Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L - 2022 | week 08

Pierre-Luc Germain

ETH Zürich

Plan for today

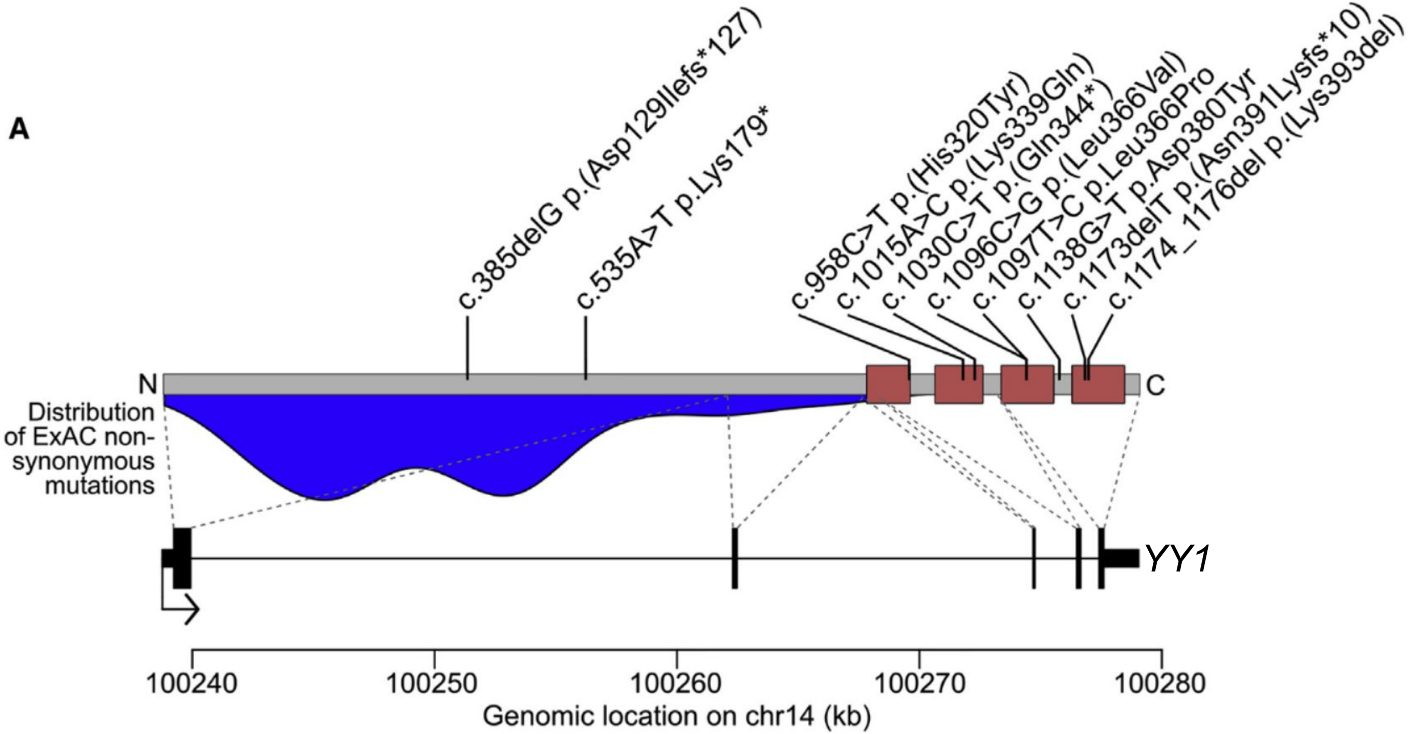
- Any question on the assignment?
- Our case for the practical: GDVS
- Normalization and differential analysis
- Discussion for course projects

Our case study today

- Lymphoblastoid cells from patients with Gabriele-de Vries syndrome (and controls)
- OMIM:
    - “Gabriele-de Vries syndrome is an autosomal dominant neurodevelopmental disorder characterized by delayed psychomotor development, variable cognitive impairment, often with behavioral problems, feeding problems, some movement abnormalities, and dysmorphic facial features. Affected individuals may also have a variety of congenital abnormalities.”
- Caused by haploinsufficiency in the *YY1* gene
- Data: YY1 ChIP-seq in mutant and control lymphoblastoid lines

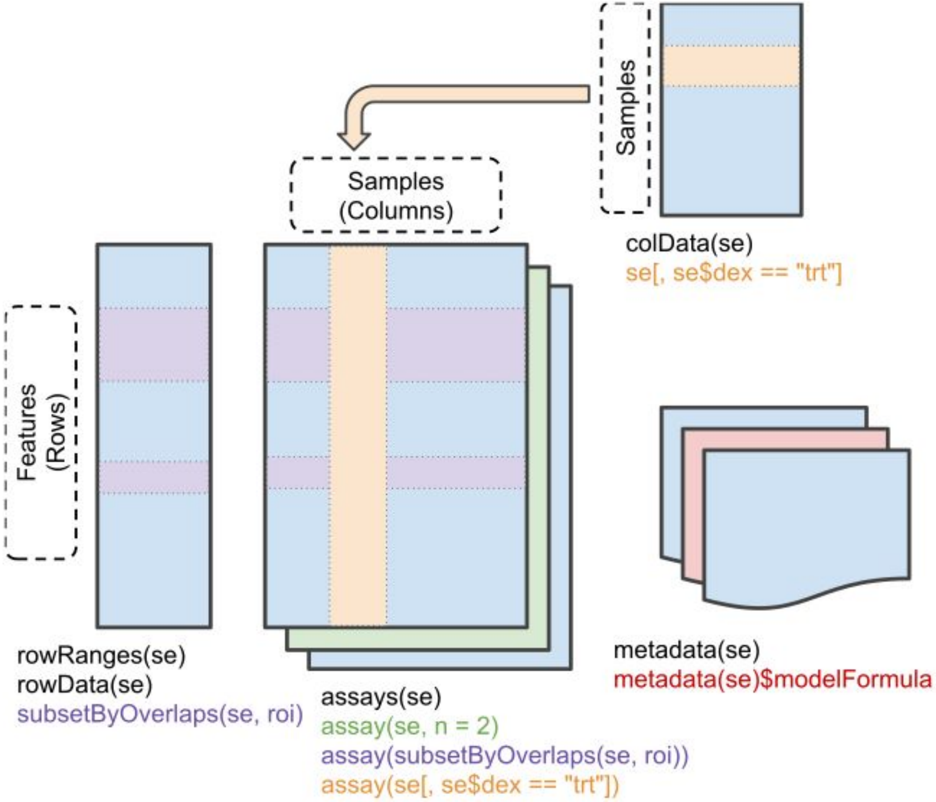
(Gabriele, Vulto-van Silfhout, Germain et al., AJHG 2017)

Our case study today



(Gabriele, Vulto-van Silfhout, Germain et al., AJHG 2017)

The SummarizedExperiment structure



colData(se)
se[, se$dex == "trt"]

rowRanges(se)
rowData(se)
subsetByOverlaps(se, roi)

assays(se)
assay(se, n = 2)
assay(subsetByOverlaps(se, roi))
assay(se[, se$dex == "trt"])

metadata(se)
metadata(se)$modelFormula

[Documentation](#)

Normalization

- Standard TMM normalization in edgeR assumes that the majority of regions don’t change across condition.
  - This does not work when there are global increases or decreases (i.e. at most sites) between conditions
- Background normalization assumes that the signal-to-noise ratio is the same across experiments
  - This works nicely for visualization, but can create artifactual differences in highly-enriched regions when the quality of the enrichment differs between experiments
- Common (or top) peak normalization (e.g. Shao et al., Genome Biology 2012) assumes that the regions that are significantly enriched in both conditions don’t change

- S3norm (Xiang et al., NAR 2020) performs both background and common peak normalization

Differential analysis

- Since the data is count-based, we can rely on the same tools that have been developed for RNAseq differential expression analysis: edgeR, limma::voom, and DESeq2
- These tools use information sharing across features to better estimate variability (see UZH's STA426 course for how these models actually work!)
- Because they use generalized linear models, they easily enable analysis of complex designs
    - See how most statistical tests can be expressed as linear models

Assignment

Until next week, come up with a preliminary plan for your project, summarizing:

1. What is the topic?
2. What data will you be using?
3. What are the analyses you wish to reproduce, or the questions you wish to answer?

This is not a final plan, but the start of a discussion!

Write that up in a Rmarkdown that you can upload to your repository.