A comprehensive comparison between reference-based and 'de novo' isoform assembly approaches

Oscar Rodriguez, Juan Carlos Triviño, Rebeca Miñambres, Sheila Zuñiga, Sonia Santillán, Mayte Gil, Reyes Claramunt, Celia Buades✉
Sistemas Genómicos, Paterna, Spain

Motivation and Objectives

RNA-seq has recently become an attractive method of choice in the studies of transcriptomes, promising several advantages compared to microarrays such as higher sensibility and reproducibility. In addition, RNA-seq offers a broader dynamic range of detection and the capability of identifying novel isoforms as well as non-translated regions that may act in regulating gene expression. The reconstruction of the transcriptome can be performed following two different approaches, a reference-based method in which reads are mapped back to a reference genome, and a 'de novo' assembly strategy where reads are compared to each other to reconstruct expressed isoforms without the need of using a reference genome.

In the present studio we provide a comprehensive comparison between these two transcriptome analysis methodologies for isoforms reconstruction based on genome annotation and isoform expression levels using a Human sample. In addition, our work provides new insights into Human isoform diversity and the composition of non-canonical isoforms.

Methods

Total RNA was extracted from a Hapmap cell line culture. Strand-specific fragment libraries were built for Illumina HiSeq2000 sequencing using a paired-end strategy. A total of 30Gbs of raw data were produced for the sample.

Following standard reference-based approaches for RNA-seq data analysis, high quality reads were mapped with Tophat (Trapnell *et al.,* 2009) against the Human reference genome GRhg37/hg19. Gene expression levels were estimated using FPKM values as given by Cufflinks (Trapnell *et al.,* 2010) and DESeq (Anders and Huber, 2010).

In the 'de novo' transcriptome reconstruction approach, two algorithms, Trinity (Grabherr *et al.,* 2011) and Oases (Schulz *et al.,* 2012), were used. Resulting isoform assemblies were merged with CAP3 (Huang and Madan, 1999) to obtain a final consensus assembly. Isoform annotation and chimera detection were based on the Human annotations available at Ensembl (http://www.ensembl.org/).

Results and Discussion

Our results showed a high correlation between the reference-based approach and the 'de novo' assembly strategy in terms of the number of detected/reconstructed isoforms and their global expression. However, both methodologies showed specific differences suggesting higher susceptibility to different technical parameters and biases depending on sequencing depth, sequencing errors and the presence of complex or large variants.

Acknowledgements

This work has been financed by the 7th Framework Program.

References

Anders, S., Huber W. (2010): Differential expression analysis for sequence count data. *Genome Biology* **11**, R106+. doi:10.1186/gb-2010-11-10-r106.

Grabherr, M. G. et al. (2011): Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnology 29, 644-652. doi:10.1038/nbt.1883.

Huang, X., Madan, A. CAP3 (1999): A DNA sequence assembly program. *Genome Research* **9**, 868-877. doi:10.1101/gr.9.9.868

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

Trapnell, C. *et al.* (2010): Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515. doi: 10.1038/nbt.1621.

Schulz, M. H., Zerbino, D. R., Vingron, M., Birney, E. Oases (2012): robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092. doi:10.1093/bioinformatics/bts094.