BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 5 2012, pages 721–723
doi:10.1093/bioinformatics/bts023

Genome analysis

Advance Access publication January 17, 2012

FX: an RNA-Seq analysis tool on the cloud

Dongwan Hong[1,2,†], Arang Rhie[1,3,4,5,†], Sung-Soo Park[6,†], Jongkeun Lee[2], Young Seok Ju[1,7], Sujung Kim[6], Saet-Byeol Yu[6], Thomas Bleazard[1], Hyun-Seok Park[3], Hwanseok Rhee[7,8], Hyonyong Chong[7,8], Kap-Seok Yang[7,8], Yeon-Su Lee[2], In-Hoo Kim[2], Jin Soo Lee[2], Jong-Il Kim[1,4,5,6,*] and Jeong-Sun Seo[1,4,6,7,8,*]

[1]Genomic Medicine Institute, Medical Research Center, Seoul National University, Seoul 110-799, [2]Cancer Genomics Branch and Research Institute and Hospital, National Cancer Center, Goyang-si 410-769, [3]Department of Computer Science and Engineering, Ewha Womans University, Seoul 120-750, [4]Department of Biochemistry, Seoul National University College of Medicine, [5]Department of Biomedical Sciences, Seoul National University Graduate School, [6]Psoma Therapeutics Inc., Seoul 110-799, [7]Macrogen Inc., Seoul 153-023 and [8]Axeq Technologies, Rockville, MD 20850, USA

Associate Editor: Prof. Ivo Hofacker

ABSTRACT

Summary: FX is an RNA-Seq analysis tool, which runs in parallel on cloud computing infrastructure, for the estimation of gene expression levels and genomic variant calling. In the mapping of short RNA-Seq reads, FX uses a transcriptome-based reference primarily, generated from ~160 000 mRNA sequences from RefSeq, UCSC and Ensembl databases. This approach reduces the misalignment of reads originating from splicing junctions. Unmapped reads not aligned on known transcripts are then mapped on the human genome reference. FX allows analysis of RNA-Seq data on cloud computing infrastructures, supporting access through a user-friendly web interface.

Availability: FX is freely available on the web at (http://fx.gmi.ac.kr), and can be installed on local Hadoop clusters. Guidance for the installation and operation of FX can be found under the 'Documentation' menu on the website.

Contact: jeongsun@snu.ac.kr; jongil@snu.ac.kr

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

Received on September 2, 2011; revised on December 28, 2011; accepted on January 8, 2012

1 INTRODUCTION

Accurately quantifying gene expression levels and identifying variants in the transcriptome is important for research into cell differentiation and disease diagnosis. Recently, various genome research groups have developed genome and transcriptome analysis software based on cloud computing technology to facilitate the analysis of large amounts of sequencing data without the purchase of computing resources. Bioinformatics cloud resources using the Hadoop Map/Reduce framework have been released, such as CloudAligner (Nguyen *et al.*, 2011), CloudBurst (Schatz, 2009), Crossbow (Langmead *et al.*, 2009) and MyRNA (Langmead *et al.*, 2010) (Supplementary Table S1). However, operation of cloud-based analytic tools can be difficult for non-expert users such as biologists and medical doctors.

In our previous study (Ju *et al.*, 2011), we found a problem inherent to RNA-Seq analysis tools that align to the human reference genome: reads coming from spliced junctions of large introns cannot be aligned due to indel (insertion and deletion) sensitivity of alignment tools. To resolve this problem, recent studies have analyzed gene expression and alternative splicing by aligning short RNA-Seq reads against previously known or predicted transcripts (Ju *et al.*, 2011; Trapnell *et al.*, 2010). Following this approach, we enhanced the mapping of short reads by aligning toward a reference composed of known genes and their isoforms. To identify unannotated transcripts, FX aligns the remaining unmapped reads onto human genome reference. When short reads align to multiple cDNA sequences from the three databases, only the hit with the highest mapping score is used (Supplementary Fig. S1). This method allowed us to profile gene expression and call variants with great accuracy (Ju *et al.*, 2011).

In this work, we implemented these methods in the user-Friendly gene eXpression analytic tool (FX), allowing RNA-Seq data analysis to begin immediately upon completion of sequencing. FX can be run by researchers without investment in their own high performance computing (HPC) at low cost using the Amazon Web Services (AWS, http://aws.amazon.com). The results output by FX are gene expression profiles, SNP calls and short indels (Supplementary Material 1). This service can be accessed through our web interface. Alternatively, FX is freely available for local distribution.

2 METHODS

In FX, each step is processed by mapping, shuffling and reducing the data over worker nodes. System configuration is shown in Supplementary Figure S2. FX splits data processing into several steps (Supplementary Fig. S3). Due to this loosely coupled step design, the researcher can run each step separately with custom filter conditions.

Preprocess: before aligning paired-end sequencing reads to the reference, the *preprocess* step converts the FASTQ file to GSNAP (Wu and Nacu, 2010) input format.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Fig. 1. The web-based user interface of FX, which calls analysis modules exploiting Amazon Web Service cloud computing resources.

Mapping on cDNA sequences: FX uses GSNAP as its default alignment tool. Alignment is done against our own cDNA reference which consists of transcripts from refGenes (∼34 000), Known Genes (∼65 000) and Ensembl (∼62 000) from NCBI, UCSC and EBI, respectively. In addition, alignment of unmapped reads is carried out on the human genome reference to detect unannotated transcripts. These transcripts are reported if they have average coverage ≥ 4 and do not overlap any genes. The preprocess and alignment steps can be omitted if the user has already aligned with a different tool yielding results in SAM format. FX is mainly intended for analysis of human RNA-Seq data. However, users can add a custom reference to analyze other species.

Bioinformatic filter conditions: the GSNAP output format and SAM format are both accepted as input. This step filters out reads that have too many mismatches (>5% of the read length) and bases with Phred quality score <20 by default (Kim *et al.*, 2009). Some (4 by default) bases are trimmed in order to avoid the ambiguity of alignment at read ends. Reads with multiple alignments to the human reference genome are eliminated.

SNP and indel identification: FX calls genomic variants (SNPs and indels) with ≥ 4 uniquely aligned reads and allele frequency $\geq 1\%$ using filter criteria in (Ju *et al.*, 2011; Kim *et al.*, 2009).

Profiling of transcript expression: in this step, bases aligned on each gene are aggregated. Similar to reads per kilobase of exon per million mapped reads (RPKM) (Mortazavi *et al.*, 2008), FX uses a concept of bases per kilobase of gene model per million mapped bases (BPKM) to normalize the expression level of each gene. We defined genes with BPKM ≥ 1 as 'expressed' (Supplementary Material 2). Users may alternatively select to view results in terms of RPKM (Supplementary Material 2).

3 RESULTS

We developed FX as described in Section 2, to run on local Hadoop systems as well as the Amazon cloud system. FX was implemented using JDK 1.6.0, designed to run on clustered computers using the Hadoop Distributed File System (HDFS) built by Apache Hadoop 0.21 (http://hadoop.apache.org). A user-friendly web interface was designed, allowing researchers to adjust the bioinformatic filter conditions as desired (Fig. 1). We analyzed the transcriptome data of a Korean individual (designated AK6) using FX. Sequencing

with Illumina Genome Analyzer IIx yielded ∼70 M paired-end reads (78 bp for each end) at a coverage of 69×. We found evidence for active transcription of 1905 genes with expression level ≥1 BPKM. The total estimated running time was 81 min using 40 Amazon EC2 instances at a cost of US $45 (Supplementary Tables S2–S5 and Supplementary Fig. S4). RNA-Seq data and analysis results are available on TIARA (Hong *et al.*, 2011) and FX's website, respectively. We expect that FX will be used widely in the RNA-Seq community due to the high accuracy of its expression profiling and its user-friendly interface to the cloud.

Funding: Green Cross Therapeutics (grant No. 0411-20080023 to J.-S.S.); Korean Ministry of Knowledge Economy (grant No. 10037410-2011-02 to J.-S.S.); Korean Ministry of Education, Science, Technology (grant No. M10305030000 to J.-S.S.; grant No. 2010-0013662 to J.-I.K.); Small Medium Business Administration (grant No. 000358780109 to H.-S.P.); Amazon Web Services (June 2011 EDU research award) in part.

Conflict of Interest: none declared.

REFERENCES

Hong,D. *et al.* (2011) TIARA: a database for accurate analysis of multiple personal genomes based on cross-technology. *Nucleic Acids Res.*, **39**, D883–D888.

Ju,Y.S. *et al.* (2011) Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat. Genet.*, **43**, 745–752.

Kim,J.I. *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.

Langmead,B. *et al.* (2009) Searching for SNPs with cloud computing. *Genome Biol.*, **10**, R134.

Langmead,B. *et al.* (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, **11**, R83.

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

Nguyen,T. *et al.* (2011) CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. *BMC Res. Notes*, **4**, 171.

Schatz,M.C. (2009) CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, **25**, 1363–1369.

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.

Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.