Wide-and-shallow vs. narrow-and-deep: comparing two transcriptome assemblies of *S. serrata*

Anish M.S. Shrestha[1,*], Crissa Ann I. Lilagan[2], and Ma. Carmen A. Lagman[2]
[1] Dept. of Software Technology, College of Computer Studies, De La Salle University, Manila
[2] Dept. of Biology, College of Science, De La Salle University, Manila
**Corresponding Author: anish.shrestha@dlsu.edu.ph*

Abstract : RNA-seq is a popular, state-of-the art technique for transcriptomic studies. An ideal RNA-seq experiment is one with many replicates, each of which is sequenced deeply. However, since cost can be a limiting factor, there is often a choice to be made between high replicate count or high sequencing depth. We explore the impact of this choice on the quality of transcriptome assemblies of two RNA-seq datasets of mangrove crabs.

Key Words: RNA-seq; transcriptome assembly; mangrove crabs

1. INTRODUCTION

RNA-seq is a modern sequencing-based technique for transcriptome profiling, and has become the de facto standard approach for measuring genome-wide gene expression and its variation across samples (Stark, Grzelak, & Hadfield, 2019). With the proliferation of high-throughput sequencing technologies, RNA-seq is being employed to study the transcriptome of an ever-expanding repertoire of organisms.

When designing an RNA-seq experiment, it is important to ensure an ample number of replicates in order to mitigate the effects of technical and biological variability (Conesa, et al., 2016). In addition, it is desirable to deeply sequence the samples in order to capture a true snapshot of the expression level of all genes, even ones with lower expression levels. However, the cost of sequencing can be a limiting factor, and for a fixed cost, there are two possible design choices: (1) *wide-and-shallow,* in which a large number of replicates are sequenced at a low depth, or (2) *narrow-and-deep*, in which a fewer number of replicates are sequenced at a high sequencing depth.

It is not immediately clear which of the two choices yields better analysis results. A higher replicate count provides a better handle to assess variability and a higher statistical power to identify differentially expressed genes. On the other hand, a higher sequencing depth means the ability to capture transcripts with low expression levels, which would otherwise be drowned out by highly expressed genes in a low-depth sample.

Here we attempt to assess the trade-off between replicate count and sequencing depth by utilizing two previously generated RNA-seq datasets of mangrove crabs (*S. serrata*) – one of which is wide-and-shallow and the other narrow-and-deep. Transcriptome assembly is the first major step in the bioinformatic analysis of non-model species like *S. serrata*, and therefore we use the quality of transcriptome assembly as assessment metric.

2. METHODOLOGY

2.1 Data

We used two RNA-seq datasets obtained from the gill tissue of *S. serrata.* Both samples had a total of roughly 95 million pairs of 100bp-long paired-end reads, which were spread across 3 replicates for the narrow-and-deep dataset, and 6 replicates for the wide-and-shallow dataset.

2.2 Bioinformatics Analysis

For each dataset, we applied the following bioinformatics procedures. First, potential rRNA-derived reads were filtered out using SortmeRNA (Kopylova, Noé, & Touzet, 2012). Next, low-quality reads were filtered out and adapter sequences trimmed using Trimmomatic (Bolger, Lohse, & Usadel, 2014). From the cleaned reads, we obtained a transcriptome assembly using Trinity (Haas, et al., 2013).

2.3 *Transcriptome Quality Metric*

We used three different kinds of metrics, described below, to measure the quality of the transcriptome assembly.

2.3.1 Contig length statistics

The statistic Nx is defined as the contig length such that x% of the nucleotides in the assembly are found in contigs of length at least Nx. To reduce exaggerated values due to long transcripts with multiple isoforms, we recompute the Nx values after choosing one longest isoform per 'gene' (as defined by Trinity).

2.3.2 *Expression-level-filtered contig length statistic*

Trinity suggests using ExN50 instead of N50 as a better measure of assembly contiguity. ExN50 is the N50 value computed based on a subset of the transcripts obtained by excluding those with low expression, such that the subset accounts for x% of gene expression. The expression levels were estimated using Rsem (Li & Dewey, 2011) based on alignments of reads to the assembly computed by Bowtie2 (Langmead, Wilks, Antonescu, & Charles, 2018).

2.3.3 *Protein-coding content*

Apart from the characterization based on contig lengths, we assessed contigs for their protein-coding content. We used Blastx to count the number of protein sequences from the latest release of SwissProt that (almost-)fully align to at least one of the contigs in the assembly.

3. RESULTS AND DISCUSSION

3.1 Length statistics

For the wide-and-shallow dataset, Trinity reported 138,331 'genes' (as defined by Trinity) spread over 217,674 transcripts. For the narrow-and-deep dataset, the numbers were 127,180 and 167,710, respectively. Contig length statistics are shown in Table 1.

The assembly for the wide-and-shallow dataset is significantly larger in terms of total bases, even after removing redundancy by choosing only one isoform per gene. All length statistics – mean, median, N50 – are favorable towards wide-and-shallow.

Table 1 Contig length statistics

	Wide-shallow (bp)	Narrow-deep (bp)
N10	5022	3616
N20	3364	2422
N30	2353	1723
N40	1627	1233
N50	1077	888
Mean	679	623
Median	356	358
Total bases	93,968,507	79,254,309

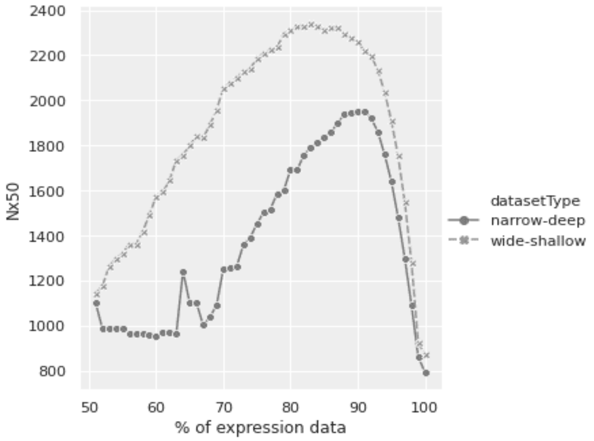

Figure 1. ExN50 value for the two datasets, for different values of x.

3.2 Expression-level-filtered contig length statistic

The ExN50 values for the two datasets are shown in Figure 1. They are consistently better for wide-and-shallow than narrow-and-deep. This is especially so around the practically more interesting 80-90 % expression cut-off value, where the two plots peak. The peak value for wide-shallow is over 2200bp, which is remarkably higher than under 2000bp for narrow-and-deep.

3.3 Protein-coding content

Table 2 summarizes the results of aligning protein sequences in SwissProt-Uniprot dataset version 2020_01 to the assembled contigs.

Of particular interest are the protein entries which align almost full length (91–100) to one of the contigs in the assembly, since these indicate almost fully assembled transcripts. Here again the assembly of wide-and-shallow handily outperforms narrow-and-deep.

Table 2 Evaluating the assemblies for protein-coding content. For proteins that align to multiple transcripts, only one with the lowest E-value is chosen.

Percentage of protein sequence length aligned	**Number of protein entries**	
	Wide-shallow	Narrow-deep
91 – 100	3775	3069
81 – 90	1406	1304
71 – 80	1020	989
61 – 70	1055	867
51 – 60	1089	974

4. CONCLUSIONS

Our results indicate that, no matter what quality metric we choose to measure transcriptome assembly quality, it is better to sample wide-and-shallow rather than narrow-and-deep. Contrary to what we initially expected, there seems to be no trade-off between replicate count and depth.

However, this is just based on just one pair of datasets and needs to be validated with more data. A limitation of our dataset is that these were obtained from different biological samples at different time points, and we have not controlled for several factors such as sex, age, temperature, sequencing device, etc.

While we only looked at transcriptome assembly quality, it is interesting to assess the impact of the two design choices on the results of differential gene expression analysis.

5. REFERENCES

Bolger, A. M., Lohse, M., & Usadel, B. (2014, 8). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., . . . Mortazavi, A. (2016, 1). A survey of best practices for RNA-seq data analysis. Genome Biology, 17. doi:10.1186/s13059-016-0881-8

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., . . . Regev, A. (2013, 7). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8, 1494–1512. doi:10.1038/nprot.2013.084

Kopylova, E., Noé, L., & Touzet, H. (2012, 12). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics (Oxford, England), 28(24), 3211–3217. doi:10.1093/bioinformatics/bts611

Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2018, 7). Scaling read aligners to hundreds of threads on general-purpose processors. (J. Hancock, Ed.) Bioinformatics, 35, 421–432. doi:10.1093/bioinformatics/bty648

Li, B., & Dewey, C. N. (2011, 8). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12. b

Stark, R., Grzelak, M., & Hadfield, J. (2019, 11). RNA sequencing: the teenage years. Nature reviews. Genetics, 20(11), 631-656. doi:10.1038/s41576-019-0150-2