A framework for benchmarking RNA-seq pipelines

Rory Kirchner[1], Brad Chapman[1], Oliver Hofmann[1], Winston Hide[1]
[1]Harvard School of Public Health, Harvard University (kirchner@hsph.harvard.edu)

Project page: http://bcbio-nextgen.readthedocs.org
Code: https://github.com/chapmanb/bcbio-nextgen and https://github.com/roryk/bcbio.rnaseq

Processing RNA-seq data to make differential expression calls requires selection of tools for filtering contamination, aligning to the genome, quantifying expression and calling differentially expressed genes. Understanding the tradeoffs between choices of tools for each step requires both a reference pipeline to test against and a set of known differentially expressed genes to use to test for accuracy. We have created a community-developed framework named bcbio-nextgen for analyzing NGS data that is easy to modify, install, and scales to thousands of samples. We have implemented a reference RNA-seq pipeline using the bcbio-nextgen framework and used a combination of test data from the Sequencing Quality Control (SEQC) project, simulated count data and simulated read data to benchmark components of the RNA-seq pipeline. We demonstrate the usefulness of the RNA-seq benchmarks by determining the best quality filtering cutoff, evaluating the performance of differential expression callers and determining appropriate sample sizes for given experimental questions. The simulator can be tuned to reflect a wide variety of experiments and can also be tuned to match a user-defined real dataset. This allows experimenters to simultaneously analyze their experiment and produce benchmarks of the differential expression callers on data similar in nature to their experiment. We hope that the bcbio-nextgen framework will be a useful tool to both analyze RNA-seq experiments and rapidly test new or updated pipeline components or callers.