American Journal of Bioinformatics Research 2015, 5(2): 21-25
DOI: 10.5923/j.bioinformatics.20150502.01

In Silico Analysis of Transcriptomes of *Catharanthus roseus* and *Rauvolfia serpentina*, Two Potent Medicinal Plants Using a Pipeline Developed from Publicly Available Tools

Lakshmi Priya P. M.[1], K. K. Sabu[2,*]

[1]Department of Bioinformatics, Union Christian College, Alwaye, India
[2]Division of Biotechnology and Bioinformatics, Jawaharlal Nehru Tropical Botanical Garden and Research Institute, Palode, India

Abstract Until recently, there was no data available on the genome sequences of medicinal plants. But now, public databases for the transcriptomes of important medicinal plants are available. But an analysis pipeline effectively combining publicly available tools is lacking which would otherwise enable in-depth analysis of the transcriptomes. In this context, we have developed an effective *in silico* analysis pipeline using tools such as FastQC, FastQ Groomer, TopHat, Cufflinks, Cuffmerge and Cuffdiff available in Galaxy platform and other tools such as DAVID, ExPASy, MetaCyc and PlantCyc. We have tested the pipeline for comparative analysis of the transcriptome of *Catharanthus roseus* (L.) G.Don and *Rauvolfia serpentina* (L.) Benth. ex Kurz, two well-known medicinal plants. This study identified genes that are similarly expressing in the roots of these plants leading to the formation of the same secondary metabolite, "Strictosidine" and also identified differentially expressing genes in the leaves of *C. roseus* and *R. serpentina* lead to the formation of different metabolites, "Vinblastine" and "Ajmaline". The findings of the study indicated that the pipeline developed is effective and helped to analyze the transcriptomes and expression data.

Keywords Transcriptome, *in silico*, Pipeline, *Catharanthus roseus*, *Rauvolfia serpentine*, Medicinal plants, NGS, Next generation sequencing

1. Introduction

Recent advances in bioinformatics has transformed all areas of biological science. The central dogma of molecular biology describes how information in genes flows into proteins through two-step process, *viz*. transcription and translation. Gene expression can be regulated at several steps including the transcription, RNA splicing, translation, and post-translational modification of a protein [1].

Out of various molecules produced through gene expression, secondary metabolites are unique in the sense that they offer diverse utilities such as drugs, flavor and fragrances, dye and pigments, pesticides, and food additives. Of the various types of these metabolites, alkaloids consist of an important group of low molecular weight nitrogen-containing organic compounds, usually with a heterocyclic structure. They are of particular interest because of their numerous biological activities including medicinal properties [2] and also proved as having important ecological functions [3]. Several studies on alkaloid-producing plants suggest that the biosynthesis and accumulation of these compounds are highly regulated process [4].

Among the alkaloid containing plants, those having monoterpenoid indole alkaloids are rich source of many pharmaceutical drugs. This class of compounds include alkaloids such as the antineoplastics vinblastine and vincristine, the antihypertensives ajmalicine and ajmaline, the antimalarial quinine. The first specific enzyme for monoterpenoid indole alkaloid biosynthesis identified was strictosidine synthase (STR) [2] in *Rauvolfia serpentina*, which condenses tryptamine and secologanin to form the first intermediate 3α (S)-strictosidine [5]. Before the characterization of the strictosidine synthase cDNA of STR, activity of the enzyme was known in different *Catharanthus* and *Rauvolfia* species, both from the Apocynaceae family [6].

Powerful bioinformatics tools are required to extract knowledge from vital amounts of information generated by high throughput genomics technologies. Studies using transcriptomes and expression profiling provide opportunities for better understanding of the plant metabolic

* Corresponding author:
sabu@jntbgri.res.in (K. K. Sabu)
Published online at http://journal.sapub.org/bioinformatics

pathways and enables analyses of the formation of plant-derived pharmaceuticals in various plant species including *R. serpentina* and *C. roseus* [7-12].

Aim of the study was to develop a suitable in silico analysis pipeline for transcriptome studies in medicinal plants and validate the same for analysis of gene expression in *Rauvolfia serpentina* and *Catharanthus roseus* using raw next generation sequence (NGS) data obtained from public databases.

2. Materials and Methods

2.1. RNA-Seq Data

RNA-Seq datasets used in this study were publicly available as single-end Illumina reads in FastQ format and retrieved from ENA Nucleotide database through Galaxy public server (usegalaxy.org). The *Catharanthus roseus* data SRR122254 (root) and SRR122251 (leaf) were 1.1Gbp and 886.4Mbp and the *Rauvolfia serpentina* data SRR125767 (root) and SRR125761 (leaf) were 989.3Mbp and 858.2Mbp respectively. Assembled transcripts of *C. roseus* and *R. serpentina* were downloaded from Medicinal Plant Genomics Resource (medicinalplantgenomics.msu.edu).

2.2. Preprocessing of RNA-Seq Dataset

C. roseus and *R. serpentina* sequence data were preprocessed before used for mapping the reads. Integrity of the dataset was verified before starting the analysis. The quality of each dataset was analyzed using FastQC tool under the NGS:QC in Galaxy. Special attention was given for quality checks with respect to per base sequence quality and per base sequence content. Low quality sequence reads were discarded to ensure that more than 70% of the retained reads had quality greater than 30. Then FastQ Groomer was used to convert the FastQ files to standard format using Phred+33 (Sanger) quality score which were used by the Galaxy for downstream processing.

2.3. Read Mapping

Reads were cleaned and mapped to the reference genome of *C. roseus* and *R. serpentina* (downloaded from plantbiology.msu.edu), which were recently sequenced, assembled and annotated [7]. The reads were mapped for each sample using the aligner TopHat under the NGS:RNA Analysis menu by selecting corresponding FastQGroomer FastQ file as RNA-Seq FastQ files. A genome was selected from the Galaxy History and the corresponding reference genome and other parameters were set as default. Using high-throughput short aligner Bowtie, Tophat aligned RNA-Seq reads to reference genome. Subsequently, the mapping results were analyzed for the identification of splice junctions between exons. RNA-Seq read alignments, among many other things, could reveal new alternative splicing events and isoforms.

2.4. Transcriptome Assembly and Expression Abundance

Quantification of gene expression from the RNA-Seq requires precise identification of isoform of each read. Cufflinks was used to assemble individual transcripts from RNA-Seq reads that have been aligned to the genome. A predicted transcriptome for each sample was created using Cufflinks under the NGS:RNA Analysis menu. A text file of BAM alignments was given to Cufflinks as input which was produced by the RNA-Seq read mapper Tophat. Bias correction (from History) was turned ON and selected the Minimum Intron length and Pre mRNA fraction as 0.05. Then the Cufflinks constructed parsimonious set of transcripts that explained the reads observed in a RNA-Seq experiment. After Assembly phase, Cufflinks quantified the gene expression level of each transfrag in the sample and also estimated transcript abundances by using a reference annotation. Gene expression levels were normalized using fragments per kilobase of exon per million mapped reads.

Cuffmerge is considered to be a 'Meta-Assembler'- as it treats the assembled transfrags the way Cufflinks treats reads, merging them together parsimoniously. The predicted transcriptomes for all samples were merged using Cuffmerge under the NGS:RNA Analysis menu. During the merging, transcripts from all the assemblies were converted to representative reads in BAM format. Cuffmerge merged transcripts that were overlapping and shared a similar exon structure (or splicing structure) to generate a longer chain of connected exons. To merge reference transcripts with sample transfrags, Cuffmerge performed reference annotation-based transcript (RABT) assembly and produced a single annotation file for use in downstream differential analysis. Once, each reads were assembled and merged, the final assembly was screened for genes and transcripts that were differentially expressed.

2.5. Differential Expression

In this study, genes were considered differentially expressed when their absolute value of $\log_2$ fold change was greater than 2 and their p value was less than 0.01. Cuffdiff calculates expression in two or more samples and examines the statistical significance of observed change in expression among them. The statistical model examined the changes that the number of reads produced by each transcript is proportional to its abundance. Cuffdiff compared BAM files generated by Tophat for each sample under the NGS:RNA Analysis menu. Gene and transcript expression level changes were reported as tabular output files containing statistics such as fold change (in $\log_2$ scale), p values (both raw and corrected for multiple testing) and gene- and transcript-related attributes such as common name and location in the genome. Cuffdiff reported additional differential analysis results which were used to identify differentially spliced or regulated genes *via* promoter switching.

2.6. Transcriptome Annotation

Functional annotation of plant transcriptomes is a difficult task due to limited availability of reference genomes in public databases. Database for Annotation, Visualization, and Integrated Discovery (DAVID; david.abcc.ncifcrf.gov) is one of the most versatile tools for functional annotation of large gene sets. DAVID provides comprehensive set of functional annotation tools to understand biological meaning behind large list of genes. Besides DAVID, tools such as ENZYME databases in ExPASy (expasy.org) which is a repository of information relative to the nomenclature of enzymes, MetaCyc (metacyc.org) as a curated database of experimentally elucidated metabolic pathways from all domains of life containing 2260 pathways from 2600 different organisms and PlantCyc (plantcyc.org) which is part of the Plant Metabolic Network (PMN) providing a broad network of plant metabolic pathway databases that contains curated information from the literature and computational analyses about the genes, enzymes, compounds, reactions and pathways involved in primary and secondary metabolism in plants were also used for the functional analysis of the *C. roseus* and *R. serpentina* transcriptomes.

3. Result and Discussion

The main objective of the study was to develop a pipeline for analysis of the transcriptome data of medicinal plants. Modern high throughput sequencers generate tens of millions of sequence reads in a single run. Before analyzing these sequences to draw biological conclusions one should always perform some simple quality control checks to ensure that the raw data quality problems which may otherwise affect the analysis during the later steps.

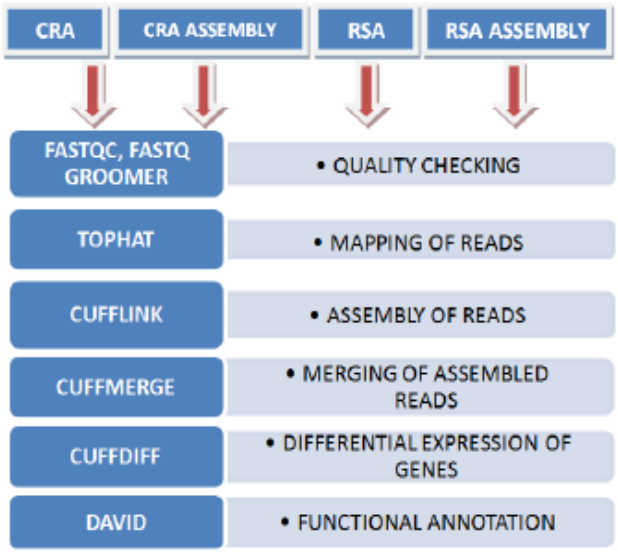

Figure 1. Data analysis pipeline for transcriptome analysis in *Catharanthus roseus and Rauvolfia serpentina*

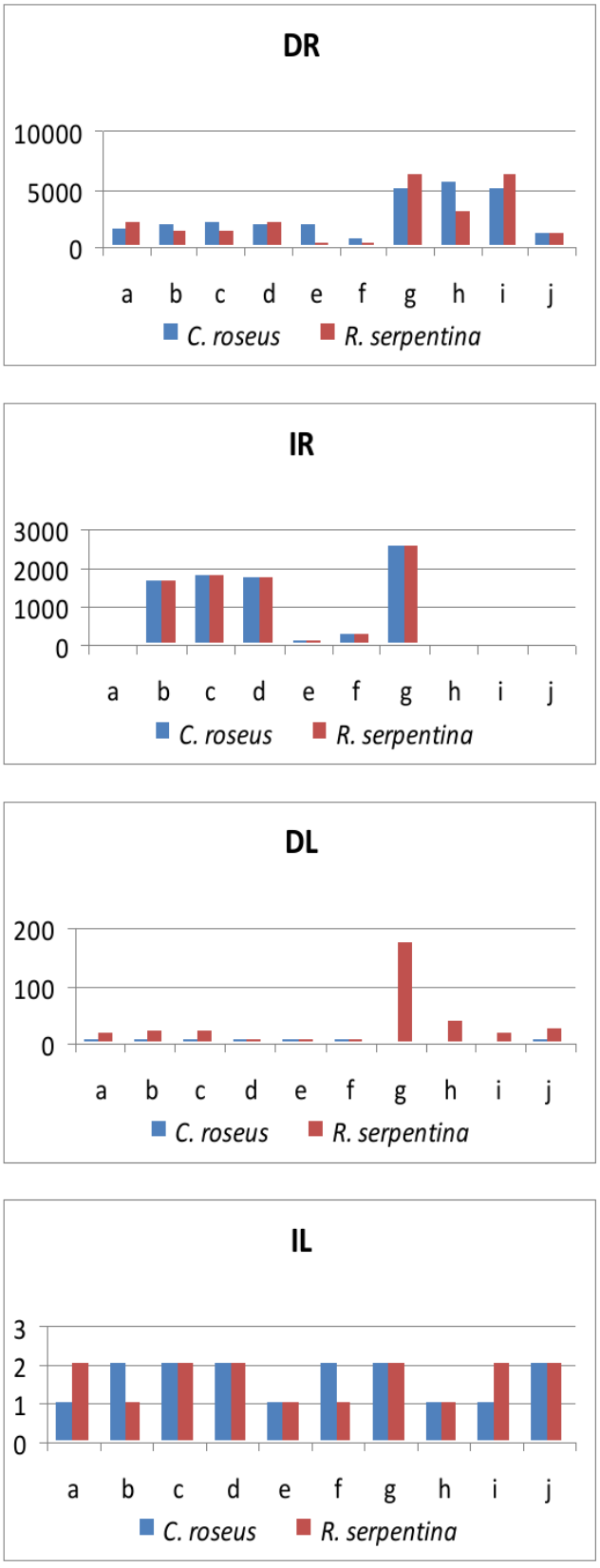

Figure 2. Differentially (D) and identically (I) expressed genes in annotated categories (a-j) of *Catharanthus roseus* and *Rauvolfia serpentina* roots (R) and leaves (L) respectively

The analysis was performed by a series of analysis modules (Figure 1). The output with respect to all critical parameters indicated that the data quality was good for further analysis. It was followed by conversion of FastQ files to Fasta format, read mapping through TopHat, transcriptome assembly using Cufflinks and Cuffmerge, differential expression using Cuffdiff through Galaxy platform. It was followed by grouping genes using gene ontology functional categories for better understanding the molecular processes, their functions and associated secondary metabolic pathways present in the two species.

Four sets of transcriptomes were analysed in this study. Two belonged to roots and leaves of *C. roseus* and the remaining two from those of *R.. serpentina*. Total of 33,312 genes from *C. roseus* and 31,118 from *R. serpentina* were assigned to functional categories. Of these, 25,304 and 22,820 genes were differentially expressed in the roots of *C. roseus* and *R. serpentina* respectively. It was interesting to note that around 7,970 identical genes were present in the roots of both species. On the other hand, only few genes could be assigned to functional categories from the leaves of the two species. For example, 17 and 315 genes were differentially expressed in the leaves of *C. roseus* and *R. serpentina* respectively. Whereas only 16 genes were identical in the leaves of the two species.

Out of the similar transcripts found in both species, ExPASy and MetaCyc database search inferred that one of them codes for an enzyme, strictosidine synthase (STR) involved in alkaloid biosynthesis. Besides, genes for DNA alkylation, binding and methylation, defense mechanism, regulation of epigenetic gene expression, etc were also identified from the roots of both species.

Out of those genes that are differentially expressed, leaf transcriptome analysis of *C. roseus* followed by ExPASy and MetaCyc database search reveled an enzyme, Tabersonine 16-hydroxylase which is involved in biosynthesis of Vindoline, a secondary metabolite used to treat a variety of cancers including Hodgkin's lymphoma, Lymphoblastic leukemias and nephroblastomas. Similarly, there were some genes present only in the leaves of *R. serpentina* such as the one codes Raucaffricine-O-beta-D-glucosidase specifically involved in alkaloid biosynthesis, including Ajmaline, an important plant-derived pharmaceutical used in the treatment of heart disorders.

The pipeline developed was successfully able to associate transcripts to annotated genes, which in turn to corresponding gene products and their targets. For example, genes for various processes like biological regulation, response to stimuli, response to stress, defense response, etc and corresponding enzymes involved were identified in the study. These findings would be very useful in studies where differential analysis is conducted in response to various stress factors like drought [14], dehydration [15], etc. It could also be applied to several situations as in [16-18] where comparative analysis of plant response to a stimulus was evaluated.

4. Conclusions

With appropriate controls for data quality, transcriptome analysis can be carried out to identify gene families [13]. By analyzing RNA-Seq data from *C. roseus* and *R. serpentina* roots and leaves using the newly developed pipeline, it was possible to identify transcripts expressing in these two important medicinal plants. It was clearly evident that there were many differentially and identically expressed genes and Strictosidine synthase, Vindoline and Ajmaline were only few examples cited to demonstrate the utility of the pipeline developed. It is expected that the *in silico* analysis pipeline developed would serve the purpose for locating genes present in different plant species.

Transcriptomic datasets from the medicinal plants would enable discovery of new biosynthetic genes involved in the production of medicinally important secondary metabolites across these and other taxa. The capability of next-generation sequencing to generate a near-complete transcriptome now opens the door for elucidating some of the most chemically prolific, genetically intractable, species of plants.

ACKNOWLEDGEMENTS

Authors thank the Director of Jawaharlal Nehru Tropical Botanical Garden and Research Institute for permission granted to carry out this work.

REFERENCES

[1] Buratowski, S. 2008. Gene Expression-Where to start? Science 322: 154-158.

[2] Kutchan, T. M. 1998. Molecular Genetics of Plant Alkaloid Biosynthesis. The Alkaloids, 50: 257-316.

[3] Mazid, M., Khan T. A., and Mohammad, F. 2011. Role of secondary metabolites in defense mechanisms of plants. Biology and Medicine 3(2): 232-249.

[4] Grothe, T., Kutchan, T. M., and Zenk, M. H. 2002. Salutaridinol 7-o-acetyltransferase and derivatives thereof. Google Patents. http://google.com/patents/WO2002101052A2?cl=en.

[5] J. Stöckigt, and M.H. Zenk. 1977. Strictosidine (isovincoside): the key intermediate in the biosynthesis of monoterpenoid indole alkaloids. J. Chem. Soc. Chem. Commun. 18: 646–648.

[6] Stöckigt J., Barleben L, Panjikar S, Loris EA. 2008. 3D-Structure and function of strictosidine synthase – the key enzyme of monoterpenoid indole alkaloid biosynthesis. Plant Physiology and Biochemistry 46(3): 340-55.

[7] Góngora-Castillo, E., Fedewa, G., Yeo, Y., Chappell, J., Della Penna, D., and Buell, C. R., 2012. Genomic Approaches for Interrogating the Biochemistry of Medicinal

Plant Species. Natural Product Biosynthesis by Microorganisms and Plants, Part C, Elsevier, p. 139–159.

[8] Liu, L.-Y. D., Tseng, H.-I., Lin, C.-P., Lin, Y.-Y., Huang, Y.-H., Huang, C.-K., Chang, T.-H., and Lin, S.-S. 2014. High-throughput transcriptome analysis of the leafy flower transition of *Catharanthus roseus* induced by peanut witches'-broom phytoplasma infection. Plant Cell Physiology 56.

[9] Miettinen, K., Dong, L., Navrot, N., Schneider, T., Burlat, V., Pollier, J., Woittiez, L., Krol, S. v. d., Lugan, R., Ilc, T., Verpoorte, R., Oksman-Caldentey, K.-M., Martinoia, E., Bouwmeester, H., Goossens, A., Memelink, J., and Werck-Reichhart, D. 2014. The seco-iridoid pathway from *Catharanthus roseus*. Nature Communications 5: 3606.

[10] Schluttenhofer, C., Pattanaik, S., Patra, B., and Yuan, L. 2014. Analyses of *Catharanthus roseus* and *Arabidopsis thaliana* WRKY transcription factors reveal involvement in jasmonate signaling. BMC Genomics 15: 502.

[11] St-Pierre, B., VaÂzquez-Flota, F., and De Luca, V. 1999. Multicellular compartmentation of *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. Plant Cell 11: 887.

[12] Verma, M., Ghangal, R., Sharma, R., Sinha, A. K., and Jain, M. 2014. Transcriptome Analysis of *Catharanthus roseus* for Gene Discovery and Expression Profiling. PLoS ONE 9: e103583.

[13] Straub, S. C. K., Bailey, C. D., Cronn, R. C., Fishbein, M., and Liston, A. 2015. Challenges of using transcriptomes for comparative gene family evolution: Examples from Apocynaceae, Botany 2015: Edmonton, Alberta, Botanical Society of America.

[14] Le, D. T., Nishiyama, R., Watanabe, Y., Tanaka, M., Seki, M., et al. 2012. Differential Gene Expression in Soybean Leaf Tissues at Late Developmental Stages under Drought Stress Revealed by Genome-Wide Transcriptome Analysis. PLoS ONE 7(11): e49522. doi: 10.1371/journal.pone.0049522.

[15] Lata, C., Sahu, P. P., and Prasad, M. 2010. Comparative transcriptome analysis of differentially expressed genes in foxtail millet (*Setaria italica* L.) during dehydration stress. Biochemical and Biophysical Research Communications 393(4): 720-727. doi: 10.1016/j.bbrc.2010.02.068.

[16] Nikiforova, V., Freitag, J., Kempa, S., Adamik, M., Hesse, H. and Hoefgen, R. (2003), Transcriptome analysis of sulfur depletion in *Arabidopsis thaliana*: interlacing of biosynthetic pathways provides response specificity. The Plant Journal, 33: 633–650. doi:10.1046/j.1365-313X.2003.01657.x.

[17] Maathuis, F. J. M., Filatov, V., Herzyk, P., C. Krijger, G., B. Axelsen, K., Chen, S., Green, B. J., Li, Y., Madagan, K. L., Sánchez-Fernández, R., Forde, B. G., Palmgren, M. G., Rea, P. A., Williams, L. E., Sanders, D. and Amtmann, A. 2003. Transcriptome analysis of root transporters reveals participation of multiple gene families in the response to cation stress. The Plant Journal, 35: 675–692. doi:10.1046/j.1365-313X.2003.01839.x.

[18] Weber, M., Trampczynska, A. and Clemens, S. 2005. Comparative transcriptome analysis of toxic metal responses in Arabidopsis thaliana and the Cd^{2+} hypertolerant facultative metallophyte Arabidopsis halleri. Plant, Cell and Environment 29: 950–963 doi: 10.1111/j.1365-3040.2005.01479.x.