Evaluating the performance of available tools for building de novo transcriptome hybrid assemblies by combining reads of different length

Jacques Lagnel, Tereza Manousaki, Costas Tsigenopoulos, Anastasia Tsagkarakou

Evalua&ng the performance of available tools for building de novo transcriptome hybrid assemblies by combining....

Conference Paper · November 2016

4 authors, including:

Jacques Lagnel
INRA French National Institute for Agricultural Research

Tereza Manousaki

Evaluating the performance of available tools for building *de novo* transcriptome hybrid assemblies by combining reads of different length

Lagnel J.[1], Manousaki T.[1], Tsigenopoulos C.[1], Tsagkarakou A.[2]

[1] Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Heraklion, Greece
[2] Hellenic Agricultural Organisation - "DEMETER", NAGREF - Institute of Olive Tree, Subtropical Crops and Viticulture, Heraklion, Greece

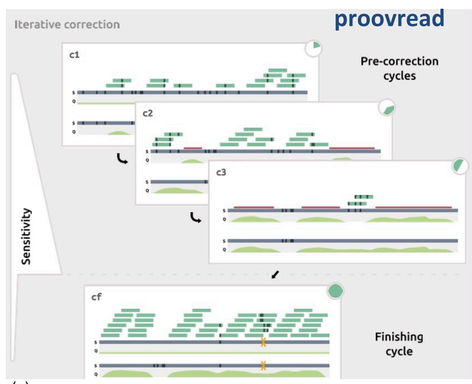
A. Background

One of the greatest avenues opened by next generation sequencing (NGS) technologies is the possibility to sequence on large scale species with no prior genomic information, the so called non-model species, at reasonable cost. This has revolutionized the way biologists approach questions, allowing the implementation of experiments previously considered impossible. However, unlike model species, non-model species lack a reference genome or transcriptome assembly, which encouraged the community to develop various strategies to deal with the challenge of constructing a reference assembly de novo. Short reads often lead to the assembly of incomplete contigs with low error rate, while longer reads improve greatly the coverage of the sequenced transcripts but are error-prone. A generally recognized need is to increase the length of used reads, but current technologies cannot offer long reads without increasing the error rate at the same time. Although the combination of short and long reads promises to bring together the advantages of both read types, i.e. long reads and low error rate, it is questionable whether this is applicable with the available tools. Here, we aim at comparing de novo assemblies incorporating reads produced by different sequencing platforms, and in particular short reads from Illumina HiSeq2000 (100 bp PE), a bit longer reads produced by Illumina MiSeq (300bp PE), and long reads produced by the PacBio technology (1500 bp). We use as test case the transcriptome sequencing of the whitefly *Bemisia tabbaci*[1] and assess the efficiency of available software developed to combine reads with different lengths. Finally, we provide basic guidelines for improving the assemblies produced for non-model species and propose future directions to deal with this challenge faced by most researchers of organismal biology.

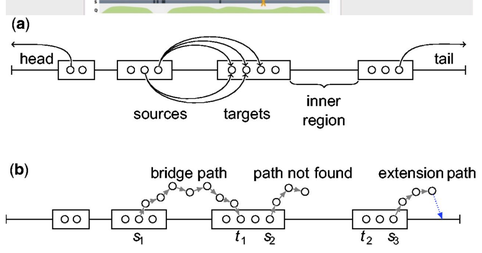
B. Hybrid assembly strategy algorithms

Hybrid assembly uses two sets of reads: the reference read set, whose error rate is assumed to be small, and the PacBio read set, which is then corrected using the reference set. Typically, the reference set contains Illumina reads.

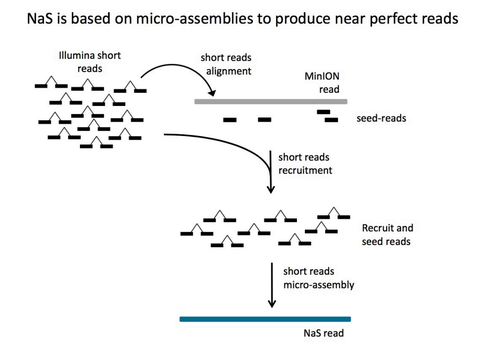
Short read alignment based proovread[2], PBcR[3]



de Bruijn graph based LoRDEC[4]



Micro-assemblies based NaS[5]



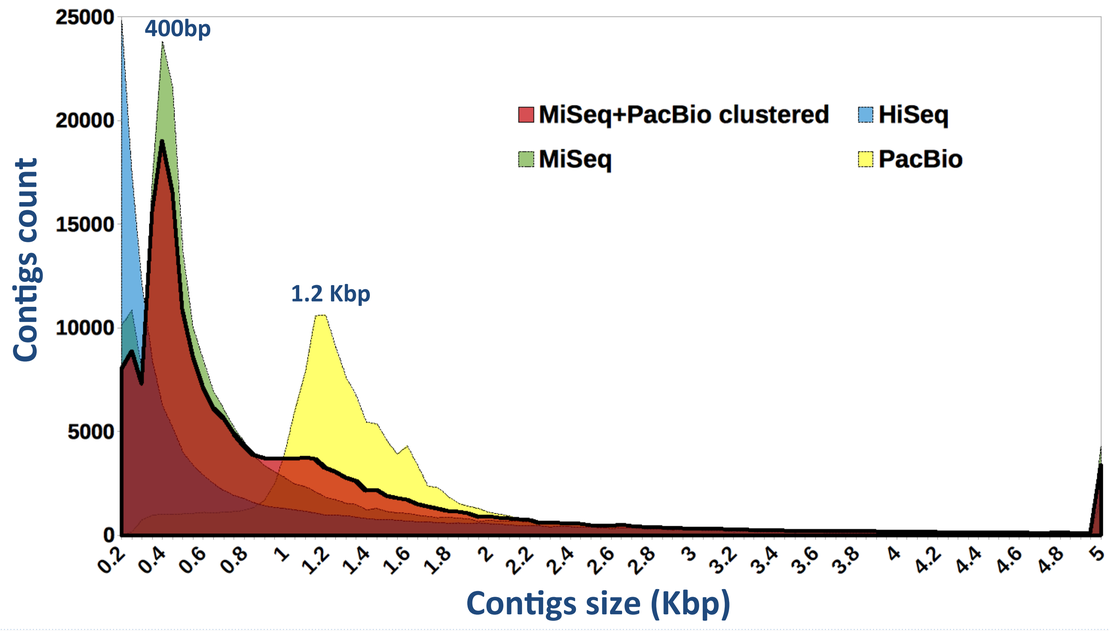
C. The experiment

Total RNA preparation (pool of 1500 adults)

Normalization of cDNA for equalization of gene representation

Library size selection	300bp	1200bp	550bp
	HiSeq2000 100 bp PE	PacBio 2 SMRT cells	MiSeq 300 bp PE
Reads count (M)/ Gbp	90.2 / 17.8	1 / 1.4	18.9 / 5.6
	de novo assembly Trinity[6]	hybrid PacBio Long Reads correction pipeline "proovread"	*de novo* assembly Strand specific Trinity

D. Results - Contigs size distribution



E. Results - Assembly summary metrics

	HiSeq*	MiSeq*	PacBio	MiSeq+PacBio*	MiSeq+PacBio clustered
# of contigs	130,801	174,811	55,023	187,208	199,783
Sum length (Mbp)	125	178	62	190	207
Average length (bp)	958	1,016	1,132	1,017	1,037
Min size (bp)	201	224	71	201	201
Max size (bp)	33,419	41,172	3,985	41,172	41,172
N50	2,031	1,586	1,233	1,574	1,448
Busco[7] (%)	85	88	36	88	89

* Trinity assembly

Conclusion

Read length is detrimental to the assembly quality. Regardless of the massive short reads used, long reads led to the greatest improvement of the assembly. The best strategy to reconstruct a de novo transcriptome assembly is to combine MiSeq and PacBio reads and cluster the contigs.

REFERENCES

1 Ilias A, et al. Transcription analysis of neonicotinoid resistance in Mediterranean (MED) populations of B. tabaci reveal novel cytochrome P450s, but no nAChR mutations associated with the phenotype. BMC Genomics. 2015;16:939.
2 Hackl T, Hedrich R, Schultz J, Foester F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics. 2014;30:3004–11.
3 Koren et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30:692–.
4 Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30:3506–14.
5 Madoui et al. Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics. 2015;16:327.
6 Grabherr et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
7 Simão et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. Oxford University Press; 2015;31:3210–2.