Bioinformatic approaches to regulatory genomics and epigenomics

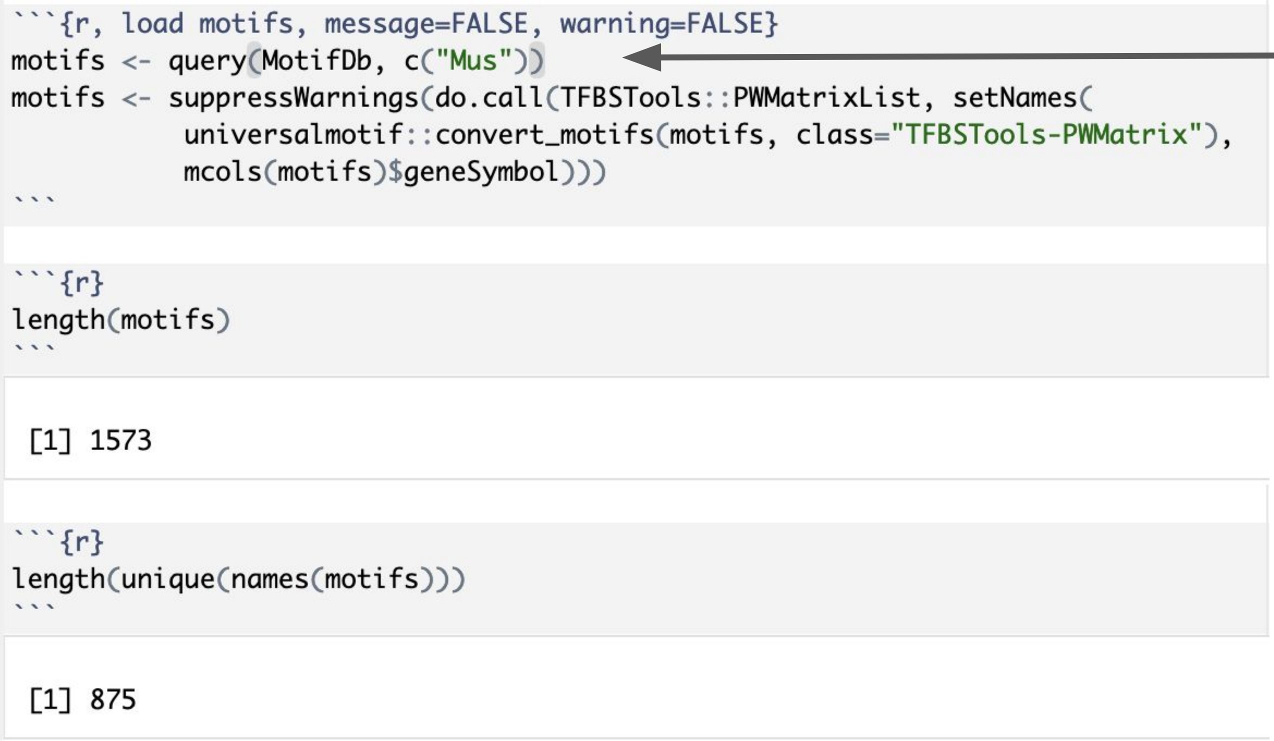
376-1347-00L | week 09

Pierre-Luc Germain

Plan for today

- Debriefing on the assignment
- Quick primer on count-based (discrete) statistics
- Our case study for the practical: GDVS
- Differential binding analysis
- Normalization methods

Debriefing: Redundant motifs



~~~
```{r, load motifs, message=FALSE, warning=FALSE}
motifs <- query(MotifDb, c("Mus"))
motifs <- suppressWarnings(do.call(TFBSTools::PWMatrixList, setNames(
           universalmotif::convert_motifs(motifs, class="TFBSTools-PWMatrix"),
           mcols(motifs)$geneSymbol)))
```
~~~

~~~
```{r}
length(motifs)
```
~~~

```
 [1] 1573
```

~~~
```{r}
length(unique(names(motifs)))
```
~~~

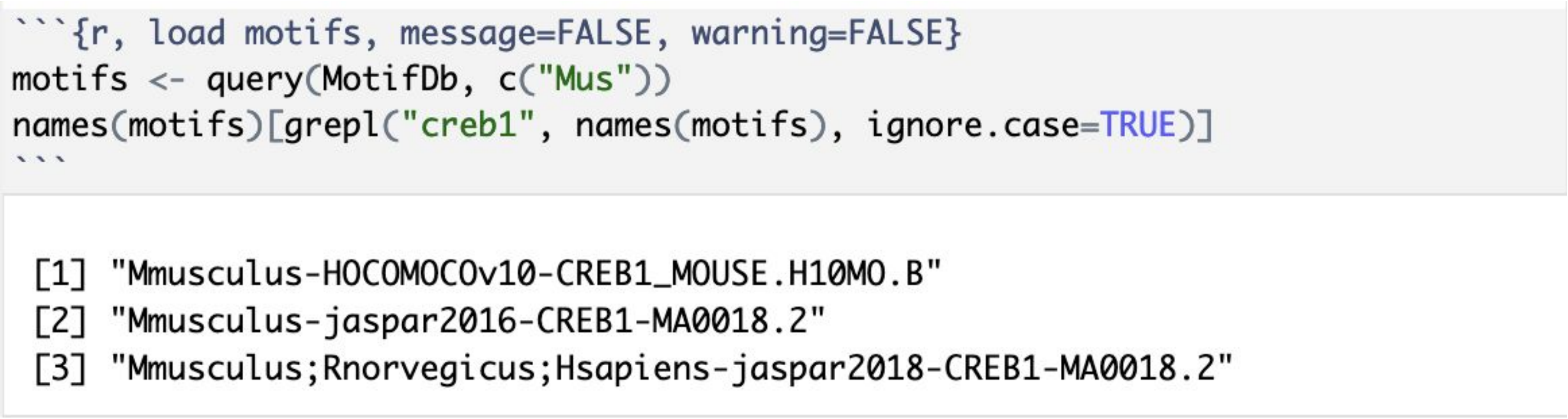
```
 [1] 875
```

if motifs are searched like this, all DBs linked via MotifDb will be queried for Mouse motifs

This leads to many duplicated motifs

Debriefing: Redundant motifs

hence the same motif might have entries in several databases, e.g. here in Jaspar and HOCOMOCO



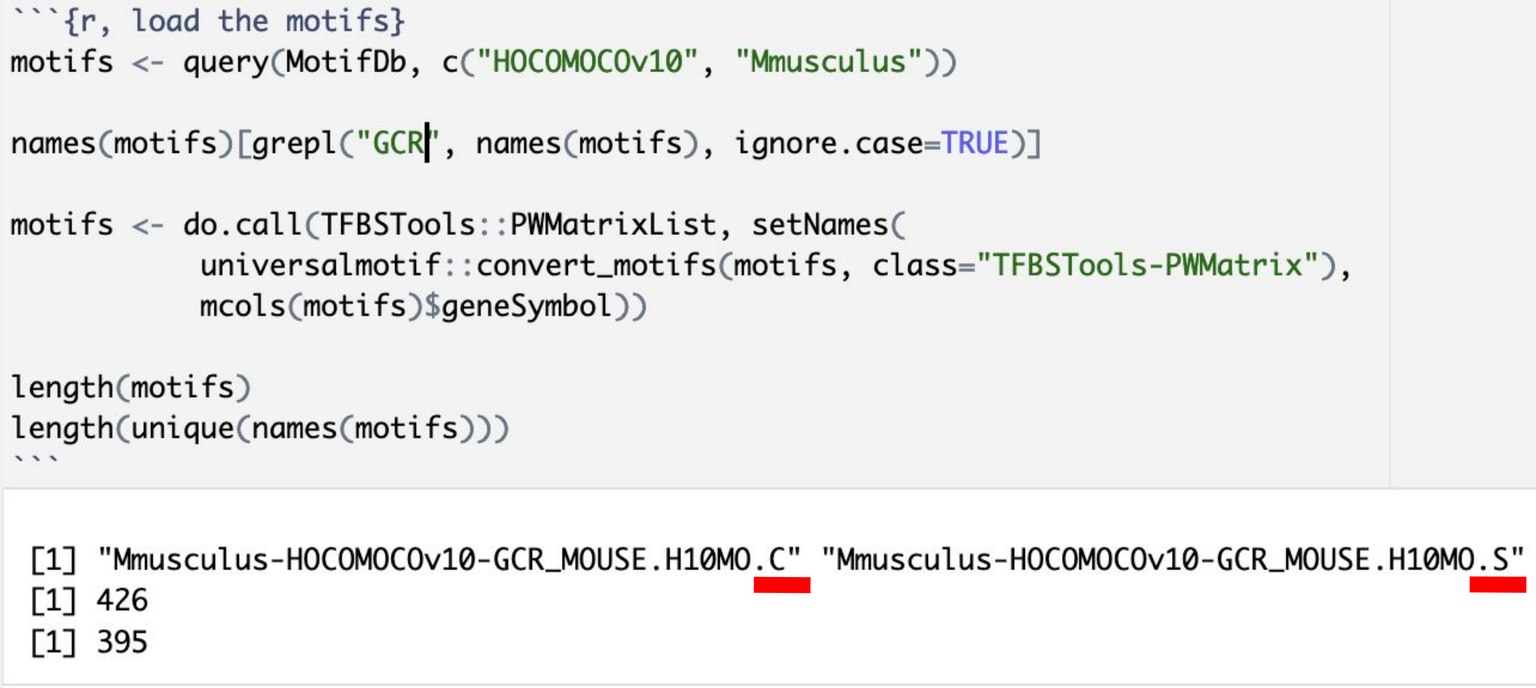
~~~
```{r, load motifs, message=FALSE, warning=FALSE}
motifs <- query(MotifDb, c("Mus"))
names(motifs)[grepl("creb1", names(motifs), ignore.case=TRUE)]
```
~~~

```
 [1] "Mmusculus-HOCOMOCOv10-CREB1_MOUSE.H10MO.B"
 [2] "Mmusculus-jaspar2016-CREB1-MA0018.2"
 [3] "Mmusculus;Rnorvegicus;Hsapiens-jaspar2018-CREB1-MA0018.2"
```

=> this will influence downstream tasks, e.g the multiple-testing correction.

Debriefing: Redundant motifs

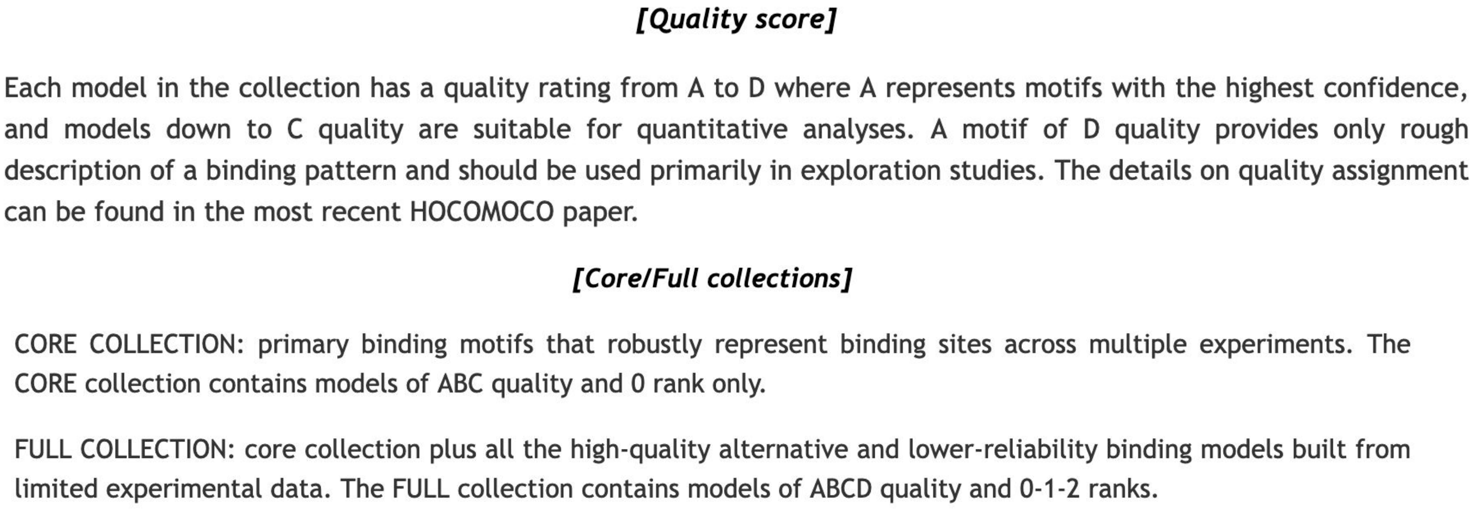
problem can persist (but less duplicates) when downloading from one DB.



~~~
```{r, load the motifs}
motifs <- query(MotifDb, c("HOCOMOCOv10", "Mmusculus"))

names(motifs)[grepl("GCR", names(motifs), ignore.case=TRUE)]

motifs <- do.call(TFBSTools::PWMatrixList, setNames(
           universalmotif::convert_motifs(motifs, class="TFBSTools-PWMatrix"),
           mcols(motifs)$geneSymbol))

length(motifs)
length(unique(names(motifs)))
```
~~~

```
 [1] "Mmusculus-HOCOMOCOv10-GCR_MOUSE.H10MO.C" "Mmusculus-HOCOMOCOv10-GCR_MOUSE.H10MO.S"
 [1] 426
 [1] 395
```

Debriefing: Redundant motifs



[Quality score]

Each model in the collection has a quality rating from A to D where A represents motifs with the highest confidence, and models down to C quality are suitable for quantitative analyses. A motif of D quality provides only rough description of a binding pattern and should be used primarily in exploration studies. The details on quality assignment can be found in the most recent HOCOMOCO paper.

[Core/Full collections]

CORE COLLECTION: primary binding motifs that robustly represent binding sites across multiple experiments. The CORE collection contains models of ABC quality and 0 rank only.

FULL COLLECTION: core collection plus all the high-quality alternative and lower-reliability binding models built from limited experimental data. The FULL collection contains models of ABCD quality and 0-1-2 ranks.

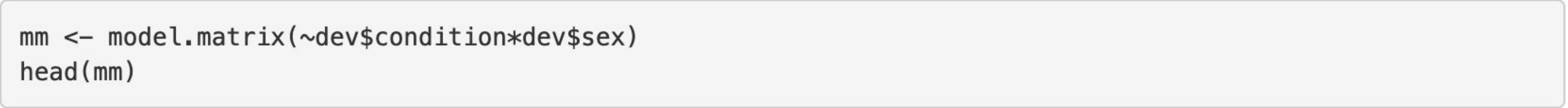
from: https://hocomoco11.autosome.org/help

P,S,M suffixes, can be P (ChI**P**-Seq), S (HT-**S**ELEX), or M (**M**ethyl-HT-SELEX), or any combination of those three for motifs found in several types of experiments.

from: https://hocomoco12.autosome.org/faq#faq

Debriefing: sex-specific responses to stress

if we want to find sex-specific responses to the stressor, we do need to include an interaction term:



```
mm <- model.matrix(~dev$condition*dev$sex)
head(mm)
```

note:
~ A*B is the same as ~ A + B + A:B

Debriefing on the assignments

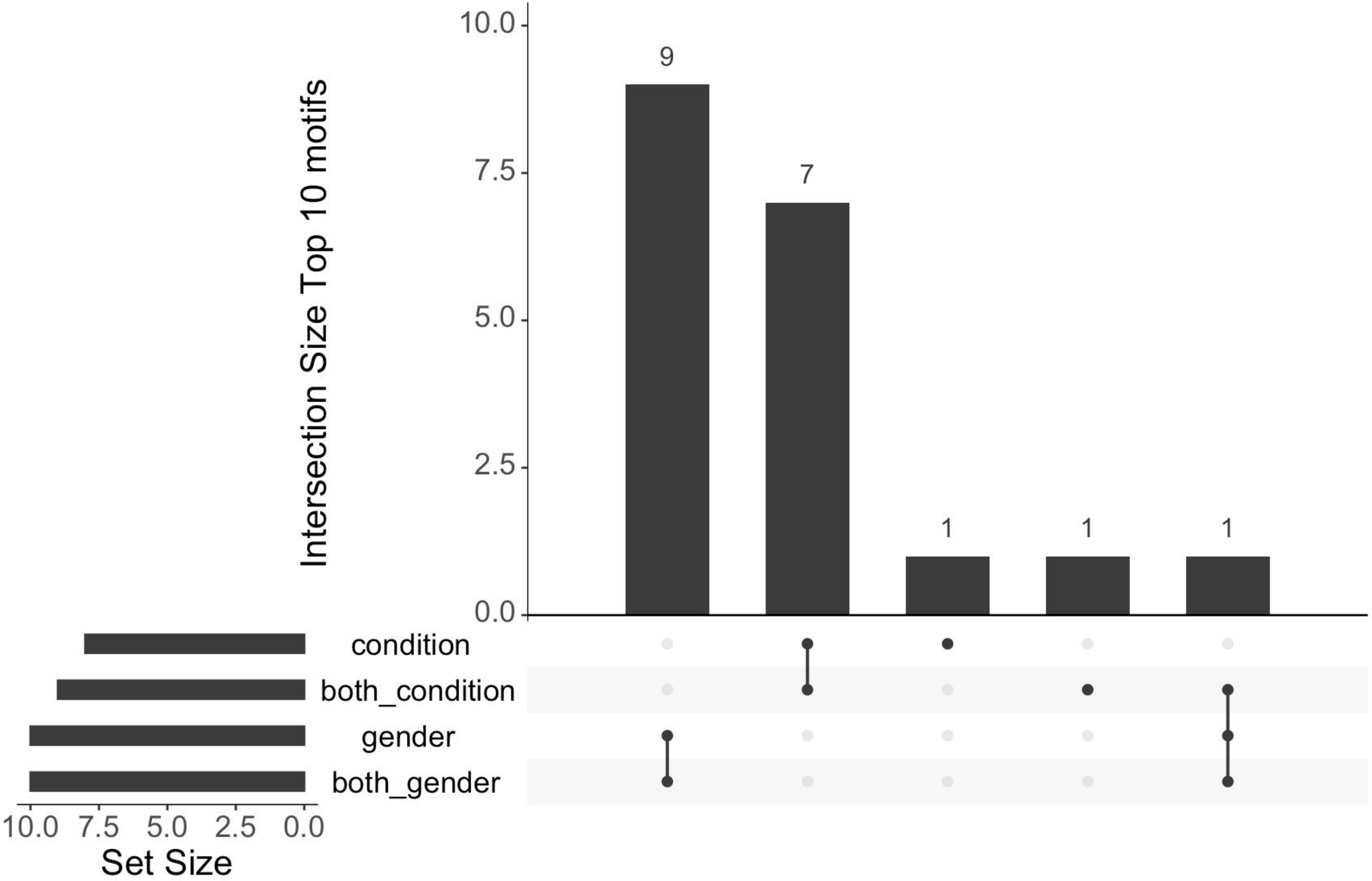
Statistical testing:

```
mm <- model.matrix(~sex+condition, ...)
```

and

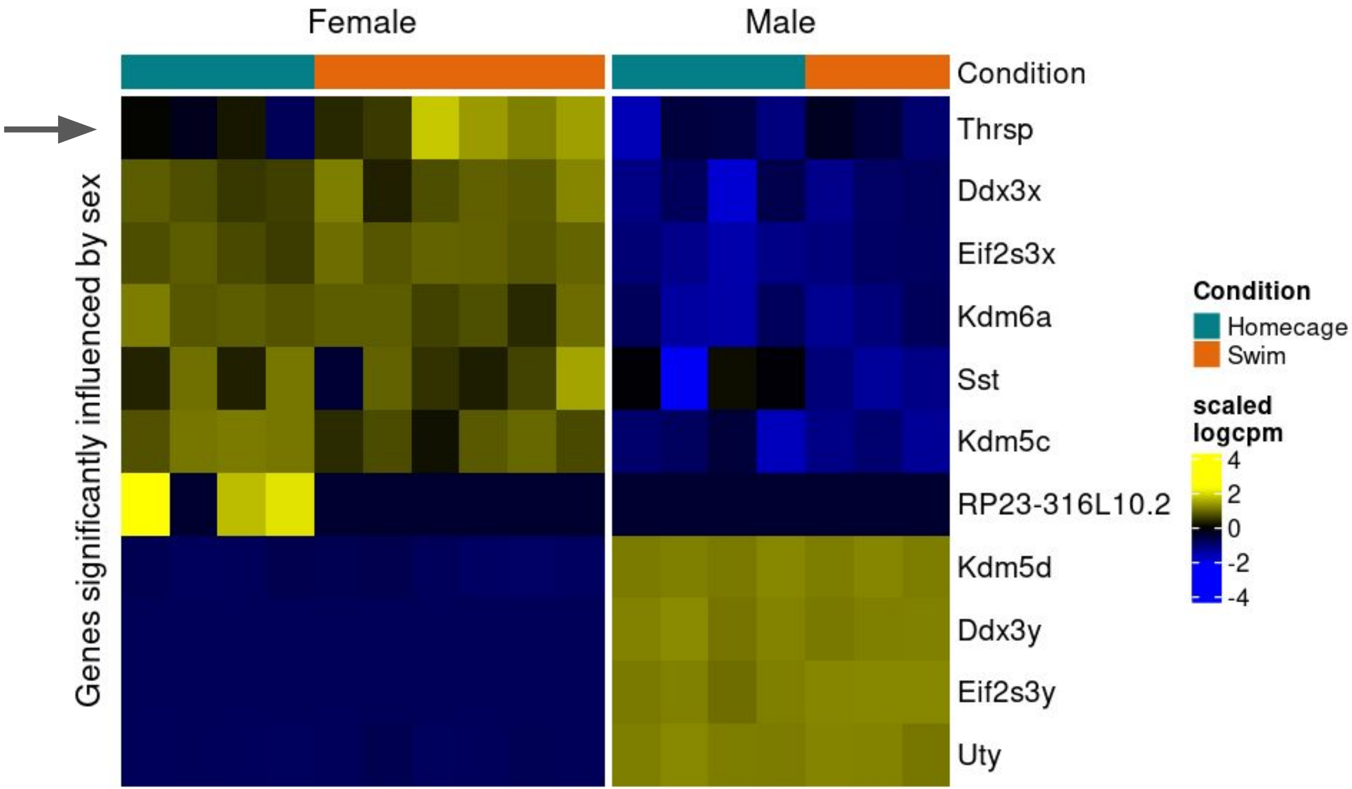
```
mm <- model.matrix(~sex, ...)

mm <- model.matrix(~condition, ...)
```

is not equivalent!


Intersection Size Top 10 motifs
10.0
7.5
5.0
2.5
0.0
9
7
1
1
1
condition
both_condition
gender
both_gender
10.0
7.5
5.0
2.5
0.0
Set Size

An example (from RNAseq)

Since some genes are stress-responsive, if the model does not include the condition the variability might appear high, leading to a failure to detect sex differences.



Count-based statistics

Count-based statistics

The **number of fragments** overlapping a given peak is not a **continuous** variable, but a **discrete** variable: it can only be non-negative integers.

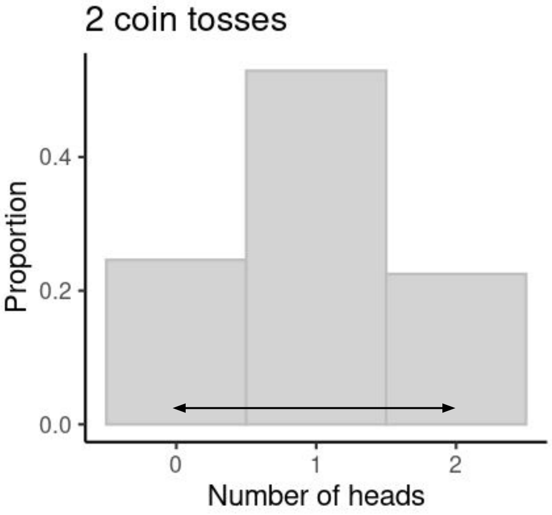
We must consider this information in the statistical analysis of count data, and so we can’t use statistics based on a normal distribution (e.g. t-tests or standard linear models).

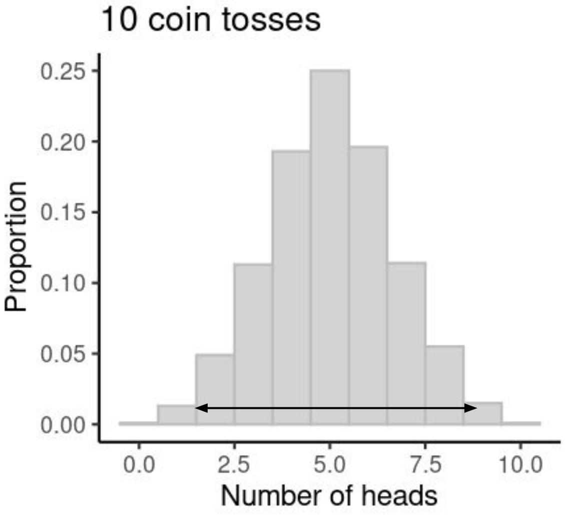
We’ll use **generalized linear models**, which can be applied on different types of data distributions.

Count-based statistics

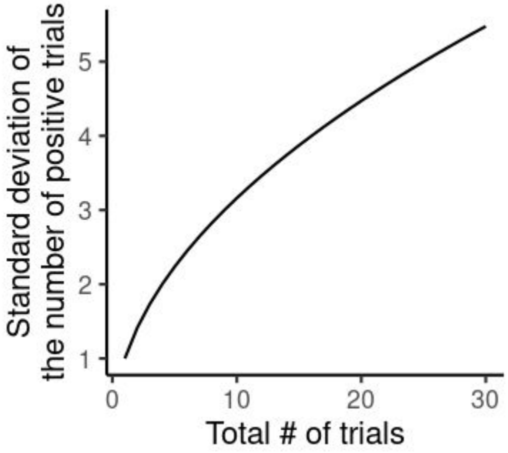
Consider coin tossing, where each toss can result in two outcomes.

Suppose each of us toss a coin X times, and we count the number of ‘heads’.





The results are Poisson-distributed, meaning that the variance is equal to the number of tosses



Count-based statistics

However, counts coming from real phenomena (more complicated than coin tossing) are typically not quite Poisson-distributed: the variance typically increases *more* than would be expected, in other words they show **over-dispersion**.

This is typically because of sources of variation other than the random sampling process, which are typically unknown.

For this reason, count statistics are most often analyzed with the **negative binomial distribution**, which includes an extra dispersion parameter.

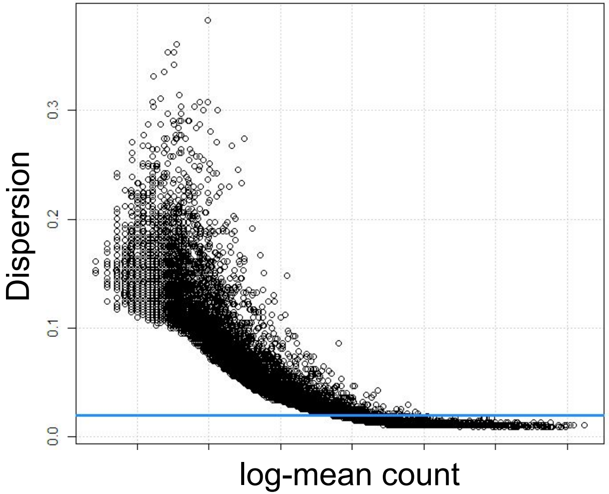
Information-sharing across features

In principle, each feature (e.g. gene or peak) we are analyzing could have a different degree over-dispersion, so that we would have to estimate this parameter separately for each of them.

Typically, we don't have enough samples to do this accurately.

RNAseq analysis packages such as [edgeR]() therefore rely on a trend between average counts and their overdispersion to moderate those estimates.

We can rely on the same packages for epigenomic data.

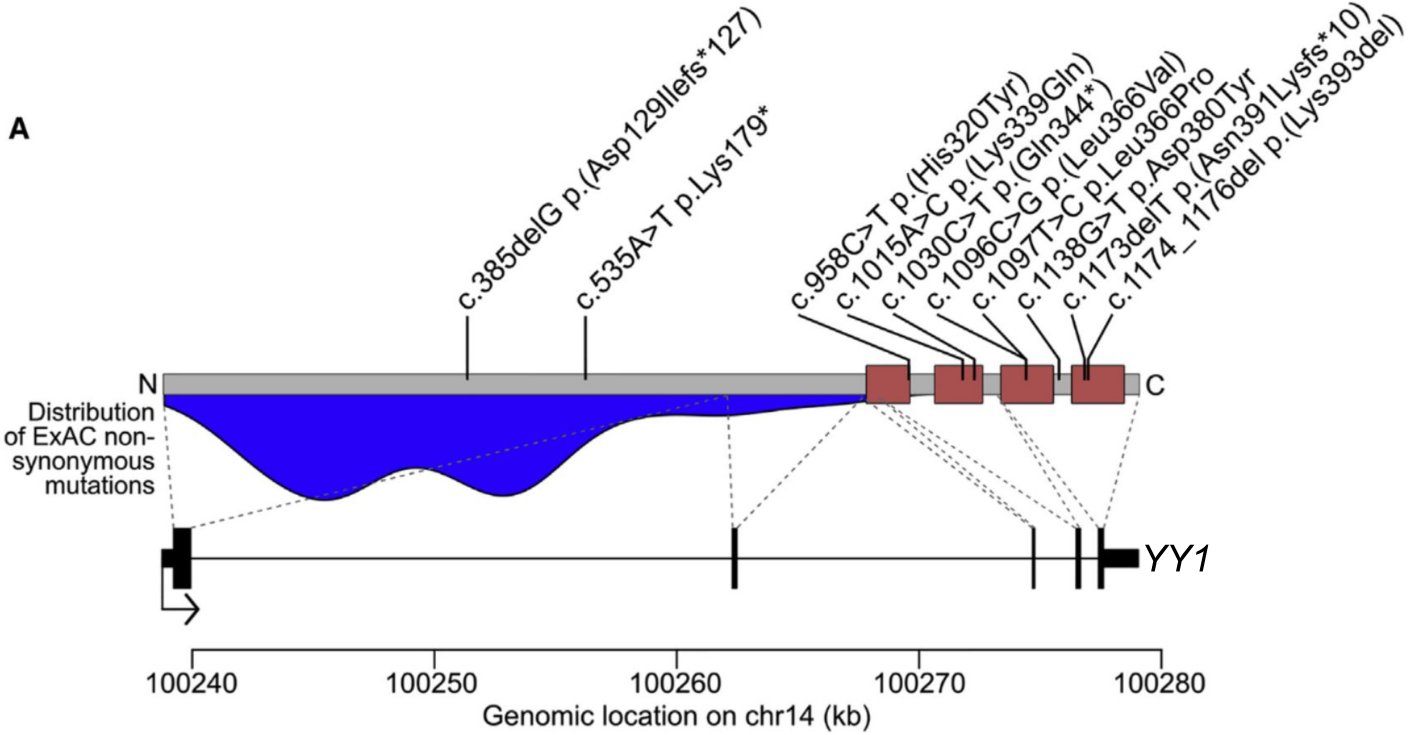


Our case study today

- 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

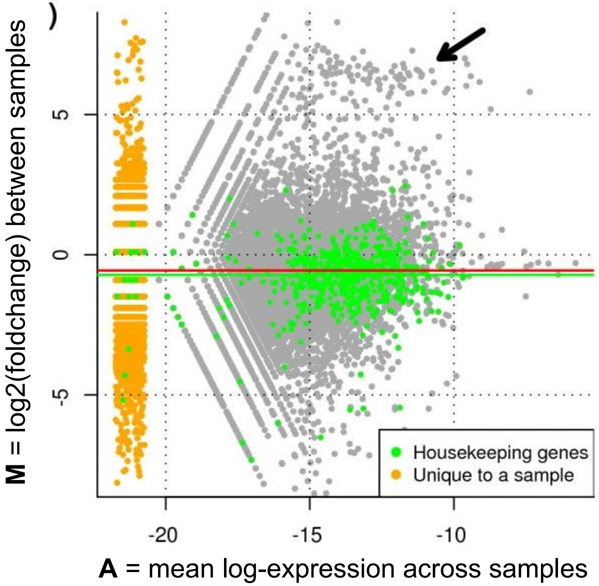


(Gabriele, Vulto-van Silfhout, Germain et al., AJHG 2017)

Practical

Normalization

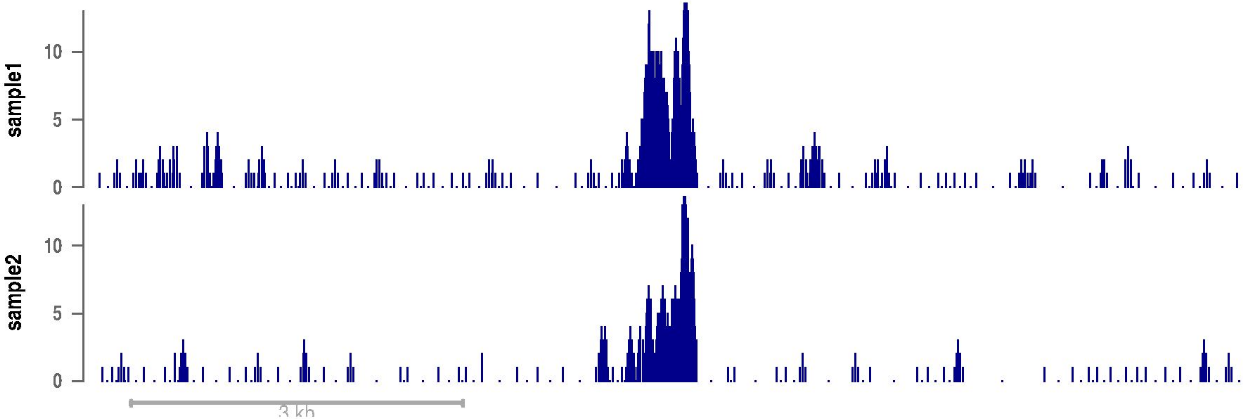
- Standard **TMM normalization** in edgeR assumes that the majority of regions don’t change across condition.



(Adapted from Robinson & Oshlack, Genome Biology 2010)

Normalization

- Standard **TMM normalization** in edgeR assumes that the majority of regions (i.e. peaks) 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 background noise is the same across experiments

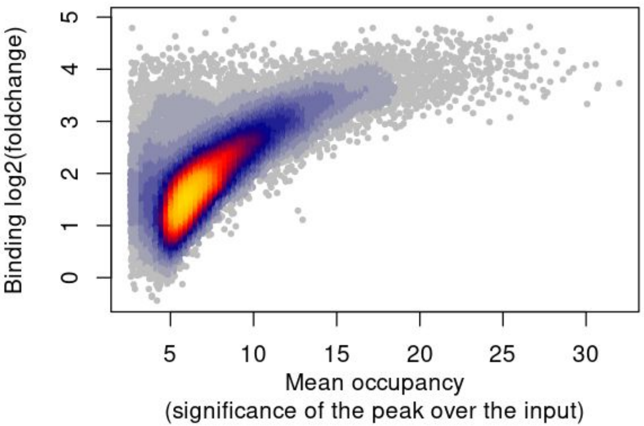


Normalization

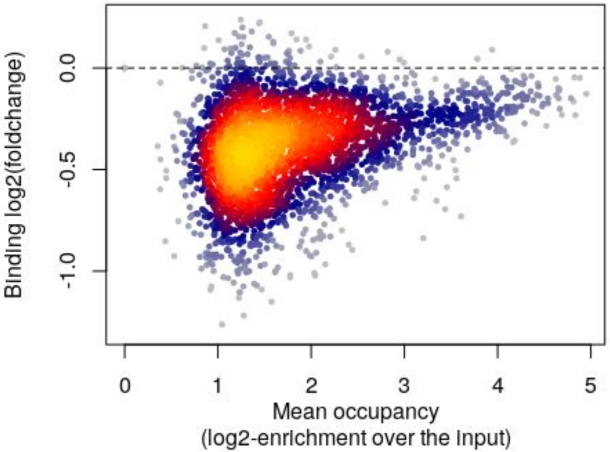
- Standard **TMM normalization** in edgeR assumes that the majority of regions (i.e. peaks) don’t change across condition.
    - This does not work when there are global increases or decreases (i.e. at most sites) between conditions
- **Background normalization** assumes that the background noise is the same across experiments
    - This often works nicely, but can create artifactual differences in highly-enriched regions when the quality of the enrichment differs a lot between experiments
- Common (or top) peak normalization (e.g. **MAnorm**, 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

When the amount/activity of a TF is changed, not all sites change in binding equally

From inactive to active nuclear receptor



YY1 haploinsufficiency

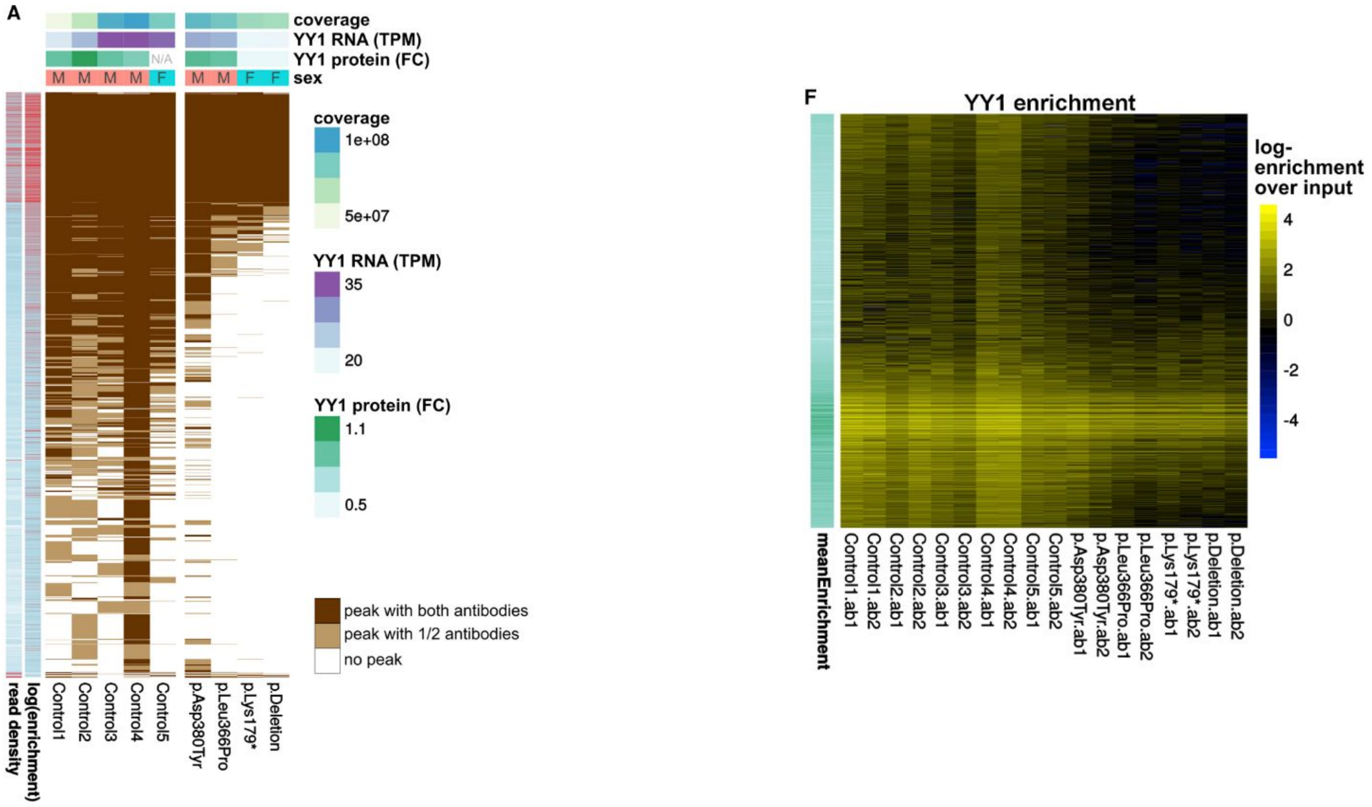


When activating a TF that wasn't active at all, the TF goes to high-occupancy sites, and these sites tend to show the largest changes in binding

When reducing a TF that's already active, the TF also tends to keep going to the high-occupancy sites, so the low-occupancy sites are those that tend to show the largest change

* co-factors can influence these dynamics

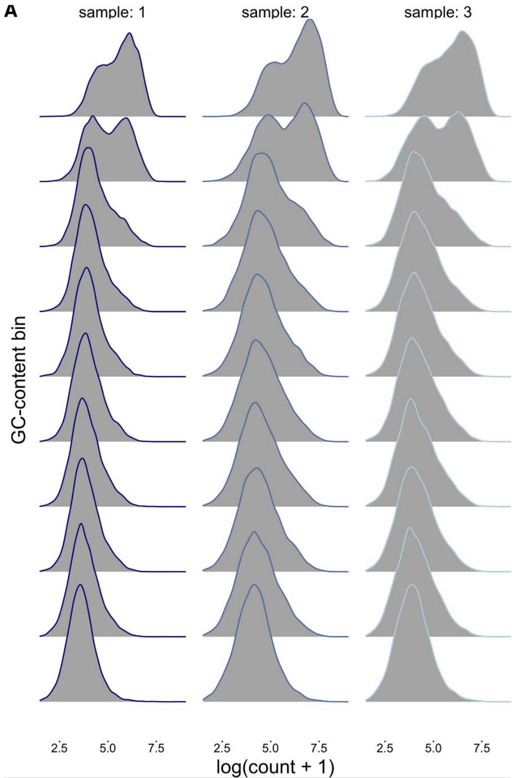
Haploinsufficiency affects especially low-occupancy regions



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

Further methods for the normalization of technical biases

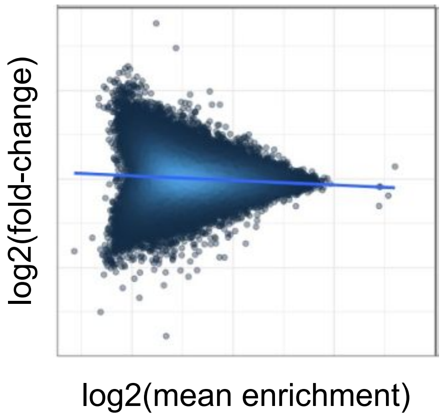
Specifically for ATAC-seq: smooth quantile normalization over GC bins



Improves most downstream tasks (differential expression, TF activity, etc.)

(adapted from Van den Berge et al., *Cell Reports Methods* 2022)

loess normalization for enrichment/concentration bias



(see e.g. the *csaw* package – Lun & Smyth, NAR 2014)

Assignment

Until next week, form teams of 2-3, and 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 render and upload to your repository.