Conference Paper

SMART-RDA: A Galaxy Workflow for RNA-Seq Data Analysis

Redi Aditama, Zulfikar Achmad Tanjung, Widyartini Made Sudania, and Tony Liwang

Plant Production and Biotechnology Division, PT SMART, Tbk., Sinar Mas Land Plaza, 2nd Tower, 10th Fl., Jl. M.H. Thamrin No. 51, Jakarta 10350, Indonesia

Abstract

RNA-seq using the Next Generation Sequencing (NGS) approach is a common technology to analyze large-scale RNA transcript data for gene expression studies. However, an appropriate bioinformatics tool is needed to analyze a large amount of transcriptomes data from RNA-seq experiment. The aim of this study was to construct a system that can be easily applied to analyze RNA-seq data. RNA-seq analysis tool as SMART-RDA was constructed in this study. It is a computational workflow based on Galaxy framework to be used for analyzing RNA-seq raw data into gene expression information. This workflow was adapted from a well-known Tuxedo Protocol for RNA-seq analysis with some modifications. Expression value from each transcriptome was quantitatively stated as Fragments Per Kilobase of exon per Million fragments (FPKM). RNA-seq data of sterile and fertile oil palm (Pisifera) pollens derived from Sequence Read Archive (SRA) NCBI were used to test this workflow in local facility Galaxy server. The results showed that differentially gene expression in pollens might be responsible for sterile and fertile characteristics in palm oil Pisifera.

Keywords: FPKM; Galaxy workflow; Gene expression; RNA sequencing.

Corresponding Author:
Tony Liwang
biotechnology@sinarmas-agri.com

Received: 11 February 2017
Accepted: 08 March 2017
Published: 26 March 2017

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ICBS Conference Committee.

OPEN ACCESS

1. Introduction

The Next-generation sequencing (NGS) has been rapidly developed in recent years, providing much cheaper and higher throughput than the earlier generation Sanger sequencing [1]. This technology allows rapid advance in many fields related to biological sciences, one of them is the RNA sequencing (RNA-seq) experiments [2]. Considered as an alternative to microarrays, RNA-seq is now used for quantitative transcriptomics and identification of novel transcripts. It is a powerful tool with a remarkably diverse range of applications, from detailed studies of biological processes at the cell type-specific level to studies of fundamental questions in many biological systems on an evolutionary timescale [3].

The computational challenges of RNA-seq data analysis are divided into three main categories: (i) read mapping, (ii) transcriptome reconstruction, and (iii) expression quantifications [4]. Read mapping process is divided into two types, unspliced and spliced alignment. Unspliced alignment uses reads and reference transcriptomes as

input. It aligns reads to a known reference transcriptome. Some examples of unspliced aligner programs are Short-read mapping package (SHRiMP) [5], Bowtie [6] and BWA [7]. On the other hand, spliced aligner uses reads and reference genome as input. TopHat is an example of widely used spliced aligners program to process RNA-seq reads [8].

The aim of transcriptome reconstruction is to define a precise map of all transcripts and isoforms that are expressed in particular samples. Cufflinks is one of the most used genome-guided assembly programs for transcriptome reconstruction that identifies novel transcripts using a known genome [9]. Unlike Cufflinks, Velvet [10] and Trans-ABySS [11] identify novel genes and transcript isoforms without a known reference genome. The final step of RNA-seq data analysis is expression quantification. Alexa-seq uses reads and transcript models to quantify gene expressions [12] while cufflinks use aligned reads to quantify transcript isoform levels. There are several differential expression programs to compare expression levels between two or more sets of transcriptomes, including Cuffdiff [13]. The programs use read alignments and transcript models to identify differentially expressed genes or transcript isoforms.

Most of the programs described above have been used to analyze RNA-seq raw data because they are powerful and open for public. However, at least three steps of command writings is needed to obtain transcript level information. This can be problematic for biologists who are not familiar with command writings and it also has a huge amount of parameters and raw data that need to be analyzed. Galaxy platform is available to simplify this kind of work [14]. It uses a web interface to cloud computing resources, providing command-line-driven tools, such as TopHat and Cufflinks, for users without UNIX skills through the web and the computing clouds.

This study aims to construct SMART RNA-seq Data Analyzer (SMART-RDA), a Galaxy workflow that can be easily applied to analyze RNA-seq raw data into differential expression information of genes. A well-known Tuxedo protocol was used as the backbone of this workflow. Few additional tools and modifications were added to turn this workflow into a more comprehensive and user-friendly tool.

2. Material and Method

2.1. Material

This study utilized two files from Sequence Read Archive (SRA) [15] derived from NCBI under ID SRX278051 and SRX278050 to test the workflow performance. These files are the RNA sequence of sterile and fertile oil palm Pisifera pollens. Each sequence contains 228.3 and 295.4×10^6 million bases, respectively. Programs used in this research are Galaxy, TopHat, Cufflinks, Cuffdiff, and CummeRbund, which were downloaded from their official websites. Galaxy tools were downloaded from Galaxy ToolShed (**https://toolshed.g2.bx.psu.edu/**).

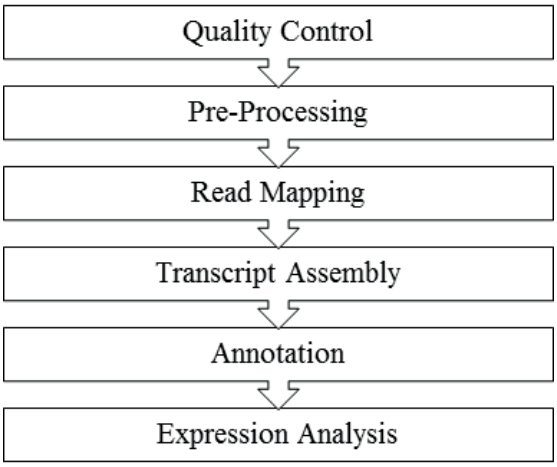

Figure 1: Workflow Steps.

2.2. Methods

This workflow was constructed based on a well-known Tuxedo protocol. The Quality Control stage was added before performing tuxedo protocol to improve the quality of reads. Detailed steps of the workflow are shown in Figure 1. Blast annotation step was added after Cuffdiff to annotate the assembled transcripts. Common tools in this workflow, like TopHat, Cufflinks, and Cuffmerge were available on the galaxy toolshed. Several tools that were not available in the galaxy toolshed were constructed using XML programming.

3. Results and Discussions

3.1. SMART-RDA Workflow

A Galaxy workflow for RNA-seq data analysis was constructed (see Figure 2) from a modified Tuxedo protocol. The first tool of this workflow is FastQ Groomer. It converted several input Fastq quality score types, like Illumina 1.8 and 454, into standard Sanger format [16]. Groomed fastq file was trimmed using FastQ Quality Trimmer. This tool trimmed the end of reads based on the aggregate value of quality scores found within a sliding window. FastQ Summary Statistics were used to create summary statistics on a fastq file before and after trimming by FastQ Quality Trimmer.

After groomed and trimmed, the reads were assembled using TopHat. Fasta file of reference genome was needed for this stage. TopHat aligned RNA-seq reads into genome-sized sequence using the ultra-high-throughput short read aligner Bowtie, and then analyzed the mapping results to identify splicing junctions between exons. The output of TopHat was BAM file called accepted hits. BAM to SAM converter was needed to convert TopHat output into SAM formatted file [17]. The conversion process of BAM to SAM allowed users to review TopHat output directly.

The aligned reads from TopHat output were assembled by Cufflinks. This program assembled reads into transcripts, estimated their abundances, and tested them for

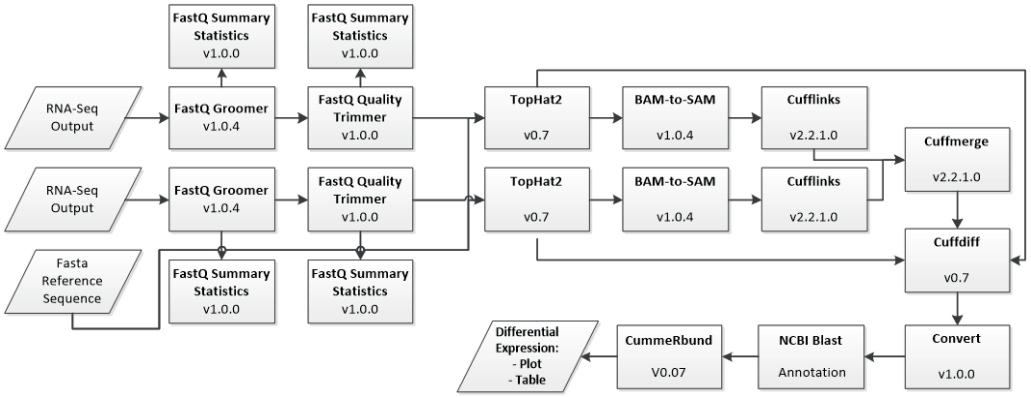

Figure 2: SMART-RDA Workflow.

differential expressions and regulations. This process produced GTF files that contain Cufflinks assembled isoforms. These files contained information of assembled reads, transcripts, and their abundances. The workflow produced two Cufflinks output files at one time because it was designed to process two different samples simultaneously. However, Cuffmerge was still needed to merge two GTF files produced by Cufflinks.

The process of finding significant changes in transcript expression was performed by Cuffdiff. In this stage, aligned reads produced by TopHat were quantified using assembled transcript produced by Cufflinks as a reference. The outputs of this step were tabular files that contained information about transcripts location, expressions, and differential analysis. The expression level of each transcript was stated in term of Fragments Per Kilobase of exon per Million fragments (FPKM). In these units, the relative abundances of transcripts were described in terms of expected biological object (fragments) observed from RNA-seq experiment.

The next stage in this workflow was to annotate each transcript previously analyzed by Cuffdiff. The annotation process used NCBI BLAST+blastn [18, 19] to search annotation from known databases. Information about locus position of transcripts from Cuffdiff outputs was used to extract transcript sequences from a reference genome. Blast results produced in this process were then combined as an annotation into Cuffdiff output. The final stage of the workflow was visualization of transcript quantification and differential analysis using cummeRbund. This tool produced several visualizations, including boxplot, scatters, distributions, and volcano plots.

3.2. Workflow performance test

SMART-RDA workflow was tested using two files from NCBI SRA of oil palm Pisifera pollen RNA-seq under ID SRX278051 (fertile) and SRX278050 (sterile). The fertile reads size was 472.6 Mb while the sterile reads was 613.2 Mb. After trimming, the size of fertile reads decreased to 295.2 Mb and the sterile reads to 486.6 Mb. This stage removed 176 low-quality sequences from fertile and 112 from sterile reads.

Tuxedo protocol has been successfully implemented by this workflow. The CummeRbund stage produced four type visualizations that represented the differential

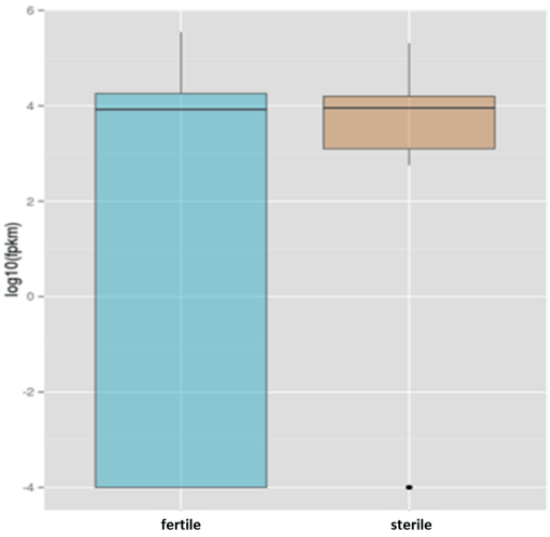

Figure 3: Boxplot of differential gene expressions of fertile and sterile oil palm Pisifera pollens.

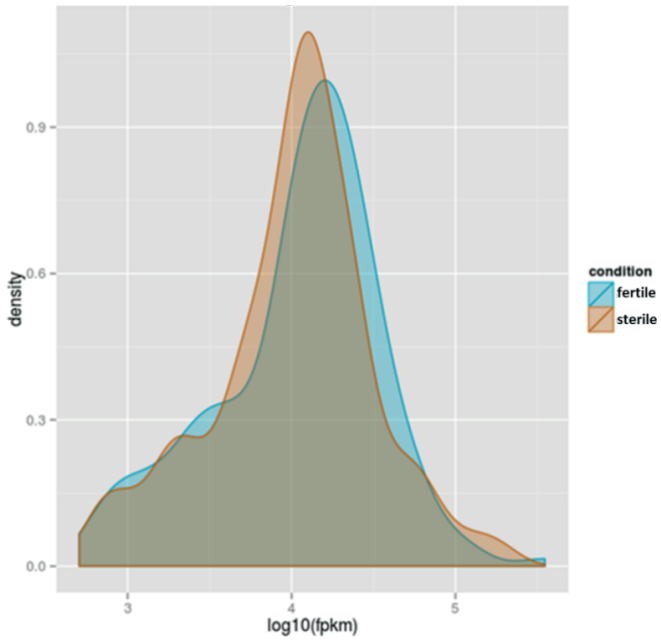

Figure 4: Density plot of differential gene expressions of fertile and sterile oil palm Pisifera pollens.

expression between two conditions. The first visualization was a boxplot of differential gene expressions (see Figure 3). Based on this plot, the average value of FPKM between fertile and sterile pollen was almost the same. In terms of range, fertile pollen had wider ranges of gene expression values.

The other type of visualization was density plot, as shown in Figure 4. This plot showed the differential gene expression patterns between fertile and sterile Pisifera pollens. The gene expression of sterile pollens was distributed in lower fpkm values, but with higher density than fertile pollens.

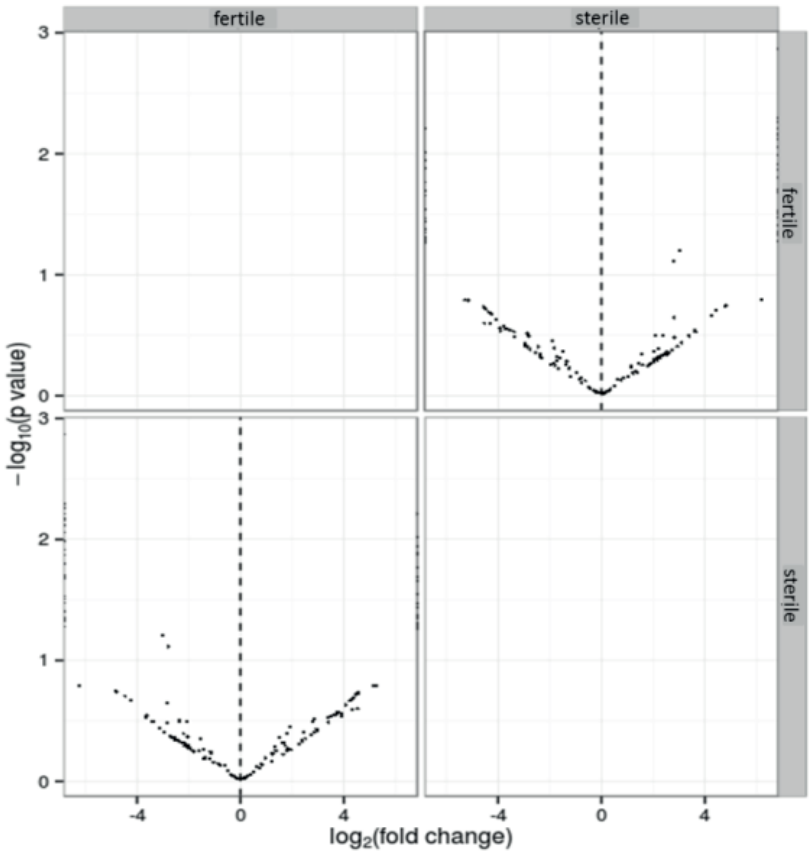

Figure 5: Volcano plot describes differential gene expression of fertile and sterile oil palm Pisifera pollens.

While boxplot and density plot described expression distributions, volcano plots (see Figure 5) described significant level of each transcript. Volcano plot showed several transcripts with higher expression in fertile than in sterile condition.

Finally, the most important result of this workflow was the list of genes with large difference in gene expressions (see Table 1.). There were ten genes with high significant expression levels. The level significance of differential expression was determined by P value. Furthermore, this data would be useful in determining the genes that played a role in palm pollens fertility.

4. Conclusions

SMART-RDA, a galaxy workflow that can be used to analyze RNA-seq data into differential expression information, was constructed based on modified Tuxedo protocol. Performance test using SRA data of fertile and sterile oil palm Pisifera pollens detected ten genes with high significant expression levels. Three visualizations and one table describing expression condition of samples were produced by this workflow, which was capable of performing differential analysis of large RNA-seq data.

Locus	Genes	FPKM		p value
		Fertile	Sterile	
EG5_Chr8:33574178-33580606	ubiquitin-conjugating enzyme [*Elaeis guineensis*]	52 905	0	0.02210
EG5_Chr5:26736162-26736842	beta-1,3-glucanase [*Elaeis guineensis*]	38 158	0	0.02395
EG5_Chr14:2597123-2598066	elongation factor 1-alpha, putative [*Ricinus communis*]	22 501	0	0.02570
EG5_Chr3:1054681-1056044	uncharacterized protein LOC100812783 [*Glycine max*]	0	13 788.6	0.02780
p5_sc00292:963590-964056	putative elicitor inducible beta-1,3-glucanase NtEIG-E76 [*Oryza sativa Japonica Group*]	0	13 672.6	0.02780
EG5_Chr3:28391917-28399716	ribosomal protein [*Elaeis guineensis*]	0	12 611.8	0.02780
EG5_Chr13:23110722-23111204	DUF246 domain-containing protein [*Medicago truncatula*]	0	22 311.8	0.02820
EG5_Chr13:27541179- 27541819	PREDICTED: elongation factor 1-alpha [*Vitis vinifera*]	0	12 161.8	0.02820
EG5_Chr1:3675588-3675755	uncharacterized protein LOC100276758 [*Zea mays*]	1844460	0	0.03080
EG5_Chr16:1554435-1555188	hypothetical protein OsI_32485 [*Oryza sativa Indica Group*]	27908	0	0.03080

Table 1: List of genes with high significance level of expressions.

Acknowledgements

The authors would like to thank the management of PT. SMART Tbk. for supporting this research and publication.

References

[1] A. Grada and K. Weinbrecht, "Next-generation sequencing: Methodology and application," *Journal of Investigative Dermatology*, vol. 133, no. 8, p. e11, 2013.

[2] V. Thakur and R. Varshney, "Challenges and strategies for next generation sequencing (NGS) data analysis," *J Comput Sci Syst Biol*, vol. 03, no. 02, pp. 040–042, 2010.

[3] L. B. B. Martin, Z. Fei, J. J. Giovannoni, and J. K. C. Rose, "Catalyzing plant science research with RNA-seq," *Frontiers in Plant Science*, vol. 4, article no. 66, 2013.

[4] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, no. 6, pp. 469–477, 2011.

[5] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno, "SHRiMP: Accurate mapping of short color-space reads," *PLoS Computational Biology*, vol. 5, no. 5, Article ID e1000386, 2009.

[6] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article no. R25, 2009.

[7] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[8] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: Discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[9] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nature Biotechnology*, vol. 31, no. 1, pp. 46–53, 2013.

[10] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney, "Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels," *Bioinformatics*, vol. 28, no. 8, Article ID bts094, pp. 1086–1092, 2012.

[11] G. Robertson, J. Schein, R. Chiu et al., "De novo assembly and analysis of RNA-seq data," *Nature Methods*, vol. 7, no. 11, pp. 909–912, 2010.

[12] M. Griffith, O. L. Griffith, J. Mwenifumbo et al., "Alternative expression analysis by RNA sequencing," *Nature Methods*, vol. 7, no. 10, pp. 843–847, 2010.

[13] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[14] B. Giardine, C. Riemer, R. C. Hardison et al., "Galaxy: A platform for interactive large-scale genome analysis," *Genome Research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[15] Y. Kodama, M. Shumway, and R. Leinonen, "The sequence read archive: Explosive growth of sequencing data," *Nucleic Acids Research*, vol. 40, no. 1, pp. D54–D56, 2012.

[16] D. Blankenberg, A. Gordon, G. Von Kuster et al., "Manipulation of FASTQ data with galaxy," *Bioinformatics*, vol. 26, no. 14, Article ID btq281, pp. 1783–1785, 2010.

[17] H. Li, B. Handsaker, A. Wysoker et al., "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[18] P. J. A. Cock, B. A. Grüning, K. Paszkiewicz, and L. Pritchard, "Galaxy tools and workflows for sequence analysis with applications in molecular plant pathology," *PeerJ*, vol. 2013, no. 1, article no. e167, 2013.

[19] C. Camacho, G. Coulouris, V. Avagyan et al., "BLAST+: Architecture and applications," *BMC Bioinformatics*, vol. 10, article no. 421, 2009.