Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L - 2022 | week 12

Pierre-Luc Germain

ETH Zürich

Plan

- Quick return on the week 10 assignment
- Debriefing on the assignment (from week 11)
- Single-cell chromatin (ATAC) data
- Student presentations and discussion

Return on the week 10 assignment

- Question:
  - “Use rGREAT enrichment analysis (from last week's practical) to characterize the DMRs. Phrase your hypothesis for the analysis (what is your enrichment analysis testing exactly?).”

- What most of you did is either:

```
ml <- signal2matrix(tracks, regions=dmr)
cl <- clusterSignalMatrices(ml)
split_regions <- split(dmr, cl)
job <- submitGreatJob(gr=split_regions[["2"]], bg=dmr, species="mm10")
...
```

or:

```
job <- submitGreatJob(gr=dmr, species="mm10")
...
```

Both are fine, the description of what the enrichments mean is critical.
An enrichment is always relative to something, and it’s necessary to spell this out when describing the results!

Debriefing on the assignment (week 11)

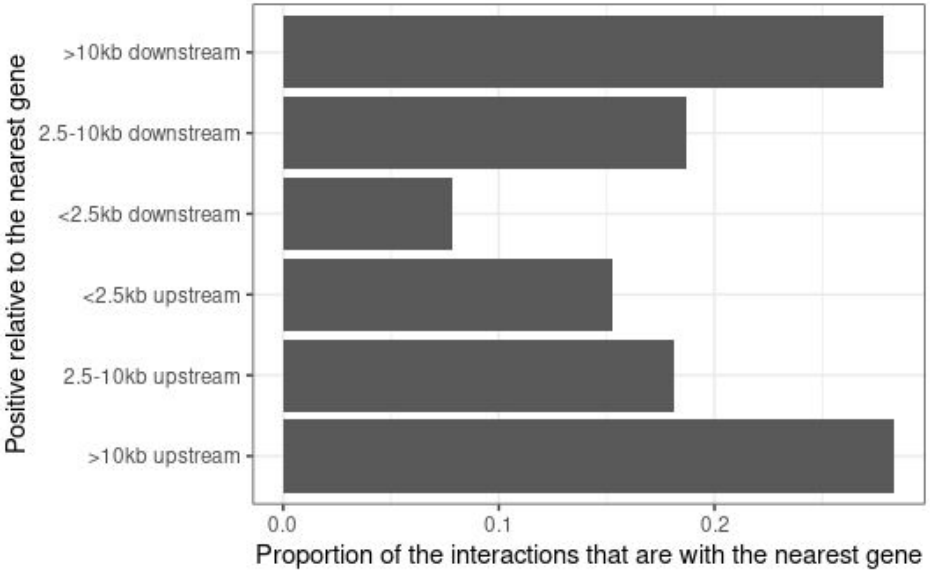
- Question:
  - “For each set [between 2.5kb and 10kb from a TSS, or more than 10kb], what proportion of the interactions are with the nearest gene?”
- Hint:
  - “beware not to count, when calculating proportions, peaks that don’t have interactions with any TSS!”

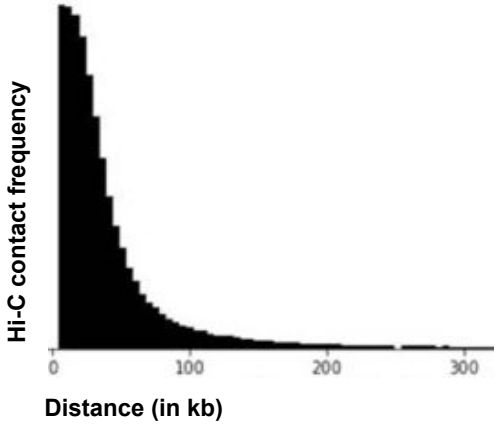
- What many of you did:

```
equal_proximal = proximal_peaks$gene_name == proximal_peaks$target
equal_proximal[is.na(equal_proximal)] = FALSE
isSame = sum(sum(equal_proximal[equal_proximal == TRUE]))
isSame/length(equal_proximal)
```

This resulted in very low proportions

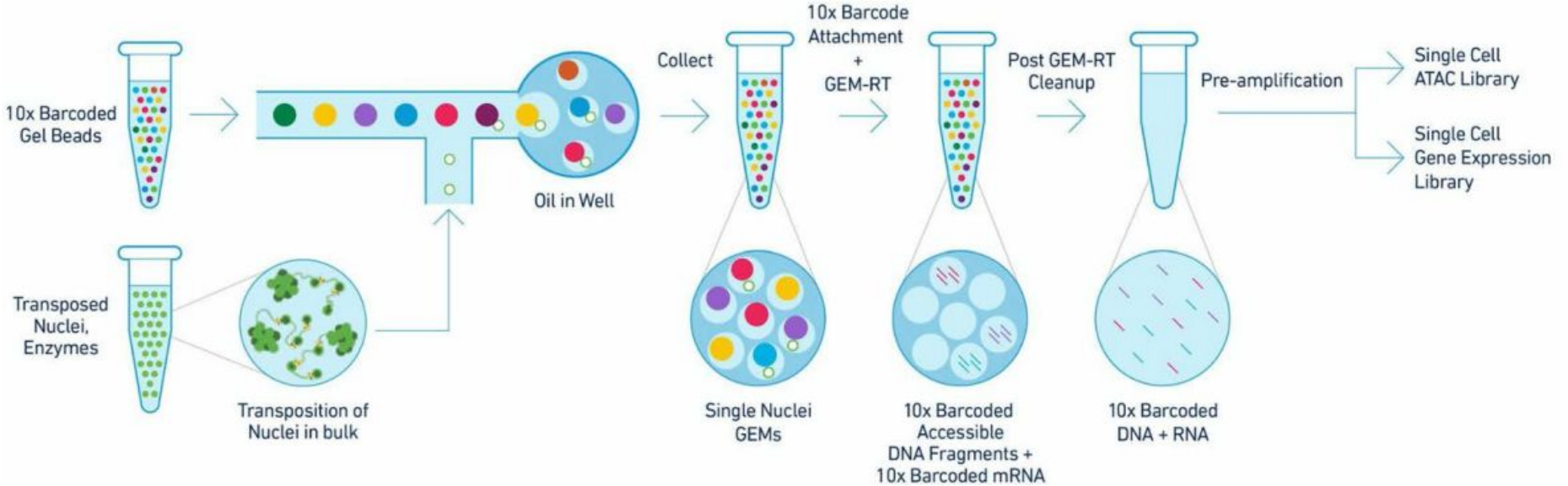
Debriefing on the assignment

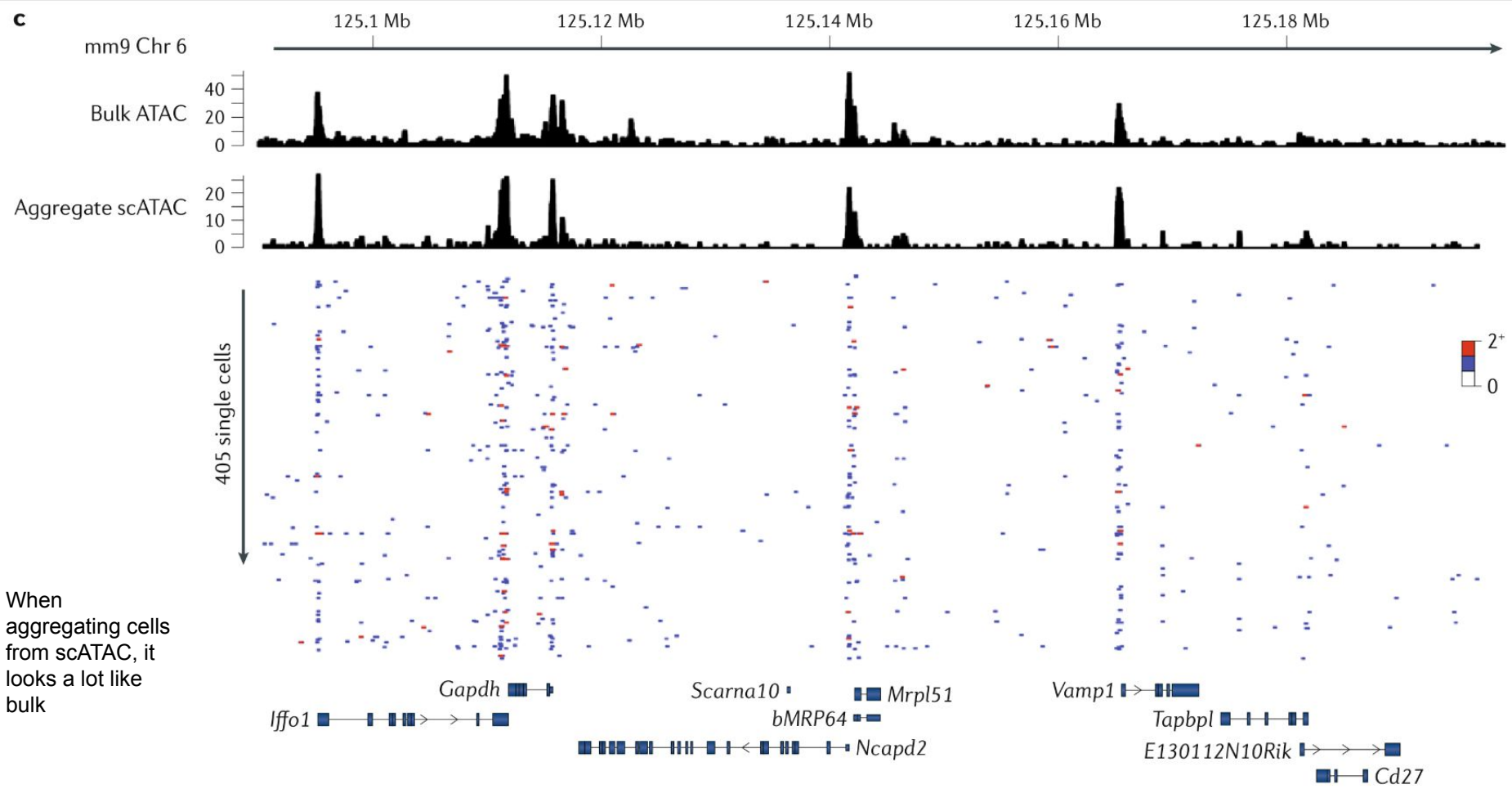




Under the null hypothesis, we expect to have more contacts from DNA regions that are closer to each other. When identifying interactions, methods look for an excess over the null hypothesis

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





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

(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
```

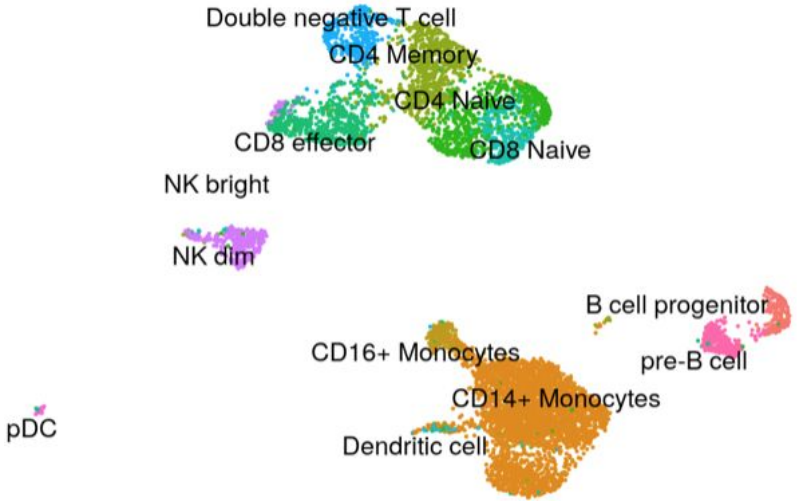
2. From this, we count the number of fragments from each cell overlapping genomic windows (either whole-genome tiles or feature-based)

```
        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:

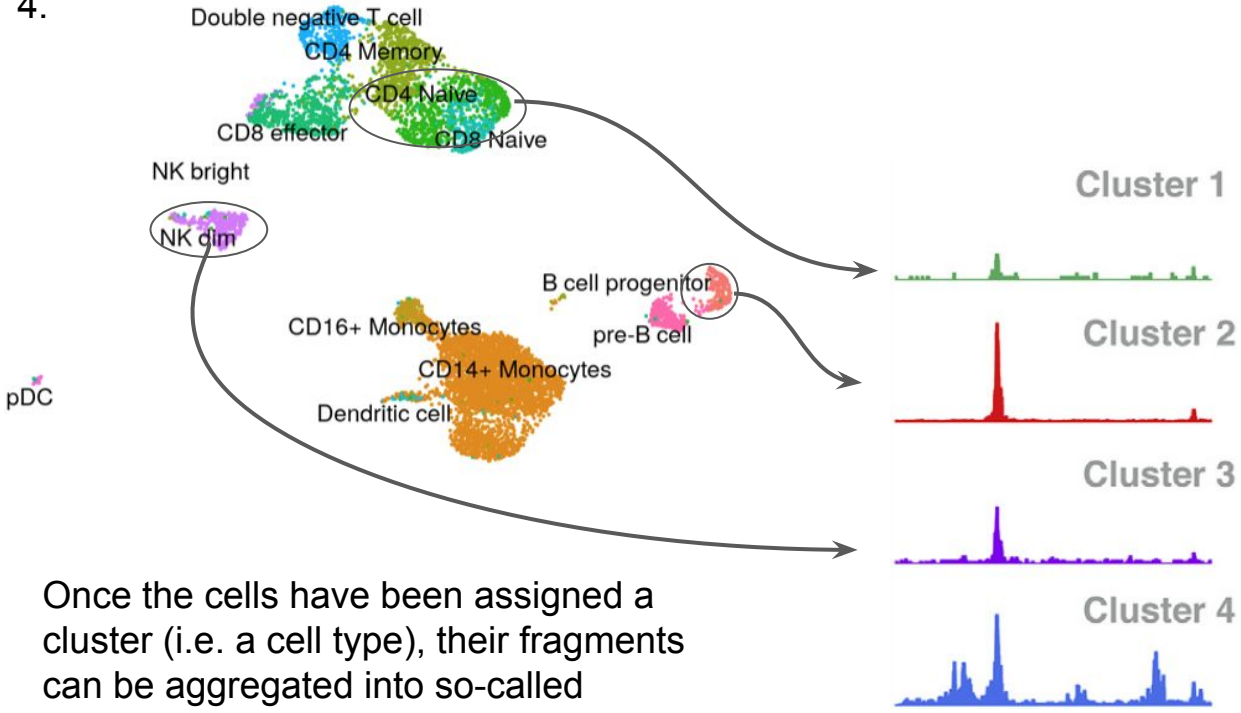


Two main pipelines with excellent documentation:

- ArchR
- Signac

Single-cell ATACseq analysis in a nutshell

4.



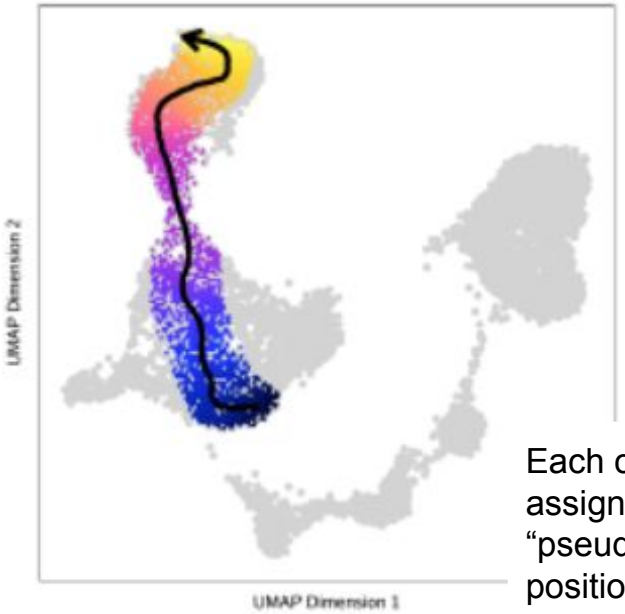
Once the cells have been assigned a cluster (i.e. a cell type), their fragments can be aggregated into so-called "**pseudo-bulk**" profiles

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!

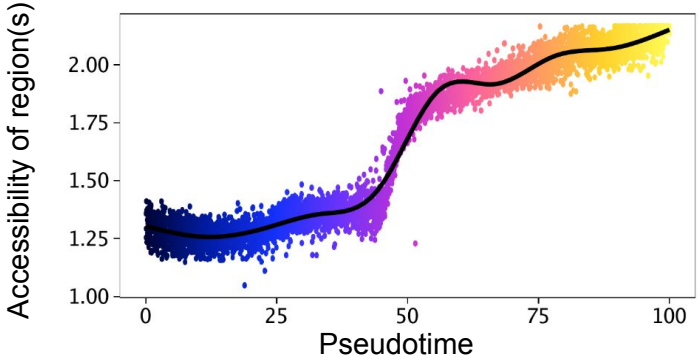
One also often do some work at the pseudo-bulk level (e.g. calling peaks) before going back to the cell-level

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

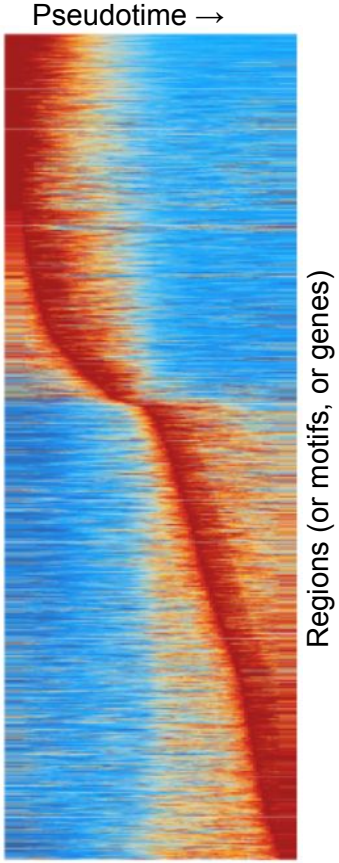
We identify “trajectories” across cells



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

Due to the very high sparsity and noise of the data, this does not 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)