Journal of
Computer Science & Systems Biology

ISSN: 0974-7230

Ali et al., J Comput Sci Syst Biol 2018, 11:3
DOI: 10.4172/jcsb.1000272

Research Article Open Access

Implementation of a Reproducible, Accessible and Transparent RNA-seq Bioinformatics Pipeline within the Galaxy Platform

Thahmina Ali[1], Baekdoo Kim[1], Carlos Lijeron[1], Changsu Dong[1], Claudia Wultsch[1,2] and Konstantinos Krampis[1,3,4*]

[1]Weill Cornell Medical College and Hunter College of The City University of New York, New York, USA
[2]American Museum of Natural History, New York, USA
[3]Department of Biological Sciences, Hunter College of The City University of New York, New York, USA
[4]Department of Physiology and Biophysics, Institute for Computational Biomedicine, Weill Cornell Medical College, Cornell University, New York, USA

Abstract

The technology of RNA sequencing (RNA-seq) has not only proven powerful in transcriptome studies but has become a key approach in translational medicine. In this work, we present a reusable and reproducible bioinformatics pipeline for processing and analyzing RNA-seq data, implemented as an automated workflow within the open-source, web-based Galaxy web platform. With this workflow, researchers with little to no training in computer programming or bioinformatics experience can perform routine, high quality RNA-seq analysis, and generate intuitive results. We evaluated our implementation approach of the RNA-Seq pipeline using cancer transcriptome data, demonstrating that it is well positioned for clinical applications providing a set of advantages over existing methods.

Keywords: Reproducible; Transparent RNA-seq; Bioinformatics; Galaxy platform

Introduction

Within the clinical laboratory setting, next generation sequencing (NGS) is being rapidly embraced as an invaluable diagnostic tool [1,2]. Recently, whole transcriptome shotgun sequencing (WTSS), also known as RNA-seq (RNA sequencing) has been developed for transcriptome profiling [3]. RNA-seq determines the abundance of gene transcripts in a RNA sample, superseding microarrays, which was the gold standard of gene expression analysis for almost a decade. Developing RNA-seq data analysis pipelines requires sophisticated bioinformatics skills as well as extensive computing capacity in order to handle large-scale genomic datasets [4,5]. Yet, a large fraction of the scientific community needing to use NGS approaches is impeded by limited to no background in informatics or programming. The lack of user-friendly bioinformatics analysis pipelines and computing tools to engage with the data, motivated our development of an automated RNA-seq pipeline within the Galaxy platform, a popular web application [6,7]. Galaxy is an open source, web-based genomics graphical workbench for accessible, reproducible and shareable data-intensive biomedical research.

Here we designed and implemented a Galaxy-based RNA-seq pipeline that performs differential expression and variant analysis. Galaxy's data processing framework provides a simple way to integrate and encapsulate a suite of open-source computational tools and datasets in a single graphical user interface (GUI). Galaxy also provides the functionality to orchestrate bioinformatics workflows by connecting the tools in a systematic, automated chain of operations structured to perform any desired biomedical research analysis. Our RNA-seq analysis pipeline is suited for applications in a clinical setting and is capable of detecting rare and low abundance transcripts, compare transcriptomes of cells and tissues treated under different and multiple conditions, search for alternative spliced variants, and detect genetic variants. Furthermore, our RNA-seq pipeline represents a standardized and centralized bioinformatics platform that can be easily managed, shared, and used by novices and experts alike.

Approach

RNA-seq workflow using galaxy

The pipeline was implemented using Galaxy's user-friendly, web-based interface, which served as a workbench to chain a set of bioinformatics tools and create an automated workflow (Figure 1). The required tools were imported from the public Galaxy toolshed. Copies of the pipeline are available both on our local instance of Galaxy and the public server. Our Galaxy-based pipeline tracks all analysis steps (e.g., data import, various analyses) via Galaxy's "history" panel, allowing researchers to extract and rerun specific stages of the pipelines multiple times. Furthermore, all Galaxy objects, such as the analysis history, data inputs or outputs, and workflows receive a web address that can be easily shared with other researchers, making our entire analysis workflow fully accessible and reproducible.

Workflow design and implementation

The applicability of our RNA-seq workflow was tested using cancer RNA-seq datasets from the ArrayExpress archive of the European Bioinformatics Institute (EBI) [8]. The datasets consisted of poly-A RNA sequencing paired-end reads and technical replicates sequenced using Illumina's HiSeq 2000 platform. Our workflow requires input read files generated by Illumina and one reference annotation file for the human genome (UCSC hg19). In total, the workflow performs forty-four analysis steps, using eleven bioinformatics tools (Figure 1). The workflow can analyze eight input datasets in parallel. For our study, we used two replicates of forward and reverse reads from

***Corresponding author:** Konstantinos Krampis, Department of Physiology and Biophysics, Institute for Computational Biomedicine, Weill Cornell Medical College, Cornell University, New York, NY, USA, Tel: 212-896-0461; E-mail: agbiotec@gmail.com

Received April 12, 2018; **Accepted** April 24, 2018; **Published** May 10, 2018

Citation: Ali T, Kim B, Lijeron C, Dong C, Wultsch C, et al. (2018) Implementation of a Reproducible, Accessible and Transparent RNA-seq Bioinformatics Pipeline within the Galaxy Platform. J Comput Sci Syst Biol 11: 195-199. doi:10.4172/jcsb.1000272

healthy patient tissue (data set 1-2, 5-6 in Figure 1) and two replicates of forward and reverse reads from prostate cancer tissue from the same patients (data set 3-4, 7-8 in Figure 1). There are four logical components in our workflow that can be described as the following: A. Quality Check (QC, Figure 1) left blue rectangle) performed using FastQC, version 0.71, which summarizes quality control checks of raw sequence data in html output format, with basic statistics for guidance in the pre-processing decisions [9]. Once high-quality reads are obtained, forward and reverse reads are remove and add from; Figure 1 right blue rectangle) with Tophat2, version 2.0.14, which uses Bowtie2, version 2.3.4.1 as its core engine and a pre-indexed human reference genome (UCSC hg19) [10]. Tophat2 provides an algorithm for identifying non-continuous mapped reads across splice junctions, in order to search for splicing signals with the main goal to a build set of possible introns in each transcript. The output from this step is a Binary Alignment Map (BAM) file of mapped exonic reads. To view the alignment file in text format, we included a BAM-SAM, version 1.0 (Sequence Alignment Map) converter tool in the pipeline [11]. The next module on our workflow is the C. Differential Gene Expression Analysis, which is includes Transcript Assembly and Transcriptome Quantification (Figure 1, red rectangles). Last, we performed the D. Variant Analysis (Figure 1, green rectangles).

First, the Transcript Assembly is performed using the Cufflinks2, version 2.2.1 software suite, to reconstruct and quantify the full set of transcripts [12]. Cufflinks2 (Figure 1, left red rectangle) uses the BAM alignment file from TopHat and a reference annotation file in Gene Table Format (GTF) to generate a transcriptome assembly. Next, assemblies are merged together using the Cuffmerge tool. To validate the assembled transcript fragments by Cufflinks, the Cuffcompare tool (Figure 1, red rectangle) is also used to compare the assembled transcripts from the merged assembly to a reference annotation genome. At this stage, any new splice variants by comparison of the assembled transcripts to the reference annotation, are saved in a new GTF file. Second, the Transcriptome Quantification is performed by the Cuffcompare tools, using the GTF file along with the alignment BAM files [13]. Next, the Cuffcompare and Cuffdiff tools calculate the expression levels in fragments per kilobase of exon per million fragments mapped (FPKM), and test for statistical significance of difference in expression levels between samples. The next step is the Variant Analysis, which applies SAMtools mpileup tool, version 2.1.3 and uses the alignment BAM files along with the reference genome files [11,14,15]. The Variant Analysis generates a "pileup" of read bases in order to identify Single Nucleotide Polymorphisms (SNPs), for each position in RNA-seq transcripts compared to the reference genome. In the final step of our bioinformatics pipeline, the variant caller tool

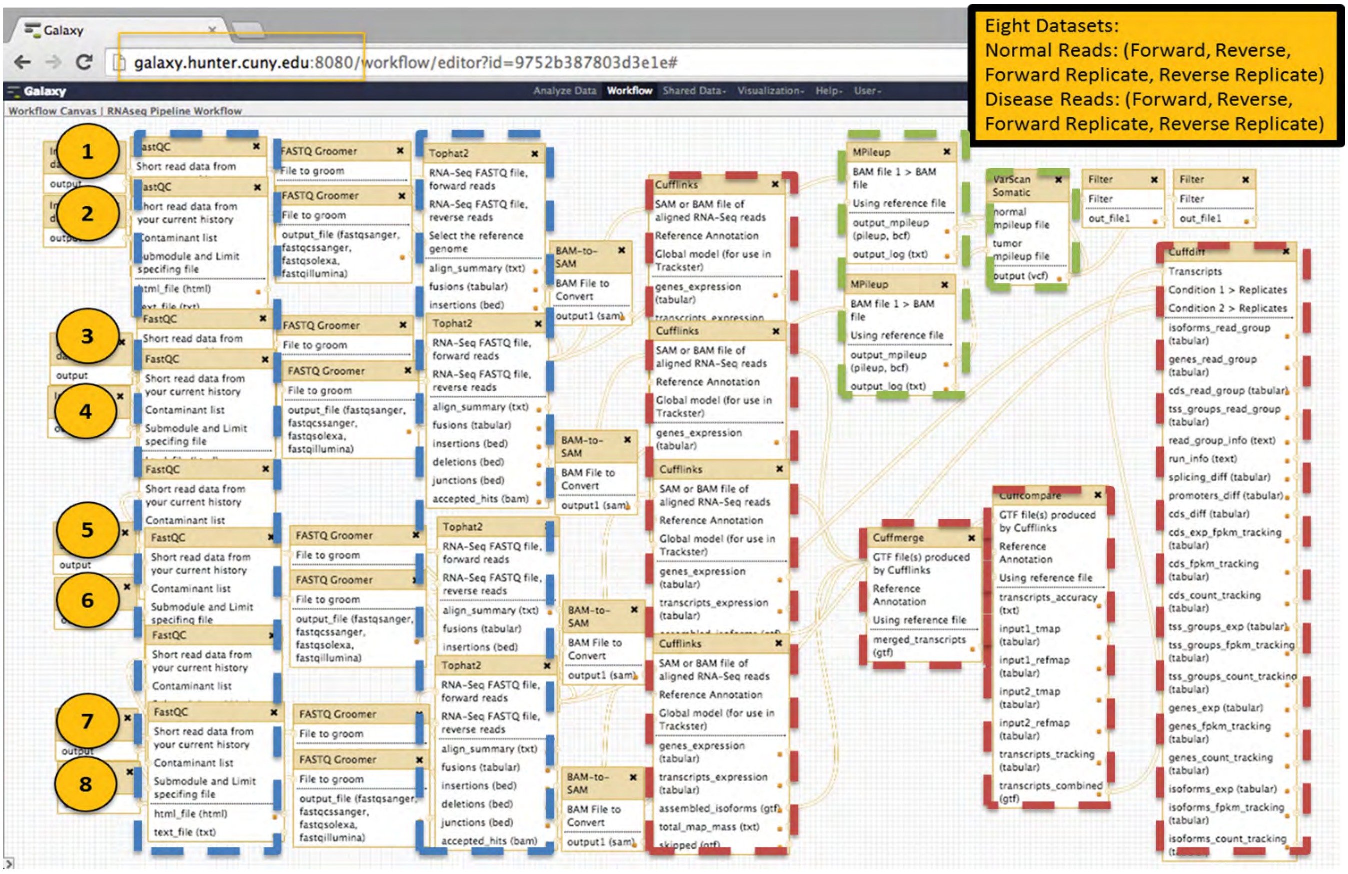
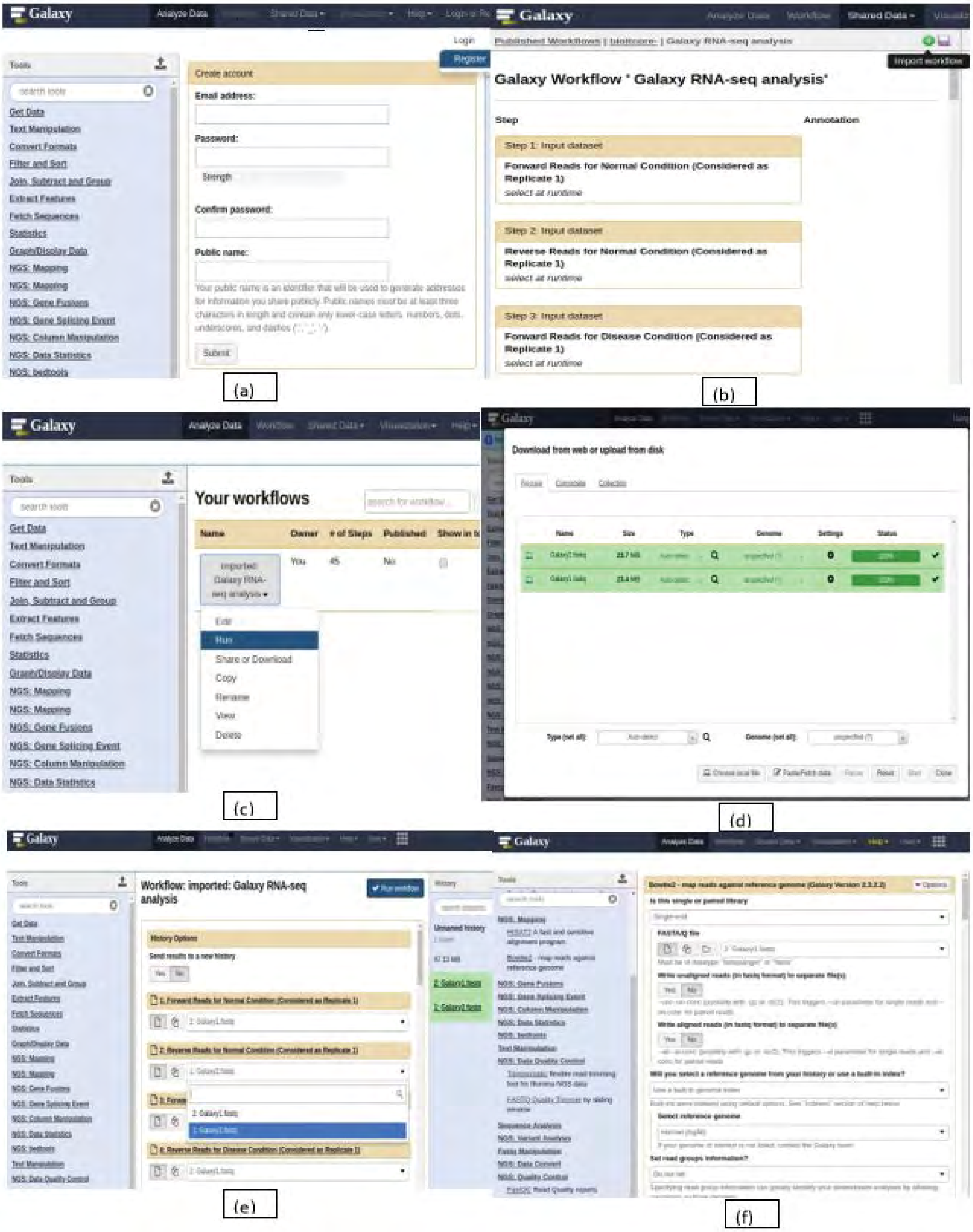

Eight Datasets:
Normal Reads: (Forward, Reverse, Forward Replicate, Reverse Replicate)
Disease Reads: (Forward, Reverse, Forward Replicate, Reverse Replicate)

Figure 1: RNA-seq analysis workflow in Galaxy: A. Quality Check (left blue rectangle), B. Alignment (right blue rectangle), C. Differential Expression Analysis (red rectangles), D. Variant Analysis (green rectangles). The input samples are normal tissue samples (1, forward; 2, reverse reads), prostate cancer samples (3, forward; 4 reverse reads), and a replicate of each of these, respectively (5,6 and 7,8).



Galaxy Workflow ' Galaxy RNA-seq analysis'

Step 1: Input dataset
Forward Reads for Normal Condition (Considered as Replicate 1)
select at runtime

Step 2: Input dataset
Reverse Reads for Normal Condition (Considered as Replicate 1)
select at runtime

Step 3: Input dataset
Forward Reads for Disease Condition (Considered as Replicate 1)
select at runtime

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least three characters in length and contain only lower-case letters, numbers, dots, underscores, and dashes ('.', '_', '-').

Figure 2: Access to tg RNA-seq analysis pipeline through Galaxy by (a). registering for an account; (b). Importing the workflow under a user's Galaxy account; (c). Running the pipeline; (d). Importing data by upload from the local drive or public database under the user's Galaxy account. All the data appear on the right side column; (e). Starting the RNA-seq pipeline by providing a set of basic parameters, and; (f). Users can choose to run a single tool from ones the list available on the left side column, independently of the RNA-seq pipeline, and without having to run the complete pipeline. Links to our local and the public Galaxy server are provided in the main text. the number of differentially expressed genes and isoforms.

VarScan, version 2.3.5 is used to ADD VERB furthermore somatic and germline mutations in the mpileup files, reporting the results in a tabular file [16].

Easy access to the RNA-seq data analysis workflow

Access to the RNA-seq workflow is very simple, and users first have to create a free account (Figure 2a) on our local instance of Galaxy and the public server (web links available on section 2.2 of the manuscript). Users can then login and import and run the workflow and data under their account (Figure 2b-2d), with a few clicks of the mouse on the Galaxy web interface. The imported data are available on the right column of the Galaxy interface, and to run the workflow users simply need to select these datasets which automatically appear on the drop-down menus in the center of the interface (Figure 2e). Furthermore, users can select a single tool from the left column of Galaxy (Figure 2f) and provide the tools parameters in the center interface similarly to the workflow.

Workflow results with RNA-seq data

Using the prostate cancer datasets from ArrayExpress, our RNA-seq bioinformatics workflow identified between 10,912 to 12,482 differentially expressed genes per sample (Table 1). The differential expression was calculated with a significance of 0.05<p-value<0.01 when comparing normal to cancer tissue. Furthermore, when including all isoforms/splice variants for these genes the number of differentially expressed transcripts ranged between 12,449 and 16,121 (Table 1, column 3) per sample, when comparing cancer versus healthy tissue. Finally, our bioinformatics pipelines also discovered up to 485 novel splice variants per sample (Table 2). These results are included in the output of the Galaxy history that reports the differentially expressed transcripts, in addition to the GTF annotation file, allowing the novel splice variants to be used in future bioinformatics analyses.

Discussion

Here we present a bioinformatics pipeline that performs RNA-seq analysis using high-throughput sequencing data. The Galaxy-based RNA-seq pipeline is user-friendly and provides an easily accessible and intuitive interface, which offers flexibility for researchers to edit the steps of the workflow. For example, the widely developed Tuxedo software suite, Bowtie, TopHat, and Cufflinks used for alignment and differential expression analysis in our workflow can be easily substituted by other tools with similar functionality available through the Galaxy ToolShed [5]. Besides customizing tools within the workflow, parameter values for each tool can be also adjusted, allowing high flexibility for the user.

Conclusions

Furthermore, the Galaxy web-based workbench offers a range of built-in functionality, for researchers collaborating on projects involving large-scale genomic sequencing projects. For example, the RNA-seq transcriptome analysis workflow used in the current study, including any modifications to the actual workflow, can be easily shared by publishing the web link on the public Galaxy server. In addition, the output datasets generated by running the workflow can be similarly shared, allowing reproducibility and transparency of the bioinformatics analysis. In conclusion, we have performed a reproducible, accessible, and transparent computational biology experiment for RNA-seq data analysis that is applicable for both research and the clinical practice, providing a novel example towards further standardization of bioinformatics analysis.

Sample ID	No. of genes	No. of isoforms	No. of novel isoforms
2	11,185	14,313	485
3	12,482	16,020	433
4	12,184	15,059	437
5	12,184	15,219	467
6	12,014	15,605	460
7	11,796	14,975	457
8	12,277	14,638	373
9	11,084	13,173	388
10	11,235	12,449	357
11	12,438	16,121	370
12	10,912	14,440	430
13	12,264	15,061	479
14	11,975	13,824	308

Table 1: RNA-seq analysis summary reports: the number of differentially expressed genes and isoforms when counting all splice variants of a gene (second and third column). The fourth column presents the number of novel isoforms and splice variants discovered per sample.

Computer server: 40 CPU, 32 GB memory	
Dataset:	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-567/samples/
Run times	Repeat 1: 18 hrs 25 min
	Repeat 2: 17 hrs 44 min
	Repeat 3: 19 hrs 28 min
Total Data Size:	40 GB

Table 2: Running times of the RNA-seq analysis during three trials with the human transcriptome from the European Bioinformatics Institute (EBI).

Acknowledgements

Funding: Work of the Krampis group was supported by a Research Centers in Minority Institutions Program grant from the National Institute on Minority Health and Health Disparities (8 G12 MD007599) of the National Institutes of Health.

References

1. Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, et al. (2013) Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nature Biotechnology 31: 1023-1031.
2. Rizzo JM, Buck MJ (2012) Key principles and clinical applications of "next-generation" DNA sequencing. Cancer Prevention Research 5: 887-900.
3. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10: 57.
4. Goncalves A, Tikhonov A, Brazma A, Kapushesky M (2011) A pipeline for RNA-seq data processing and quality assessment. Bioinformatics 27: 867-869.
5. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7: 562-578.
6. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11: 86.
7. Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, et al. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research 44: 3-10.

8. Ren S, Peng Z, Mao JH, Yu Y, Yin C, et al. (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Research 22: 806-821.

9. Andrews S (2010) FastQC: A quality control tool for high throughput sequence data.

10. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions,deletions and gene fusions. Genome Biology 14: 36.

11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

12. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28: 511-515.

13. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, et al. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology 31: 46-53.

14. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987-2993.

15. Li H (2011) Improving SNP discovery by base alignment quality. Bioinformatics 27: 1157-1158.

16. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 22: 568-576.