Improving automated de-novo transcriptome definition in non-model organisms by integrating manually defined gene information

Ester Feldmesser, Shilo Rosenwasser, Assaf Vardi, Shifra Ben-Dor✉
Weizmann Institute of Science, Rehovot, Israel

Motivation and Objectives

Non-model organisms are of great ecological and economic significance, consequently the understanding of their unique metabolic pathways by investigating their gene expression profiles is essential. The bloom-forming alga *Emiliania huxleyi* is a cosmopolitan unicellular photoautotroph that plays a prominent role in the marine carbon cycle. Its intricate calcite coccoliths account for a third of the total marine CaCO3 production, making it highly susceptible to future ocean acidification.

The advent of next generation sequencing (NGS) technologies and corresponding bioinformatics analysis tools has allowed the definition of transcriptomes in non-model organisms and automated transcriptome assemblies have become common over time. Several methods that integrate de-novo assembly together with genome based assembly have been proposed for non-model organisms (Martin and Wang, 2011). Yet, there are many open challenges in defining genes, where genomes are not available or incomplete. The available genome assembly of *E. huxleyi* is a draft and was constructed from 454 reads in one round of assembly. A large number of available unassembled reads, numerous repeats and duplications, as well as holes in the genome, indicated that the genome alone would not provide a good basis for building transcripts.

In spite of the high numbers of transcriptome assemblies that have been performed, quality control of the transcript building process is rarely performed, if ever. To test and improve the quality of the automated transcriptome definition, we used 63 manually defined and curated genes, several of them experimentally validated. After each step in the automated definition pipeline, the presence of the manually defined genes was checked, allowing troubleshooting of missing genes and improving our pipeline. To the best of our knowledge, this is the first time that an automated transcript definition is subjected to quality control using manually defined and curated genes and thereafter the process is improved.

Methods

Three different approaches were applied in parallel to the automated definition of *E. huxleyi* transcripts, two of them utilizing the read data. The first was de-novo assembly using CLC Assembly Cell (http://www.clcbio.com/products/clc-assembly-cell/) and then CAP3 (Huang and Madan, 1999) to remove redundancy. The second was a genome-based alignment, in which the reads of each sample were aligned separately to the genome using TopHat (Trapnell *et al*, 2009). After the alignment, Cufflinks and Cuffcompare (Trapnell *et al*, 2010) were applied to all the TopHat outputs to define transcripts. In the third approach, available *E. huxleyi* ESTs were clustered using TGICL (http://compbio.dfci.harvard.edu/tgi/software/).

In parallel, genes were manually defined. Protein sequences of the target genes from human, Arabidopsis and yeast, were compared to the *E. huxleyi* genome on the JGI genome website. Hits were inspected to see if any transcript or EST evidence was available. If there were ESTs available, they were assembled into transcripts, and compared to the predictions available. When (NGS) reads became available, they were used to correct and improve the gene definition. If more than one hit were retrieved, each successive hit was also checked to see if it was truly an independent hit, representing a family member, or a duplication, which was then classified as real or artificial. If no ESTs were available as an anchor for a predicted transcript, then a combination of reads (if available), prediction based on blast hits and the JGI predictions were used to construct a transcript. If there was no genomic hit, searches were performed against *E. huxleyi* ESTs, in order to identify sequences that might not have been mapped to the genome. Transcripts were then constructed and extended as far as possible by running successive blasts.

The improvements added to the automatic pipeline after troubleshooting missing manually defined genes were: (1) The Partek (http://www.partek.com) software was applied to find regions to which reads were aligned, but where

Cuffcompare transcripts were not defined, (2) Artificially fused transcripts were split using in-house developed PERL scripts and (3) At the end, two different clustering algorithms were applied to the collection of potential transcripts, TGICL that strongly removed redundancy but loses genes and CAP3 that does not lose genes, but leaves redundancy in the collection. The two approaches were integrated.

Results and Discussion

The final transcriptome collection included 75092 transcripts. The transcript lengths ranged from 301 to 34193 base pairs (bp), 34680 of the transcripts have a length of more than 1000 bp and 23993 are between 500 and 1000 bp. Open reading frames (ORFs) covered the entire transcript in 44% of the transcripts and in approximately 70% of the transcripts, the ORFs covered more than 80%. A high percentage of the reads (80%) were successfully mapped to the transcript collection.

The inter-play between the automated pipeline and the quality control using manually defined genes indicated which additional processes were required to improve the transcriptome definition. In the first assessment of the transcriptome quality, presence of the 63 genes in the three transcript definition approaches was examined. Four of the genes had no coverage in the RNA-Seq, and were not expected to be found in the read-based arms of the assembly. In the genome based transcript collection, of the 59 possible genes, eleven genes were missed. In the de novo assembly, twelve genes were missed.

E. huxleyi has a very high percentage of non-canonical splice junctions, and relatively high rates of intron read-through, which caused unique issues with the currently available tools. While individual tools missed genes and artificially joined overlapping transcripts, combining the results of several tools improved the completeness and quality considerably. The final collection, created from the integration of several quality control and improvement rounds, was compared to the manually defined set both on the DNA and protein level. 61 transcripts and 47 proteins were found, an improvement of 20% versus any of the read-based approaches alone.

References

Huang, X, and Madan A. (1999) CAP3: A DNA sequence assembly program. *Genome Res* **9**(9), 868-877.

Martin, J A and Wang Z. (2011) Next-generation transcriptome assembly. *Nat Rev Genet* **12**(10): 671-682. doi:10.1038/nrg3068

Trapnell, C., L. Pachter, and S. L. Salzberg. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111. doi:10.1093/bioinformatics/btp120

Trapnell, C, Williams BA, Pertea G, Mortazavi A, Kwan G et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5): 511-515.doi: 10.1038/nbt.1621