Current Plant Biology 11–12 (2017) 41–45

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/

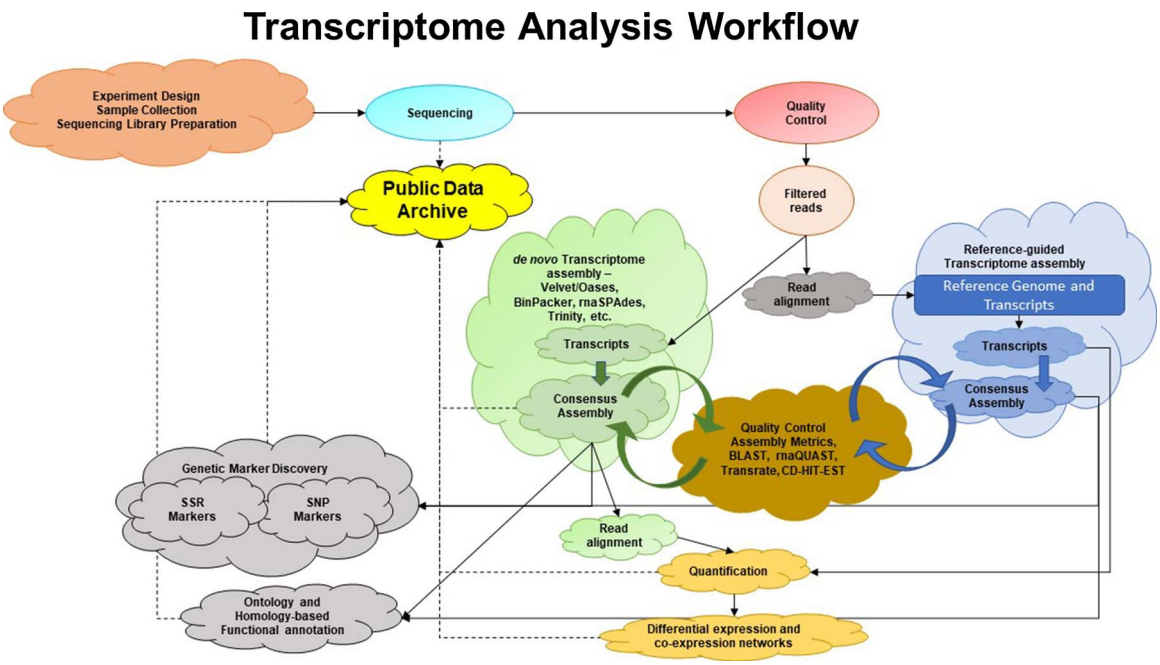

Fig. 1. Workflow showing various components of a transcriptome assembly and analysis.

2.3.2. SPAdes

SPAdes [23] (http://bioinf.spbau.ru/spades) was first released in 2012 and was intended for single-cell transcriptome studies, fungal, and other small genomes. Using a modified version of the de Bruijn graph, SPAdes leverages mate-pair information to *de novo* assemble the genome. rnaSPAdes (http://cab.spbu.ru/software/rnaspades/) was added to the pipeline in 2015 with aim to address the uneven coverage depth that are native to RNA-Seq data. Unlike SPAdes, rnaSPAdes has been used for assembling transcriptomes of several species regardless of their genome size. Assembly optimization is represented by the output of rnaSPAdes, where a researcher is given options of what assembly to use in downstream analyses. Three assemblies are produced; an assembly that contains all assembled transcripts, an assembly that contains only long and highly expressed transcripts, and an assembly that contains only short and low expressed transcripts. Such data may be used at the researcher's discretion.

2.3.3. Trinity

Trinity [24] released in 2013 uses the de Bruijn graph algorithm to assemble *de novo* transcriptome using short reads. Trinity combines three independent software packages to process RNA-Seq reads: The Inchworm package assembles the RNA-Seq reads into transcripts; Chrysalis clusters the assembled transcripts and constructs de Bruijn graphs for each cluster; and finally, Butterfly analyses the graphs and produces the full-length transcripts. Trinity has a very detailed protocol [25] and is constantly maintained and supported on GitHub (https://github.com/trinityrnaseq/trinityrnaseq/wiki). Trinity also provides built-in metrics to assess quality of assembled transcripts—in particular the ExN50.

2.3.4. BinPacker

BinPacker was released in 2016 [26]. Unlike the previously mentioned assemblers that employ the de Bruijn graph to construct transcripts, BinPacker incorporates coverage information to construct the splicing graph. Each splicing graph is a representative of all the alternative splicing transcripts for each locus, based on the assumption that each splicing graph represents one expressed transcript. The bin-packing algorithm aims to optimize the edge-path-cover for each splicing graph to recover the set of transcripts that can be assembled through overlapping sequence reads and junctions.

3. Evaluation of *de novo* transcriptome assembly

Following the *de novo* transcriptome assembly, the standard quality metrics for the assemblies is determined by the “N50” length defined as the shortest sequence length at 50% of the genome. Although the “N50” metric is popularly used in assessing genome assemblies, it is hardly appropriate for transcriptome assemblies that represent a dynamic measurement of the RNA abundance in the spatial and temporal snapshot of the biological sample [27]. The “ExN50” is more appropriate for transcriptome assemblies. The “ExN50” examines the top most highly expressed transcripts that represent 50% of the total normalized expression data. This takes into account the dynamic nature of transcriptomes and requires transcript abundance estimation in order to be calculated. The “ExN50” can be easily calculated if used within the Trinity protocol. This assessment is followed by additional steps. To capture and count all reads that map to the assembled transcripts, it is important to align sequence reads back to the assembled transcriptome. Typically, a large majority (70–90%) of all reads should map back to the assembly. Furthermore, collection of basic assembly metrics such as shortest/longest transcript lengths, mean/median transcript lengths, and number of transcripts is required. These metrics can be compared directly to cDNA reference sets of closely related species and may suggest some insights into the RNA processing in the biological sample and the percent coverage of the gene set represented in the assembly. It is also recommended that researchers should BLAST [28] the assembled transcriptome to closely related species. Starting with available references at the genus level, researchers can expect the percent identity of the assembled transcriptome to decrease as comparisons are made at the family and order level. For good measure, an outgroup should be chosen and will be expected to have the lowest level of percent identity.

I addition to the above-mentioned metrics, the following software applications can be used to filter the artefacts and improve the transcriptome assembly quality.

3.1. rnaQUAST

The rnaQUAST software package [29] is a quality evaluation tool that can compare various assembly approaches when a reference genome is available. The assembled transcriptome is aligned to the reference genome to calculate simple metrics that represent

completeness, correctness levels of the assembly, and estimating the percent gene coverage. When a reference genome is not available, rnaQUAST can use the plant Benchmarking Universal Single-Copy Orthologs (BUSCO) gene set [30] and GeneMarkS-T [31] as a reference. The rnaQUAST can also compare different transcriptome assemblies at once and provide quality metrics as output.

3.2. TransRate

The TransRate software package [32] (https://github.com/Blahah/transrate) takes the sequenced reads and the transcript assembly as input to show the potential artefacts of *de novo* transcriptome assembly. TransRate can also assess transcriptomes built from different assemblers and analyze against proteins or transcripts from a related species and inspect the alignments. Ultimately, TransRate outputs a suite of metrics that informs the researchers on choosing between assemblers, parameters, and optimized assembly.

3.3. CD-HIT-EST

When assembling a *de novo* transcriptome it can be difficult to identify redundantly assembled isoforms due to varying expression levels. CD-HIT-EST [33] (https://github.com/weizhongli/cdhit) clusters similar sequences in an assembled transcriptome and outputs a set of non-redundant representative sequences. In short CD-HIT-EST takes the canonical sequence of a produced cluster to remove sequences above a specified identity threshold correct the bias within a given assembled transcriptome.

4. Perspective on archiving the raw and generated data

The advancements in next-generation sequencing technology also created computational challenges in how the data is deposited, formatted and remains accessible for re-use by the community of researchers. Both, funding agencies and the peer-reviewed scientific journals require depositing raw genomics data generated by next-generation sequencing platforms to public data repositories (see Table 1). The European Molecular Biology Laboratory – European Bioinformatics Institute hosts the ArrayExpress (https://www.ebi.ac.uk/arrayexpress/submit/overview.html) [34,35] and uses Annotare [36] an online tools for the transcriptome data submission, whereas the National Center for Biotechnology Information hosts the Sequence Read Archive (SRA) [37]. Each repository requires a certain level of metadata that includes use of Ontologies for plants [8] and those from the OBO Foundry [9] describing submitted datasets. These repositories make data publicly available for reuse and re-analysis and check for reproducibility without referring to an associated paper. The raw data could be used for re-analysis as updates in the reference genome or transcriptome assembly and analysis software often happen in future that may result in improvement of the assemblies [34].

In addition to depositing raw data at ArrayExpress and SRA, researchers should consider submitting generated data from various analyses to open source repositories (see Table 1). This includes transcriptome assemblies, normalized and/or log-transformed gene expression data (with appropriate statistics measure) used for actual analysis, alignment data, functional annotations, genetic markers, *etc*. The aligned data can be submitted to the Track Hub [38], thus allowing users to anonymously upload their aligned data to a reference genome (when available) accessible from UCSC and Ensembl genome browsers [39,40]. Researchers may also share their data with Model organism database or clade specific databases. Many emerging public data repositories like CyVerse (http://www.cyverse.org/data-store) [41,42], DataDryad (http://datadryad.org/), Track Hub (https://trackhubregistry.org/about), and Harvard Dataverse (https://dataverse.harvard.edu/) allow users to submit various types of data and issue DOIs. This secondary data can also be used for testing new algorithms or analytical/visualization tools, or for curation and annotation of public datasets and for teaching/training purposes.

5. Conclusion

In the scope of this review, we presented some examples of *de novo* transcriptome assemblers and example scripts for researchers to use (https://github.com/Jaiswal-lab/Transcriptome_Assembly_Scripts). Upon assembling the transcriptome, we provide strategies and methods for researchers to benchmark the quality of their assemblies.

The high quality transcriptome assemblies will reduce bias and erroneous results when the assembly is used in downstream analyses such as identification of differentially expressed transcripts, or for mining potential genetic markers (*e.g.* simple sequence repeats; SSR markers) [1,2]). Finally, we discussed the many options for researchers to formally deposit raw and generated data to repositories that issue DOIs as mentioned above and ensure proper citation of authors when data is used.

The massive amount of data generated by next-generation sequencing is continuously growing. The decreasing cost of sequencing has lowered the barriers for conducting such research and future platforms such as PacBio and Ion Torrent will only add to the amount of data generated for transcriptome based studies. It is imperative that institutions have a contingency plan to securely store the data, allow for easy access, and ensure proper credit is given to authors when available data is used in future studies.

Funding

This work was supported by the National Science Foundation award [IOS-1127112].

Conflict of interest statement

The authors have no conflict of interest.

References

[1] S.E. Fox, M. Geniza, M. Hanumappa, S. Naithani, C. Sullivan, J. Preece, V.K. Tiwari, J. Elser, J.M. Leonard, A. Sage, et al., De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat Triticum monococcum, PLoS One 9 (2014) e96855.
[2] S.E. Fox, J. Preece, J.A. Kimbrel, G.L. Marchini, A. Sage, K. Youens-Clark, M.B. Cruzan, P. Jaiswal, Sequencing and de novo transcriptome assembly of Brachypodium sylvaticum (Poaceae), Appl. Plant Sci. 1 (2013) 1200011.
[3] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Sci. Data 3 (2016) 160018.
[4] A.F. Adam-Blondon, M. Alaux, C. Pommier, D. Cantu, Z.M. Cheng, G.R. Cramer, C. Davies, S. Delrot, L. Deluc, G. Di Gaspero, et al., Towards an open grapevine information system, Hortic. Res. 3 (2016) 16056.
[5] A. Brazma, Minimum Information About a Microarray Experiment (MIAME) – Successes, Failures, ChallengesThescientificworldjo, Sci. World. J. 9 (2009) 420–423.
[6] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, et al., Minimum information about a microarray experiment (MIAME) – toward standards for microarray data, Nat. Genet. 29 (2001) 365–371.
[7] (FGED), F.g.D.S. (2012), *MINSEQE. http://fged.org/projects/minseqe/*.
[8] L. Cooper, A. Meier, M.A. Laporte, J.L. Elser, C. Mungall, B.T. Sinn, D. Cavaliere, S. Carbon, N.A. Dunn, B. Smith, et al., The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics, Nucleic Acids Res. (2017), http://dx.doi.org/10.1093/nar/gkx1152.
[9] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (2007) 1251–1255.
[10] J.Z. Levin, M. Yassour, X.A. Adiconis, C. Nusbaum, D.A. Thompson, N. Friedman, A. Gnirke, A. Regev, Comprehensive comparative analysis of strand-specific RNA sequencing methods, Nat. Methods 7 (2010) 709–U767.
[11] J.R. Miller, S. Koren, G. Sutton, Assembly algorithms for next-generation sequencing data, Genomics 95 (2010) 315–327.
[12] S.W. Hartley, J.C. Mullikin, QoRTs: a comprehensive toolset for quality control and

data processing of RNA-Seq experimentsBMC Bioinformatics, BMC Bioinf. 16 (2015) 224.
[13] X. Li, A. Nair, S. Wang, L. Wang, Quality control of RNA-seq experiments, Methods Mol. Biol. 1269 (2015) 137–146.
[14] I. Kozarewa, Z.M. Ning, M.A. Quail, M.J. Sanders, M. Berriman, D.J. Turner, Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G plus C)-biased genomes, Nat. Methods 6 (2009) 291–295.
[15] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120.
[16] N.A., J. and N., F.J. (2017) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files Software.
[17] L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments, Bioinformatics 28 (2012) 2184–2185.
[18] C. Del Fabbro, S. Scalabrin, M. Morgante, F.M. Giorgi, An extensive evaluation of read trimming effects on Illumina NGS data analysis, PLoS One 8 (2013).
[19] Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, et al., SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads, Bioinformatics 30 (2014) 1660–1666.
[20] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S.D. Jackman, K. Mungall, S. Lee, H.M. Okada, J.Q. Qian, et al., De novo assembly and analysis of RNA-seq data, Nat. Methods 7 (2010) 909–912.
[21] D.R. Zerbino, E. Birney, Velvet: Algorithms for de novo short read assembly using de Bruijn graphs, Genome Res. 18 (2008) 821–829.
[22] M.H. Schulz, D.R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, Bioinformatics 28 (2012) 1086–1092.
[23] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (2012) 455–477.
[24] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q.D. Zeng, et al., Full-length transcriptome assembly from RNA-seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644-U130.
[25] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, et al., De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512.
[26] J.T. Liu, G.J. Li, Z. Chang, T. Yu, B.Q. Liu, R. McMullen, P.Y. Chen, X.Z. Huang, BinPacker: packing-based de novo transcriptome assembly from RNA-seq data, PLoS Comput. Biol. 12 (2016).
[27] S.T. O'Neil, S.J. Emrich, Assessing de novo transcriptome assembly metrics for consistency and utility, BMC Genomics 14 (2013).
[28] D.W. Mount, (2007) Using the Basic Local Alignment Search Tool (BLAST). CSH protocols, 2007, pdb top17.
[29] E. Bushmanova, D. Antipov, A. Lapidus, V. Suvorov, A.D. Prjibelski, rnaQUAST: a quality assessment tool for de novo transcriptome assemblies, Bioinformatics 32 (2016) 2210–2212.
[30] F.A. Simao, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212.
[31] S. Tang, A. Lomsadze, M. Borodovsky, Identification of protein coding regions in RNA transcripts, Nucleic Acids Res. 43 (2015) e78.
[32] R. Smith-Unna, C. Boursnell, R. Patro, J.M. Hibberd, S. Kelly, TransRate: reference-free quality assessment of de novo transcriptome assemblies, Genome Res. 26 (2016) 1134–1144.
[33] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[34] I. Papatheodorou, N.A. Fonseca, M. Keays, Y.A. Tang, E. Barrera, W. Bazant, M. Burke, A. Fullgrabe, A.M. Fuentes, N. George, et al., Expression Atlas: gene and protein expression across multiple studies and organisms, Nucleic Acids Res. (2017), http://dx.doi.org/10.1093/nar/gkx1158.
[35] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y.A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, et al., ArrayExpress update–simplifying data submissions, Nucleic Acids Res. 43 (2015) D1113–D1116.
[36] R. Shankar, H. Parkinson, T. Burdett, E. Hastings, J. Liu, M. Miller, R. Srinivasa, J. White, A. Brazma, G. Sherlock, et al., Annotare:–a tool for annotating high-throughput biomedical investigations and resulting data, Bioinformatics 26 (2010) 2470–2471.
[37] R. Leinonen, H. Sugawara, M. Shumway, International Nucleotide Sequence Database, C, The sequence read archive, Nucleic Acids Res. 39 (2011) D19–21.
[38] M.K. Tello-Ruiz, S. Naithani, J.C. Stein, P. Gupta, M. Campbell, A. Olson, S. Wei, J. Preece, M.J. Geniza, Y. Jiao, et al., Gramene 2018: unifying comparative genomics and pathway resources for plant research, Nucleic Acids Res. (2017).
[39] J. Casper, A.S. Zweig, C. Villarreal, C. Tyner, M.L. Speir, K.R. Rosenbloom, B.J. Raney, C.M. Lee, B.T. Lee, D. Karolchik, et al., The UCSC genome browser database: 2018 update, Nucleic Acids Res. (2017), http://dx.doi.org/10.1093/nar/gkx1020.
[40] P.J. Kersey, J.E. Allen, A. Allot, M. Barba, S. Boddu, B.J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, C. Grabmueller, et al., Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species, Nucleic Acids Res. (2017), http://dx.doi.org/10.1093/nar/gkx1011.
[41] B.L. Joyce, A.K. Haug-Baltzell, J.P. Hulvey, F. McCarthy, U.K. Devisetty, E. Lyons, Leveraging CyVerse resources for de novo comparative transcriptomics of underserved (non-model) organisms, J. Vis. Exp.: JoVE (2017).
[42] U.K. Devisetty, K. Kennedy, P. Sarando, N. Merchant, E. Lyons, Bringing your tools to CyVerse discovery environment using docker, F1000Research 5 (2016) 1442.