Comparative assessment of computational methods for quantifying mammalian transcript isoforms from RNA-Seq data

Naoki Nariai[1] nariai@megabank.tohoku.ac.jp
Osamu Hirose[2] hirose@se.kanazawa-u.ac.jp
Kaname Kojima[1] kojima@megabank.tohoku.ac.jp
Kazuko Ueno[1] ueno@megabank.tohoku.ac.jp
Masao Nagasaki[1] nagasaki@megabank.tohoku.ac.jp

[1] Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Seiryo-machi, Aoba-ku, Sendai, Miyagi, 980-8575, Japan
[2] Faculty of Electrical and Computer Engineering, Institute of Science and Engineering, Kanazawa University, Kakuma, Kanazawa, Ishikawa, 920-1192, Japan

Keywords: transcript isoform abundance estimation, RNA-Seq, variational Bayesian inference

Many human genes express multiple transcript isoforms through alternative splicing, which greatly increases diversity of protein function. Although RNA sequencing (RNA-Seq) technologies have been widely used in measuring amounts of transcribed mRNA, accurate estimation of transcript isoform abundances from RNA-Seq data is challenging because reads often map to more than one transcript isoforms or paralogs whose sequences are similar to each other. We evaluate several computational methods to estimate transcript isoform abundances from RNA-Seq data. Here, we evaluate performance with TopHat and Cufflinks [1], RSEM [2] and TIGAR [3]. Compared to RSEM, TIGAR can handle gapped alignments of reads against reference sequences so that it allows insertion or deletion errors within reads, and it also optimizes the number of transcript isoforms by variational Bayesian inference through an iterative procedure. On simulated datasets, TIGAR by variational Bayesian inference outperformed the comparable quantification methods in inferring transcript isoform abundances, and at the same time its rate of convergence was faster than that of the expectation maximization algorithm. On real RNA-Seq data of human cell line samples, it was shown that prediction result with TIGAR was more consistent among technical replicates than those of other methods.

[1] Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature Protocols*, 7(3):562-578, 2012.
[2] Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N., RNA-Seq gene expression estimation with read mapping uncertainty, *Bioinformatics*, 26(4):493-500, 2010.
[3] Nariai, N., Hirose, O., Kojima, K., and Nagasaki, M., TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference, *Bioinformatics*, 29(18):2292-2299, 2013.