Improvement of RNA-Seq precision with MapAl

Paweł P. Łabaj¹, Bryan E. Linggi², H. Steven Wiley², and David P. Kreil¹,³

¹ Chair of Bioinformatics, Boku University Vienna, Austria
² EMSL, Pacific Northwest National Laboratory, Richland, U.S.A.
³ School of Life Sciences, The University of Warwick, U.K.
e-mail: mapal2012@boku.ac.at

RNA-Seq for quantitative profiling

We have compiled an ultra-deep, replicated RNA-Seq reference set. This now allows a transcript by transcript analysis of reproducibility, a necessary complement to characterizations of bias [1]. It turns out that expression estimates for most genes are very noisy [2]. Measurement precision, however, determines the power of analyses to reliably identify relevant signals or changes, independent of whether replicates are employed or not [3].

We here introduce MapAl, a novel tool allowing the reliable assessment of double the total number of transcripts compared to other popular pipelines.

Benefits of exploiting gene models

MapAl is much more sensitive in the identification of known splice junctions. It is able to find 2–3 times as many reads falling on exon-exon junctions as TopHat.

These reads often play a key role in identifying the expression of a particular splice-form, and therefore determine:

• splice-form identification rate
• transcript-specific measurement precision

Fig 1. Workflows for RNA-Seq expression profiling. The established TopHat [4] pipeline (a) is compared with the newly proposed approach with MapAl [5] (b).

Fig 2. Statistics of identified and reliably measured transcripts [2].

- exploiting gene models already at the alignment stage allows identification over 100,000 known transcripts
- expression levels of about 57,000 of them could be measured reliably (relative error <20%)
- combined with transcripts of genes discovered de novo about 70,000 of them could be measured reliably

Fig 3. Distribution of measurements errors

In general for MapAl we observe:

• smaller maximum errors
• larger number of transcripts that could be measured with low errors

Fig 4. Cumulative distributions of the standard deviation for alternative technologies and data processing protocols.

References