Research Article

A Transcriptome Post-Scaffolding Method for Assembling High Quality Contigs

Mingming Liu,[1] Zach N. Adelman,[2] Kevin M. Myles,[2] and Liqing Zhang[1]

[1] *Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA*
[2] *Fralin Life Science Institute, Virginia Tech, Blacksburg, VA 24060, USA*

Correspondence should be addressed to Liqing Zhang; lqzhang@cs.vt.edu

Received 27 January 2014; Revised 18 April 2014; Accepted 4 May 2014; Published 28 May 2014

Academic Editor: Giancarlo Mauri

With the rapid development of high throughput sequencing technologies, new transcriptomes can be sequenced for little cost with high coverage. Sequence assembly approaches have been modified to meet the requirements for *de novo* transcriptomes, which have complications not found in traditional genome assemblies such as variation in coverage for each candidate mRNA and alternative splicing. As a consequence, *de novo* assembly strategies tend to generate a large number of redundant contigs due to sequence variations, which adversely affects downstream analysis and experiments. In this work we proposed TransPS, a transcriptome post-scaffolding method, to generate high quality, nonredundant *de novo* transcriptomes. TransPS shows promising results on the test transcriptome datasets, where redundancy is greatly reduced by more than 50% and, at the same time, coverage is improved considerably. The web server and source code are available.

1. Introduction

The rapid development of the next generation sequencing technologies has catalyzed the development of new genome assembly tools able to handle the volume and complexity of the resulting data. Despite the advantages of next generation sequencing technologies, the length of the sequence generated by these modern instruments is considerably short (~100–300 bp), which poses challenge to sequence assembly algorithms. As with short read genome assembly, transcriptome assembly needs to connect short and sometimes low quality reads. However, transcriptome assembly is even more difficult than genome assembly due to the complication of factors such as highly variable sequencing depth, strand specificity, and transcript variants [1].

There are three typical transcriptome assembly strategies: the reference based strategy, the *de novo* strategy, and the hybrid strategy that combines both reference based and *de novo* strategies. Widely used transcriptome assembly tools include Cufflinks [2] for reference based assembly and Trinity [3] and Oases [4] for *de novo* assembly. These *de novo* assemblers are very sensitive to sequencing errors/polymorphisms, which result in considerable redundancy in the output contigs. Paired-read sequencing technology can help reduce the number of contigs, as the expected distance between read pairs can be used to place contigs in their likely order and orientation. Some assembly tools, such as Trinity, do not include a scaffolding step, while most others provide a scaffolding option only as a built-in function which cannot be independently controlled or effectively used to reduce the number of redundant contigs.

Few studies focus on independently scaffolding a preassembled transcriptome. SSPACE [5] is a tool to scaffold preassembled genome contigs using paired end data. Scaffolding translation mapping (STM) [6] is an optimization method of *de novo* transcriptome assembly by scaffolding contigs and reads. In this work, we developed transcriptome post-scaffolding (TransPS) to scaffold preassembled transcriptome from any desired assembler using a reference species to improve transcriptome coverage and reduce config redundancy. TransPS follows a similar framework as STM, whereby BLASTX coordinates are used to guide the scaffolding process. However, TransPS has several advantages over STM. TransPS is computationally much faster (<3 sec with

Table 1: Data sets description.

Organism	Original number of contigs	Preassembled methods	Reference	Source
D. frontalis [9]	20,966	Newbler 2.5	*Tribolium castaneum*	GAFI01 (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GAFI01)
D. ponderosae [10]	20,540	Newbler 2.5	*Tribolium castaneum*	GAFX01 (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GAFX01)
D. citri [11]	27,821	Velvet v.1.0.19, Oases v.0.1.19	*Acyrthosiphon pisum*	GACJ01 (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GACJ01)
I. ricinus [12]	8,685	Abyss v.1.3.4, Trinity v.2012-06 Cap3 v.1999	*Ixodes scapularis*	GADI01 (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=GADI01)

20 k contigs on an Intel *i*7-2600 machine) and easily adapted for parallel implementation. More importantly, STM relies on CAP3 [7] program to merge two partially overlapping contigs into a single scaffold, which has several drawbacks. For example, because only two sequences are used, CAP3 has no way to break ties between mismatched bases, resulting in sequence artifact at the junctions that can confound downstream analyses.

TransPS overcomes this by using a more precise user-tunable strategy to determine the joint region between two contigs. TransPS is easy to use and can greatly reduce the redundancy in the original contig set. It reduced the number of original contigs by at least 50% for our test datasets, while the qualities of the scaffolded contigs are greatly enhanced in terms of coverage. The web server and source code are available at https://bioinformatics.cs.vt.edu/zhanglab/transps/.

2. Materials and Methods

2.1. Data Set Description. The TransPS pipeline was evaluated by four datasets taken from published NCBI Transcriptome Shotgun Assembly (TSA) projects, including *Dendroctonus frontalis* (southern pine bark beetle, TSA record: GAFI01), *Dendroctonus ponderosae* (black hills beetle, TSA record: GAFX01), *Diaphorina citri* (Asian citrus psyllid, TSA record: GACJ01), and *Ixodes ricinus* (sheep Tick, TSA record: GACI01). Table 1 shows the corresponding reference sequences and preassembled methods used for each organism and other details.

2.2. Post-Scaffolding. Input to TransPS includes the set of preassembled contigs and the BLASTX search results containing the alignment of the contigs to the set of reference amino acid sequences. Figure 1 shows an overview of the procedure. The post-scaffolding procedure follows an align-layout-consensus structure, consisting of three major stages. First, original contigs were used to search for their best alignments to the reference amino acid sequence database using BLASTX and categorized into three groups (accepted, unused, and scaffolding) in terms of the alignment results. This alignment procedure is summarized in Figure 1 above the dashed line. Second, contigs in the scaffolding group are placed in the right order and orientation according to their aligned coordinates to the reference. Third, those nonredundant contigs (contigs in the scaffolding group) matching to the same reference sequences are scaffolded/assembled into one supercontig. Layout stage and consensus stage are shown below the dashed line in Figure 1. The algorithm for scaffolding is illustrated in the supplementary material (see Algorithm 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/961823). Basically, the contigs are scaffolding in terms of two cases as shown in supplementary Figure S1, overlapping contigs and nonoverlapping contigs. Please see supplementary material for the detailed procedure.

The output of TransPS includes contigs in three different groups, accepted contigs (*A*), scaffolded contigs (*S*), and unused/redundant contigs (*R*). The contigs in *R* could be the real redundancy due to genetic variants or different transcripts due to alternative splicing. In order to provide users with detailed account of how the assembly of a particular config is done, a matching map between the original contigs and the corresponding reference protein sequences used in the scaffolding algorithm is also provided.

3. Results

Table 2 shows the number of scaffolds from the original contigs that match at least one reference sequence. The average number of configs per scaffold is between 2 and 3. The contigs in the redundant set (*R*) account for most of the original contigs (% Reduction). Especially for *D. citri*, the redundant sequence researches 90%. Please see Table 3 for detailed distribution of original contigs in three different groups. The majority of the sequences fall into the unused group, which means that these sequences are redundant for scaffolding. However, the sequences in the unused/redundant group may contain paralog or alternative splicing sequences of the corresponding sequences used for scaffolding.

The new scaffolded contigs are compared with the original contigs by measuring the coverage ratio against the matched reference protein sequence. Coverage ratio was calculated as the matched percentage between the scaffolded (or original) contig and the matched reference protein sequence

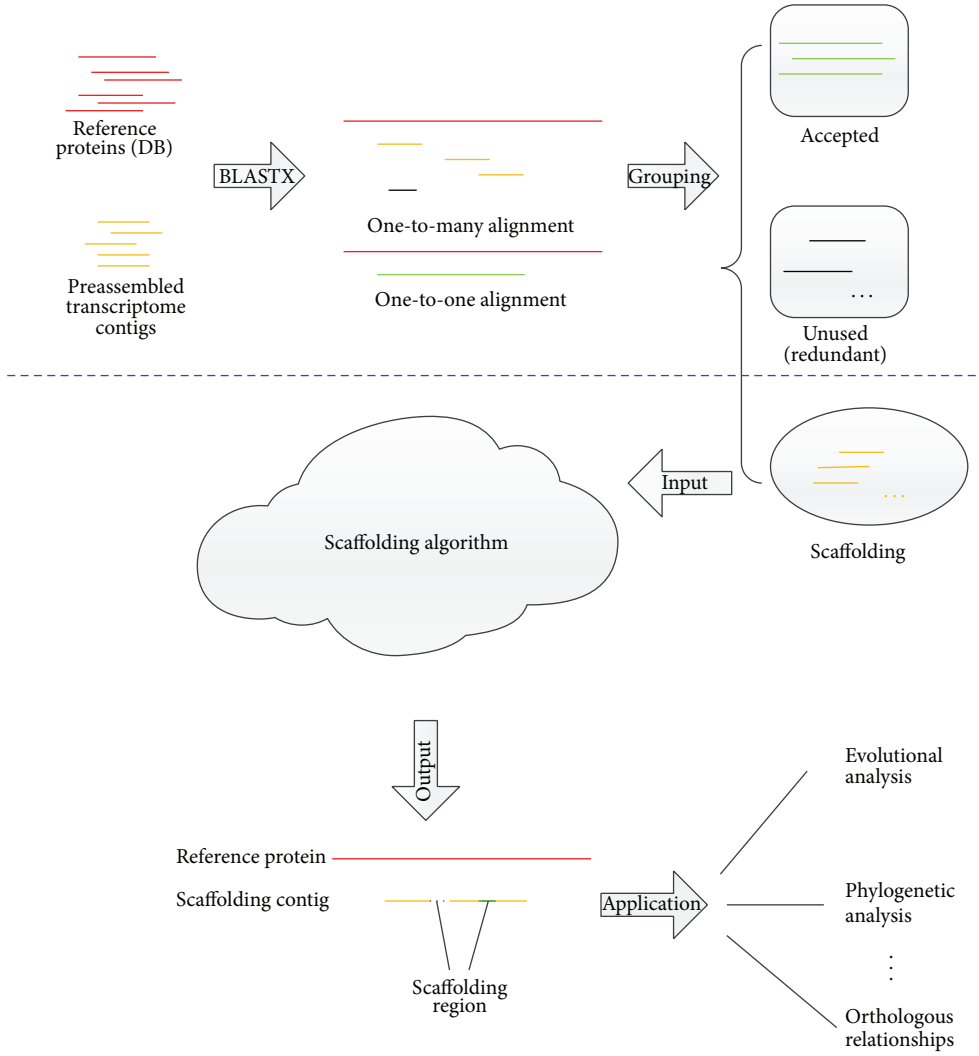

Figure 1: An overview of TransPS.

Table 2: Number of contigs used for scaffolding from different organisms.

Organism	Number of matched original contigs	Number of scaffolds	Number of contigs/scaffold	% Reduction
D. frontalis	15,095	496	2.17	70
D. ponderosae	16,457	433	2.15	74
D. citri	16,732	165	2.21	90
I. ricinus	7,787	71	2.04	57

Number of contigs/scaffold: average number of original contigs per scaffold.
% Reduction: percentage of "redundant" contigs removed.

Table 3: Distribution of original contigs in different groups.

Organism	Number of contigs for scaffolding	Number of contigs accepted	Number of contigs unused (redundant)	Total
D. frontalis	1,075	3,900	10,120	15,095
D. ponderosae	930	3,835	11,692	16,475
D. citri	365	1,525	14,842	16,732
I. ricinus	145	3,275	4,365	7,787

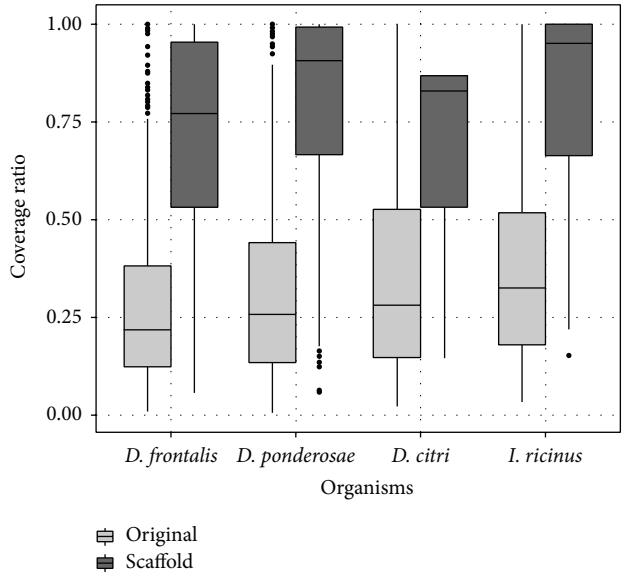

Figure 2: Coverage comparison between scaffolds and original contigs for different data sets.

in a global sequence alignment. Global sequence alignments between protein sequences and nucleotide sequences were performed using NAP [8]. The redundant and accepted sequences in the original transcriptome set were removed from comparison. As shown in Figure 2, the scaffolded contigs have a much higher coverage ratio compared with the original ones and the median of the scaffolded group is significantly higher than the original set of contigs in all the tested datasets.

4. Conclusion

This paper proposed a referenced-based transcriptome post-scaffolding method that considerately reduces the redundancy of the original *de novo* assembled contigs while greatly increasing the coverage ratio. As more genomes become available in the public databases, the more likely it is that there will be a species with a complete sequenced genome closely related to a new genome that one is interested to study for its transcriptomes. In this case, a post-assembly step takes advantage of the information in existing sequenced genomes that can better guide the *de novo* assembly process.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to thank Dr. Lenwood S Heath for help with the code and discussions.

References

[1] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics*, vol. 12, no. 10, pp. 671–682, 2011.

[2] R. Adam, H. Pimentel, C. Trapnell, and L. Pachter, "Identification of novel transcripts in annotated genomes using RNA-seq," *Bioinformatics*, vol. 27, no. 17, pp. 2325–2329, 2011.

[3] M. G. Grabherr, B. J. Haas, M. Yassour et al., "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology*, vol. 29, no. 7, pp. 644–652, 2011.

[4] H. S. Marcel, D. R. Zerbino, M. Vingron, and E. Birney, "Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels," *Bioinformatics*, vol. 28, no. 8, pp. 1086–1092, 2012.

[5] M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, "Scaffolding pre-assembled contigs using SSPACE," *Bioinformatics*, vol. 27, no. 4, pp. 578–579, 2011.

[6] Y. Surget-Groba and J. I. Montoya-Burgos, "Optimization of de novo transcriptome assembly from next-generation sequencing data," *Genome Research*, vol. 20, no. 10, pp. 1432–1440, 2010.

[7] X. Huang, "A contig assembly program based on sensitive detection of fragment overlaps," *Genomics*, vol. 14, no. 1, pp. 18–25, 1992.

[8] X. Huang and J. Zhang, "Methods for comparing a DNA sequence with a protein sequence," *Bioinformatics*, vol. 12, pp. 497–506, 1996.

[9] C. I. Keeling, M. M. S. Yuen, N. Y. Liao et al., "Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest," *Genome Biology*, vol. 14, no. 3, p. R27, 2013.

[10] C. I. Keeling, H. Henderson, M. Li et al., "Transcriptome and full-length cDNA resources for the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major insect pest of pine forests," *Insect Biochemistry and Molecular Biology*, vol. 42, no. 8, pp. 525–536, 2012.

[11] J. Reese, S. L. Johnson, W. B. Hunter et al., "Characterization of the Asian citrus psyllid transcriptome," *Journal of Genomics*, vol. 2, pp. 54–58, 2012.

[12] A. Schwarz, B. M. von Reumont, J. Erhart et al., "De novo Ixodes ricinus salivary transcriptome analysis using two different next generation sequencing methodologies," *The FASEB Journal*, vol. 27, no. 12, pp. 4745–4756, 2013.