Fast annotation-agnostic differential expression analysis
Leonardo Collado-Torres¹, ², Andrew E. Jaffe², Jeffrey T. Leek³
¹Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health,
²Lieber Institute for Brain Development, Maltz Research Laboratories

Introduction
Since the development of high-throughput technologies for probing the genome we have been interested in finding differences across groups that could potentially explain the phenotypic differences we observe. In other words, methods for generation of hypothesis at a large scale where we try our best to remove artifacts. The traditional tools have focused on the transcriptome and are highly dependent on existing annotation. Frazee et al¹ developed a statistical framework to find candidate Differentially Expressed Regions (DERs) without relying on annotation. We have implemented a modified version of this approach that is faster in order to handle larger data sets and whose total processing time is comparable to other tools such as DESeq (Anders et al, Genome Biology 2010).

DERfinder

Results
Time and memory needed:
- 20 samples
  - Load & filter data: 10 cores with mclapply 1 hr 15 min, 177 GB
  - Make models: 20 min, 52 GB
  - Analysis: 30 permutations, 4 cores each chr, total 59 mins
    - chr1 41 min, 46 GB
    - Merging: 30 min, 22 GB
    - Report: 27 min, 17 GB
    - Total wallclock time: 3 hr 46 min

A richer data set: 69 samples
- Load raw data: each chr, total 3 hr 28 min
  - chr1 1 hr 28 min, 38 GB
    - Merge: 1 hr 7 min, 67 GB
    - Filter data: each chr, total 12 min
      - chr1 12 min, 10 GB
    - Merging: 1 hr 6 min, 46 GB
    - Report: 2 hr 29 min, 45 GB
    - Total wallclock time: 9 hr 3 min

Conclusions
Goal accomplished: from BAM files to annotated candidate DERs in less than a day!
Comparable time versus other methods (DESeq, ...)

Open questions/todo:
- Reduce memory requirements.
- When to merge regions?
- How to adjust for coverage?

References
3. https://github.com/lcolladotor/derfinder

LCT is supported by CONACyT México and LIBD.