BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 6 2014, pages 891–892
doi:10.1093/bioinformatics/btt649

Gene expression

Advance Access publication November 8, 2013

RNAseqViewer: visualization tool for RNA-Seq data

Xavier Rogé[1] and Xuegong Zhang[1,2,*]

[1]MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/Department of Automation and [2]School of Life Sciences, Tsinghua University, Beijing 100084, China

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: With the advances of RNA sequencing technologies, scientists need new tools to analyze transcriptome data. We introduce RNAseqViewer, a new visualization tool dedicated to RNA-Seq data. The program offers innovative ways to represent transcriptome data for single or multiple samples. It is a handy tool for scientists who use RNA-Seq data to compare multiple transcriptomes, for example, to compare gene expression and alternative splicing of cancer samples or of different development stages.

Availability and implementation: RNAseqViewer is freely available for academic use at http://bioinfo.au.tsinghua.edu.cn/software/RNAseqViewer/

Contact: zhangxg@tsinghua.edu.cn

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

Received on July 17, 2013; revised on October 16, 2013; accepted on November 5, 2013

1 INTRODUCTION

The development of next-generation sequencing allows large-scale transcriptome studies, using massively parallel sequencing of complementary DNA reverse transcribed from RNA (RNA-Seq) technologies (Mak, 2011; Wang *et al*., 2009). Scientists now face the challenge of analyzing the unprecedented amount of available sequencing data. In addition to the use of automated tools, such analysis needs human review and interpretation of the data to gain insight into them, verify consistency of programs' results, understand intricate cases and structures and reveal underlying patterns. As data files generated by automated tools for RNA-Seq analysis are formatted for computational treatments and can exceed gigabytes, they are improper for easy direct human reading. This calls for effective visualization tools, capable of displaying big datasets in intuitive ways.

Genomic visualization software can be divided into two categories: web-based genome browsers like those from UCSC and Ensembl (Kent *et al*., 2002; Spudich *et al*., 2007; Wang *et al*., 2013) and standalone software including the Integrative Genomics Viewer and Savant (Fiume *et al*., 2010; Robinson *et al*., 2011). Only the latter category is adapted to the visualization of the large locally stored datasets usually manipulated in RNA-Seq analysis, as web upload is prohibitive. However, existing solutions are limited to a few tracks visualized simultaneously, and it is not possible to compare the multiple samples of medium or large RNA-Seq experiments.

*To whom correspondence should be addressed.

Here we introduce RNAseqViewer, a desktop program to visualize the data from the RNA-Seq analyzing process, for single or multiple samples. By focusing on expression of genes and transcript isoforms, the program offers innovative ways to present the transcriptome data in a quantitative and interactive manner.

2 IMPLEMENTATION

RNAseqViewer is a cross-platform program implemented in the language Python (http://www.python.org/). It makes use of the package PySide, a binding to the graphical library Qt (http://qt-project.org/wiki/PySide), and the programs Samtools and Tabix for reading Sequence Alignment/Map (SAM), BAM and gene transfer format (GTF) files (Li, 2011; Li *et al*., 2009).

3 PROGRAM OVERVIEW

3.1 Supported data types

RNAseqViewer supports seven types of data widely used in RNA-Seq analysis (see Table 1). The program can read the results of RNA mappers like TopHat (Trapnell *et al*., 2009), namely, read alignments in SAM/BAM format and splicing junction sites in BED format. A third major type consists in transcriptomes, assembled by programs like Cufflinks (Trapnell *et al*., 2010) in GTF format. Reference sequence and genome annotations can be loaded from files, respectively, in FASTA and RefFlat or GTF format. The program also supports Wiggle format to visualize dense numeric data and BED format for any other type of tracks.

3.2 Graphs

We designed the graphs with three objectives in mind: (i) display as much information as possible while keeping the resulting graph simple, (ii) offer compact views to allow dozens of

Table 1. Types of data and graphs supported by RNAseqViewer

Data type	Format	Views
Reads alignments	SAM/BAM	Mapped reads, read coverage, heatmap-like view
Splicing junctions	BED	Junction lines, junction reads
Transcriptomes	GTF	Expanded view, collapsed view
Genome annotations	RefFlat, GTF	Expanded view, collapsed view
DNA sequence	FASTA	Sequence
Numeric data	Wiggle	Histogram
Other tracks	BED	Expanded view, collapsed view

tracks to be seen simultaneously and (iii) keep the views intuitive for biologists in regard to current visualization standards.

Therefore RNAseqViewer offers different types of graphs for each type of data (see Table 1 and Supplementary Figs S2 and S3). They can be chosen by the users according to their needs and they automatically adapt to the scale with a semantic zoom approach.

Many graphs use colors as a mean to display more data without overloading the graph and in a compact way. Colors represent scores like read count or FPKM (Fragments per kilobase of exon per million fragments mapped) in heatmap-like graphs, or they are used to highlight significant details like potential single nucleotide polymorphisms. Additional data are displayed in tool tips or in an optional frame.

For all graphs, intronic regions can optionally be hidden. This is especially useful for genes with long introns and short exons. In this mode, the exons of these genes can be viewed on a single screen without being shrunk, hence offering a better view of data mapped to the exons.

3.3 User interface

Seamless browsing of the datasets is highly desirable to gain an extensive insight into the data. To achieve this goal, RNAseqViewer offers a dynamic user interface (see Supplementary Fig. S1) and preloads data next to the current view so that they can be displayed immediately when browsing around.

Users can either use the mouse or the keyboard to navigate along the chromosomes in a Google Maps-like fashion. Panning is achieved by dragging the tracks, and zooming is achieved by scrolling with the mouse wheel. Alternatively users can use the interface's navigation buttons or the keyboard shortcuts. Search for gene name and coordinates input allow direct access to specific locations.

3.4 Memory and speed performances

Special attention has been given to the memory management so that gigabyte-sized datasets can be visualized without exceeding the memory limits or impacting speed performance. RNAseqViewer only loads in memory the data that are immediately necessary and those that are close to the current view.

A specific strategy is applied for file formats usually chosen for big datasets. Before visualizing files in BAM, GTF or FASTA format, an index file is created using Samtools or Tabix. Thanks to the index, data do not need to be fully loaded in memory, yet they are remotely accessible in a short time.

4 USE EXAMPLE

Among other datasets, RNAseqViewer has been tested with the data deposited in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE22260 (Kannan *et al.*, 2011). The dataset contains 36-nt pair-end reads sequenced with Illumina GAII platform from 20 prostate cancer tumors and 10 matched normal tissues. We mapped the reads on the reference human genome hg18 with TopHat and used the result for *de novo* transcriptome assembly with Cufflinks. Figure 1 and Supplementary Figure S3 show screenshots of the data loaded in RNAseqViewer on a Linux system with a RAM of 1 GB.

Funding: This work is supported in part by the National Basic Research Program of China (2012CB316504), the Hi-tech Research and Development Program of China (2012AA020401) and NSFC grant 91010016.

Conflict of Interest: none declared.

Fig. 1. Screenshots of RNAseqViewer. The visualized dataset contains 36-nt pair-end reads from 20 prostate cancer tumors and 10 matched normal tissues. The reads have been mapped to the reference human genome hg18 with TopHat. (**A**) Heatmap-like view of gene expression of the 20 tumor samples (top 20 lines) and 10 control samples (bottom 10 lines) for two isoforms of the gene HDAC2. Colors depend on the FPKM of the sample represented by the line and the exon represented by the column. Intronic regions are hidden. The graph was generated from the 30 BAM files resulting from reads mapping. (**B**) Gene expression of one sample: read coverage, splicing junctions, with colors depending on the junction reads count, and transcripts, with colors depending on FPKM. The two first graphs were generated from TopHat output (BAM and BED file) and the latter from cufflinks output (GTF file)

REFERENCES

Fiume,M. *et al.* (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, **26**, 1938–1944.

Kannan,K. *et al.* (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proc. Natl Acad. Sci. USA*, **108**, 9172–9177.

Kent,W.J. *et al.* (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.

Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Mak,H.C. (2011) Next-generation sequence analysis. *Nat. Biotechnol.*, **29**, 45–49.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Spudich,G. *et al.* (2007) Genome browsing with Ensembl: a practical overview. *Brief. Funct. Genomic Proteomic*, **6**, 202–219.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

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.

Wang,J. *et al.* (2013) A brief introduction to web-based genome browsers. *Brief. Bioinform.*, **14**, 131–143.

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