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

Leonardo Collado-Torres[1], Alyssa Frazee[1], Andrew Jaffe[2], Sarven Sabunciyan[3], Jeffrey T. Leek[1]

[1]Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health,
[2]Department of Biostatistics at JHSPH and Lieber Institute for Brain Development,
[3]Department 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),
3. identify-then-annotate Frazee et al (2013), *derfinder*.

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)

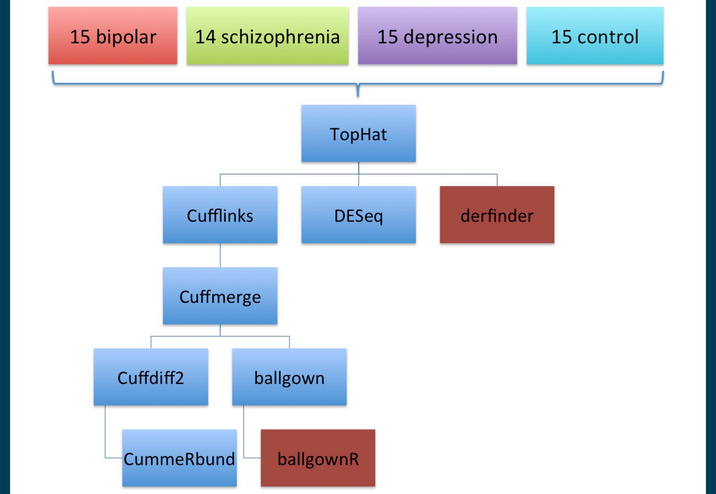
Objectives

- Compare leading methods.
- Improve *derfinder.*

Tools used

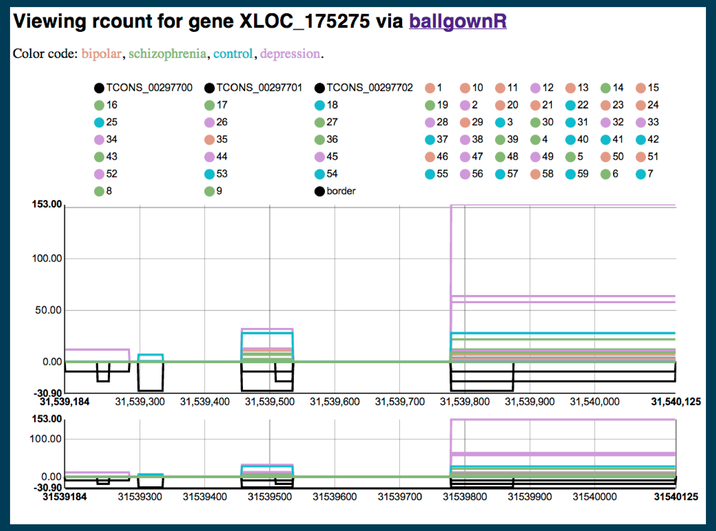
The project has been a combination of reducing hard disk requirements (e.x. ~2TB down to 317 GB), reducing memory load (e.x. 75 to 2.5 GB), reducing input/output (e.x. storing medians instead of re-calculating per permutation), and reducing wallclock computing time (e.x. 9 to 3 hrs).

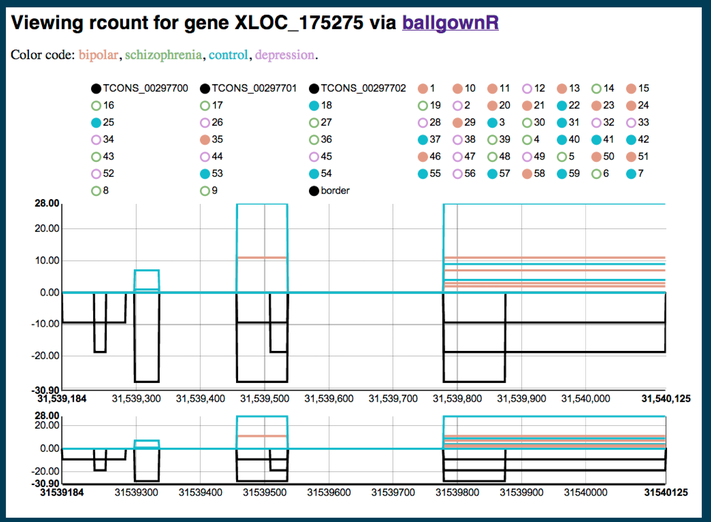
- Extensive use of *enigma2* for parallelizing when possible.
- *IRanges* for reducing the memory load.
- *Rsamtools* for faster processing of alignment files.
- Interactive visualization (D3) via *clickme*.



Order of execution relationships between the main tools.

Results so far





To do

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

References

1. Frazee, A. S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek (2013). Differential expression analysis of rna-seq data at single base resolution.
2. https://github.com/alyssafrazeee/derfinder
3. https://github.com/lcolladotor/ballgownR-devel

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

LCT is supported by CONACyT and R01HG006102.