DE NOVO TRANSCRIPTOME ASSEMBLY OF SOLiD SEQUENCING DATA IN CUCUMIS MELO

Purru Supriya[1]and K V Bhat[2]

[1]Division of Bioinformatics, Indian Agricultural Research Institute, Pusa campus, New Delhi 110012, India
aarush.supriya@gmail.com
[2] National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi 110012, India
kvbhat@nbpgr.ernet.in

ABSTRACT

As sequencing technologies progress, focus shifts towards solving bioinformatic challenges, of which sequence read assembly is the first task. In the present study, we have carried out a comparison of two assemblers (SeqMan and CLC) for transcriptome assembly, using a new dataset from Cucumis melo. Between two assemblers SeqMan generated an excess of small, redundant contigs where as CLC generated the least redundant assembly. Since different assemblers use different algorithms to build contigs, we followed the merging of assemblies by CAP3 and found that the merged assembly is better than individual assemblies and more consistent in the number and size of contigs. Combining the assemblies from different programs gave a more credible final product, and therefore this approach is recommended for quantitative output.

KEYWORDS

De novo assembly, Transcriptome, Contig, RNA-Seq

1. INTRODUCTION

Cucurbitaceae is an important family in the plant kingdom, whose importance is just after *Graminae*, *Leguminosae* and *Solanaceae*. Over the past several years, the genome sequencing of many crops in *Cucurbitaceae* has been completed, such as cucumber [1], muskmelon [2] and watermelon [3].

Abiotic stress factors negatively impact the agricultural production systems world over. Genetic enhancement of agricultural species for abiotic stress tolerance has not met with much success. This information is lacking for most of the economically important traits such as moisture stress in melons. In the present study we used transcriptome data of musk melon for assembly and these set of assembled transcripts allows for initial gene expression studies.

RNA-seq is cost-economic and time-saving, particularly compared to traditional expressed sequence tag (EST) sequencing and it can generate transcriptome data for non-model species by means of incomplete genome information [4]. In addition to profiling gene expression, RNA-seq has shown powerful applications in areas, such as cataloguing of non-coding RNAs, investigation of the transcriptional structure of genes and splicing patterns and the study of posttranscriptional modification and mutations [5]. Conventionally, transcriptome projects have been continued to run on Sanger dideoxy-sequenced expressed sequence tags. However, the second-generation sequencing technologies provide much higher throughput than Sanger sequencing at a lower cost per base, these new technologies are progressively more used. Massively parallel sequencing platforms, such as the Illumina, Inc. Genome Analyzer, Applied Biosystems SOLiD System, and 454 Life Sciences (Roche) GS FLX, have provided an unprecedented increase in DNA sequencing throughput. At present, these technologies generate high-quality short reads from 25 to 500 bp in length, which is considerably shorter than the capillary-based sequencing technology.

The three platforms offer a variety of experimental approaches for characterizing a transcriptome, including single-end and paired-end cDNA sequencing, tag profiling, methylation assays, small RNA sequencing, sample tagging to permit small sub sample identification, and splice variant analyses [6]. RNA-seq has high dynamic range of detection *i.e.* very low and very high abundance transcripts can be detected with RNA-seq while microarrays lack sensitivity to detect genes expressed at either high or low levels. Using this technique several genes were detected that are expressed during berry development in *Vitis vinifera* [7].

In parallel with the technological improvements that have increased the throughput of the next generation short read sequencers, many algorithmic advances have been made in *de novo* assemblers for short read data. Prior to the development of transcriptome assembly computer programs, transcriptome data were analyzed primarily by mapping on to a reference genome. Though genome alignment is a robust way of characterizing transcript sequences, this method is disadvantaged by its inability to account for incidents of structural alterations of mRNA transcripts, such as alternative splicing [8]. A number of assembly programs are available today. Although these programs have been commonly successful in assembling genomes, transcriptome assembly presents some distinctive challenges. High sequence coverage for a genome may possibly indicate the occurrence of repetitive sequences, for a transcriptome, they may indicate abundance. In addition, unlike genome sequencing, transcriptome sequencing can be strand-specific, due to the possibility of both sense and antisense transcripts. Finally, it can be difficult to reconstruct and tease apart all splicing isoforms [9].

2. MATERIALS AND METHODS

2.1. Materials

In the present study we used transcriptome data of *Cucumis melo* var.*agrestris.* The accession of the material used was selected after an initial screening of several accessions for moisture stress tolerance. Plants were grown under a rain-out shelter, with normal irrigation conditions and moisture stress conditions where plants were subjected to moisture stress at 30 days after sowing without giving irrigation in case of stress condition. Fresh leaves were collected from plants 60 days after treatment and immediately submerged in RNA*later* for storage without jeopardizing the quality or quantity of RNA. RNA*later* eliminates the need to immediately process tissue specimens or to freeze samples in liquid nitrogen for later processing. Total RNA was extracted from leaves using TRIZOL RNA isolation protocol followed by cDNA preparation using Life Technologies cDNA synthesis kit. ABI SOLiD sequencing platform was used to sequence cDNA samples. A total of 47,035,393 and 45,152,235 high quality unique reads of transcriptomic data for control sample and stress sample were used in this study (Table 1).

Table 1. Read Statistics

Description	**Control**	**Stress**
Raw reads	control.csfasta	stress.csfasta
Raw reads	control_QV.qual	stress_QV.qual
No of reads	47,035,393	45,152,235

2.2. Assembly

De novo or reference-independent strategy is used to directly assemble transcripts by finding overlaps between the reads. This strategy is applied when a reference genome is not available or is poorly annotated. A number of transcriptome assembly programs have been developed like Trans-ABySS [10], Multiple-k [11], Rnnotator [12], Oases [13], Trinity [14], SOAPdenovo2 [15] and SSP [16].

There are two fundamental approaches in algorithms for short-read assemblers: overlap graphs and de Brujin graphs. Most established assemblers that were developed for sanger reads follow the overlap-layout-consensus paradigm. They compute all pair-wise overlaps between these reads and capture this information in a graph. Each node in the graph corresponds to a read, and an edge denotes an overlap between two reads. The overlap graph is used to compute a layout of reads and a consensus sequence of contigs. De Brujin graphs reduce the computational effort by breaking reads into smaller sequences called k-mers, where parameter k denotes the length in bases of these sequences. The de Brujin graph captures overlaps of length k-1 between these k-mers and not between the actual reads. By reducing the entire data set down to k-mer overlaps the de Brujin graph reduces the high redundancy in short-read data sets. In this study, we used three assemblers: CLC Genomics work bench, DNA STAR's SeqMan NGen and CAP3 for hybrid assembly. For each assembler, we used the default parameters suggested for transcriptome assembly. These assemblers differ in the algorithms used and how they treat individual reads. SeqMan and CAP3 use variations of the Overlap- Layout-Consensus (OLC) strategy where as CLC uses de Bruijn graph path finding. Table 2 shows the Features of assembly programmes compared in this study.

Table 2. Features of assembly programmes compared in this study

Assembler	**Type**	**Splits reads**	**Author**	**URL**
CLC	de Bruijn graph	Yes	CLC	http://www.clcbio.com/
SeqMan NGen	OLC	No	DNA STAR	http://www.dnastar.com/t-products-seqman- ngen.aspx

2.3. Comparison of assemblers

In the present study first we used CLC Genomics Workbench. Its *de novo* assembly algorithm offers extensive support for a variety of data formats, including both short and long reads and mixing of paired reads. Parallelly, *de novo* assembly was also run on DNA STAR's SeqMan NGen. SeqMan NGen is groundbreaking sequence assembly software that has the ability to assemble any size transcriptome quickly and accurately. It assembles data from all next-generation sequencing platforms.

2.4. Merging assemblies to improve credibility

Once the tens to hundreds of thousands of short reads have been produced, it is important to correctly assemble these to estimate the sequence of all the transcripts. Most transcriptome assembly projects use only one program for assembling sequencing reads, but there is no evidence that the programs used to date are optimal. Different algorithms are used in different assembly programs to derive final contigs. These algorithms may model different portions of the transcriptome with different accuracies [17]. We combined two assemblies at a time by treating their contigs as pseudo-reads and assembled using CAP3.

CAP3 assembly algorithm consists of three major phases. In the first phase, 5'and 3' poor regions of each read are identified and removed. Overlaps between reads are computed. False overlaps are identified and removed. In the second phase, reads are joined to form contigs in decreasing order of overlap scores. Then, forward–reverse constraints are used to make corrections to contigs. In the third phase, a multiple sequence alignment of reads is constructed and a consensus sequence along with a quality value for each base is computed for each contig. Base quality values are used in computation of overlaps and construction of multiple sequence alignments [18].

3. Results and Discussion

CLC was the fastest assembler when compared to DNA STAR and used the least amount of memory. The CLC assembly cell user manual [19] states that it achieves this remarkable speedup over other assemblers by utilizing de Bruijn graphs rather than the traditional OLC paradigm, by efficiently using multiple cores, and by optimising low level machine code. De Bruijn graph assemblers split reads into overlapping k-mers and utilize short reads with high coverage efficiently to make large contigs possible, but for transcriptome datasets where different contigs may have varying and often low coverage, they are unlikely to be optimal. DNA STAR generated more no. of contigs and shorter contigs over all (Table 3).

Table 3. Results of CLC and SeqMan NGen

Statistics	CLC		SeqMan NGen	
	Control	Stress	Control	Stress
No. of contigs produced	564	586	32,546	36,628
N50	283	284	740	748
Maximum transcript length	1116	1198	1051	1136
Minimum transcript length	200	200	60	98

CLC generated the least redundant assembly, a direct consequence of using a de Bruijn graph algorithm. The CLC assembly was poorer, signifying that this assembly is more fragmented than others. We improved the assembly by merging two assemblies at a time using a traditional OLC assembler (CAP3). Merging of assemblies performed with different programs is a frequently used approach in genome assembly projects, especially those that employ multiple sequencing technologies. Application of this strategy to the problem of *de novo* transcriptome assembly appears particularly useful. After merging the initial assemblies from CLC and DNASTAR, CAP3 generated larger transcripts (Table 4). Figure 1 shows CAP3 assembly statistics.

Table 4: CAP3 assembly statistics

Statistics	Control	Stress
No. of transcripts produced	12859	13448
N50	904	900
Maximum transcript length	2665	3039
Transcripts > 2000bp	35	34
1000<transcripts<2000bp	1973	1995
500<transcripts<1000	3859	3868
Transcripts<500	6992	7551

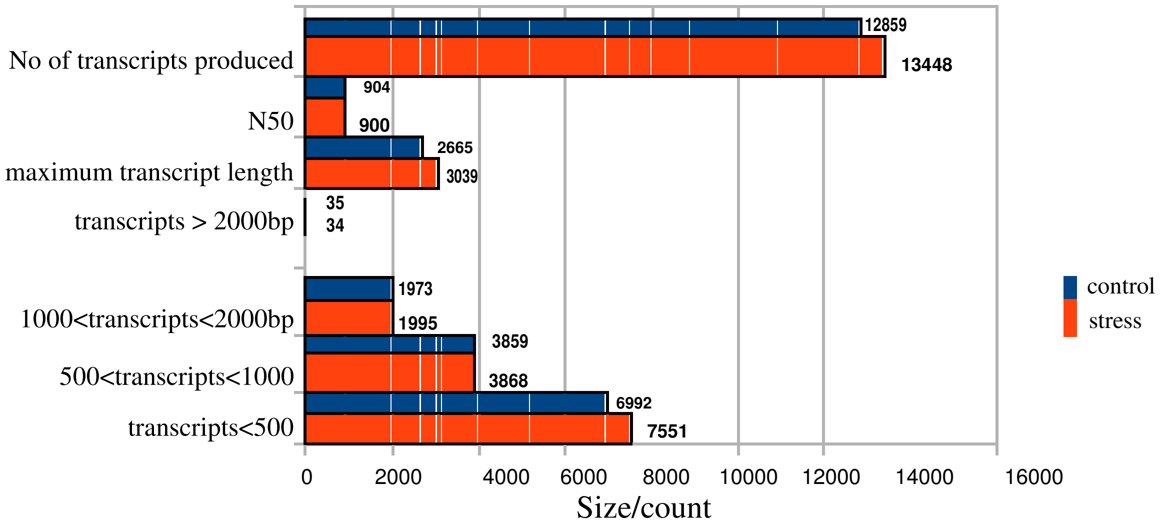

Figure 1. CAP3 Assembly Statistics

4. CONCLUSION

We compared two assembly programs in this study in which DNA STAR uses OLC assembly strategy where as CLC uses de Bruijn graphs. The CLC *de novo* assembler is clearly a step in the right direction because its de Bruijn graph algorithm achieves reasonable results in very little time on large datasets. Perhaps the way ahead is to use the de Bruijn graph approach for transcripts with high coverage and the OLC approach for transcripts with low coverage. Hybrid assembly strategy delivers robust contigs from intermediate assemblies produced by current programs, and this strategy is likely to be of utility in deriving the best assemblies from future programs as well. These set of assembled transcripts for both the samples will be functionally annotated in order to find out the differentially expressed genes which will greatly aid in development of melon cultivars with tolerance to moisture stress.

ACKNOWLEDGEMENTS

This research was supported by NBPGR, IASRI and IARI, New Delhi, India.

REFERENCES

[1] Huang SW, Li RQ, Zhang ZH, Li L and Gu XF (2009) "The genome of the cucumber, Cucumis sativus L". Nature Genetics, 41: 1275-1281.

[2] Jordi GM, Andrej B, Walter S, Michael B, Gisela M, Gonzalez V, Elizabeth H, Francisco C, Luca C, Ernesto L, Tyler A, Salvador CG, Blanca J, Joaquin C, Pello Z, Gonzalez D, Luis RM, Marcus D, Lei D, Miguel AT, Belen LG, Marta M, Luming Y, Yiqun W, Arcadi N, Tomas MB, Miguel AA, Fernando N, Belen P, Toni G, Guglielmo R, RodericG, Josep C, Pere A and Pere (2012) "The genome of melon (Cucumismelo L.)". Proceedings of National Academy of Sciences, 109: 11872–11877.

[3] Guo S, Zhang J, Sun H, Salse J and Lucas WJ (2012) “The draft genome of watermelon (Citrulluslanatus) and resequencing of 20 diverse accessions”. Nature Genetics, 45: 51-58.

[4] Wang Z, Gerstein M, Snyder M (2009) “RNA-Seq: a revolutionary tool for transcriptomics”. Nature Reviews Genetics, 10: 57–63.

[5] Westermann AJ, Gorski SA, Vogel J (2012) “Dual RNA-seq of pathogen and host”. Nature Reviews Microbiology, 10: 618 630.

[6] Wall KP, Jim LM, Andre SC, Abdelali B, Erik W, Haiying L, Lena L, Lynn PT, Yi H, John EC, Hong M, Stephan CS, Douglas ES, Pamela SS, Naomi A and Claude WP (2009) “Comparison of next generation sequencing technologies for transcriptome characterization”. BMC Genomics, 10: 347.

[7] Zenoni S, Ferrarini A, Enrico G, Luciano X, Marianna F, Giovanni M, Diana B, Mario P and Massimo D (2010) “Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-Seq”. Plant Physiology, 152: 1787–1795.

[8] Inanç B, Shaun DJ, Cydney BN, Jenny Q, Richard V, Greg S, Ryan DM, Yongjun Z, Martin H, Jacqueline ES, Doug EH, Joseph MC, Randy DG, Marco AM and Steven JM (2009) “De novo transcriptome assembly with ABySS”. Bioinformatics, 25: 2872–2877.

[9] Martin J.A., Wang Z. (2011). "Next-generation transcriptome assembly". Nature Reviews Genetics, 12: 671– 682.

[10] Robertson D, Schein J, Chiu R, Corbett R, Field M, Mungall K, Lee S, Hisanaga MO, Jenny QQ, Malachi G, Anthony R, Nina T, Timothee C, Yaron SB, Richard N, Simon KC, Rong S, Richard V, Baljit K, Anna-Liisa P, Angela T, Yong JZ, Richard AM, Martin H, Marco AM, Steven JM, Pamela AH and Inanc B (2010) “De novo assembly and analysis of RNA-seq data”. Nature Method, 7: 909-912.

[11] Surget-Groba Y and Montoya-Burgos JI (2010) “Optimization of de novo transcriptome assembly from next-generation sequencing data”. Genome Research, 20: 1432-1440.

[12] Martin J, Vincent MB, Zhide F, Xiandong M, Matthew B, Tao Z, Sherlock G, Michael S, Zhong W (2010) “Rnnotator: an automated de novotranscriptome assembly pipeline from stranded RNA-Seq reads”. BMC genomics, 11: 663.

[13] Schulz MH, Zerbino DR, Martin V and Ewan B (2012) “Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels”. Bioinformatics, 28: 1086-1092.

[14] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson Da, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N andRegev A (2011) “Full-length transcriptome assembly from RNA-Seq data without a reference genome”. Nature Biotechnology, 29(7): 644–652.

[15] Luo R, Binghang L, Yinlong X, Zhenyu L, Weihua H, Jianying Y, Guangzhu H, Yanxiang C, Qi P, Yunjie L, Jingbo T, Gengxiong W, Hao Z, Yujian S, Yong L, Chang Y, Bo W, Yao L, Changlei H, David WC, Siu-Ming Y, Shaoliang P, Zhu X, Guangming L, Xiangke L, Yingrui L, Huanming Y, Jian W, Tak-Wah L and Jun W (2012) “SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler”. Giga Science, 1: 18.

[16] Safikhani Z, Mehdi S, Hamid P and Changiz E (2013) “SSP: An interval integer linear programming for de novotranscriptome assembly and isoform discovery of RNA-seq reads”. Genomics, 102: 507–514.

[17] Kumar S and Blaxter ML (2010) “Comparing de novo assemblers for 454 transcriptome data”. BMC Genomics, 11: 571.

[18] Huang X and Madan A (1999) “CAP3: A DNA Sequence Assembly Program”. Genome Research, 9: 868-877.

[19] http://www.clcbio.com/files/usermanuals/CLC_Genomics_Workbench_User_Manual.pdf

Authors

1. Purru Supriya
   Ph.D Research Scholar
   Division of Bioinformatics
   Indian Agricultural Research Institute
   Pusa campus
   New Delhi 110012, India.

2. K V Bhat
   Principal Scientist
   National Bureau of Plant Genetic Resources
   Pusa Campus
   New Delhi 110012, India