AuPairWise: biologically focused RNA-seq quality control using co-expression

Ballouz, S. and Gillis, J.

Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Boulevard Woodbury, NY 11797, USA.

A principal claim for RNA-sequencing has been greater replicability, typically measured in sample-sample correlations of gene expression levels. Replicability of transcript abundances in this way will provide misleading estimates of the replicability of conditional variation, which is what is of interest in most expression analyses. Heuristics which implicitly address this problem have emerged in quality control measures to obtain ‘good’ differential expression results. However, these methods involve strict filters such as discarding low expressing genes or using technical replicates to remove discordant transcripts, and are costly or simply ad hoc.

Instead, we show that gene-level replicability is a more useful metric, and demonstrate that it can be modeled in a co-expression framework, using known co-expressing gene pairs as pseudo-replicates instead of true replicates. We use this as a quality control metric: by modelling the effects of noise that perturbs a gene’s expression, we can then measure the aggregate effect of this perturbation on these co-expressing gene-pairs or ‘housekeeping interactions’. We find that perturbing expression by only 5% within its usual range of values is readily detectable (AUROC~0.73), suggesting this test is extraordinarily sensitive. In addition to making the software readily available (github.com/sarbal/AuPairWise), we have adapted the test to optimize RNA-seq alignment with the STAR aligner tool. Our findings suggest that more stringent parameters at the read mapping stage (e.g., minimum alignment scores) would have a modestly positive impact, making the post-hoc filtering done for high-expressing or high fold-changes a more intuitive part of direct quality control.