Differential expression RNA-seq analysis with a large data set from brain samples

Leonardo Collado-Torres¹, Alyssa Frazee¹, Andrew Jaffe², Sarven Sabunciyan³, Jeffrey T. Leek¹

¹Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health,
²Department of Biostatistics at JHSPH and Lieber Institute for Brain Development,
³Department of Pediatrics, The Johns Hopkins School of Medicine

Introduction

Differential expression analysis from RNA-seq data can be done with three types of methods:

1. annotate-then-identify (DESeq, edgeR),
2. assemble-then-identify (Cuffdiff2),

We have a unique large data set (59 samples) where we can compare these methods. Running derfinder involves:

- Aligning with TopHat: 20 cores, ~12 hrs per sample
- Merging samples by chromosome (250 mi x 59 max)
- Filtering by row statistics (e.x. at least 1 column > 5)
- HHM by chunks due to memory limits (by 100 000)
- P-values by permutations (10-20 per chr)

Objectives

- Compare leading methods.
- Improve derfinder.

Tools used

The project has been a combination of reducing hard disk requirements (e.x. ~2TB down to 317 GB), reducing memory load (e.x. 75 to 2.5 GB), reducing input/output (e.x. storing medians instead of re-calculation per permutation), and reducing wallclock computing time (e.x. 9 to 3 hrs).

- Extensive use of enigma2 for parallelizing when possible.
- IRanges for reducing the memory load.
- Rsamtools for faster processing of alignment files.
- Interactive visualization (D3) via clickme.

Results so far

- Reduce the computation requirements for derfinder.
- Design visualizations that allow us to distinguish artifacts from results.
- Implement batch correction on RNA-seq data.

References

2. https://github.com/alyssafrazeee/derfinder

Work supported by the Stanley Medical Research Institute: samples and sequencing.

LCT is supported by CONACyT and R01HG006102.