BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 18 2011, pages 2598–2600
doi:10.1093/bioinformatics/btr441

Sequence analysis

Advance Access publication July 27, 2011

RseqFlow: workflows for RNA-Seq data analysis

Ying Wang[1,2,*], Gaurang Mehta[3], Rajiv Mayani[3], Jingxi Lu[1], Tade Souaiaia[1], Yangho Chen[1], Andrew Clark[4], Hee Jae Yoon[4], Lin Wan[1], Oleg V. Evgrafov[4], James A. Knowles[4,*], Ewa Deelman[3,*] and Ting Chen[1,*]

[1]Department of Biological Sciences, USC, 1050 Childs Way, Los Angeles, CA 90089, USA, [2]Department of Automation, Xiamen University, China, [3]USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA and [4]USC Keck School of Medicine, Department of Psychiatry, 1501 San Pablo St., Los Angeles, CA 90033, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: We have developed an RNA-Seq analysis workflow for single-ended Illumina reads, termed RseqFlow. This workflow includes a set of analytic functions, such as quality control for sequencing data, signal tracks of mapped reads, calculation of expression levels, identification of differentially expressed genes and coding SNPs calling. This workflow is formalized and managed by the Pegasus Workflow Management System, which maps the analysis modules onto available computational resources, automatically executes the steps in the appropriate order and supervises the whole running process. RseqFlow is available as a Virtual Machine with all the necessary software, which eliminates any complex configuration and installation steps.

Availability and implementation: http://genomics.isi.edu/rnaseq

Contact: wangying@xmu.edu.cn; knowles@med.usc.edu; deelman@isi.edu; tingchen@usc.edu

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

Received on March 13, 2011; revised on May 23, 2011; accepted on July 23, 2011

1 INTRODUCTION

With the recent technological breakthroughs in the study of the whole transcriptome (Marioni *et al*., 2008; Mortazavi *et al*., 2008), RNA-Seq datasets have been generated to decipher, characterize and quantify the set of all RNA molecules produced in cells. Read-alignment tools (Chen *et al*., 2009; Langmead *et al*., 2009; Li and Durbin, 2009) and gene expression-estimation tools (Jiang and Wong, 2009; Srivastava and Chen, 2010) have been developed to analyze millions of reads generated from mRNA sequencing and to determine the expression level. However, the majority of these tools have focused on single functions, resulting in the frequent incompatibility of data interfaces among different tools. Therefore, integrating these tools into an overall computation, or, a workflow, is critical for downstream RNA-Seq data analysis.

Recent works on RNA-seq analysis pipelines include inGAP (Qi *et al*., 2010), RSEQtools (Habegger *et al*., 2011) and ArrayExpressHTS (Goncalves *et al*., 2011). Each of them offers a subset of common RNA-Seq analytical functions that are organized and managed by specific software management system. In this article, we present an RNA-Seq analysis workflow, RseqFlow, which attempts to integrate more analytical functions than the previous tools, and at the same time, be flexible and easy to use. In particular, this workflow is organized by the Pegasus Workflow Managment System (Deelman *et al*., 2005, 2007; http://pegasus.isi.edu), which supports large-scale workflows on diverse environments in a scalable and reliable fashion (Callaghan *et al*., 2010). Users can easily acquire analysis results with just one command, instead of in a step-by-step manner and inputting commands and arguments for each module. We also provide a downloadable Virtual Machine (VM) image that allows users to run the RseqFlow easily using different computational resources, including a local laptop, a workstation, high-performance computer clusters, national Grids or computational Clouds (Juve *et al*., 2010), thus eliminating any complex software installation and workflow configuration steps.

2 FEATURES AND FUNCTIONS

2.1 Functions of RNA-Seq analysis workflow

The RseqFlow workflow (Supplementary Figure S1) contains several modules and features. The major modules include mapping reads to genome and transcriptome references, performing quality control (QC) of sequencing data (described in Supplementary Section S2), generating files for visualizing signal tracks based on the mapping results, calculating gene expression levels, identifying differentially expressed genes, calling coding SNPs (described in Supplementary Section S4) and producing MRF and BAM files (described in Supplementary Section S5). The modules in workflow can be extended for more functions, or replaced by other methods. Some of the unique features of RseqFlow are specified in the following.

Mapping reads: the current version includes two alignment tools, Bowtie (Langmead *et al*., 2009) and PerM (Chen *et al*., 2009) as options. The workflow consists of two separate tracks for mapping, one for the male genome with chromosome Y, and the other for the female genome without chromosome Y. Reads are mapped to both the genome and the transcriptome reference because alignments to the genome can locate transcripts that are not annotated in the transcriptome database, while alignments to the transcriptome can identify splice junctions. The mapping results are merged to output a unique-mapped read set and a multi-mapped read set for

*To whom correspondence should be addressed.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

downstream analysis. The detail of the merging can be found in the Supplementary Section S1.

Calculating expression levels: expression levels of genes, exons and splice junctions are calculated based on the reads aligned to the transcriptome. For genes and exons, we calculate RPKM (Number of reads Per Kilobases per Million mapped reads) values (Mortazavi *et al*., 2008), and for splice junctions, we calculate RPM (Number of reads Per Million mapped reads) values. We further implemented three choices of RPKM values, using different strategies to treat the multi-reads that are mapped to multiple genome/gene regions. These include the elimination of multi-reads (RPKM_Uniq), the random assignment of multi-reads (RPKM_Random) and the elimination of both multi-reads and un-mappable transcript regions (RPKM_UM). An un-mappable transcript region is a sub-sequence of a transcript that is identical, or almost identical, to sub-sequences of other genes or other genome sub-sequences. We pre-compute libraries for un-mappable transcript regions based on gender, read length and gene annotation databases. Among the three RPKM formulas, we found that RPKM_UM gave the most accurate gene expression levels and is therefore the default option in our workflow. Details are described in the Supplementary Section S3.

Identifying Differentially Expressed (DE) genes: we implemented the negative binomial model proposed in DESeq (Anders and Huber, 2010) to compute differentially expressed genes. For datasets with replicates, we use the negative binomial model to compute *P*-values for differentially expressed genes. For datasets without replicates, *P*-values for exons are computed first and then combined into a single value using the Fisher probability test.

Flexibility of the workflow: the workflow allows substitution of existing functions with other methods, and can be easily extended to include more analytical functions. In addition, it can be revised for other species, using the details described in the Supplementary Section S3.1. The first version of RseqFlow was built for the human genome, using human genome Hg19 and the transcriptome database of Gencode v3. Recently, we built a workflow for Rhesus, which can be downloaded from our website (http://pegasus.isi.edu).

2.2 Workflow with Pegasus and virtual machine

RseqFlow uses Pegasus to develop and manage workflow execution (Deelman *et al*., 2005). Pegasus manages task execution and data movement efficiently, and allows users to track the provenance of the workflow execution results, including the modules/codes used, the machine on which the modules ran, the parameters used for the analysis and the datasets and their sizes. If failures occur during execution, Pegasus tries to automatically recover from them by resubmitting the failed tasks or by re-planning the failed part of the workflow.

To reduce the time necessary to begin using our workflow, we have created an RseqFlow VM capable of running on Windows, Linux and Macintosh without complex software installation and workflow configuration steps. This VM allows users to run RseqFlow either on a local machine or on Amazon EC2. Instructions on running the VM on a local machine with sample datasets as well as user datasets are provided on our website at (http://genomics.isi.edu/rnaseq). The VM can also be used as a submit host to schedule jobs on remote resources such as clusters, grids and clouds. This advanced setup does require several additional configuration steps, and will be automated in a later release.

RseqFlow consists of a total of 155 computing tasks and 31 data management tasks when the input sample data is split into 64 parts to map to the genome, in parallel. This is a configurable parameter for the workflow. RseqFlow has been tested in a cluster environment as well as on our VM. Our test cluster was a 118 nodes Linux cluster with 54 dual hexcore Intel x86_64 Xeon 2.66 Ghz, 24 GB RAM nodes and 64 dual quadcore Intel x86_64 Xeon 2.3 Ghz, 16 GB RAM nodes. The input dataset including the sample dataset in fastq format and reference data is about 29 GB in size. The parallel workflow runs on our test hardware in 187 mins. Without splitting the data it takes ~784 mins to run on the same cluster, thus resulting in a 440% improvement in runtime. During execution the workflow uses a max of 72 cores in parallel. The workflow uses 53 GB scratch space on a shared file system across all the nodes and saves 37 GB of output data.

3 CONCLUSIONS

In summary, we have developed RseqFlow, a workflow containing an interacting set of RNA-Seq analytic functions. For operational efficiency, modules are formalized into a workflow and managed by Pegasus Workflow Management System in a VM environment. Although the current version of RseqFlow is built on human genome Hg19 and Gencode, it is easy to re-compile the workflow for other species. The analysis modules can be easily customized because the inputs to the modules are in standard data format. Users can also substitute existing functions with other tools. And the workflow is open to further function extension, and we plan to expand the workflow by adding functions for novel alternative splicing identification and other required analysis.

ACKNOWLEDGEMENTS

We also wish to thank our collaborators in the TAHBD for helpful discussions.

Funding: NIH/NIMH 5 RC2 MH090047-01, Transcriptional Atlas of Human Brain Development (TAHBD). Center of Excellence of Genome Sciences: Genomic Analysis of the Genotype-Phenotype Map (NIH/HG 2 P50 HG002790-06).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Callaghan,S. *et al*. (2010) Scaling up workflow-based applications. *J. Comput. Syst. Sci.*, **76**, 428–446.

Chen,Y. *et al*. (2009) PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, **25**, 2514–2521.

Deelman,E. *et al*. (2005) Pegasus: a framework for mapping complex scientific workflows onto distributed systems. *Sci. Program. J.*, **13**, 219–237.

Deelman,E. *et al*. (2007) Pegasus: mapping large-scale workflows to distributed resources. In *Workflows for e-Science*, Part III, Springer, London, pp. 376–394.

Goncalves,A. *et al*. (2011) A pipeline for RNA-seq data processing and quality assessment. *Bioinformatics*, **27**, 867–869.

Habegger,L. *et al*. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, **27**, 281–283.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Juve,G. *et al*. (2010) Data sharing options for scientific workflows on Amazon EC2. In *Proceedings of the 2010 ACM/IEEE International Conference for High*

Performance Computing, Networking, Storage and Analysis, IEEE Computer Society Washington, DC, USA.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Marioni,J. *et al*. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome Res.*, **18**, 1509–1517.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Method*, **5**, 621–628.

Qi,J. *et al*. (2010) inGAP: an integrated next-generation genome analysis pipeline. *Bioinformatics*, **26**, 127–129.

Srivastava,S. and Chen,L. (2010) A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res.*, **38**, e170.