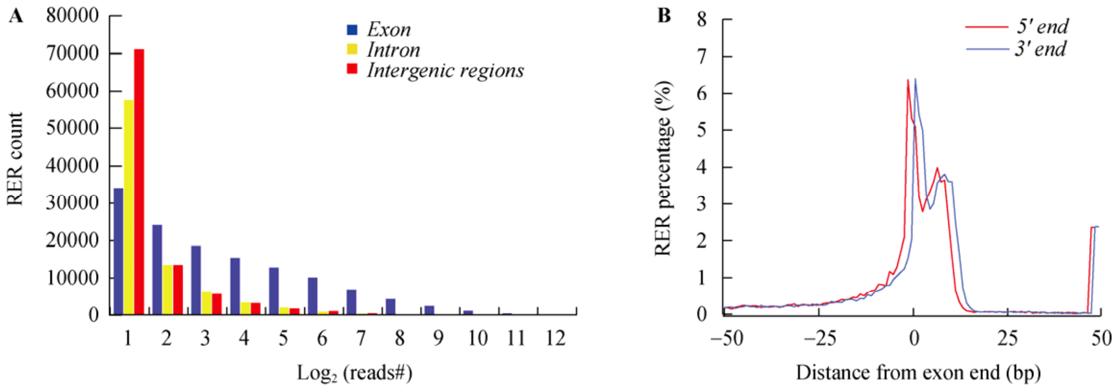

Figure 1 Summary of read-enriched regions (RERs) across the mouse genome. **A**. Distribution of read counts within RERs demonstrates possible transcription in previously non-annotated regions. **B**. Deviation between exon ends and corresponding RER boundaries. The minus numbering indicates RERs are shorter than known exons, while the positive numbering indicates RERs are longer than known exons. The apparent shortness of both first 5′ and last 3′ ends is possibly caused by transcript degradation.

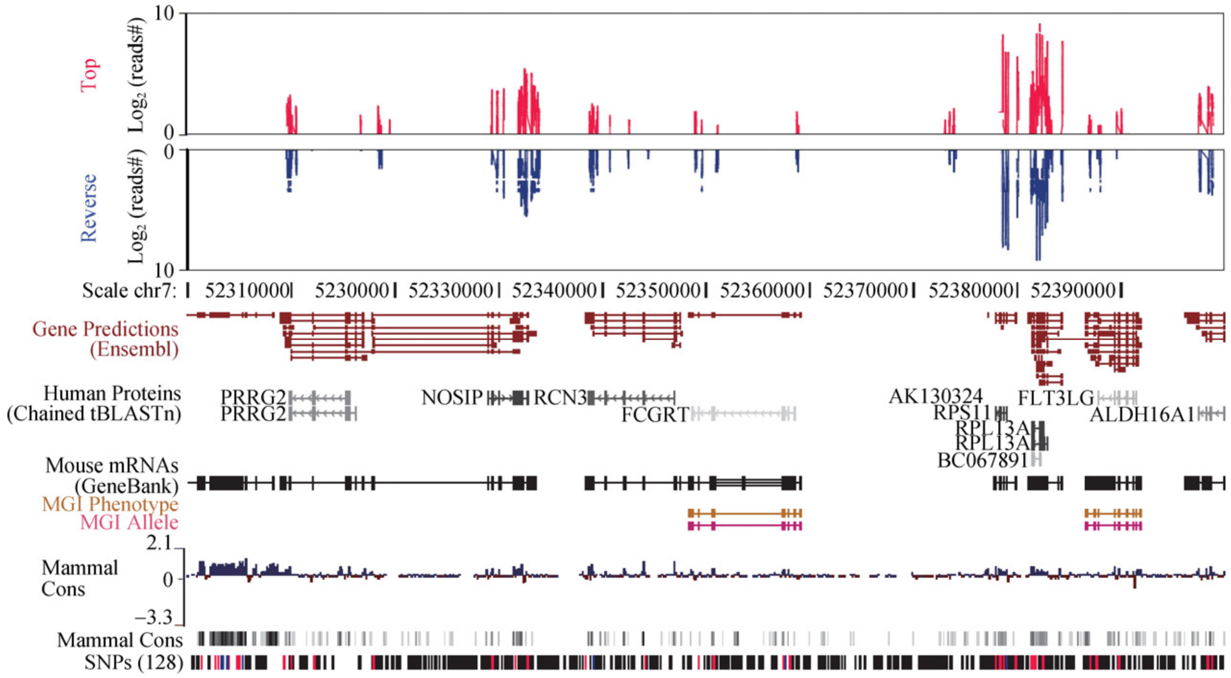

Figure 2 Transcriptome features of mouse blastomere illustrated by DGE data based on annotation available. Read distribution on chr 7 (56900000-57200000) in upper panel (Top/Reverse) was shown using sequencing data obtained from mouse blastomere. Boundaries of RERs were generated using Gene2DGE based on the read distribution. Improved annotation of gene models and novel transcriptions were also illustrated.

Furthermore, we detected a certain proportion of RERs (25,904, 26% of all RERs identified), which are located outside of the annotated regions in this transcriptome dataset. Among them, 13,291 intronic RERs (about 13% of all) were identified in 4,449 genes, indicating possible additional exons. Interestingly, about 1,016 genes (23%) only had transcripts detected in the intronic regions but not in the exonic regions. As a result, we renewed 4,788 gene models, which account for 39% of 12,277 transcribed genes in total. The remaining 12,613 RERs are located in the intergenic regions, suggesting possible presence of novel genes outside of the existing gene models.

Conclusion

Here we have developed an exploratory tool, Gene2DGE, which can be employed to determine the

RERs and improve genome annotation. The package and methods can be applied to analyze other sources of any mapped short read counts from RNA-seq data, such as results of sequencing by AB SOLiD platform and Illumina Solexa platform. Moreover, Gene2DGE can be used on any personal computer with a low requirement for computer memory capacity, since data are processed for all chromosomes in order with one chromosome at a time.

In this study, we provide an example of Gene2DGE usage to illustrate its application in transcriptome analysis. Gene2DGE has also been applied to analyze datasets from other mouse tissues or tissues from other species (data not shown). All these results indicate that Gene2DGE is a suitable tool for the renewal of known gene models by *ab initio* prediction in transcriptome dissection.

Acknowledgements

This work was supported by the National Nature Science Foundation of China (Grant No. 81171184, 31060139 and 30871384), Nature Science Foundation of Jiangxi Province (Grant No. 20114BAB215019), Department of Health of Jiangxi Province (Grant No. 20111209) and Technology Pedestal and Society Development Project of Jiangxi Province (Grant No. 2010BSA09500 and 20111BBG70009-1).

Authors' contributions

XT and LD designed the study and performed the majority of data analysis. DZ, JL, YW, QZ, XL and GL collected the dataset and participated in data analysis and visualization. XT, LD and LS supervised the project and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that no competing interests exist.

References

1 Wang, Z., *et al.* 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63.

2 Eveland, A.L., *et al.* 2010. Digital gene expression signatures for maize development. *Plant Physiol.* 154: 1024-1039.

3 Lai, Y. 2010. Differential expression analysis of Digital Gene Expression data: RNA-tag filtering, comparison of *t*-type tests and their genome-wide co-expression based adjustments. *Int. J. Bioinform. Res. Appl.* 6: 353-365.

4 Laporta, J., *et al.* 2011. Short communication: expression and alternative splicing of POU1F1 pathway genes in preimplantation bovine embryos. *J. Dairy Sci.* 94: 4220-4223.

5 Shang, H., *et al.* 2011. Identification and characterization of alternative promoters, transcripts and protein isoforms of zebrafish R2 gene. *PLoS One* 6: e24089.

6 Cloonan, N., *et al.* 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5: 613-619.

7 Sultan, M., *et al.* 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321: 956-960.

8 Wilhelm, B.T., *et al.* 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453: 1239-1243.

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

10 Wang, L., *et al.* 2010. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26: 136-138.

11 Tuch, B.B., *et al.* 2010. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One* 5: e9317.

12 Mortazavi, A., *et al.* 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628.

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