Fast annotation-agnostic differential expression analysis

Leonardo Collado-Torres[1], [2], Andrew E. Jaffe[2], Jeffrey T. Leek[1]

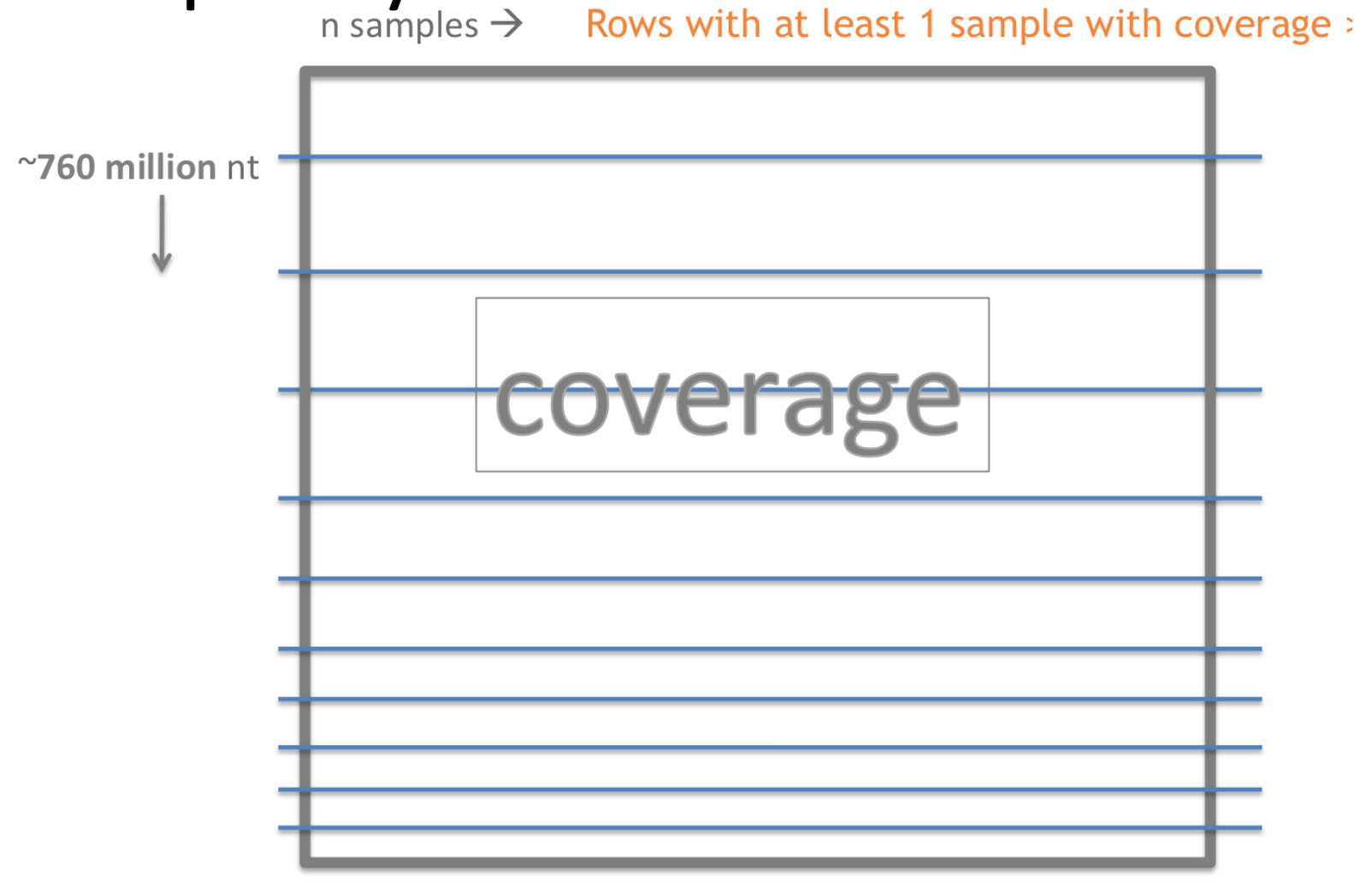
[1]Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health,
[2]Lieber Institute for Brain Development, Maltz Research Laboratories

Introduction

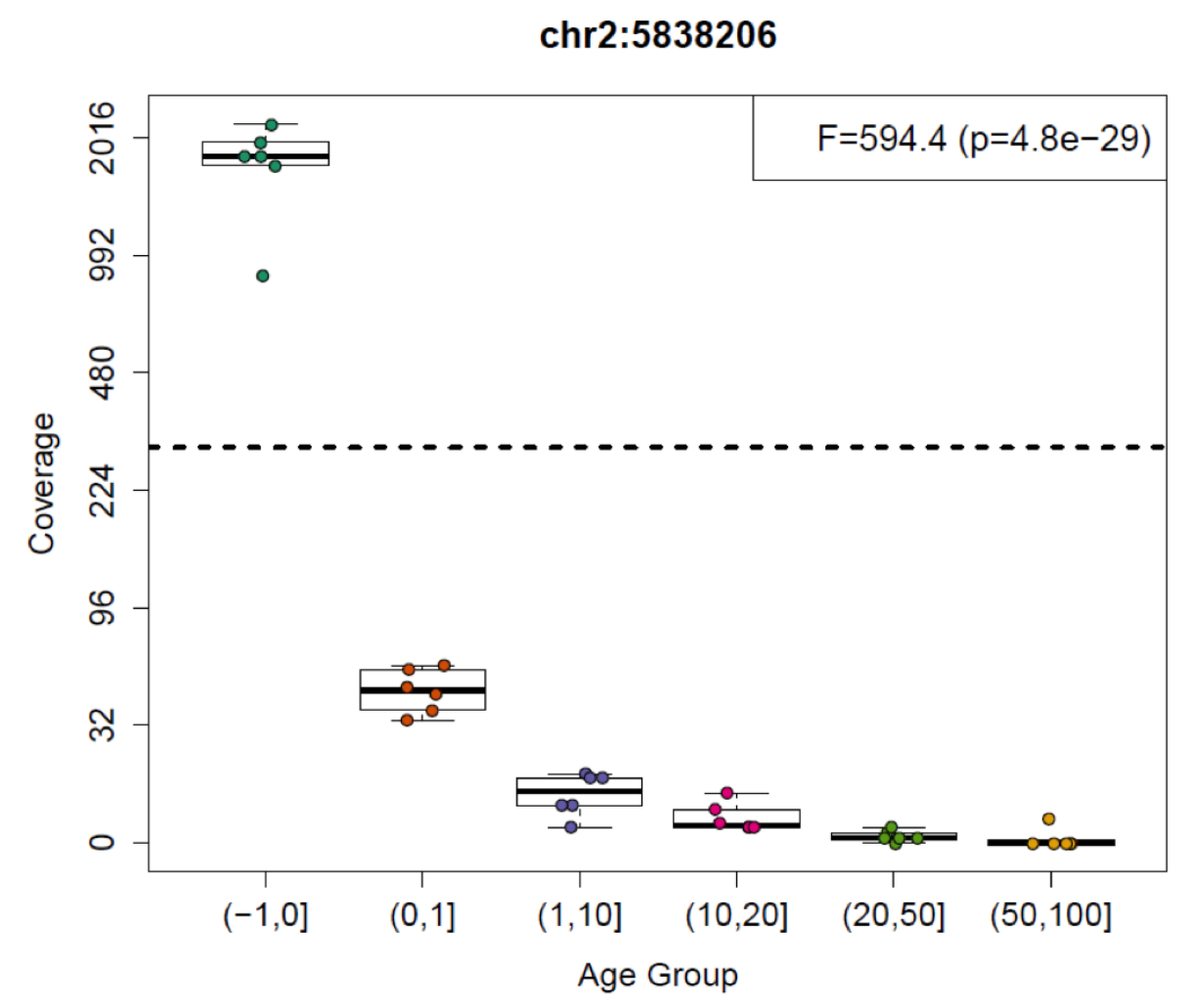
Since the development of high-throughput technologies for probing the genome we have been interested in finding differences across groups that could potentially explain the phenotypic differences we observe. In other words, methods for generation of hypothesis at a large scale where we try our best to remove artifacts. The traditional tools have focused on the transcriptome and are highly dependent on existing annotation. Frazee et al[1] 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

Split by chromosome and filter

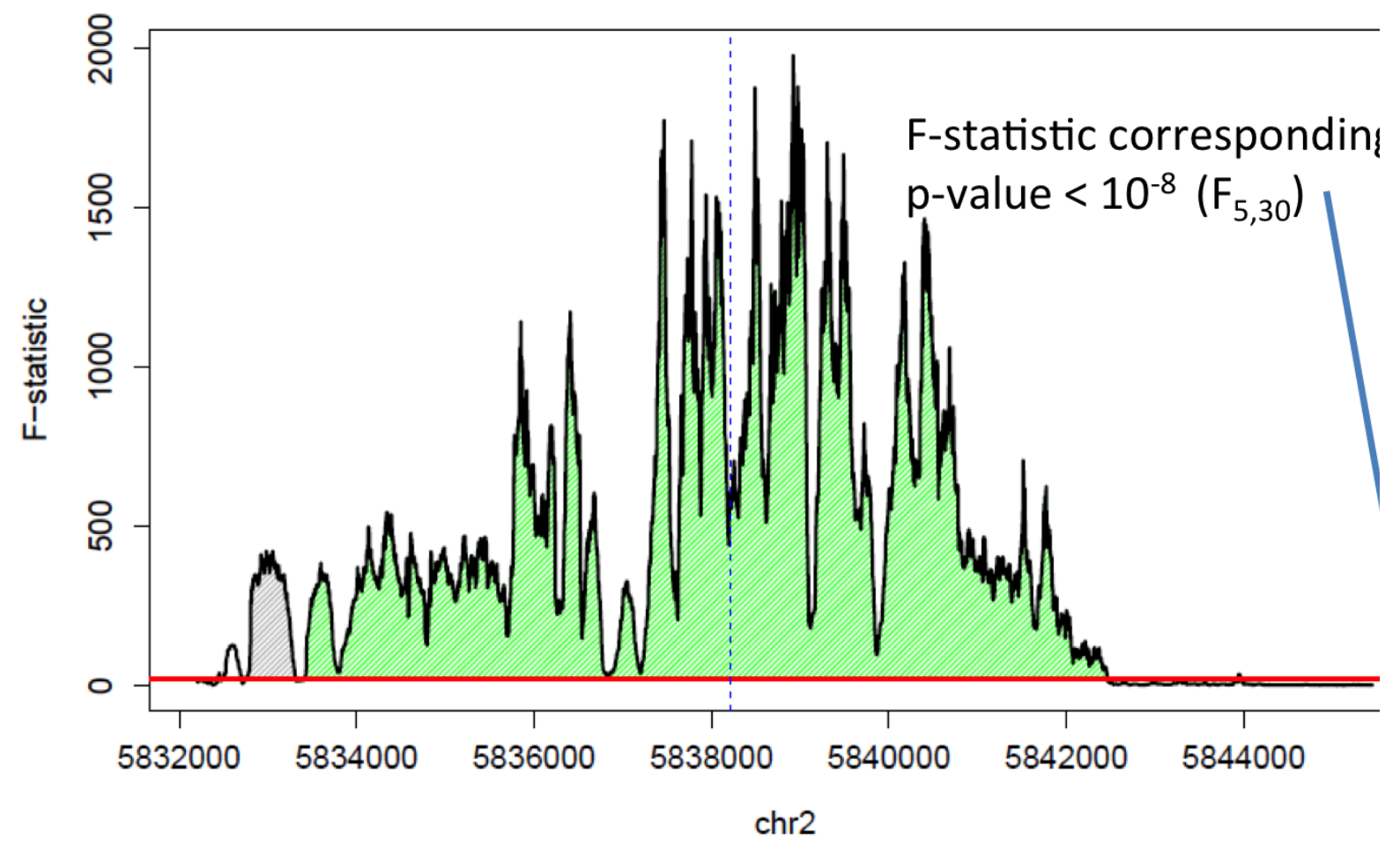


Test for base-level DE



Adapted from @andrewejaffe

Threshold on F-statistics



Adapted from @andrewejaffe

How can we make it fast?

- Avoid Input/Output as much as possible
- Work by chromosome
- Reduce memory
  - Run Length Encoding (IRanges::Rle)
    0000111111222 = (0, 1, 2)
    (4, 6, 3)
- Use multiple cores (parallel::mclapply)
  - Split data to use cores efficiently
- Calculate F-stats using Rcpp (Has + and -)

F-statistic at each base-pair

$$Y_{ij} = \log_2\left(\text{Coverage}_{ij} + 32\right)$$

$$i = 1, \ldots, 760e6; \quad j = 1, ..., \text{Nsamples}$$

- Null model

$$Y_{ij} = \beta_0 + \beta_1 \text{CoverageAdjustment}_j$$

- Alternative Model

$$Y_{ij} = \beta_0 + \beta_1 \text{CoverageAdjustment}_j + \beta_2 \text{Group}_j$$

Q-values: qvalue::qvalue

Permute model matrices and find null regions for all chromosomes.

$$k = 1, \ldots, nDERs; \quad W_k := \text{width of region } k$$

$$\text{Area}_k = \sum_{l=1}^{W_k} \text{F-statistic}_l$$

$$M = \text{number of null regions across all chrs}$$

$$\text{p-value}_k = \frac{\sum_{m=1}^{M} \text{I}\left(\text{NullArea}_m > \text{Area}_k\right) + 1}{M + 1}$$

Results

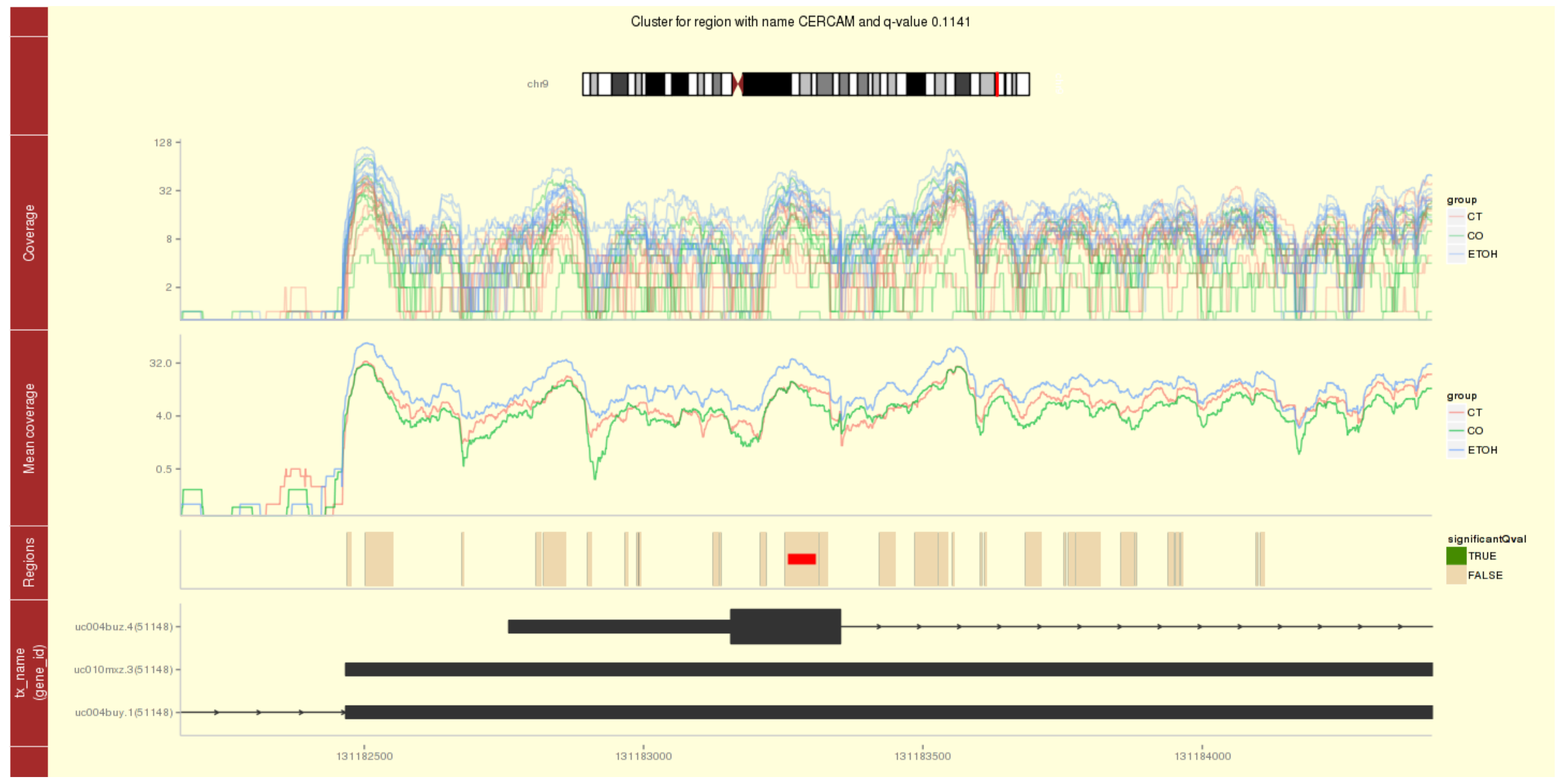
Time and memory needed:

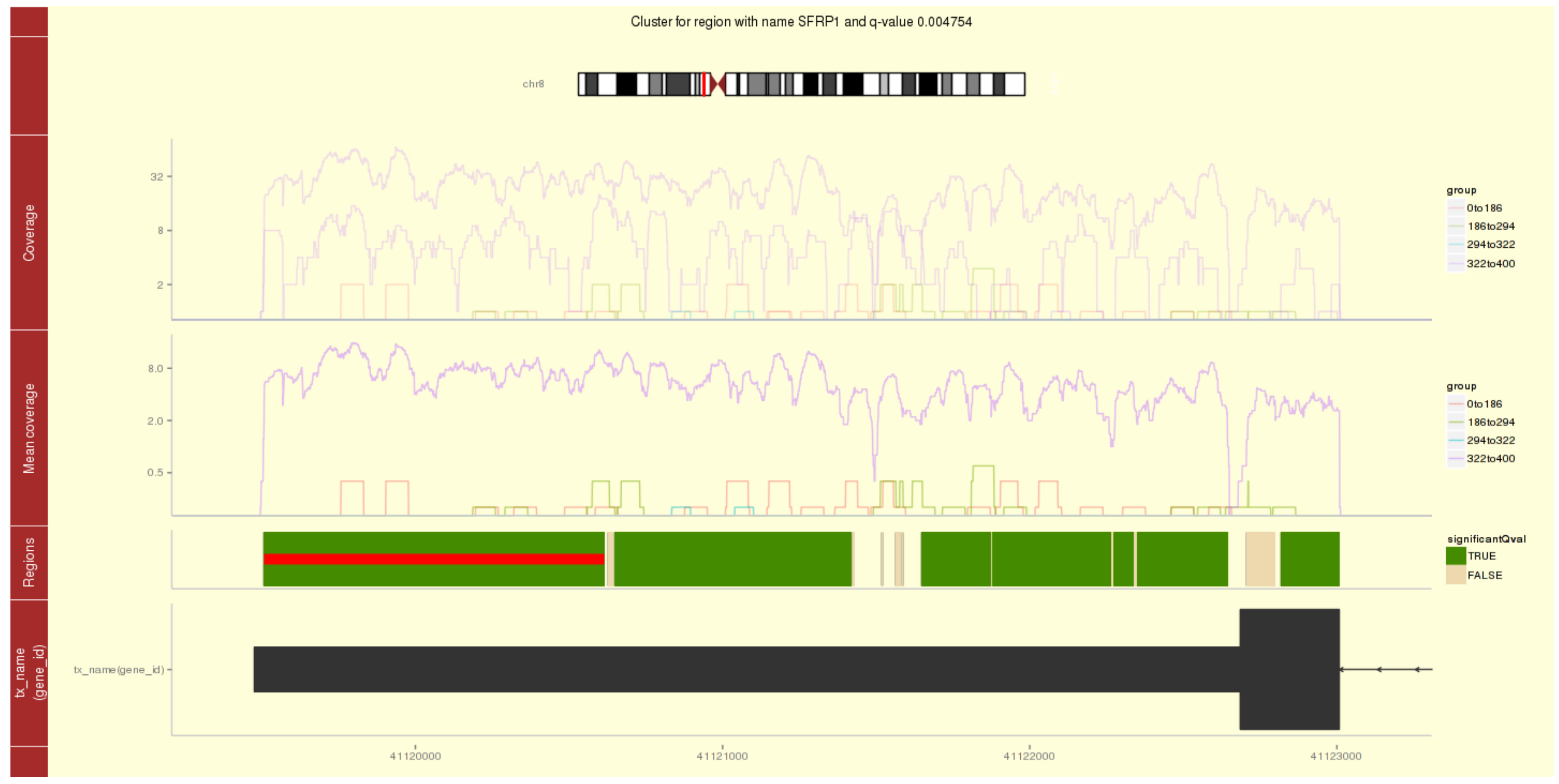
20 samples

- Load & filter data: 10 cores with mclapply
1hr 15min, 177 GB
- Make models: 20 min, 52 GB
- Analysis: 10 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 1hr 28 min
  - chr1 1hr 28 min, 18 GB
  - Merge 1hr 7 min, 67 GB
- Filter data: each chr, total 12 min
  - chr1 12 min, 10 GB
  - Merge 1hr, 62 GB
- Make models: 1 hr 49 min, 234 GB
- Analysis: 0 permutations, 8 cores each chr, 52 min
  - chr1 49 min, 258 GB, had to run twice
- Merging: 1 hr 6 min, 46 GB
- Report: 1hr 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

1. A. Frazee, S. Sabunciyan, K. D. Hansen, R. A. Irizarry, and J. T. Leek (2013). Differential expression analysis of rna-seq data at single base resolution, *Biostatistics*, in review.
2. L. Collado-Torres, A. E. Jaffe, J. T. Leek (2013). Manuscript *in preparation*.
3. https://github.com/lcolladotor/derfinder

LCT is supported by CONACyT México and LIBD.