Proc. Assoc. Advmt. Anim. Breed. Genet. 22:545-548

MULTI-ASSEMBLER PIPELINE FOR THE *DE NOVO* TRANSCRIPTOME ASSEMBLY ON NON-MODEL ORGANISMS: THE CASE OF THE BLACK TIGER PRAWN (*Penaeus monodon*)

R. Huerlimann[1,2], L. Gordon[1,3], J. Goodall[1,4], K. Siemering[1,3], N. Wade[1,4], M. Sellars[1,4], G.E. Maes[1,5] and D.R. Jerry[1,3]

[1] ARC Research Hub for Advanced Prawn Breeding
[2] Centre for Sustainable Tropical Fisheries and Aquaculture, James Cook University, Townsville, Queensland, Australia
[3] Australian Genome Research Facility, Melbourne, Victoria, Australia
[4] CSIRO Agriculture and Food, Brisbane, Queensland, Australia
[5] Laboratory of Biodiversity and Evolutionary Genomics, KU Leuven, Leuven, Belgium

SUMMARY

A high-quality transcriptome is important for genome annotation and differential gene expression studies, but a comprehensive transcriptome assembly for non-model species like prawns is still challenging. Most assemblies are carried out in a single assembler; however, recent publications have shown that while different assemblers produce a shared core of contigs, they each also produce unique contigs. Using the transcriptome assembly of the black tiger prawn (*Penaeus monodon*) as an example, we merged the assemblies generated by four transcriptome assemblers, and incorporated newly published best practices into a novel pipeline. This multi assembler approach produces an improved, less redundant assembly which is also transferable to other non-model species. Therefore, in contrast to older approaches, using multiple assemblers improves assemblies by using the strengths of different assemblers, while decreasing their weaknesses.

INTRODUCTION

Complete transcriptomes are an important resource that can be used for differential gene expression studies (Wang et al. 2009), genome annotation (Saha et al. 2002), and more recently genome scaffolding (Song et al. 2016), among other applications. The two main methods for transcriptome assembly are genome guided and *de novo*. A genome guided transcriptome assembly is computationally simpler, but depends on the completeness of the reference genome and is impeded by sequencing errors and isoforms (Grabherr et al. 2011). In contrast, the *de novo* approach is used when no reference genome is available, but is computationally more complicated, especially for large data sets. While model species often have a variety of genomic resources available, these are by definition lacking for non-model species.

Recently the number of transcriptome assemblers has exploded from the limited number that was available ten years ago. These various assemblers have different strengths and weaknesses, resulting in contigs that are unique to a specific tool (Smith-Unna et al. 2016). Trinity, one of the most popular assemblers (cited in 2865scientific articles based on Web of Science as of January 2017), can assemble most transcripts including different isoforms, or recent gene duplications (Grabherr et al. 2011), although with the drawback of the final transcriptome often including a large number of misassembled contigs. Another bias in transcriptome assembly is introduced by sequencing errors or increased heterozygosity due to sequencing multiple individuals, both leading to more fragmented assemblies.

Recently, MacManes (2016) published recommendations for the transcriptome assembly of non-model species, suggesting to only sequence tissues from one individual, which is not always possible (for example for small organisms), and to use Rcorrector to reduce sequencing errors.

BUSCO (Simão et al. 2015) and TransRate (Smith-Unna et al. 2016) are used to assess the final transcriptome and remove low coverage reads. Transfuse has been made available (github.com/cboursnell/transfuse), which can merge multiple transcriptome assemblies from different individuals or different assemblers using reads to improve the final assembly. Therefore, the aim of this study is to compare the assembly of individual samples using four assemblers (Trinity, Bridger, BinPacker and IDBA-tran) merged using Transfuse with a single assembly in Trinity.

MATERIALS AND METHODS

Samples were collected from five different individuals of *Penaeus monodon* (3 female, 2 males). Two replicates each of the following tissues were sent for sequencing: eyestalk, female gonad, male gonad, gills, haemolymph, hepatopancreas, muscle and stomach. One sample each from gills, haemolymph and stomach failed the library preparation, resulting in 13 successfully sequenced samples. Sequencing was carried out at the Australian Genome Research Facility in Melbourne, Australia, on a HiSeq 2500 using a 125 bp paired-end, strand-specific, ribo-minus protocol. On average, 20 million reads were obtained per sample with an average of 91% bases ≥Q30.

Two assembly approaches were used: one assembling all samples collectively in a single assembler (single assembly, Fig. 1a) and the other where each sample was assembled individually in four assemblers (multi assembly, Fig. 1b). The transcriptome generally followed the recommendations of MacManes (2016). For both approaches, the individual samples were collectively error corrected using RCorrector version 1.0.2 (Song et al. 2015).

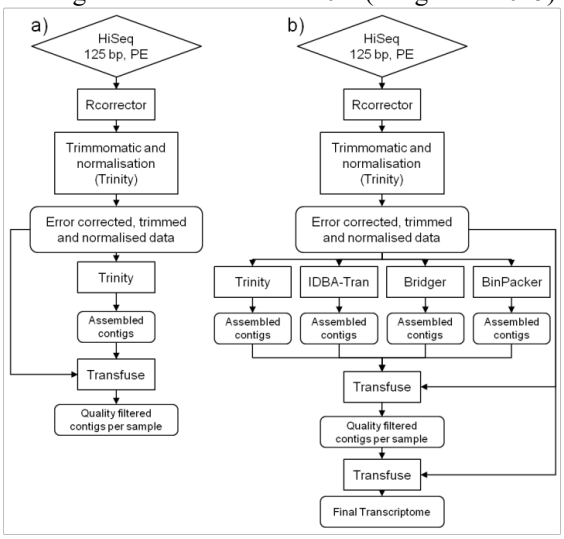

Figure 1. Assembly pipeline for a) single assembler approach and b) multi assembler approach

Using Trinity 2.2.0 (Grabherr et al. 2011), adapter and bases with a Phred score <2 were trimmed with trimmomatic (Bolger et al. 2014) and reads were in-silico normalised. For the single assembly, the 13 samples were concatenated and assembled in Trinity. The multi assembly was carried out for each sample individually in Trinity 2.2.0, BinPacker 1.0 (Liu et al. 2016), Bridger r2014-12-01 (Chang et al. 2015) and IDBA-Tran 1.1.1 (Peng et al. 2013). For IDBA-Tran the k60 transcriptome was used for downstream processing. For both approaches transfuse version 0.5.0 (https://github.com/cboursnell/transfuse) was used to remove redundant contigs, and also merge

the individual assemblies for the multi assembly approach. For the multi assembly, the transcriptomes of the four assemblers were merged by sample using transfuse in a first round. In a second round, samples were then merged with transfuse into a final transcriptome.

The two final assemblies were annotated using Blast2Go (Conesa et al. 2005) against the SwissProt database (Boeckmann et al. 2003) downloaded on 12. January 2017. The quality of both assemblies was assessed with BUSCO version 1.2 (Simão et al. (2015) using the arthropod set and TransRate version 1.0.3 (Smith-Unna et al. 2016).

RESULTS AND DISCUSSION

The aim of this study was to compare the use of multiple assemblers (multi: Trinity, BinPacker, Bridger, IDBA-Tran) on individual samples with a combined approach using only one assembler (single: Trinity). When comparing the two approaches, the multi assembly resulted in a more manageable number of contigs and lower duplication levels; however, at the price of completeness (Fig. 2a). The number of fragmented contigs was comparable in both approaches.

The raw assembly in Trinity resulted in 280,846 contigs, with 85% of the arthropod Benchmarking Universal Single-Copy Orthologs (BUSOs) complete, of which 24% were duplicated (Fig. 2a). After merging with Transfuse, this was reduced to 212,526 contigs with C:83%[D:24%] and 36,086 contigs annotated with SwissProt. In contrast, the sum of the contigs of all samples in the four assemblers added up to 2,412,355 contigs. Merging the individual assemblies by sample reduced the total number of contigs in the 13 samples combined to 392,349 with a C:70%[D:11] (Fig. 2a). The second round of merging of the individual samples into a final transcriptome resulted in 73,406 contigs with C:70%[D:10%] of which 17,885 contigs were annotated with SwissProt. The single assembly resulted in 10,470 unigenes (29% of annotated contigs), while the multi assembly resulted in 8,450 unigenes (47% of annotated contigs), with 7071 shared unigenes (Fig. 2b). The BUSCO analysis and unigene comparison shows that while the single assembly approach produces more annotated contigs, most of these contigs are duplicated.

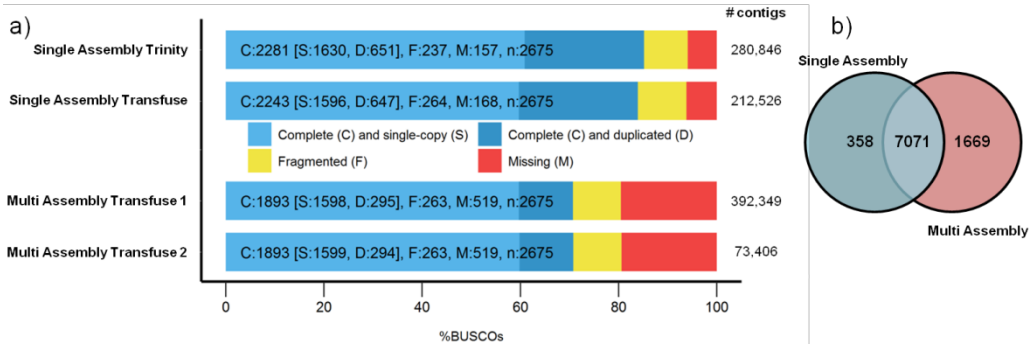

Figure 2. a) Benchmarking Universal Single-Copy Orthologs (BUSOs) values and #contigs for the two approaches using single and multiple assemblers. b) Venn diagram showing number of shared and unique genes identified in Blast2Go

Table 1. Quality assessment using TransRate. Scores and percentages derived from mapping reads to the assembly

	Assembly Score	# of contigs	Assembly Size	N50	Percent mapping	Percent bases uncovered	Percent contigs low covered
Single	0.48	212,526	171.1 Mb	1571	81.8	35.6	80.2
Multi	0.36	73,406	65.5 Mb	1687	82.7	18.2	36.9

Comparing the TransRate mapping scores of the two assemblies strategies, the multi assembly exhibited higher support for the contigs. While the single assembly has a slightly higher assembly score of 0.48 compared to 0.36 in the multi assembly, the percentages of reads mapping to the transcriptome (81-83%) and N50 values (1500 bp to 1687 bp) were comparable (Table 1). However, the multi assembly had a lower proportion of bases that were not covered by reads (18.2% compared to 35.6%) and fewer contigs with low read coverage reads (36.9% compared to 80.2%).

Compared to two other multi-tissue decapods assemblies, the present assembly lies between the assemblies of the two freshwater crayfish *Astacus astacus* (Theissinger et al. 2016) and *Cherax quadricarinatus* (Tan et al. 2016). The *A. astacus* assembly combined four tissues (abdominal muscle, hepatopancreas, ovaries and green glands) and used Trinity only for the assembly. This resulted in 158,649 non-redundant contig and 45,415 contigs after filtering for lowly expressed transcripts, with a BUSCO score of C:64%[D:27%] and a TransRate assembly score of 0.20. In contrast, the *C. quadricarinatus* assembly combined five tissues (heart, kidney, hepatopancreas, central nerve cord, and testis) from a single individual and used both Trinity and IDBA-Tran for the assembly and merged the contigs using Corset (Davidson et al. 2014). This resulted in 180,635 contigs between Trinity and IDBA-Tran, and a final assembly of 44,525 contigs, with a BUSCO score of C:74%[D:7%]

Based on these results, using multiple assemblers in conjunction with a merging software like Transfuse highly reduces the number of contigs to a more realistic number by removing redundant contigs. However, while the multi assembler approach in this study also reduced the over-inflation of contigs commonly found in Trinity, it came at the cost of completeness of the assembly. While older approaches to transcriptome assembly relied on a single assembler, the field is now moving towards using multiple assemblers which improves assemblies by using the strengths of different assemblers, while decreasing their weaknesses.

REFERENCES

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C. and Phan I. (2003). *Nucleic acids research* **31**: 365-370.
Bolger A. M., Lohse M. and Usadel B. (2014). *Bioinformatics*: btu170.
Chang Z., Li G., Liu J., Zhang Y., Ashby C., Liu D., et all. (2015). *Genome biology* **16**: 1.
Conesa A., Götz S., García-Gómez J. M., Terol J., et all. (2005). *Bioinformatics* **21**: 3674-3676.
Davidson N. M. and Oshlack A. (2014). *Genome biology* **15**: 1.
Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I., Adiconis X., Fan L., Raychowdhury R. and Zeng Q. (2011). *Nature biotechnology* **29**: 644-652.
Liu J., Li G., Chang Z., Yu T., Liu B., et all. (2016). *PLoS Comput Biol* **12**: e1004772.
MacManes M. D. (2016). *bioRxiv*: 035642.
Peng Y., Leung H. C., Yiu S.-M., Lv M.-J., et all. (2013). *Bioinformatics* **29**: i326-i334.
Saha S., Sparks A. B., Rago C., Akmaev V., Wang C. J., Vogelstein B., Kinzler K. W. and Velculescu V. E. (2002). *Nature biotechnology* **20**: 508-512.
Simão F. A., Waterhouse R. M., Ioannidis P., et all. (2015). *Bioinformatics*: btv351.
Smith-Unna R., Boursnell C., Patro R., et all. (2016). *Genome research*: gr. 196469.196115.
Song L. and Florea L. (2015). *GigaScience* **4**: 1.
Song L., Shankar D. S. and Florea L. (2016). *The Plant Genome*.
Tan M. H., Gan H. M., Gan H. Y., Lee Y. P., Croft L. J., Schultz M. B., Miller A. D. and Austin C. M. (2016). *Organisms Diversity & Evolution* **16**: 185-200.
Theissinger K., Falckenhayn C., Blande D., Toljamo A., Gutekunst J., Makkonen J., Jussila J., Lyko F., Schrimpf A. and Schulz R. (2016). *Marine genomics*.
Wang Z., Gerstein M. and Snyder M. (2009). *Nature reviews genetics* **10**: 57-63.