Bioinformatic approaches to regulatory genomics and epigenomics

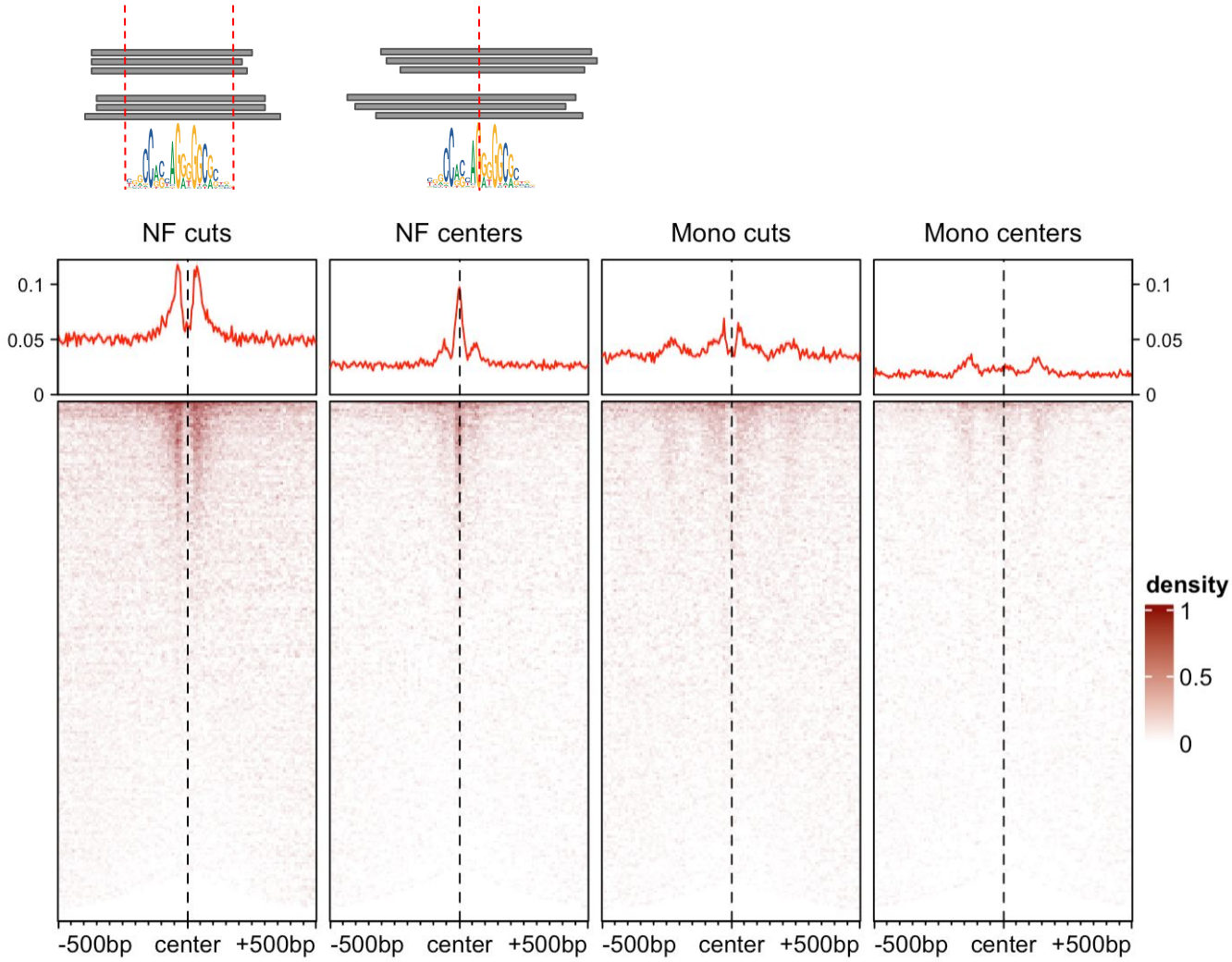
376-1347-00L | week 08

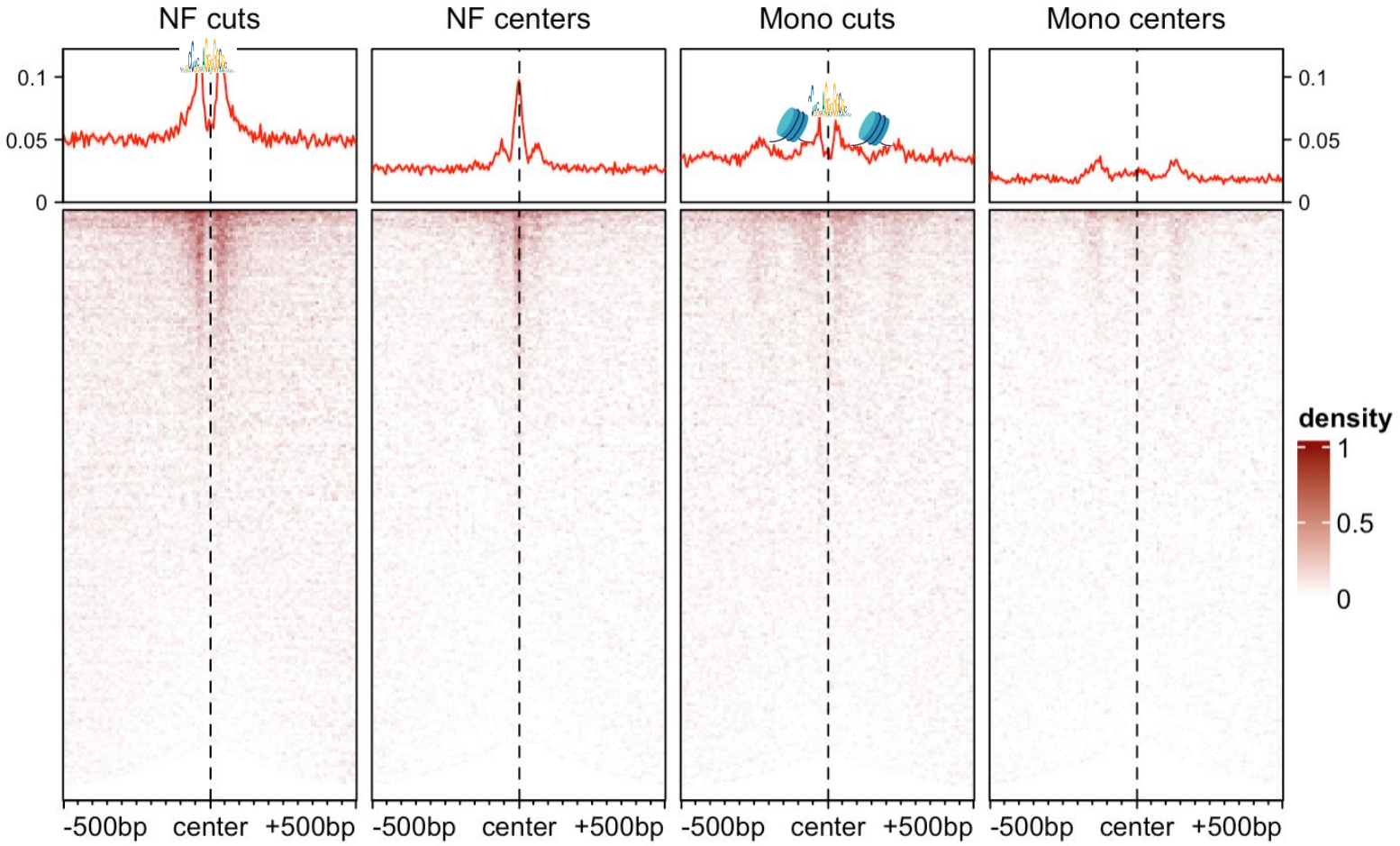
Pierre-Luc Germain

ETH Zürich

Today’s plan

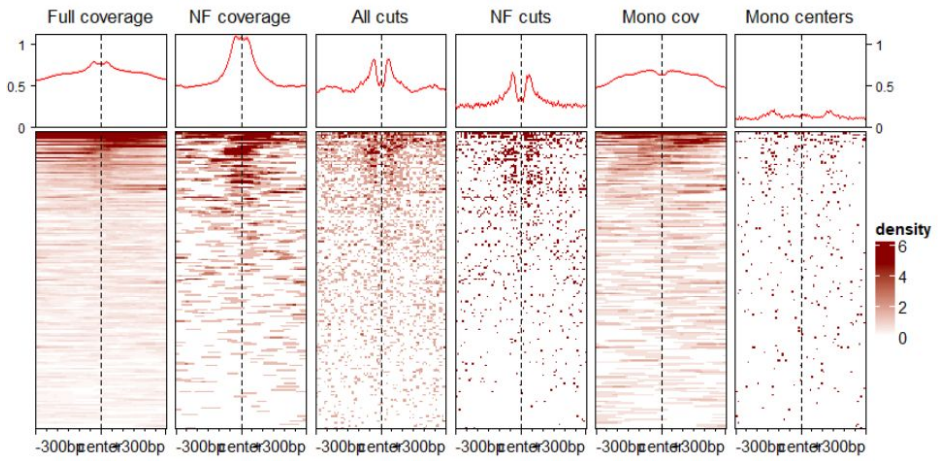
- Debriefing on the assignment & follow-up on last week
- Motif accessibility analysis
  - chromVAR, and working with SummarizedExperiment objects
- Differential motif accessibility analysis

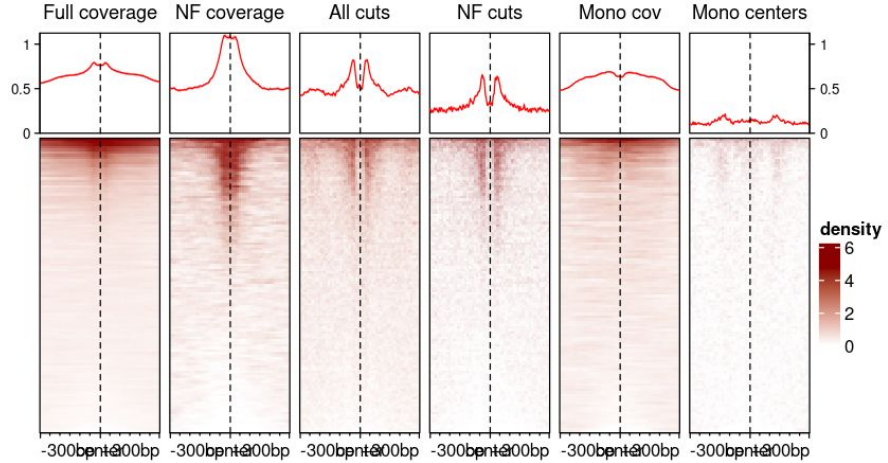

NF cuts
NF centers
Mono cuts
Mono centers
0.1
0.05
0
-500bp
center
+500bp
density
1
0.5
0


NF cuts
NF centers
Mono cuts
Mono centers
0.1
0.05
0
density
1
0.5
0
-500bp
center
+500bp

Debriefing on the assignments: install the *magick* package

For some, the exact same code gave much uglier heatmaps:





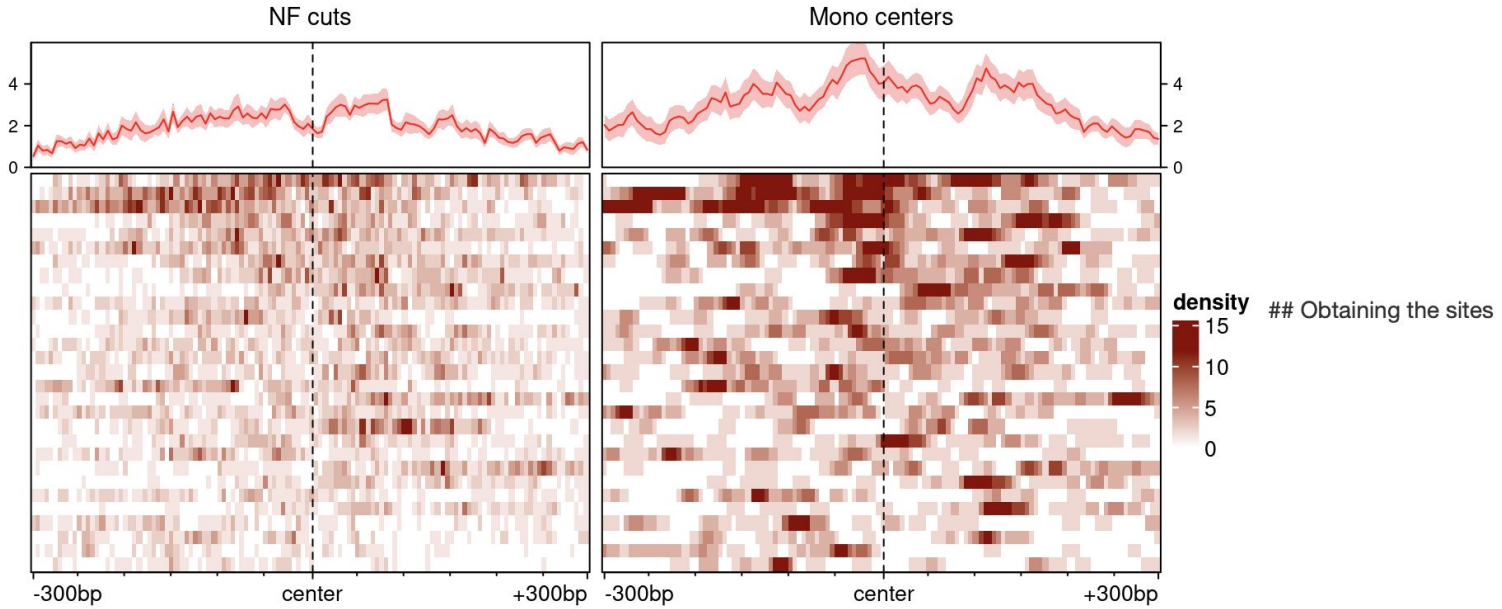
This happens because the size of the heatmap requires a rasterization, which is optimally done by the *magick* R package (on the right), but in the absence of that package it falls back to very suboptimal methods.

Solution: `BiocManager::install("magick")`

Debriefing on the assignments: plotting arguments

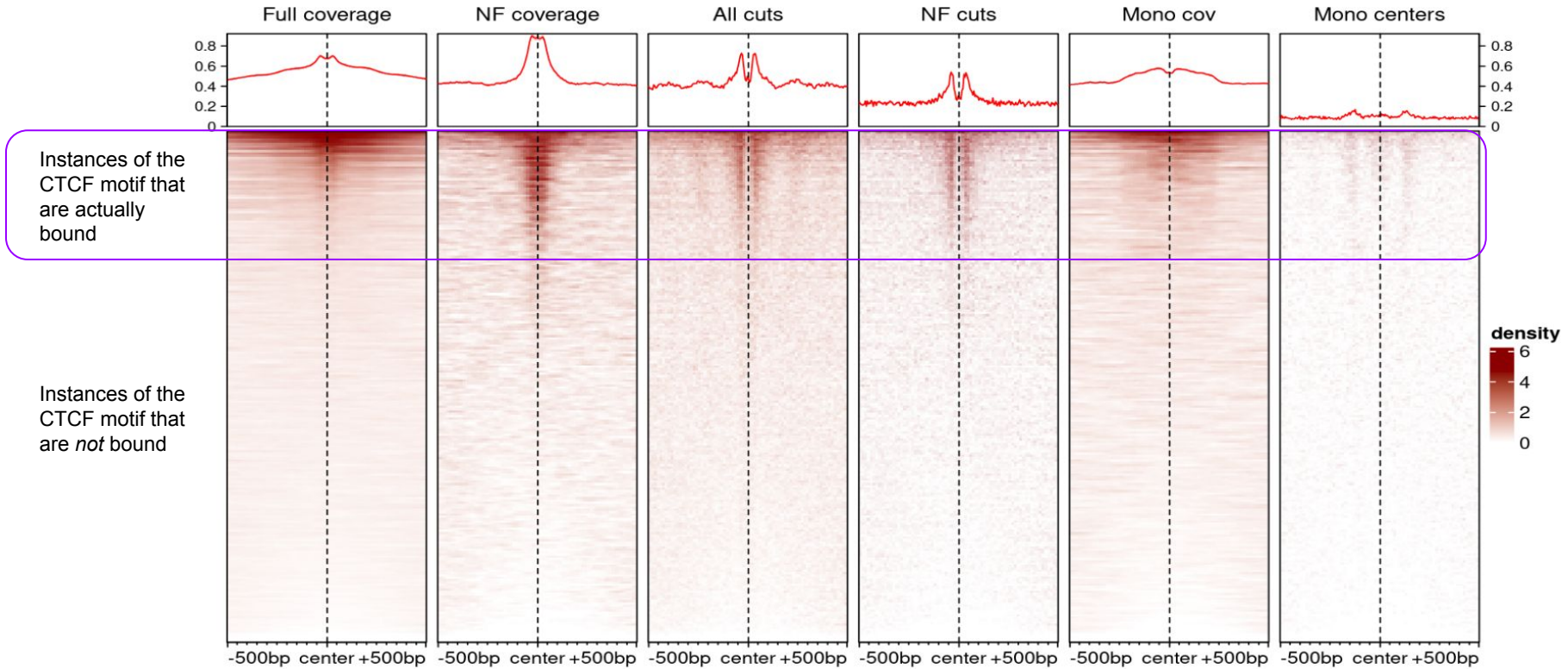
```
plotEnrichedHeatmaps(smb, trim=0.95, colors = c("white","darkred"), minRowVal=15)
```

=> After **filtering** for a minimal count of 15 **per row** (i.e. region), there might be no / too little signal across the **motif** for some **TF**s for the respective experiment. This leads to too few being included in the plot.



Subsetting to more relevant regions

Signal around occurrences of the CTCF motif in mESC



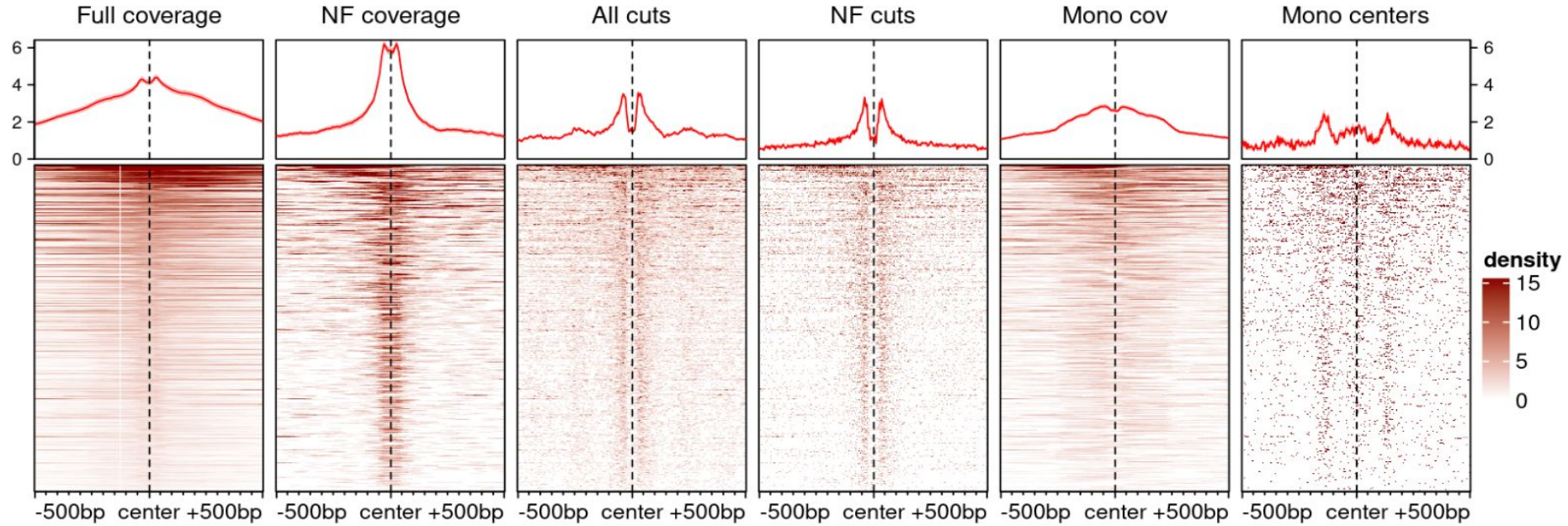
While this is useful to get an overall picture, most of the time we want to concentrate on the actual sites bound

Subsetting to more relevant regions

To concentrate on these actual binding events, we can simply restrict the object to those within ATAC peaks:

```
o <- o[overlapsAny(o,peaks), ]
```

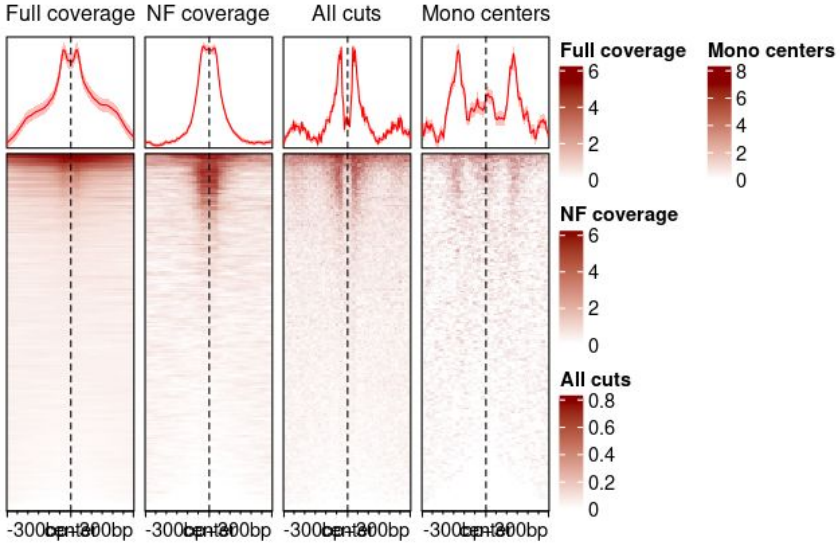
Signal around occurrences of the CTCF motif *in ATAC-seq peaks*



Subsetting to more relevant regions

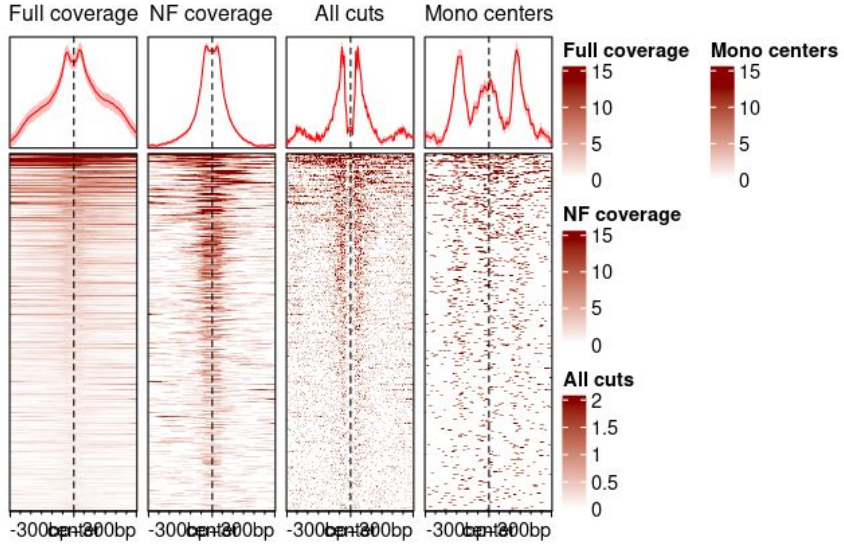
In the absence of the peaks, an alternative is to focus on regions with a higher signal:

```
plotEnrichedHeatmaps(sm, multiScale = TRUE,
                     colors=c("white","darkred"))
```



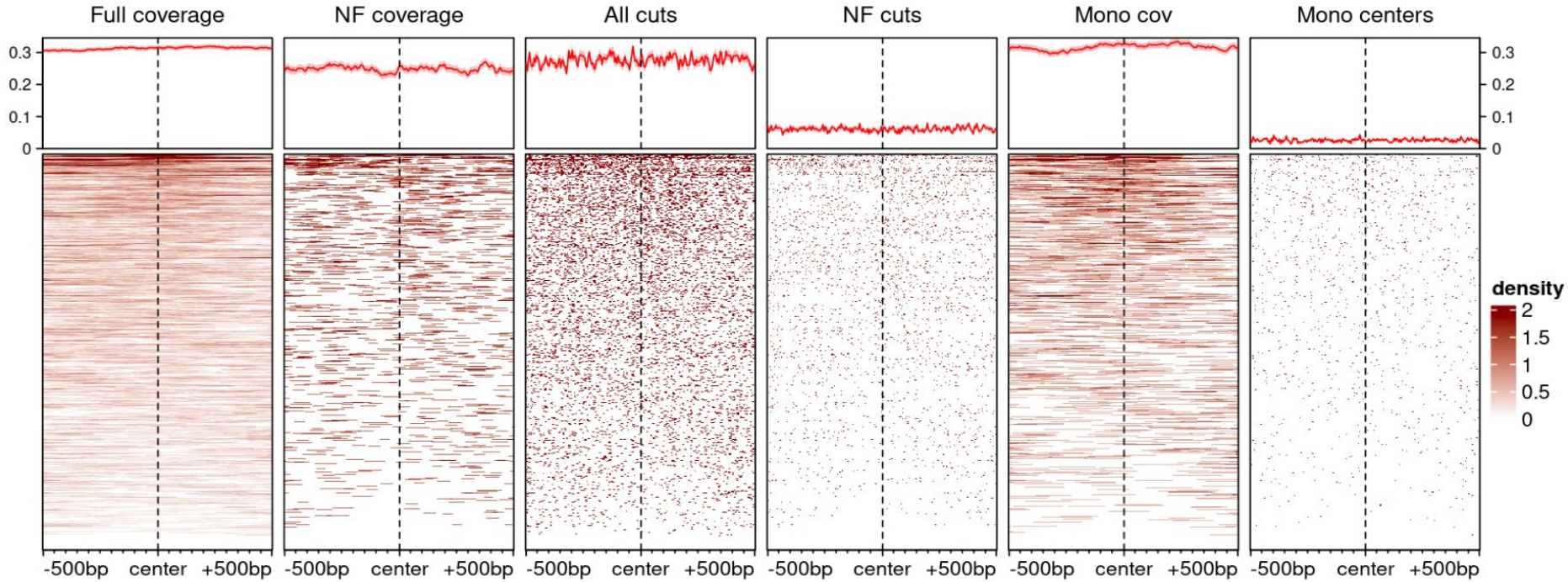

```
# subset to the 1000 regions with the
# largest enrichment:
sm.top <-
 sm[head(order(-rowMeans(score(sm))),1000),]

plotEnrichedHeatmaps(sm, multiScale = TRUE,
                     colors=c("white","darkred"))
```

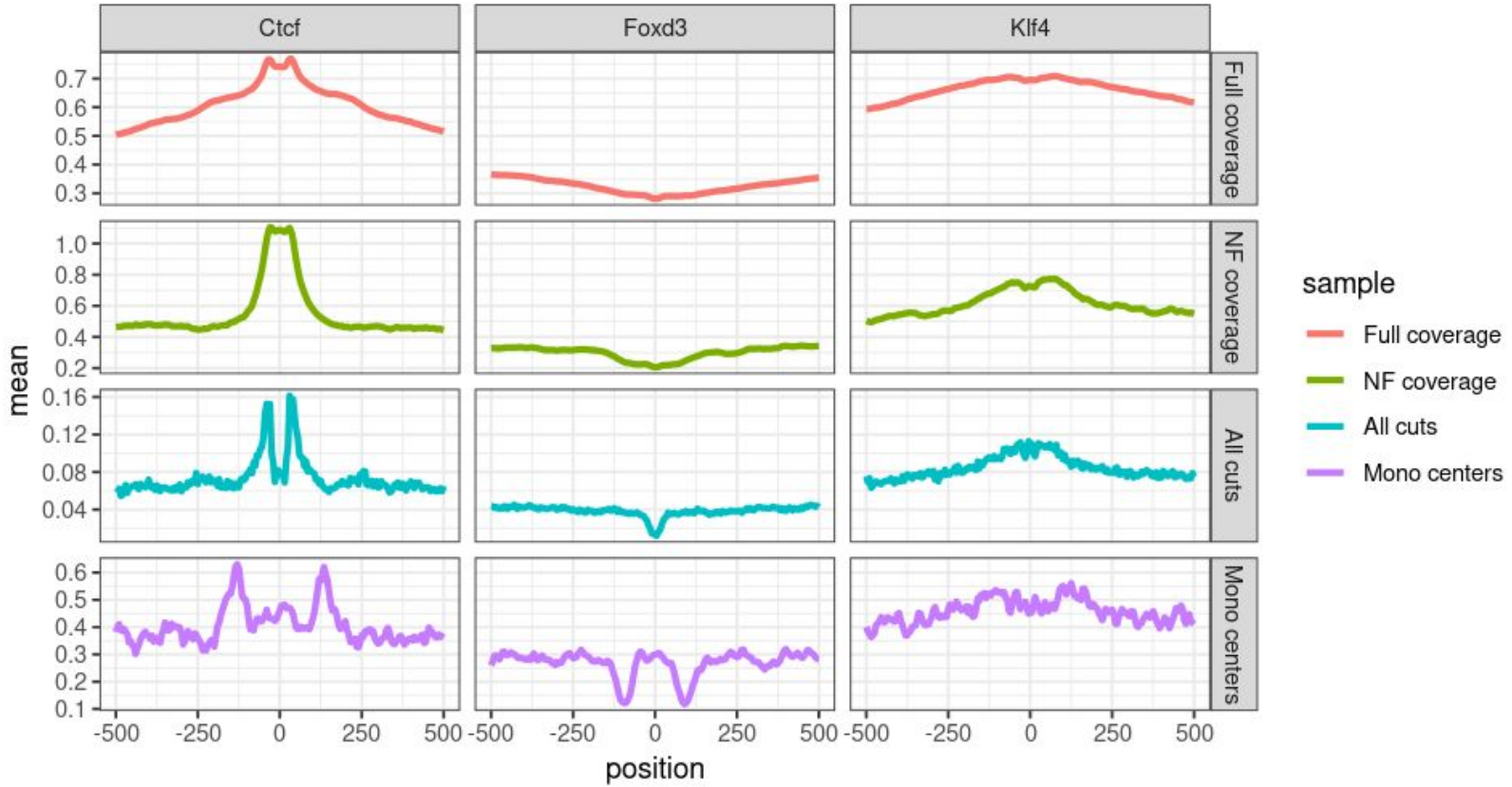


Follow-up on the assignment

Signal around occurrences of the GATA1 motif in mESC, where the factor is *not* expressed



Follow-up on the assignment



Differential motif accessibility analysis

What if, rather than looking at one motif at a time, we could simply quantify the accessibility/activity of every motif, and compare that across samples/conditions?

Methods have been developed for this, which can be grouped into two families:

Compute per-peak accessibility fold-changes between conditions

Test for motifs that have a skewed fold-change distribution

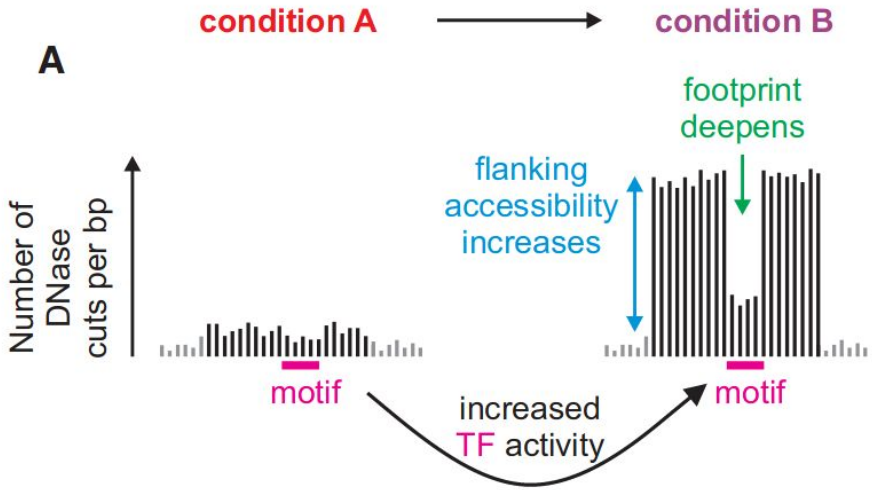
(e.g. monaLisa)

For each motif, compute per-sample activity scores,

then perform a differential analysis on those scores across conditions

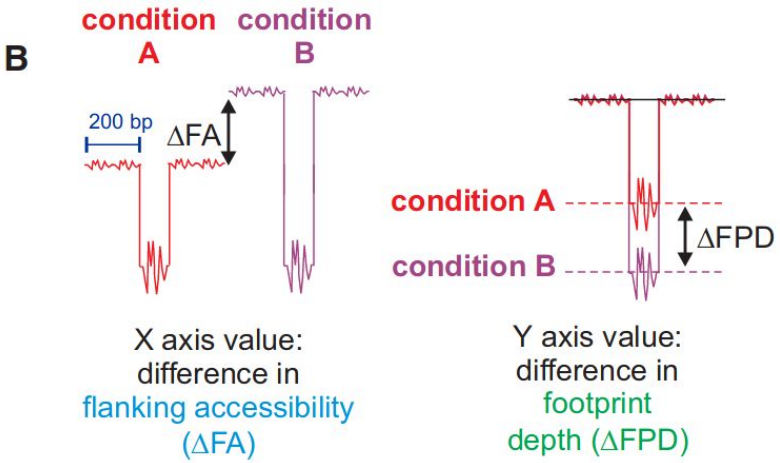
(e.g. chromVAR)

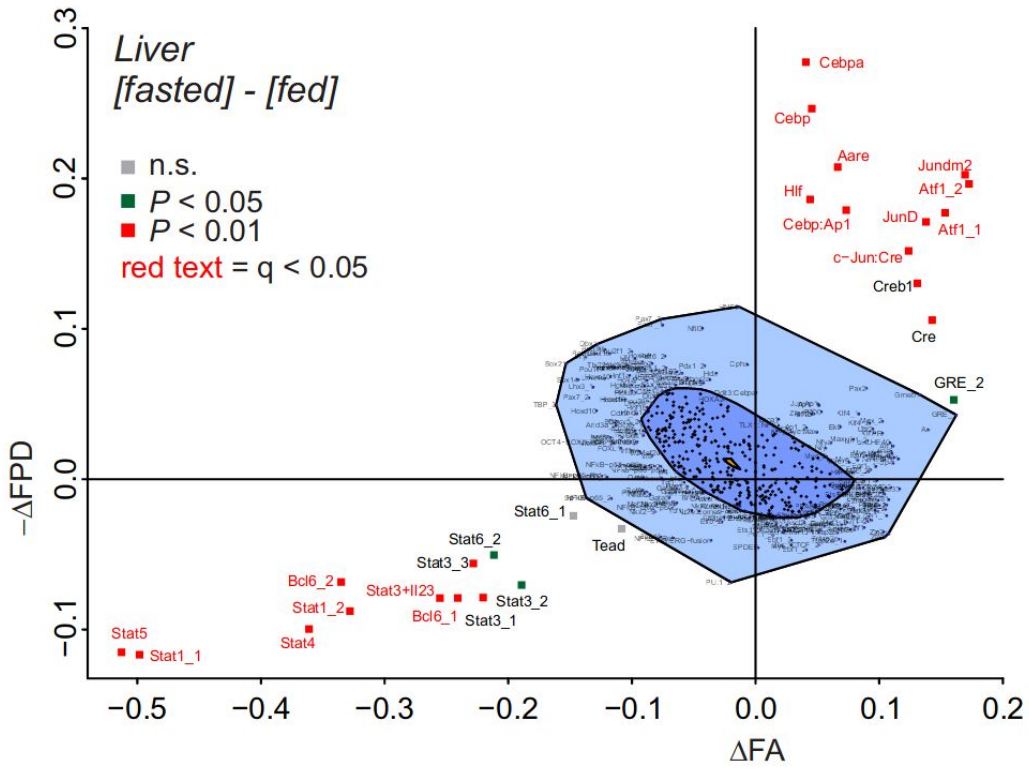
Estimating TF activity from accessibility and footprints



(Baek, Goldstein and Hager, Cell Reports 2017)

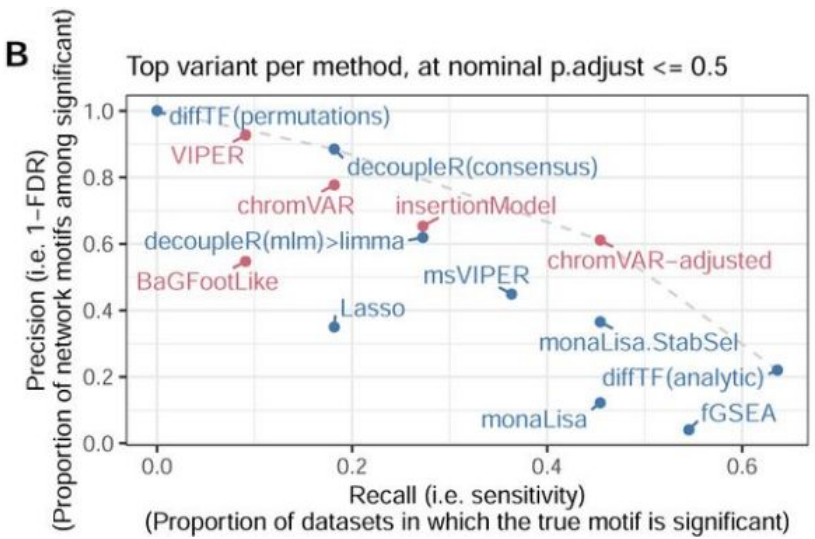
Estimating TF activity from accessibility and footprints

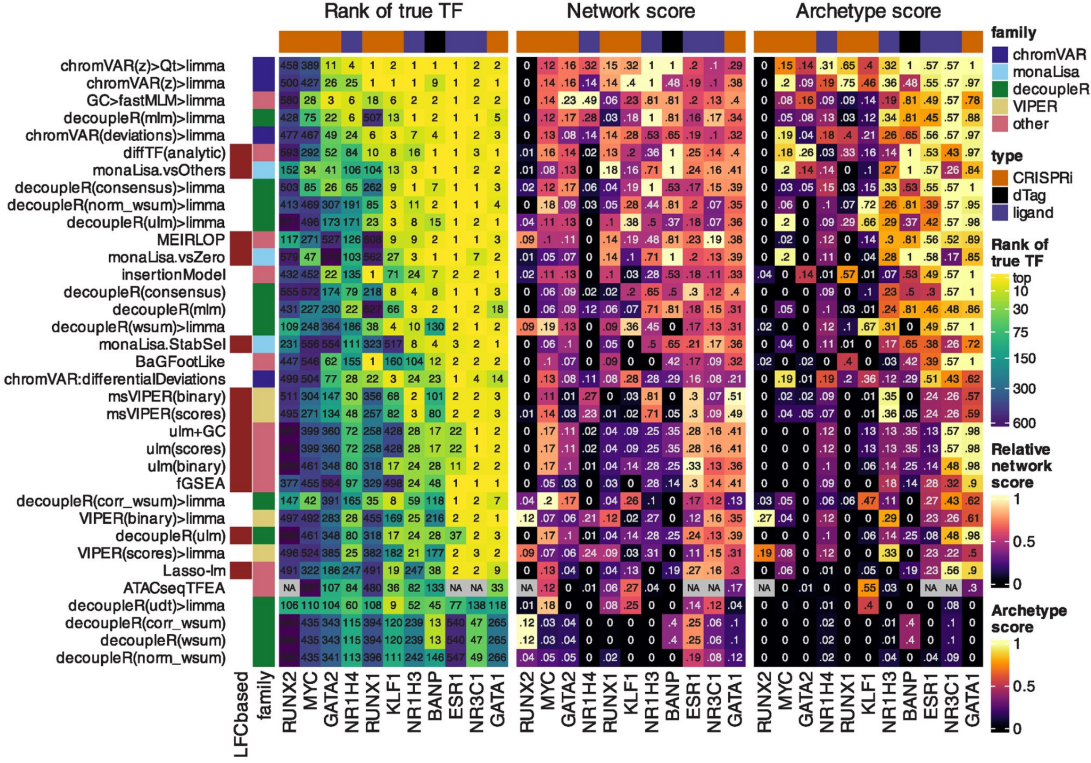




(Baek, Goldstein and Hager, Cell Reports 2017)

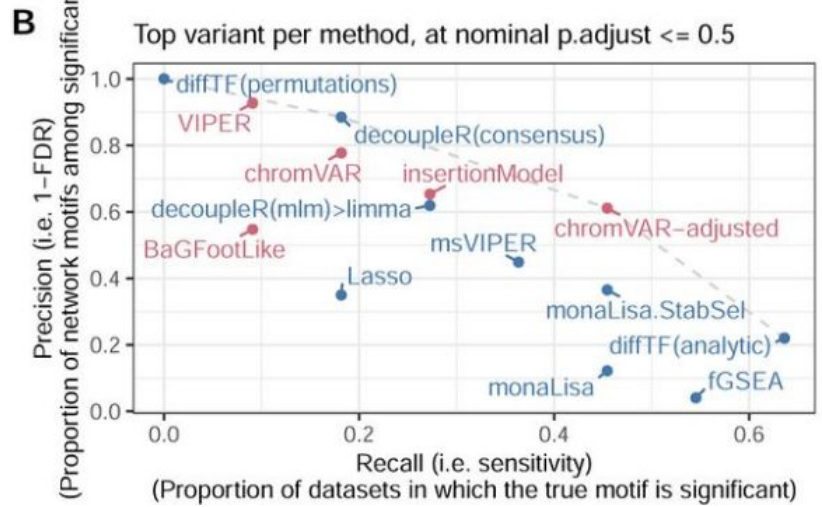
Benchmark of differential TF activity inference methods





(Gerbaldo, Sonder et al., PLOS Comp Bio 2024)

Simple tweaks on a chromVAR analysis currently provide the best method for identifying differentially-regulated motifs/TFs

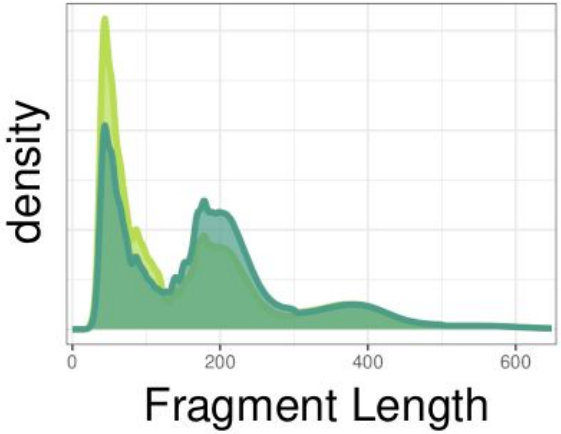


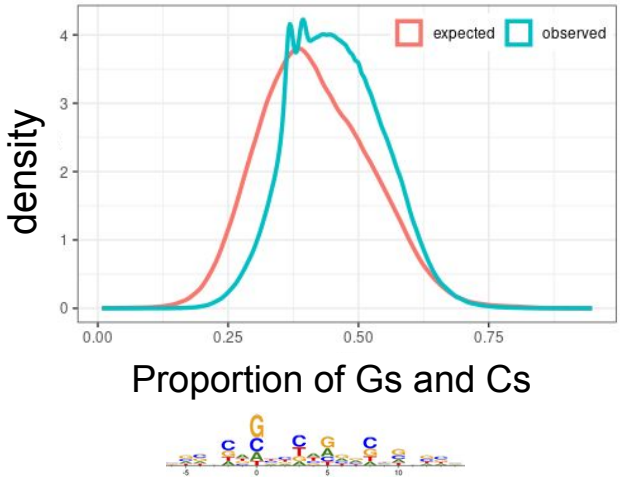
Changes:

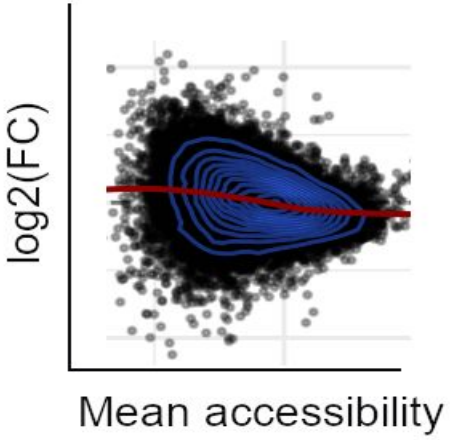
- Make sure that your peaks are resized to a comparable size
- Compute chromVAR deviations using a much larger number of backgrounds (e.g. 1000)
- Use limma's moderated statistics instead of standard t-tests

(Gerbaldo, Sonder et al., PLOS Comp Bio 2024)

Three inter-related types of technical variations in ATACseq data:







Fragment length bias

GC bias

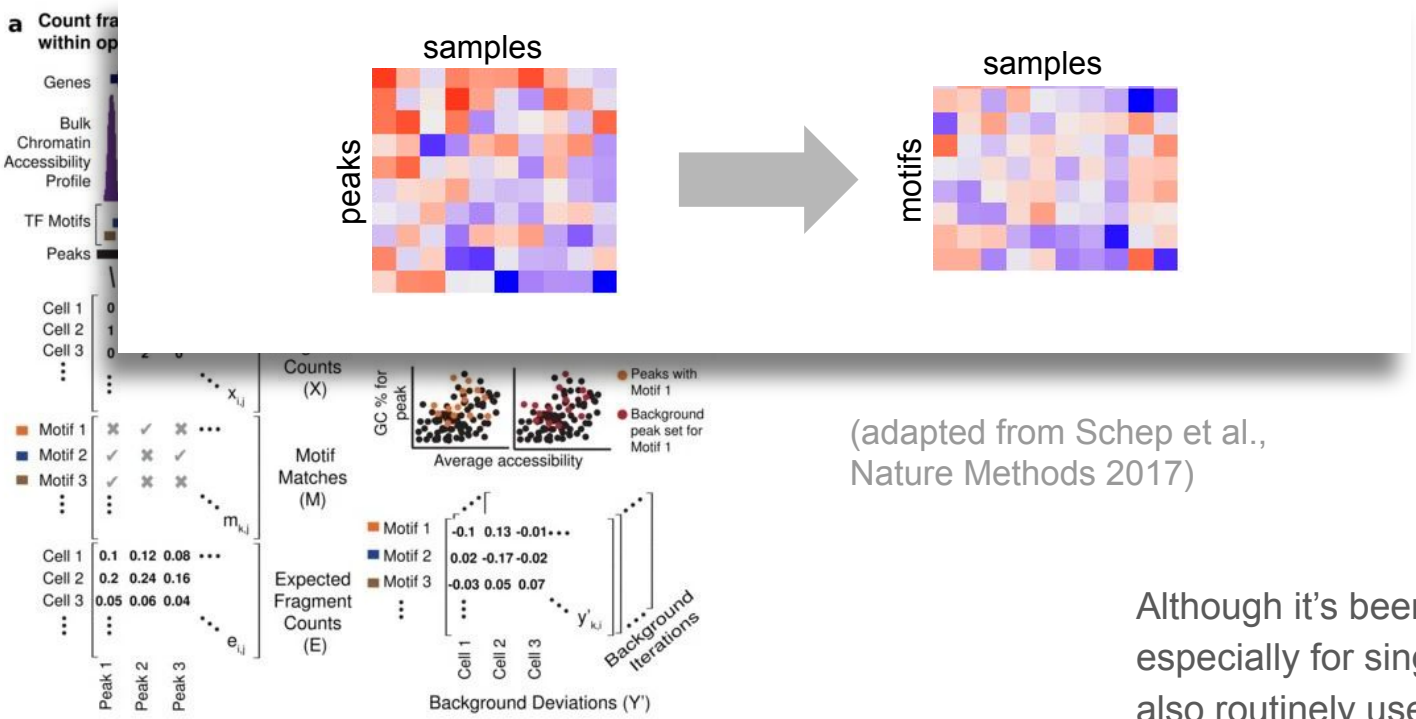
Enrichment bias

Shorter fragments typically come from regulatory elements, which are both **GC-rich** and **highly-accessible**

ChromVAR

documentation

chromVAR uses a simpler (and considerably faster) method, which essentially sums the counts of peaks that contain each motif, and compares this to a null distribution of similar peaks that don't.



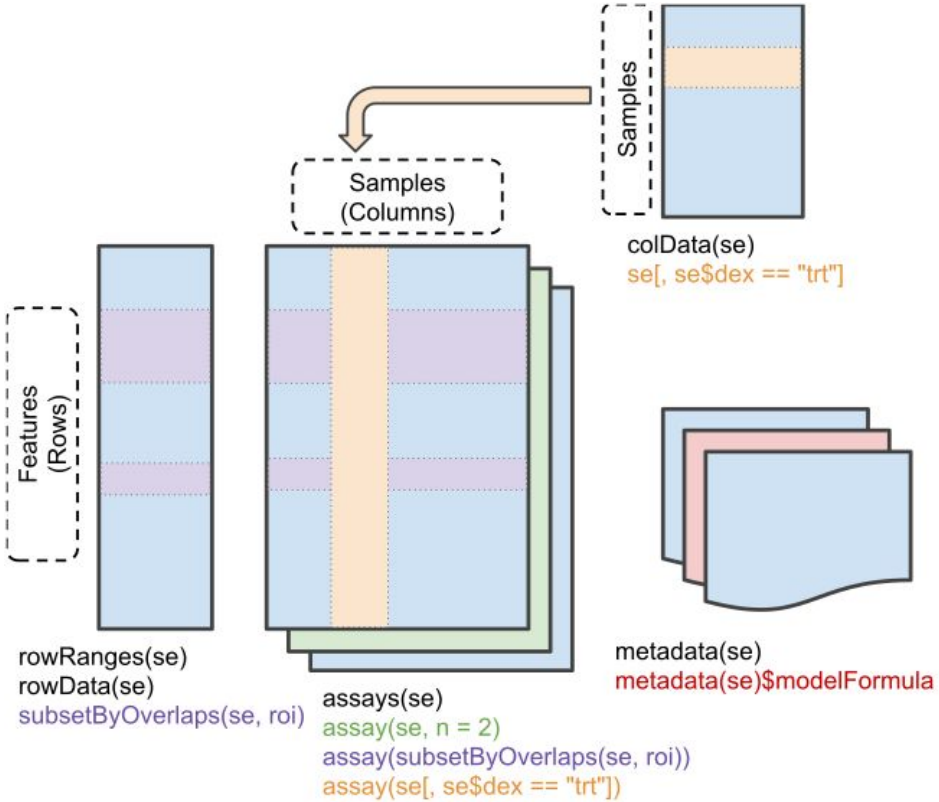
(adapted from Schep et al., Nature Methods 2017)

Although it's been developed especially for single-cell data, it's also routinely used for bulk.

Practical:

Motif accessibility analysis with chromVAR

The SummarizedExperiment structure

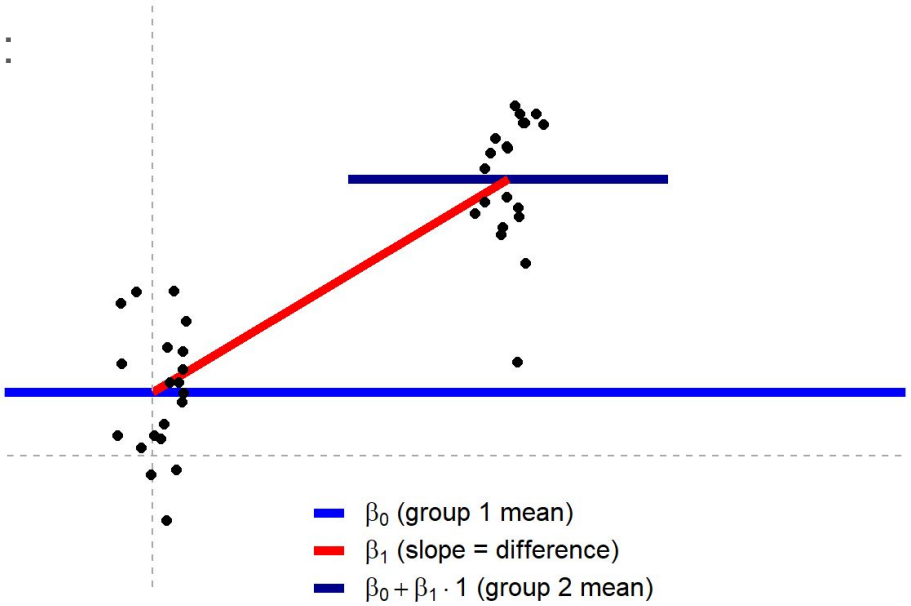


Documentation

Working with linear models

Most of the common statistical tests can be formulated as linear regression problems

Consider the *t*-test :



(Taken from an excellent explanation by Jonas Kristoffer Lindeløv)

$y \sim \beta_0 + x * \beta_1$

Where x=0 for group 1
and x=1 for group 2

Working with linear models - some simple examples

- For a two group comparison:
  - `~group` (equivalent to `~1+group`)

- Comparing between two groups, correcting for the effect of sex:
  - `~sex+group` → then we can still decide to test for the effects of the group by dropping that coefficient

- Finding sex-specific effects
  - `~sex*group` → equivalent to `~1+sex+group+sex:group`

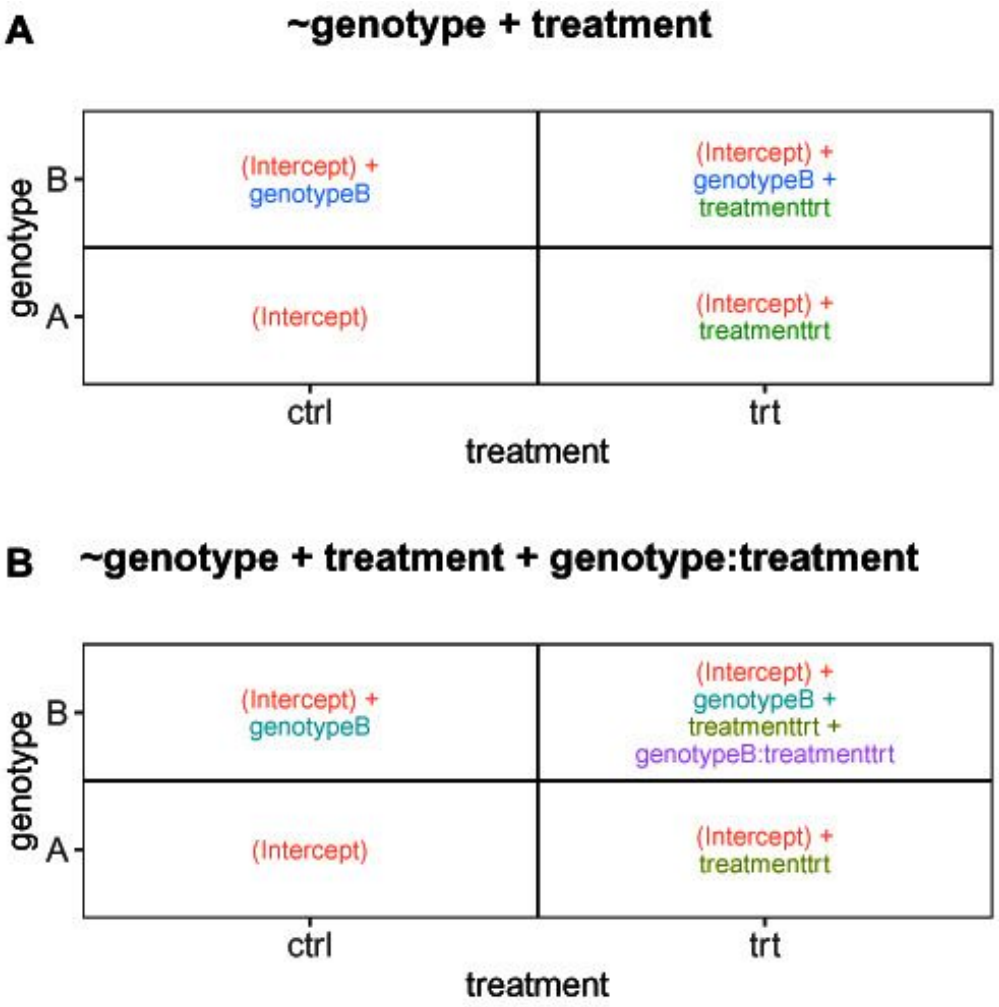
Working with linear models

Suppose

two genotypes (A & B)

and

treatment (trt) vs
control (ctrl)



(Soneson et al., *f1000* 2020)

Working with linear models - coefficients & contrasts

Testing coefficients:

```
~ genotype + treatment + genotype:treatment
```

The model matrix has the following coefficients:

```
intercept, genotypeB, treatmentTrt,
genotypeB:treatmentTrt
```

Asking for the effect of the treatment in genotype A amounts to testing coefficient `treatmentTrt`

Testing with contrasts:

Define ‘group’ as the combination of genotype & treatment, i.e. A_ctrl, B_ctrl, A_trt, B_trt

```
~ 0 + group
```

The model matrix has the following coefficients:

```
A_ctrl, B_ctrl, A_trt, B_trt
```

Asking for the effect of the treatment in genotype A amounts to testing the contrast: `A_trt - A_ctrl`

equivalent

For more information on the use of contrasts, see for instance the limma user guide

Assignment

- Download (a subset of) ATAC-seq peak counts in the hippocampus upon stress (already in SummarizedExperiment format) :
  - https://ethz-ins.org/content/mouse_mm38_hippocampus.peakCounts.SE.rds
  - (the data is from the mouse ensembl GRCm38 genome – you should already have the genome, e.g. from week 6)
- Using this object, perform a chromVAR motif analysis, and run 2 differential motif accessibility analyses, respectively:
  - comparing stressed (denoted ‘FSS’ – forced swim stress) and control animals
  - comparing male and female animals
- For each analysis, report the top most significant motifs, plot a heatmap of the normalized accessibility scores across the samples for those motifs, and write a short paragraph interpreting the results.

Start to think about your course project…

- In teams of 2-4
- Due on the 2nd of july

- The project can for example be:
  - Re-producing the analyses from a publication (in a critical fashion)
  - Analyzing new data (e.g. yours or in collaboration with a group)
  - Exploring the differences between a given set of TFs
  - Anything you think of which involves competences developed in the course

- Some project ideas are in the #general slack channel