Current Plant Biology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Tools for building *de novo* transcriptome assembly☆

Matthew Geniza[a,b], Pankaj Jaiswal[a,*]

[a] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR-97331, USA
[b] Molecular and Cellular Biology Graduate Program, Oregon State University, Corvallis, OR-97331, USA

ARTICLE INFO

Keywords:
Transcriptiome
De novo assembly
Gene expression
RNA-Seq
Differential gene expression
Plant gene expression
Plant Ontology
Gene Ontology
De novo transcriptome assembly
Genome annotation
Genetic marker identification
Plant Reactome
Velvet Oases
SPAdes
Trinity
BinPacker
RNA QUAST
TransRate
CD-HIT-ES

ABSTRACT

The availability of RNA-Seq method allows researchers to capture the spatial or temporal profile of transcriptomes from various types of biological samples. The transcriptome data from a species can be analyzed in the context of its sequenced genomes or closely related genome to score biological sample-specific transcript isoforms, novel transcribed regions and to refine gene models including identification of new genes, in addition to the differential gene expression analysis. However, many plant species of importance currently lack a sequenced genome or a closely related reference genome and thus, rely on the *de novo* methods for generating transcript models and transcriptome assemblies. Here we describe various tools used for *de novo* transcriptome assembly and discuss the data management practices and standards.

1. Introduction

The primary focus of any transcriptomic study is to provide an in-depth comparative analysis of the spatial and temporal profile of expressed genes and abundance of various transcripts between various samples. The biological sample may be selected for studying a specific stage or a body part of an organism in the context of its development or in response to a specific treatment or simply to build a transcriptome atlas of an organism. RNA-Seq is currently the method of choice for transcriptome studies: it requires miniscule quantities of input RNA/can be applied in the single cells to whole organism level; produces the low levels of background signal and informs about the abundance of transcripts; and allows study of gene expression in a species with or without a reference genome.

The RNA-Seq technology and the various software applications used for *de novo* transcriptome assemblies have particularly opened an avenue for studying non-model organisms for which a sequenced references genome is unavailable. The *de novo* transcriptome assembly may be used to align sequence reads from the same or another experiment to determine differential gene expression and to explore the genetic diversity. For example, the assembly may be used for mining potential genetic markers [1,2]. Generating *de novo* transcript assemblies for model plants like Arabidopsis, rice and maize is still useful for discovering new transcript isoforms of existing annotated genes, alternative splicing events, and novel transcribed genes from a plant variety, or in response to specific treatment.

In this manuscript, we highlight some *de novo* transcriptome assemblers that are commonly used for short-read based, reference-free or *de novo* based approaches and provide commented example scripts that contain mirrored README instructions (https://github.com/Jaiswal-lab/Transcriptome_Assembly_Scripts) for those specified assemblers, and discuss best practices when evaluating transcriptome assemblies generated from raw sequencing data from an RNA-Seq experiment. Secondly, we will provide examples of repositories that researchers may use to archive their raw and generated data in standardized formats to promote data sharing, open access, reuse and re-analysis under the

☆ This article is part of a special issue entitled “Genomic resources”.

* Corresponding author.
E-mail address: jaiswalp@science.oregonstate.edu (P. Jaiswal).

https://doi.org/10.1016/j.cpb.2017.12.004
Received 6 December 2017; Received in revised form 16 December 2017; Accepted 16 December 2017

FAIR data principles [3,4].

2. Generating *de novo* transcriptome assembly

2.1. Experimental design, metadata and biocomputing

For any scientific study, it is important to have a sound experimental design. In order to maintain high quality and reproducibility, always follow the compliance with Minimum Information about a Microarray Experiment (MIAME) [5,6] or Minimum Information about a high-throughput nucleotide SEQuencing Experiment (MINSEQE) [7] standards. It is suggested that researchers should use plant materials obtained from the single seed descend to have consistency in the genotype under study and avoid contamination, and should include at least three biological replicates for each sample type including the controls. The meta data associated with each sample should utilize appropriate Ontologies [8,9] to describe the organismal body part, growth and developmental stage, phenotype, treatments and growth conditions such as temperature and photoperiod cycles, relative humidity, type of soil, and whether the plants were grown in field or under any type of controlled environment chambers. Also, the accession ID, variety name and when available genotype of the organism should be listed.

As it pertains for *de novo* transcriptome assembly, researchers will want to consider obtaining sequence data from various Illumina platforms, that is generated in the form of paired-end (PE) or single-end (SE) reads. SE reads are cost-effective and generally appropriate for *de novo* transcriptome assembly when reference sequenced resources are available from the same or closely related species. In the event where genomic and sequence resources for a specified organism are limited or not available, PE reads are highly recommended because they preserve information on transcript directionality [10]. In the absence of a reference genome for a plant species, a *de novo* transcriptome assembly is generated to construct the full-length transcripts [11]. *De novo* assembly is typically memory intensive—depending on the organism/species, number of reads used in the command. For plant samples, we have routinely observed assembly processes use anywhere from 256Gb to 1500 Gb (1.5 Tb) of RAM. If researchers do not have institutional access to High Performance Computing (HPC) resources, they have an option to use various cyber-infrastructure listed in Table 1. The ability to run software on these infrastructures is not limited to assemblies—these resources have the capability to run a whole RNA-Seq study workflow (Fig. 1).

2.2. Quality control of raw reads before transcriptome assembly

The raw data output from the sequencing platform is in the form of FASTQ files containing the sequence reads for each replicate sample. The sequence headers and additional files may carry information on the base calls, number of reads, SE/PE and the read quality. An RNA-Seq study may produce hundreds of millions of reads per sample, not all reads are perfect. Thus the data need further quality checks and filtering [12,13]. Due to biases in the amplification process *via* PCR [14] in the sequencing workflow and potential AT or GC rich repetitive region of the transcriptome, sequencing errors are introduced. The accepted error rate for the Illumina platform is approximately 1% or 1/100 bases. Researchers may also choose to trim potentially incorrectly called bases in reads using tools such as Trimmomatic [15], Sickle [16] RSeQC [17] and those provided by the Illumina Inc., however, there is always the debate of potentially trimming good data [18].

2.3. Assembly of de novo transcriptome

Fig. 1 shows a typical RNA-Seq study workflow and some of the most popular *de novo* transcriptome assembly software used frequently by researchers. Additional softwares such as SOAPdenovo-Trans [19] and Trans-AbySS [20] are also use routienly. Users can access these programs *via* publicly available online platforms (Table 1) or install their appropriate licensed copies on the local infrastructure. We usually try to run at least two different application to build a consensus assembly from the list of the following assemblers:

2.3.1. Velvet/Oases

Originally released in 2008, Velvet [21] (https://github.com/dzerbino/velvet) was developed to create *de novo* genome assembly using short read technology. Utilizing the de Bruijn graph to assemble short reads, Velvet can also take paired end data to resolve repeat regions. To assemble transcriptomes *de novo*, Oases [22] (https://github.com/dzerbino/oases) uses the assembly produced by Velvet and clusters the contigs into loci. Similar to Velvet, Oases can use paired-end read data to construct transcript isoforms. Oases was developed to deliver resolution of alternative splicing events at the individual transcript isoforms and efficient merging of multiple transcript isoforms. The merging of multiple assemblies allows creation of a single consensus gene model that represents gene loci on the genome. Furthermore, fine tuning of assembled transcripts can be done by optimizing parameters using additional tools available at https://github.com/tseemann/VelvetOptimiser.

Table 1
List of available resources.

Cyber-infrastructures for bioinformatics analyses	
Resource	URL
CyVerse	http://www.cyverse.org/
Galaxy	https://usegalaxy.org/
GenePattern	http://software.broadinstitute.org/cancer/software/genepattern#

Data repositories		
Data type	Repository	URL
Raw sequence reads	EBI ArrayExpress	https://www.ebi.ac.uk/arrayexpress/submit/overview.html
Raw sequence reads	NCBI-SRA	https://www.ncbi.nlm.nih.gov/sra
Transcriptome assemblies, annotation, markers, *etc.*	European Nucleic Archive	https://www.ebi.ac.uk/ena/submit
All data generated from the experiment	CyVerse	http://www.cyverse.org
All data Generated from the experiment	Dryad digital repository	http://datadryad.org
All data Generated from the experiment	Harvard Dataverse	https://dataverse.harvard.edu
Transcriptome assembly	TSA	https://www.ncbi.nlm.nih.gov/genbank/tsa/
Aligned data	Track hub	http://trackhubregistry.org/