Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 13

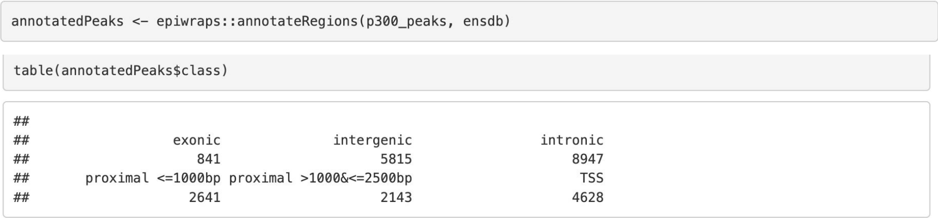
Pierre-Luc Germain

ETH Zürich

Plan for today

- Debriefing on the assignment
- Single-cell epigenomic data (scATAC-seq)
- Chromatin and disease
- Our lines of research
- Evaluation scheme for the course project

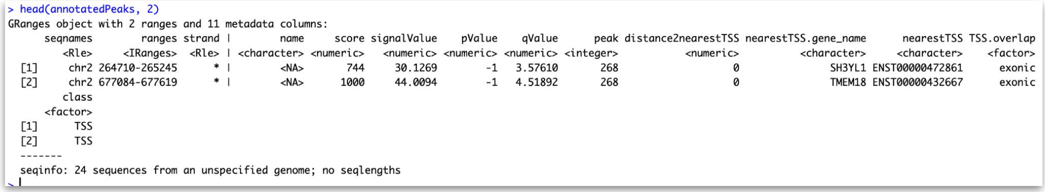
Debriefing: W12



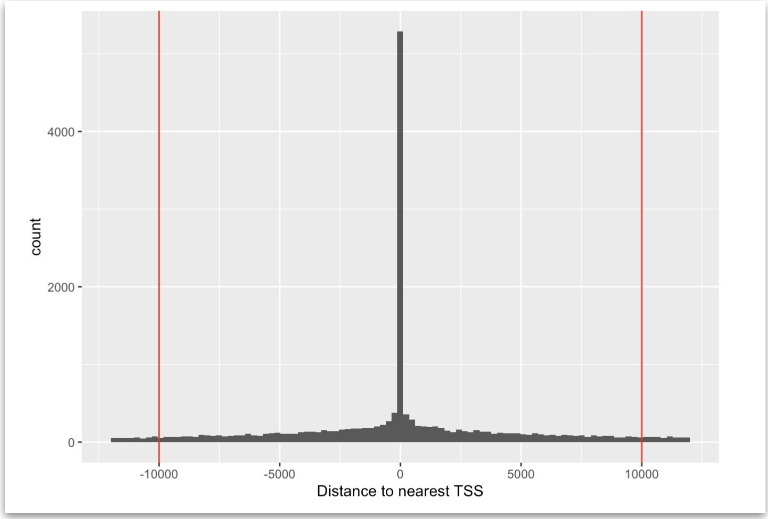
```
annotatedPeaks <- epiwraps::annotateRegions(p300_peaks, ensdb)
```

```
table(annotatedPeaks$class)
```

```
## 
##                 exonic              intergenic                intronic 
##                    841                    5815                    8947 
##       proximal <=1000bp proximal >1000&<=2500bp                     TSS 
##                   2641                    2143                    4628 
```

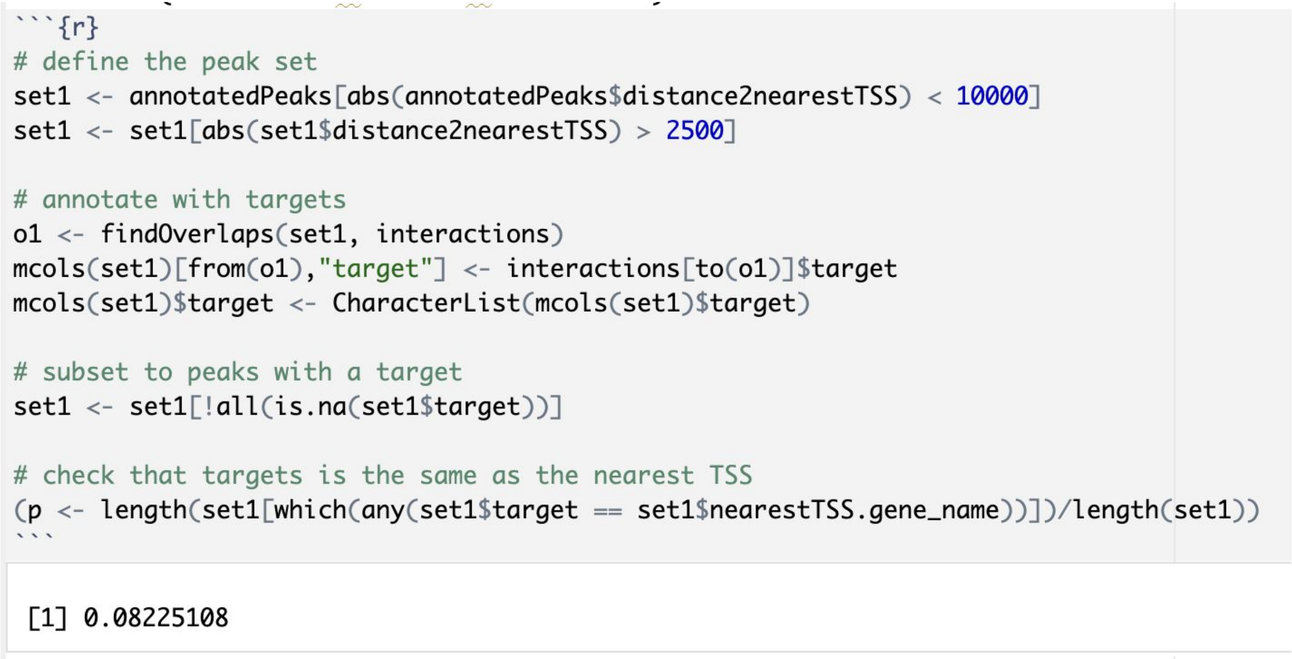


```
> head(annotatedPeaks, 2)
GRanges object with 2 ranges and 11 metadata columns:
     seqnames        ranges strand |        name     score signalValue    pValue    qValue      peak distance2nearestTSS nearestTSS.gene_name      nearestTSS TSS.overlap
        <Rle>     <IRanges>  <Rle> | <character> <numeric>   <numeric> <numeric> <numeric> <integer>           <numeric>          <character>     <character>    <factor>
  [1]    chr2 264710-265245      * |        <NA>       744     30.1269        -1   3.57610       268                   0               SH3YL1 ENST00000472861      exonic
  [2]    chr2 677084-677619      * |        <NA>      1000     44.0094        -1   4.51892       268                   0               TMEM18 ENST00000432667      exonic
        class
     <factor>
  [1]     TSS
  [2]     TSS
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
> 
```



Debriefing W12: Finding the proportions

P300 peaks between 2.5kb and 10kb from a TSS

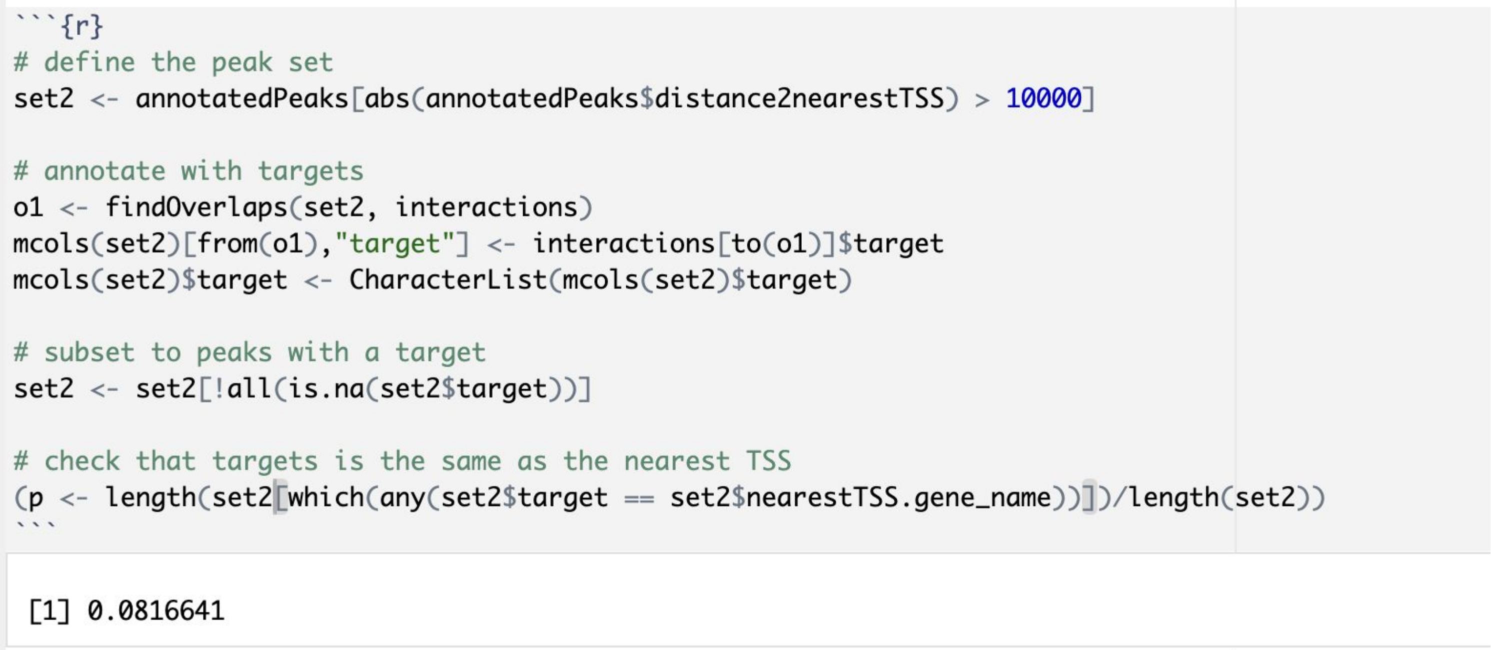


~~~
```{r}
# define the peak set
set1 <- annotatedPeaks[abs(annotatedPeaks$distance2nearestTSS) < 10000]
set1 <- set1[abs(set1$distance2nearestTSS) > 2500]

# annotate with targets
o1 <- findOverlaps(set1, interactions)
mcols(set1)[from(o1),"target"] <- interactions[to(o1)]$target
mcols(set1)$target <- CharacterList(mcols(set1)$target)

# subset to peaks with a target
set1 <- set1[!all(is.na(set1$target))]

# check that targets is the same as the nearest TSS
(p <- length(set1[which(any(set1$target == set1$nearestTSS.gene_name))])/length(set1))
```
~~~

```
[1] 0.08225108
```

Debriefing W12: Finding the proportions

P300 peaks more than 10kb from a TSS

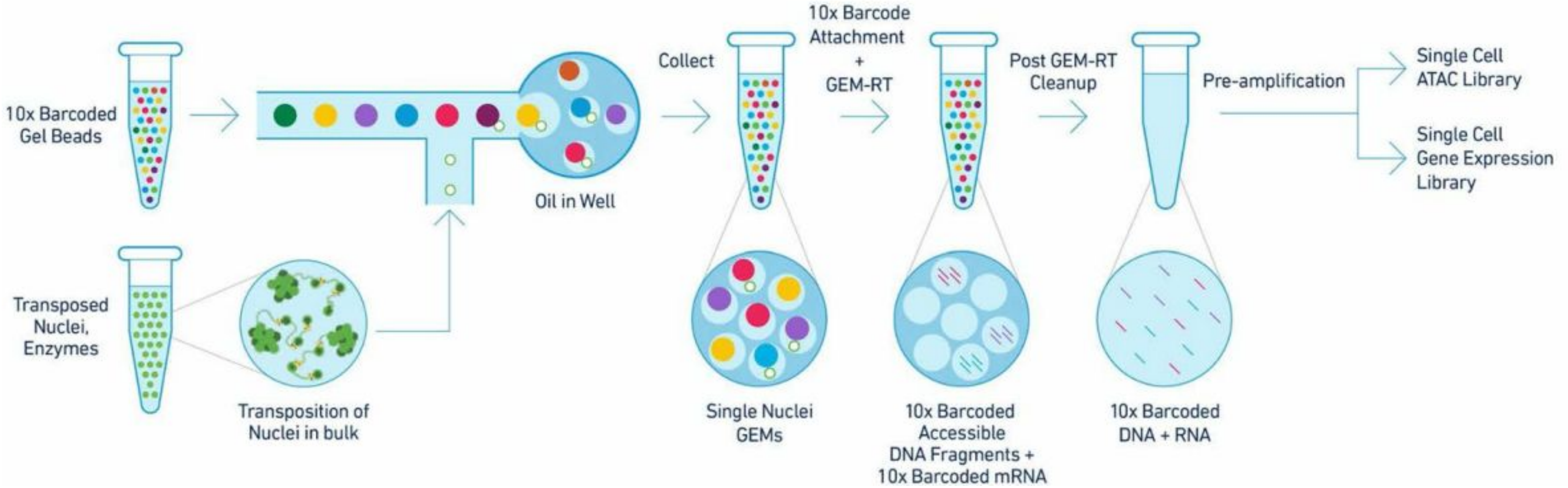


~~~
```{r}
# define the peak set
set2 <- annotatedPeaks[abs(annotatedPeaks$distance2nearestTSS) > 10000]

# annotate with targets
o1 <- findOverlaps(set2, interactions)
mcols(set2)[from(o1),"target"] <- interactions[to(o1)]$target
mcols(set2)$target <- CharacterList(mcols(set2)$target)

# subset to peaks with a target
set2 <- set2[!all(is.na(set2$target))]

# check that targets is the same as the nearest TSS
(p <- length(set2[which(any(set2$target == set2$nearestTSS.gene_name))])/length(set2))
```
~~~

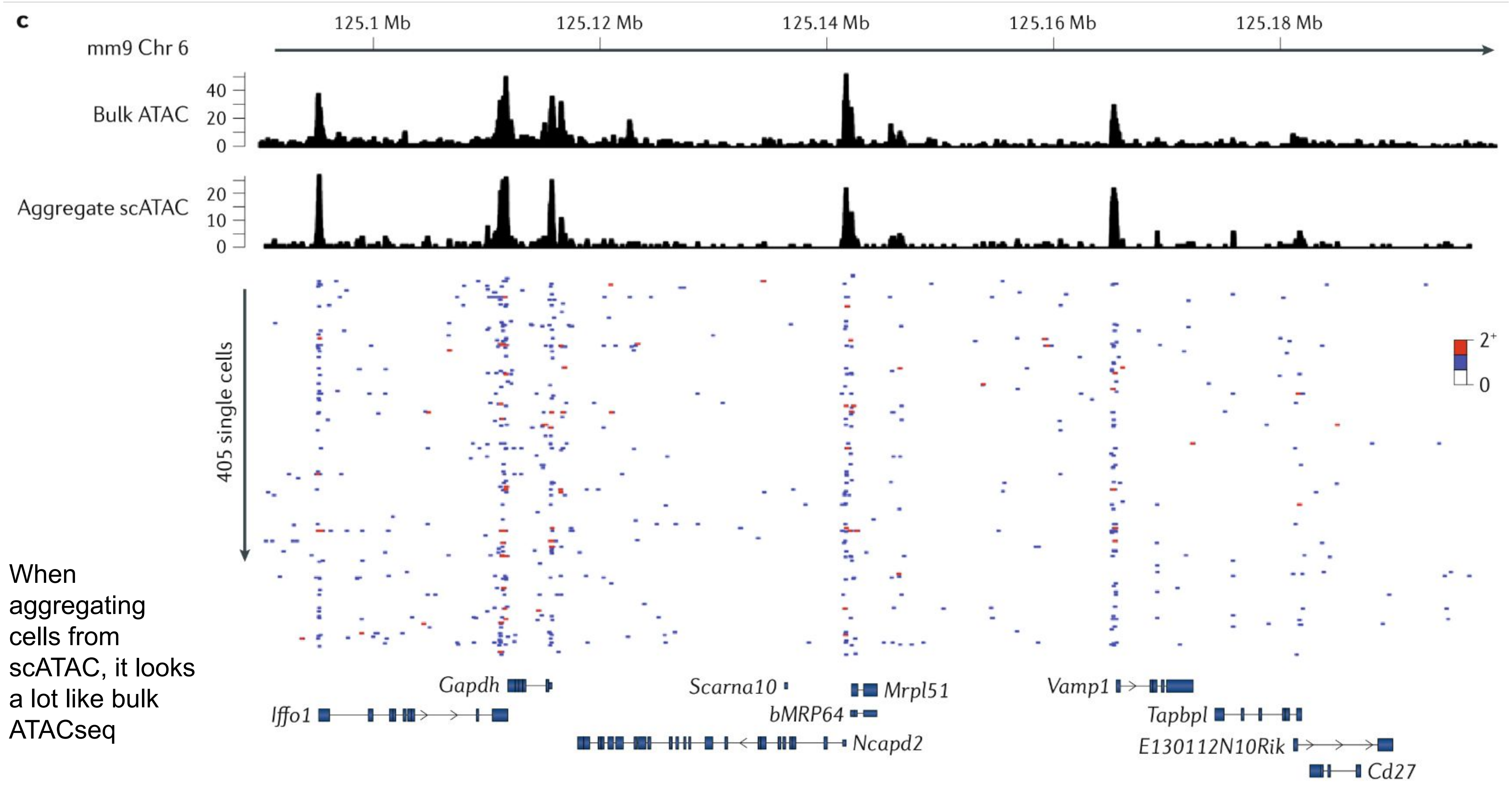
```
[1] 0.0816641
```

Single-cell *-omics

Single-cell ATAC-seq (and multi-omics)


10x Barcoded Gel Beads
Transposed Nuclei, Enzymes
Transposition of Nuclei in bulk
Oil in Well
Collect
10x Barcode Attachment + GEM-RT
Single Nuclei GEMs
Post GEM-RT Cleanup
10x Barcoded Accessible DNA Fragments + 10x Barcoded mRNA
Pre-amplification
10x Barcoded DNA + RNA
Single Cell ATAC Library
Single Cell Gene Expression Library

10x Genomics



When aggregating cells from scATAC, it looks a lot like bulk ATACseq

(Mezger et al, Nat Comm 2018)

Single-cell ATACseq analysis in a nutshell

1. The output of the genome alignment of the data is a “fragment file”, a bed-like file containing the coordinates of each fragment and the associated cell (barcode)

```
#chr    start   end     cell_barcode
chr1    10066   10536   TCAAGCAGTGCGCATC-1
chr1    10073   10278   TCAAGACGTCTGATTG-1
chr1    10073   10305   CGTTCCACAGCGTAGA-1
chr1    10079   10315   TTCAACTTCCGAGAGA-1
chr1    10085   10278   TCGTTCGCATAGGCGA-1
chr1    10091   10303   AGCGTGCTCCCATAGA-1
```

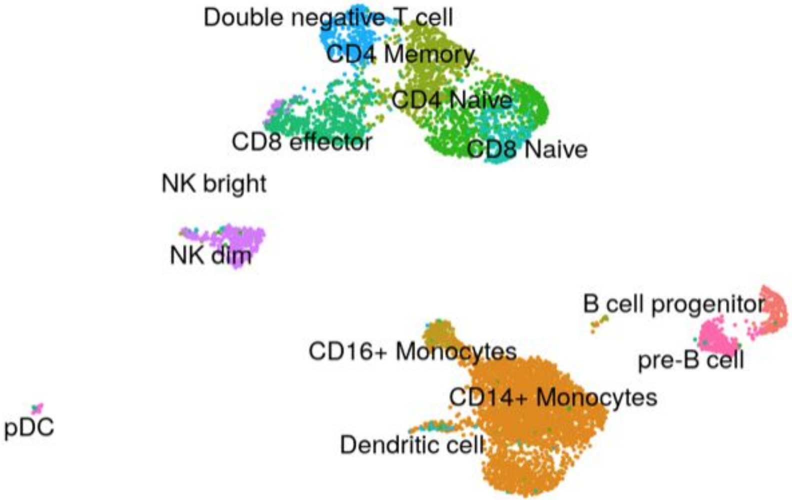
2. From this, we count the number of fragments from each cell overlapping genomic windows (either whole-genome tiles or feature-based)

```
        cell1 cell2 cell3 cell4 cell5 ...
window1   1     0     0     0     0
window2   0     0     0     0     1
window3   0     0     1     0     0
window4   0     0     0     0     0
window5   0     1     0     0     0
...
```

Can be a matrix with hundreds of thousands windows…

(filtering…)

3. Normalization, dimensional reduction (e.g. TF-IDF + LSI), and clustering of the cells:

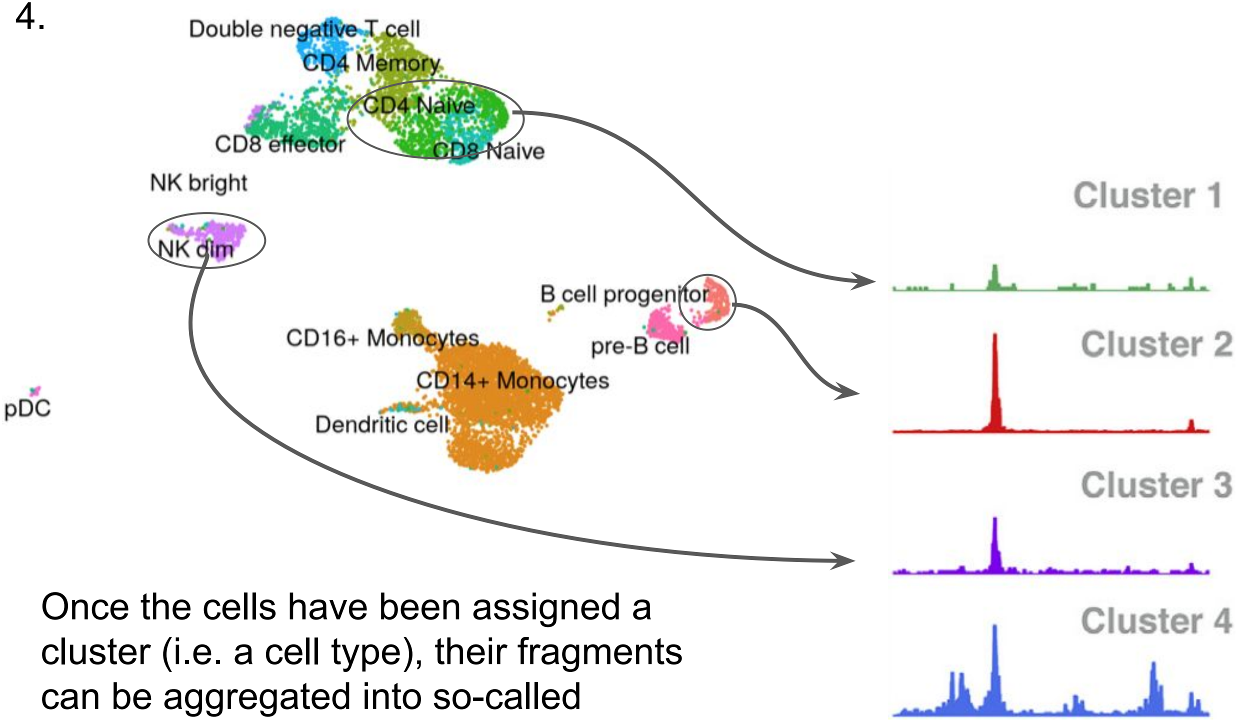


Three main pipelines with good documentation:

- Signac
- snapATAC2 (python)
- ArchR (feature-rich, scalable, but suboptimal – ref)

Single-cell ATACseq analysis in a nutshell

4.



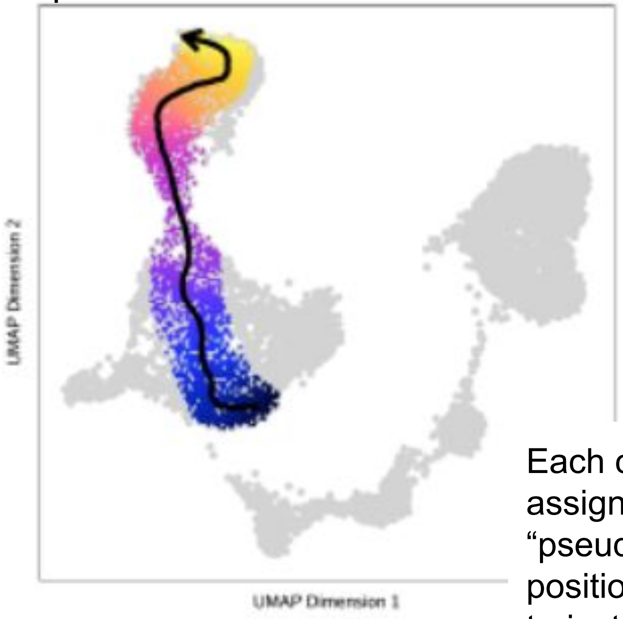
Once the cells have been assigned a cluster (i.e. a cell type), their fragments can be aggregated into so-called “**pseudo-bulk**” profiles

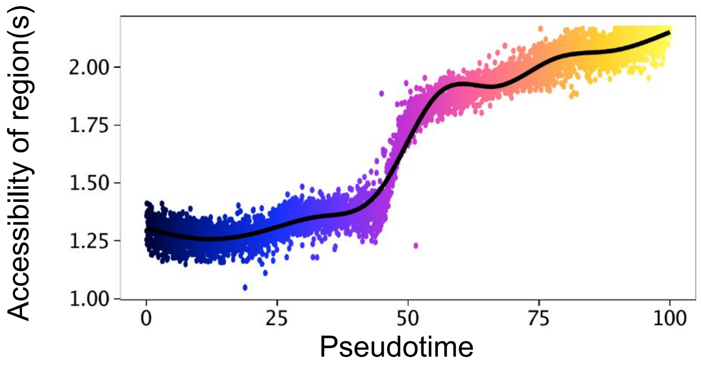
From this point on, the data is pretty much like traditional (bulk) ATACseq data, meaning that you can apply all the tools you’re familiar with, but it’s cell-type-specific!

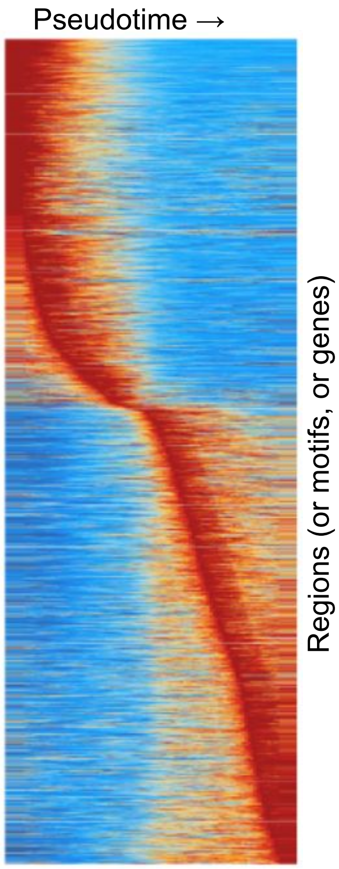
One also often do some work at the pseudo-bulk level (e.g. calling peaks) before going back to the cell-level

Single-cell ATACseq analysis – doing more at the cell-level

We identify “trajectories” across cells, representing for example a process of differentiation







Each cell can be assigned a “pseudotime”, i.e. it’s position along the trajectory

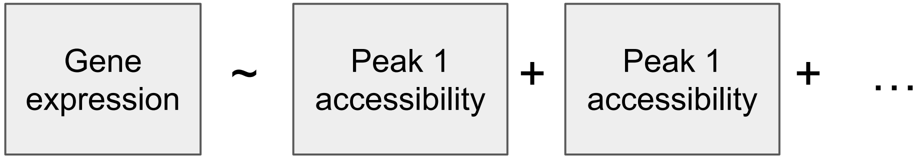
We can then track the accessibility of regions of interest across this “pseudotime”

(Adapted from ArchR documentation)

Single-cell ATACseq analysis – doing more at the cell-level

Because we have so many cells, we can use the correlation between the accessibility at distal regulatory elements (i.e. enhancers) and putative TSS to know what genes these regulate

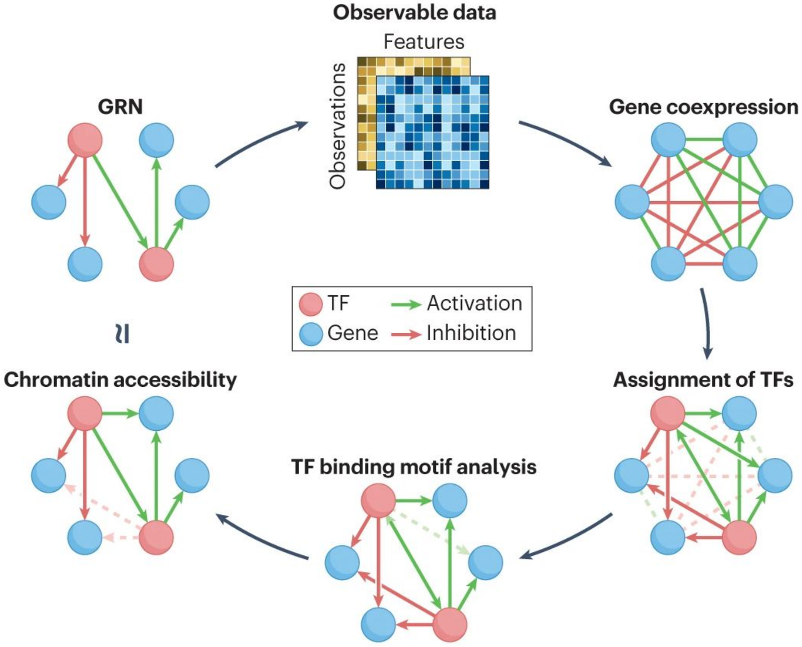
Using multimodal (ATAC + RNA) data, predict:



Gene expression ~ Peak 1 accessibility + Peak 1 accessibility + …

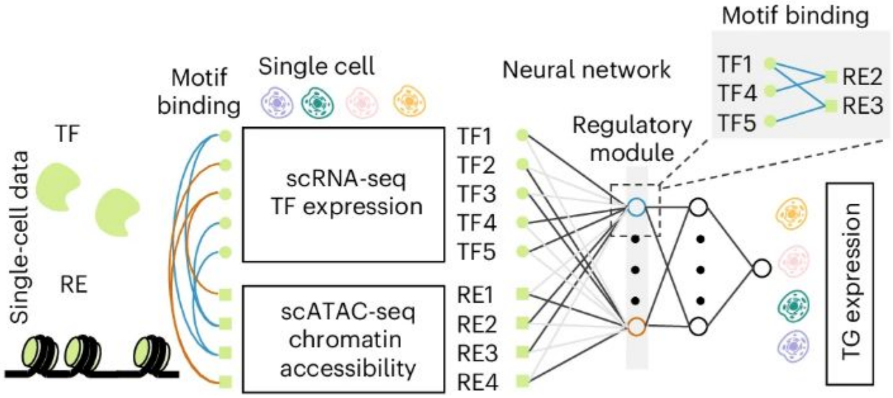
Due to the very high sparsity and noise of the data, this does not always work well. What does work well, however, is to first aggregate together groups of cells (*meta-cells*) that are highly similar, and then test correlations across meta-cells (Pliner et al., 2018; Persad et al., 2023)

Gene regulatory network inference



(adapted from Badia-i-Mompel et al., *Nat Rev Gen* 2023)

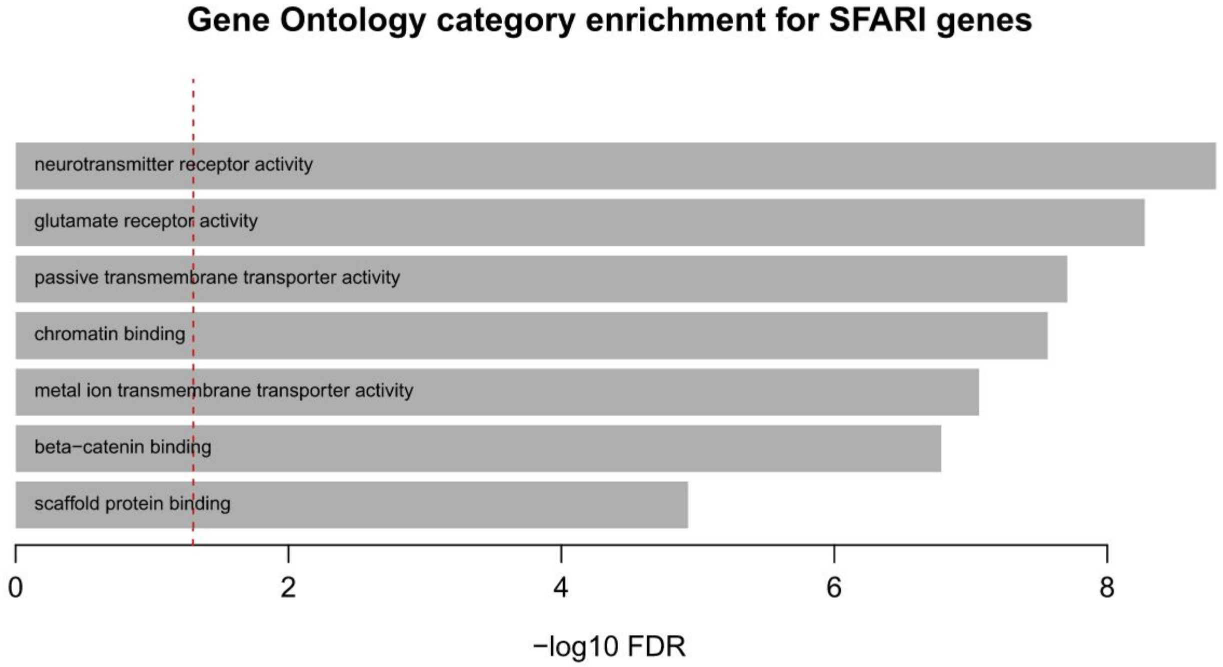
Example method : LINGER

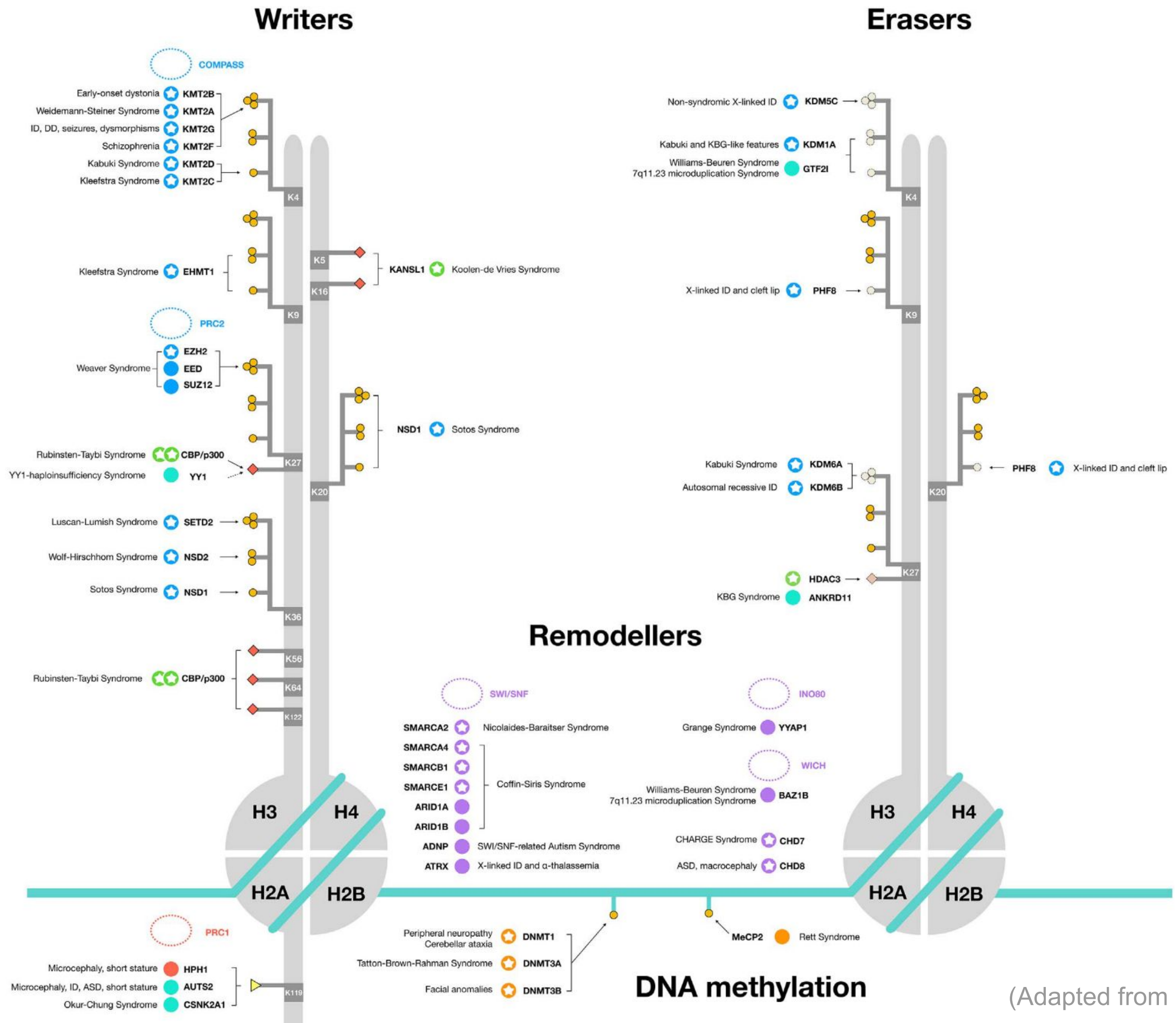


(adapted from Yuan & Duren, *Nat Bio* 2024)

Chromatin and disease

Autism-associated genes are enriched for chromatin-binding




Writers
COMPASS
Early-onset dystonia KMT2B
Weidemann-Steiner Syndrome KMT2A
ID, DD, seizures, dysmorphisms KMT2G
Schizophrenia KMT2F
Kabuki Syndrome KMT2D
Kleefstra Syndrome KMT2C
K4
Kleefstra Syndrome EHMT1
K9
K5
K16
KANSL1 Koolen-de Vries Syndrome
PRC2
EZH2
Weaver Syndrome EED
SUZ12
NSD1 Sotos Syndrome
Rubinsten-Taybi Syndrome CBP/p300
YY1-haploinsufficiency Syndrome YY1
K27
K20
Luscan-Lumish Syndrome SETD2
Wolf-Hirschhom Syndrome NSD2
Sotos Syndrome NSD1
K36
Rubinsten-Taybi Syndrome CBP/p300
K56
K64
K122
H3
H4
H2A
H2B
PRC1
Microcephaly, short stature HPH1
Microcephaly, ID, ASD, short stature AUTS2
Okur-Chung Syndrome CSNK2A1
K119
Erasers
Non-syndromic X-linked ID KDM5C
Kabuki and KBG-like features KDM1A
Williams-Beuren Syndrome 7q11.23 microduplication Syndrome GTF2I
K4
X-linked ID and cleft lip PHF8
K9
PHF8 X-linked ID and cleft lip
Kabuki Syndrome KDM6A
Autosomal recessive ID KDM6B
K20
K27
HDAC3
KBG Syndrome ANKRD11
Remodellers
SWI/SNF
SMARCA2 Nicolaides-Baraitser Syndrome
SMARCA4
SMARCB1
SMARCE1
ARID1A
ARID1B
Coffin-Siris Syndrome
ADNP SWI/SNF-related Autism Syndrome
ATRX X-linked ID and α-thalassemia
INO80
Grange Syndrome YYAP1
WICH
Williams-Beuren Syndrome 7q11.23 microduplication Syndrome BAZ1B
CHARGE Syndrome CHD7
ASD, macrocephaly CHD8
H3
H4
H2A
H2B
Peripheral neuropathy Cerebellar ataxia DNMT1
Tatton-Brown-Rahman Syndrome DNMT3A
Facial anomalies DNMT3B
MeCP2 Rett Syndrome
DNA methylation

Nearly all epigenetic modifier genes are associated with neurodevelopmental syndromes

(Adapted from Gabriele et al. 2018 10.1016/j.pnpbp.2017.12.013)

Chromatin and cancer

- Mutations within chromatin remodelling complexes are estimated to affect 10-20% of all cancers, typically leading to more “open” chromatin; in most of other cancers, the machinery is indirectly affected
- “A well-known characteristic of almost all tumours is global hypomethylation and concurrent abnormal hypermethylation at localized sites such as Cpg islands”

(Zhao et al., Nat Rev Cancer 2021)

- Some cancers (e.g. infant ependymoma) don’t show relevant DNA changes, but large epigenetic alterations
- Cancerous phenotype can be induced by (mutations in) the surrounding tissue in models, in the absence of mutation of the cells themselves (see Maffini et al, J Cell Sci 2004)
- Histone acetylases (and histone deacetylase inhibitors) are having surprising success as anti-cancer drugs (they however also affect important non-chromatin pathways, such as p53 and Nfkb)

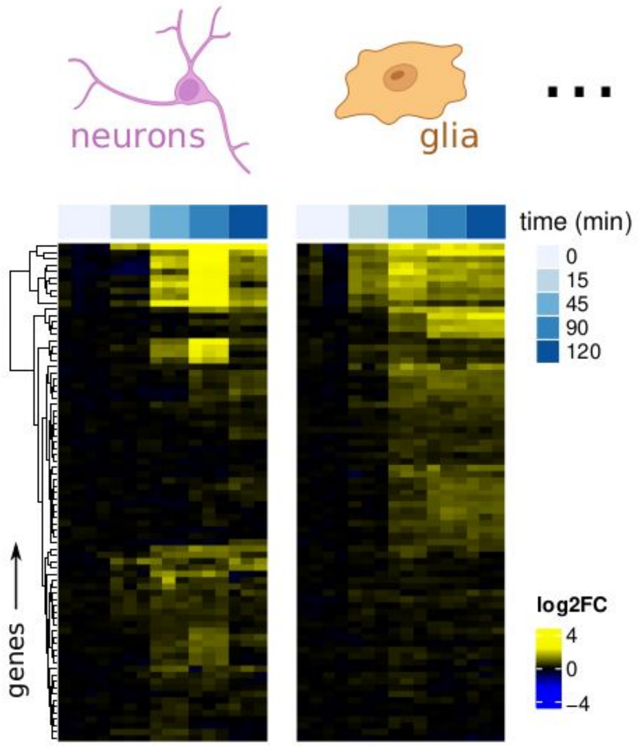
Our lines of research

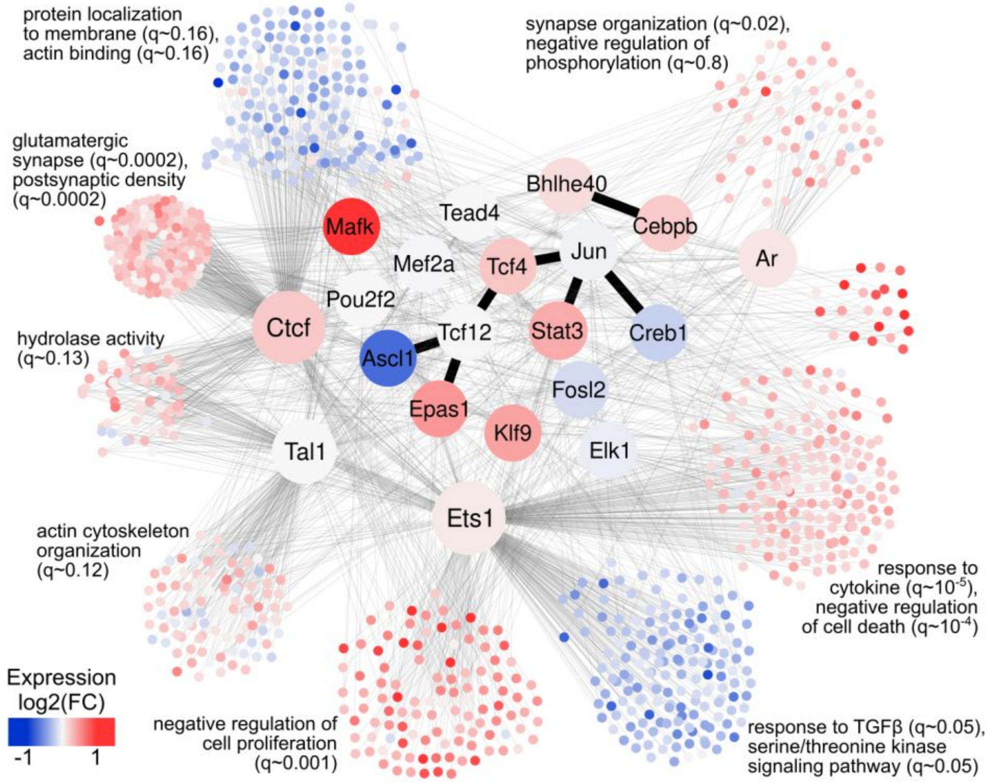
Understanding the brain’s gene expression response to stress

- How is the (gene expression) response to stress in the brain distinguish from that to normal brain activity?
- Can we decompose this response into the contributions of different inter-cellular pathways (e.g. synaptic, hormonal)?
- How much of this response is attributable to cells simply maintaining homeostasis (e.g. metabolism, oxidative stress, etc.) in the face of intense activity?

The response of different cell types to similar stimuli is partially overlapping

Can we explain these similarities and differences in terms of combinations of TF bindings?




protein localization to membrane (q~0.16), actin binding (q~0.16)
synapse organization (q~0.02), negative regulation of phosphorylation (q~0.8)
glutamatergic synapse (q~0.0002), postsynaptic density (q~0.0002)
hydrolase activity (q~0.13)
actin cytoskeleton organization (q~0.12)
response to cytokine (q~10^{-5}), negative regulation of cell death (q~10^{-4})
response to TGFβ (q~0.05), serine/threonine kinase signaling pathway (q~0.05)
negative regulation of cell proliferation (q~0.001)
Bhlhe40
Tead4
Mafk
Cebpb
Jun
Mef2a
Tcf4
Ar
Pou2f2
Ctcf
Tcf12
Stat3
Creb1
Ascl1
Fosl2
Epas1
Klf9
Tal1
Elk1
Ets1
Expression log2(FC)
-1
1

(von Ziegler, et al., 2022)

Underlying computational challenges

- How can we best analyze that kind of data?
    - In particular ATAC-seq and single-cell data
- How can we get a good idea of where TF bind in different cell types, in the absence of experimental data for most TF/celltype combinations?
- How can we best make sense of distal regulatory elements and what they regulate?
- Given a gene expression signature (e.g. of a condition/disease) and transcriptional networks, how can we best infer which TFs have a differential activity?

Course project

Grading and expectations

- 50% of the grade is based on **weekly exercices**
  - Exercices should be **submitted via github**, by thursday noon the following week
  - The best half of the exercises will make up the grade

- 50% of the grade is based on the **project** (groups of max 3 persons)
  - The project can be either:
    - Re-producing the analyses from a publication (in a critical fashion)
    - Analyzing new data (e.g. yours or in collaboration with a group)
  - The project *must be discussed and approved in advance*
  - The expected outputs of the project are:
    - a report (e.g. ~10-15 pages) with embedded full code and figures, and including an introduction and discussion of the results
  - Deadline: before the end of the day on **July 3rd**

Evaluation scheme for the course project

- Format and formal requirements (1/6)
  - Rendered markdown (html or pdf), proper scientific references, figure legends, etc.
- Introduction/conclusion (1/6)
  - More is not necessarily better: ask yourself what background would your fellow students need to understand your problem, analyses and observations
- Analysis
  - **Correctness**, i.e. lack of mistakes (1/6)
  - **Adaptability/creativity**, i.e. whether you could adapt (e.g. the tools seen in class) to your purposes (1/6)
  - **Appropriateness**, i.e. whether you used the right tool/visualization to address a question (1/6)
  - **Interpretation**, i.e. whether you correctly describe and interpret your figures and analysis results (1/6)
- Extra 1/6 (summing to 7 of max 6 points) for difficulty / going beyond the expectations (to not penalize those who undertook something harder)

A few (optional) tips for preparing your project report

- It’s often easier to split a project into different markdown documents for different steps
    - e.g. one that downloads the data, another processes it, another that answers specific questions of makes the figures with which you’ll tell your story
    - if you want to go deeper, workflowr is a great rmarkdown organization and versioning framework

- If you’re wondering how to do something in rmarkdown, consult this online book
    - See especially:
    - code chunk options
    - the section about references and bibliography
      (most reference management software, e.g. Zotero and the likes, can export references in the *bibtex* format required by rmarkdown)