A Complete Pipeline of Free Bioinformatics Tools for De Novo Transcriptome Assembly and SSR Primer Design

D. N. U. Naranpanawa[1,2], C. H. W. M. R. B. Chandrasekara[1], P. C. G. Bandaranayake[1], A. U. Bandaranayake[3]*

[1]Agricultural Biotechnology Centre, Faculty of Agriculture, University of Peradeniya, Peradeniya 20400, Sri Lanka
[2]Postgraduate Institute of Science, University of Peradeniya, Peradeniya 20400, Sri Lanka
[3]Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya 20400, Sri Lanka
** asithab@pdn.ac.lk*

During the past few decades, next-generation sequencing technologies have grown exponentially in terms of throughput, speed and reduction of sequencing cost. This has revolutionized the field of genomics, allowing the production of vast datasets. However, methods and software requirements for analyzing this data to interpret correct biological meaning are not experiencing the same growth rate. One such limitation is the unaffordable price of commercially available bioinformatics software. Hence, only a small fraction of genomes and transcriptomes have been completely assembled and annotated. Lack of reference genomes for comparative assembly lead to computationally more challenging *de novo* assembly. In addition, obtaining an assembly is a complex process that require many steps by using several complex tools. Due to this, beginners in bioinformatics might find analysis procedures too complicated and time-consuming with the associated learning-curve. Therefore, in order to aid novice biologists in assembling sequence data, and to bridge the bottleneck in computational biology and bioinformatics, we present a complete pipeline of freely available bioinformatics software for *de novo* transcriptome assembly. This pipeline was developed by combining several individual software through user-friendly shell scripts. To test the pipeline, we used Illumina HiSeq paired-end RNA-seq reads from four oil-producing *Santalum album* (sandalwood) tree samples from a published study. The raw data were first filtered for low quality reads, trimmed for adapters and normalized. Assembly was performed with Trinity *de novo* assembler. The quality of the assembly was tested with BUSCO, Bowtie2 and TransRate, and indicated to be high quality. In order to further validate the accuracy of the assembly, we used the assembled transcriptome to identify gene-specific Simple Sequence Repeat (SSR) markers. Primers were designed for eight *S. album* oil biosynthetic genes and two control genes, which were validated in the laboratory with respective samples. All primers amplified successfully, confirming the designed workflow. Furthermore, five SSR markers polymorphic among tested sandalwood accessions are potential markers to be utilized in sandalwood breeding programs. To the best of our knowledge, this is the first attempt of developing a user-friendly, validated assembly pipeline with free bioinformatics software and tools, provided with detailed documentation.

Key words: *De novo* assembly, Transcriptome assembly, Bioinformatics, SSR primer design, Assembly pipeline