ICBAA2017-33

DUAL RNA-SEQ DATA ANALYSIS PIPELINE IN PLANT HOST-PATHOGEN STUDY

Suhaila Sulaiman, Nur Syamimi Yusoff and Lee Yang Ping

Felda Global Ventures Research and Development Sdn. Bhd., FGV Innovation Centre (Biotechnology), PT. 23417 Lengkuk Teknologi, 71760 Bandar Enstek, Negeri Sembilan, Malaysia
suhaila.s@feldaglobal.com

Abstract: Over the recent years, a wide-ranging application of transcriptome analysis has been implemented in deciphering host-pathogen interaction in plants to understand the fundamental process in plant disease incidence. The emerging sequencing technology enables researchers to perform dual RNA-seq, which is an application process to capture the gene expression changes in both host and pathogen simultaneously. In parallel to the advancement of cutting edge sequencing technology, the data analysis is vital to understand the generated data from biological sight. In this study, we implement a computational pipeline of dual RNA-seq data analysis to explore the potential underlying in the dual RNA-seq data in identifying the candidates of genes that might contribute in plant host-pathogen interaction. The pipeline is composed of a series of bioinformatics approaches which delineated into five phases. The pipeline starts with data pre-processing that involves RNA-seq data quality assessment and filtration to obtain high quality data for consequence steps. It is next followed by transcriptome analysis whereby the transcripts are separated between host and pathogen, mapped and assembled, and followed by differential expression profiling. In the third phase, the assembled transcripts are functionally annotated and encoded proteins are predicted. The predicted proteins are then entitled to a sub-pipeline of secretome identification to predict the secreted protein candidates in the pathogen. In the final phase, the data analysis pipeline is finalised by homology-based analysis that is partaking to spot the candidates of genes that possibly involved in host-pathogen interaction. Through execution of this sequence-based data analysis pipeline, we anticipate the prospective influence of structure-level data analysis in the study of host-pathogen interaction in plant.

Keywords: dual RNA-seq, host-pathogen

INTRODUCTION

The interaction of host-pathogen is complex due to the fact that obtaining a comprehensive understanding of the interactions is very challenging. For instance, to understand a disease caused by pathogen in a plant, a thorough knowledge is required to identify the expressed genes in different conditions and their expression might imitate to how they respond during an infection. It is common to apply RNA sequencing (RNA-seq) application onto a single species to observe the gene expression in a single species of interest (Westermann, Gorski, & Vogel, 2012). However, in host-pathogen interaction study, a more sophisticated approach is required because two interacting organisms are being studied at the same time. Therefore, the independent transcriptome approach might not sufficient to illustrate a clear indication of expression in both organisms at the same time. Fortunately, the next generation sequencing technology is now enable researchers to perform dual RNA-seq, which is a refined approach of RNA-seq that allow the capture of both transcriptomes in two interacting species without physically

separating cells or RNA (Westermann et al., 2012; Wolf et al., 2017). Thus, the associated gene expression changes in both host and pathogen are now can be profiled simultaneously during the invasion of plant pathogen into the host. The revolutionised application of dual RNA-seq is applied to map the transcriptional networks that mediate host-pathogen interactions. The advanced approach of dual RNA-seq also reflects to the need of more comprehensive data analysis to turn out the raw data into beneficial information. *In silico* analyses are essential in the data analysis to distinguish the species-specific transcripts, as well as in exploring the profile of genes expression in both species. In this study, a pipeline to analyse the resultant data from dual RNA-seq data in plant host-pathogen study is demonstrated, as opposed to independent analysis step taken in most research. This pipeline describes a series of transcriptome data mining to sequence analysis of potential candidate of genes that might relevant in host-pathogen interaction.

METHODS

Raw data pre-processing

The RNA-seq raw data generated by sequencing machine is evaluated prior to other analyses using FASTQC program (Andrews, 2010). In this phase, reads that are not meet the quality criteria are removed from the dataset using customised PERL scripts. This includes reads with base quality less than Q_{phred} 20 and less than 50 bp in length. Besides, vector and adaptor sequences are filtered out to avoid foreign reads from the dataset. Reads that remain in the dataset are called high quality reads and subjected to further analysis.

Transcriptome analysis

Reads from the host and pathogen are distinguished using BLASTN program (Altschul, Gish, Miller, Myers, & Lipman, 1990) against available database of both organisms. After separation, the reads are assembled using one of two available approaches: reference or *de novo* assembly. When reference genome is available, reference assembly is performed by mapping reads to the genome using STAR mapping tool (Dobin et al., 2013) and assembly *via* Cufflinks tool (Trapnell et al., 2010a). Otherwise, in the absent of reference genome, a *de novo* assembly is done by Trinity tool (Grabherr et al., 2011). In each case of host and pathogen, the assembled transcripts are joined using Cuffmerge program (Trapnell et al., 2010b) to produce a reference annotation. This reference annotation is applied in the expression profiling analysis using differential expression analysis tools such as Cuffdiff (Trapnell et al., 2013) and edgeR (Robinson et al., 2010).

Annotation: genome and transcriptome

The assembled transcripts of both datasets are annotated using a combination of programs. Genes encoded by the transcripts are predicted using CodingQuarry (Testa et al., 2015). Predicted genes are annotated via a series of functional annotation analysis in BLAST2GO program (Conesa et al., 2005). On another note, transcripts are annotated using TRAPID (Van Bel et al., 2013), an online tool for *de novo* assembled transcripts data.

Secretome mining

The presence of signal peptide sequences are predicted using SignalP 4.1 server (Nielsen, 2017) and sigcleave program in EMBOSS package (Rice et al., 2000) reports on signal cleavage sites in a protein sequence. Transmembrane helices region in proteins are screened using TMHMM server (Krogh et al.,

2001). The subcellular location of eukaryotic protein is predicted using TargetP 1.1 server (Emanuelsson et al, 2000).

Homology-based analysis

Target of known genes of interest are retrieved using UniProtKB/Swiss-Prot database (Consortium, 2017). The genes are searched against predicted proteins of host/pathogen that are pre-formatted into a BLAST database. With the cut-off E-value of minimum $1e^{-5}$ and at least 30% identity level, a match that covers at least 50% of the query sequence is considered significant and subjected for further analysis.

RESULTS AND DISCUSSION

The data analysis pipeline is delineated into five phases approach: raw data pre-processing, transcriptome analysis, genome or transcriptome annotation, secretome mining and homology-based analysis (Figure 1). The pipeline is powered by open source tools and public databases that well-known as data repository for biological data analysis.

Dual RNA-seq contains a mixed of transcriptomes from both species (plant and pathogen). Prior to sequencing, one should emphasize on the sequencing depth that is required to ensure a whole representation can covers the most of the transcriptome in both species (Westermann et al., 2012). The sequenced raw reads are pre-processed in the first step to eliminate bad quality reads and foreign sequences. This is important to ensure that the consequence analysis will only consist of high quality reads to avoid noise or false positive in the analysis (Kukurba & Montgomery, 2015; Mazzoni & Kadarmideen, 2016).

The high quality reads that passed quality assessment contains the mixture of reads from both host and pathogen. The transcripts are segregated using mapping approach via homology search. At this point, available databases of both organisms are utilised to support in the separation process. This may involve genome and transcriptome databases which can be obtained from public databases or in-house research data. Remarkably, a limitation of reference database that been used might lead to high number of unmapped reads. However, this limitation could be replenished by performing a de novo assembly of those unmapped reads to identify the origin of the reads, either from host or pathogen.

Otherwise, mapping and assembly process is performed using reference-based approach when there is reference genome available (Dobin et al., 2013; Trapnell et al., 2010). Based on the assembled data, genes that expressed in different samples will be determined via differential expression using tools, such as Cuffdiff (Trapnell et al., 2013) and edgeR (Robinson et al., 2010). Up-regulated and down-regulated genes in the plant and pathogen will be obtained with corresponding expression level (in FPKM value and fold change), that is useful in spotting genes that are expected to be over- or under-expressed during certain conditions in respective organisms.

In this pipeline, the gene candidates are annotated using a series of annotation process in BLAST2GO (Conesa et al., 2005). The annotation acts as early indicator of the genes function which might describe in the gene ontology (GO) terms. For instance, genes that involve in host-pathogen interaction may associated to GO terms of GO:00044419 (involved in defense response), GO:0043657 (cellular component: host cell) and GO:0009405 (pathogenesis)

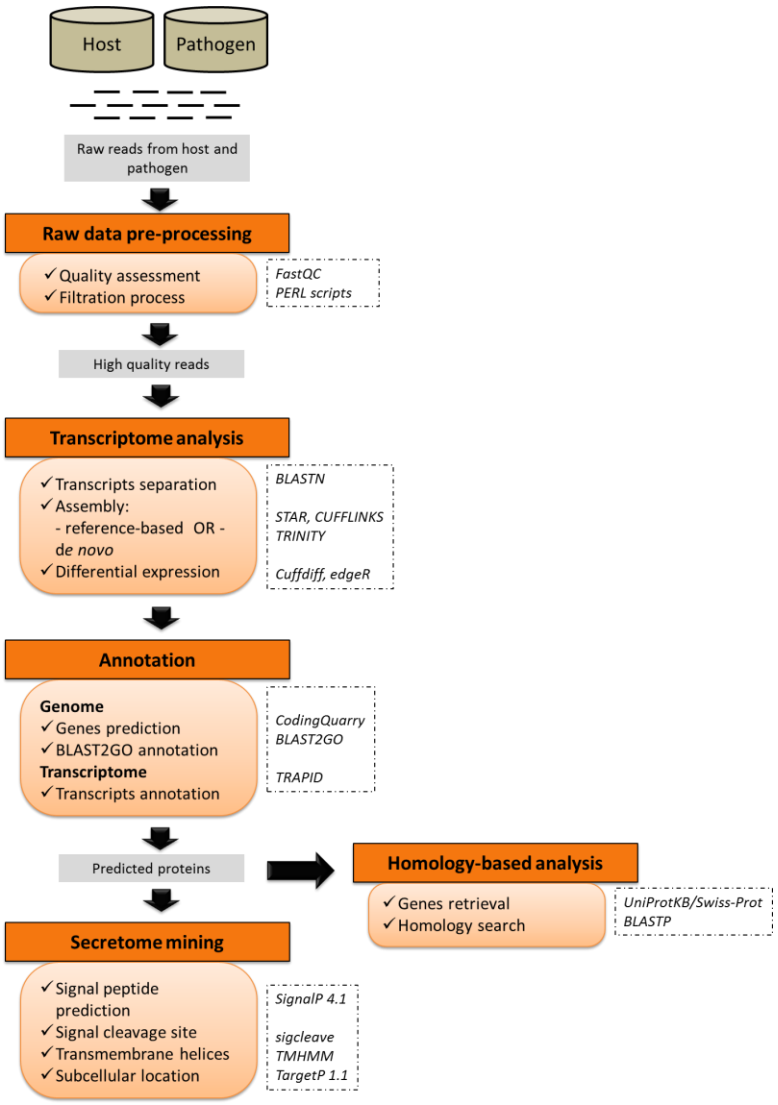

Figure 1. Pipeline for analysis of dual RNA-seq data in plant host and pathogen study. The pipeline is described into five phases (orange boxes), input/output files are shown in grey-coloured box. Implemented tools are listed in the dotted-border boxes.

CONCLUSIONS

Conclusions should include (1) the principles and generalisations inferred from the results, (2) any exceptions to, or problems with these principles and generalisations, (3) theoretical and/or practical implications of the work, and (5) conclusions drawn and recommendations.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. http://doi.org/10.1016/S0022-2836(05)80360-2.

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. *Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. Retrieved from citeulike-article-id:11583827.

Conesa, A., Gotz, S., Garciaa-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. http://doi.org/10.1093/bioinformatics/bti610.

Consortium, T. U. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. Retrieved from http://dx.doi.org/10.1093/nar/gkw1099.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. http://doi.org/10.1093/bioinformatics/bts635.

Emanuelsson, O., Nielsen, H., Brunak, S., & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology*, *300*(4), 1005–1016. http://doi.org/10.1006/jmbi.2000.3903.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–52. http://doi.org/10.1038/nbt.1883.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. http://doi.org/10.1006/jmbi.2000.4315;

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969. http://doi.org/10.1101/pdb.top084970.

Mazzoni, G., & Kadarmideen, H. N. (2016). Computational Methods for Quality Check, Preprocessing and Normalization of RNA-Seq Data for Systems Biology and Analysis BT - Systems Biology in Animal Production and Health, Vol. 2. In H. N. Kadarmideen (Ed.), (pp. 61–77). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-43332-5_3.

Nielsen, H. (2017). Predicting Secretory Proteins with SignalP BT - Protein Function Prediction: Methods and Protocols. In D. Kihara (Ed.), (pp. 59–73). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4939-7015-5_6.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics : TIG*, *16*(6), 276–277.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. http://doi.org/10.1093/bioinformatics/btp616.

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, *16*(1), 170. http://doi.org/10.1186/s12864-015-1344-4.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotech*, *31*(1), 46–53. Retrieved from http://dx.doi.org/10.1038/nbt.2450.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, *28*(5), 511–515. http://doi.org/10.1038/nbt.1621.

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., & Vandepoele, K. (2013). TRAPID: an efficient online tool for the functional and comparative analysis of de novoRNA-Seq transcriptomes.

Genome Biology, *14*(12), R134. http://doi.org/10.1186/gb-2013-14-12-r134.

Westermann, A. J., Gorski, S. A., & Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nature Reviews Microbiology*, *10*, 618.

Wolf, T., Kämmer, P., Brunke, S., & Linde, J. (2017). Two's company: studying interspecies relationships with dual RNA-seq. *Current Opinion in Microbiology*, *42*(Supplement C), 7–12. http://doi.org/https://doi.org/10.1016/j.mib.2017.09.001.