Single-cell H3K4me1 in bone marrow

Jake Yeung

2018-12-19

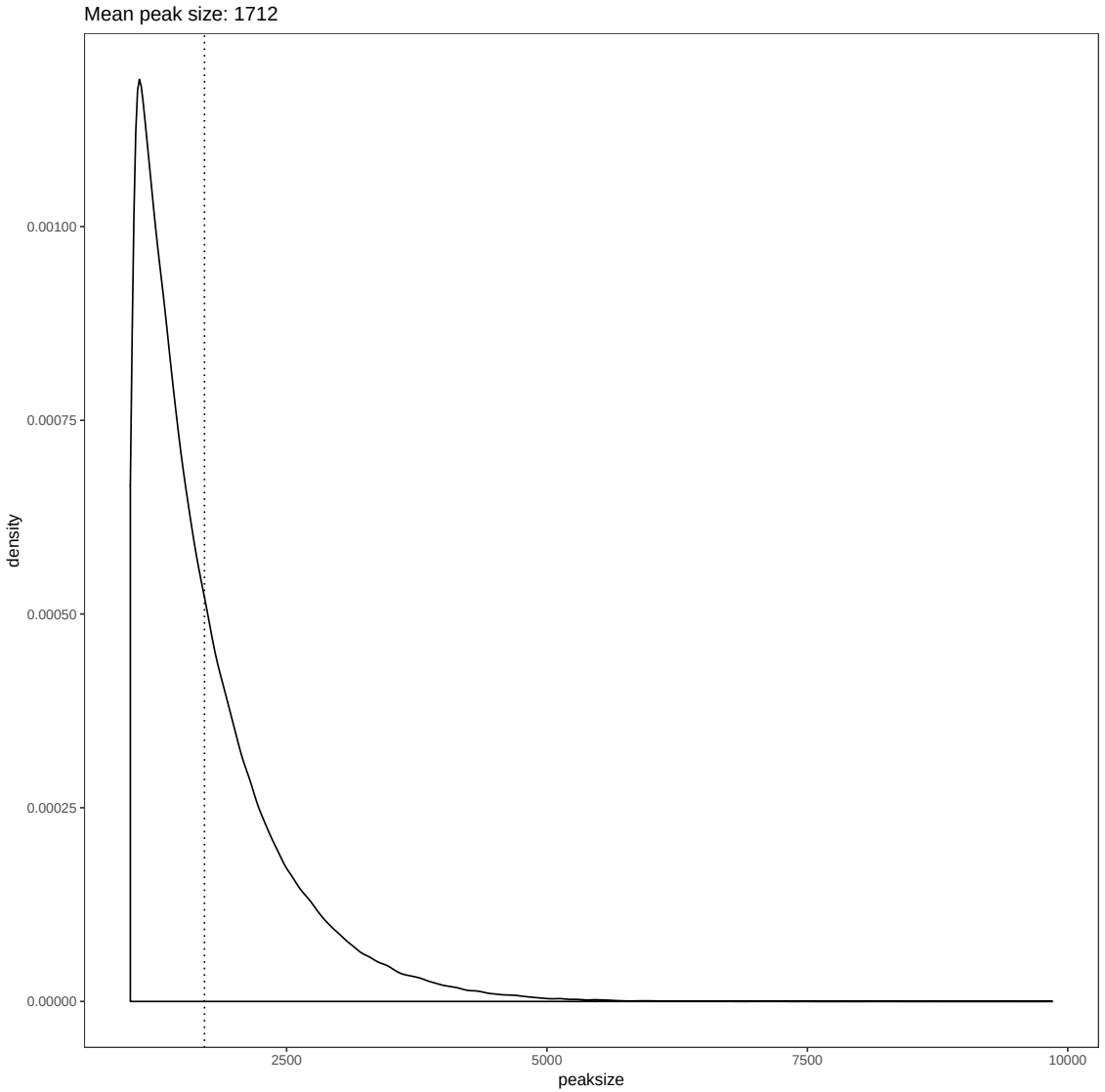
Introduction

Refining the scChiC analysis

- Merge across 4 bone marrow datasets (648 cells total, I filtered for cells with 10000 total reads, which is probably an unnecessary step).
- Call broad peaks that are at least 1kb or larger (this seems to consolidate small peaks and give a broader distribution of peak sizes).
- Filter peaks that overlap with "blacklisted" regions.
- Further filter peaks that have unexpectedly large counts within the peak.

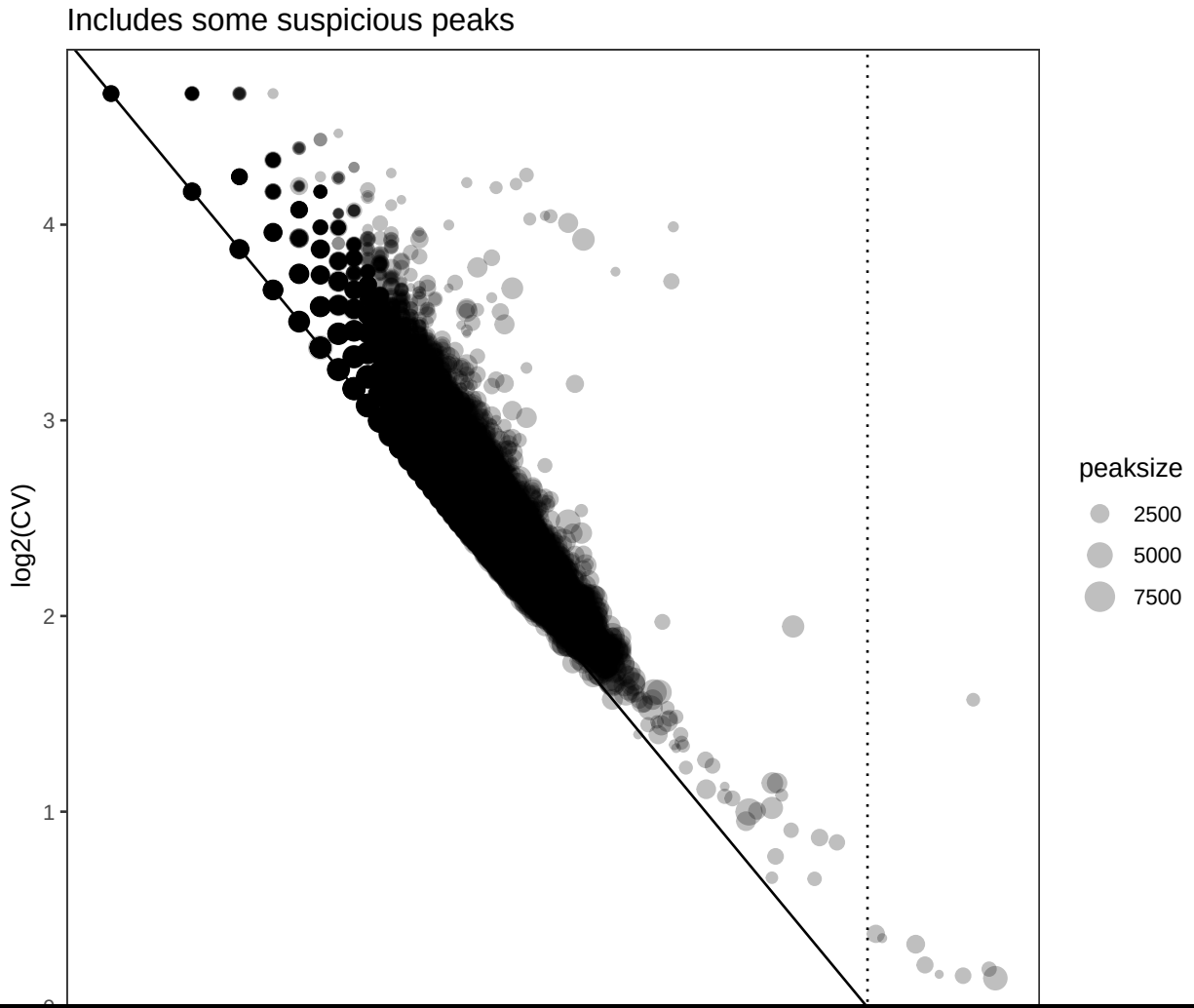
Results

Broad peak calling finds distribution of peak sizes



Maybe a few suspicious peaks remaining even after filtering for blacklist regions?

Largest mean is 3. I remove peaks with mean > 1.



Top peak after filtering is located in gene body of Erdr1 (mean=0.77)

Erdr1: erythroid differentiation regulator 1. So maybe keep that peak.

I haven't thrown out 4 peaks in chrM yet.

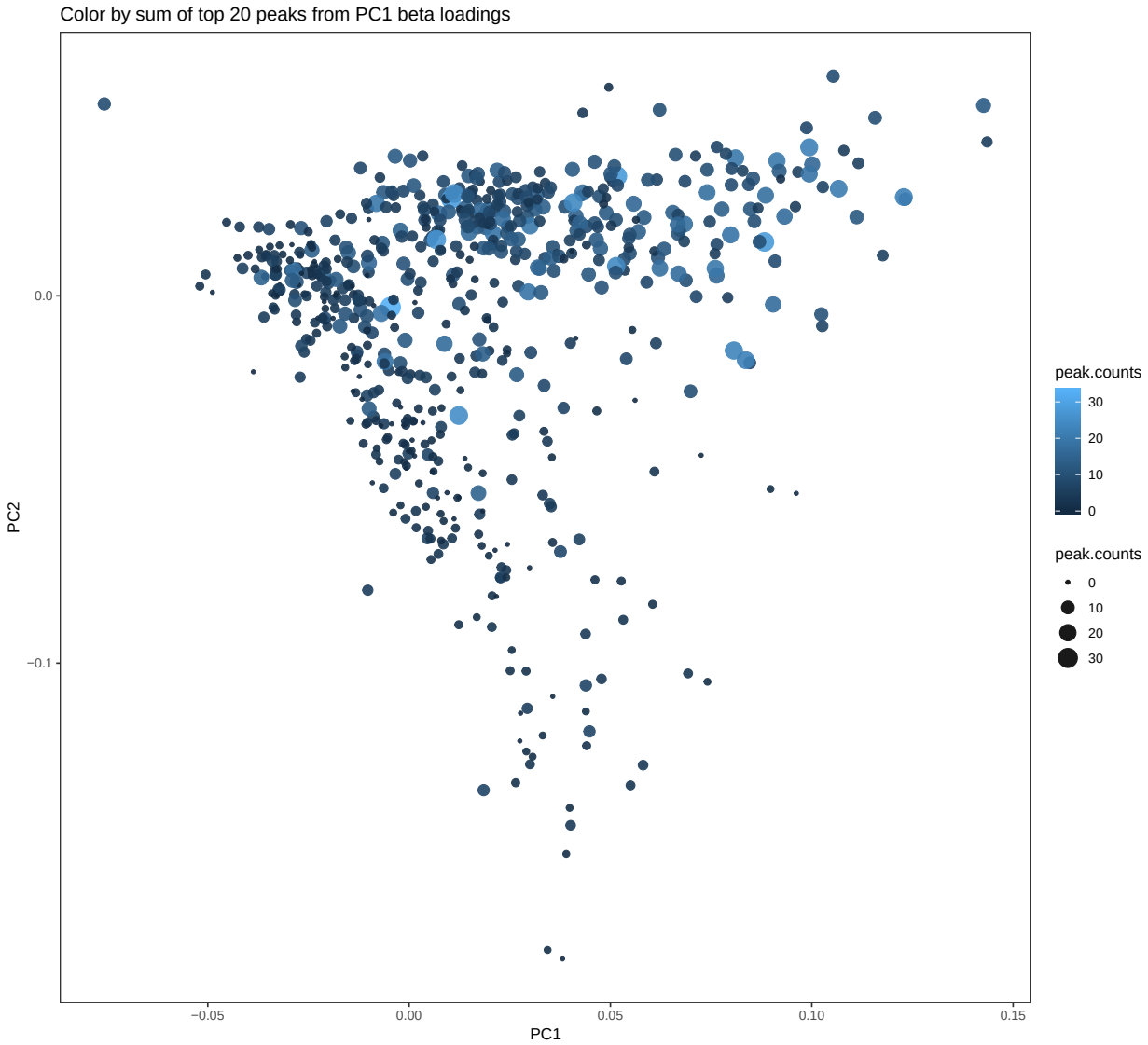
```
## # A tibble: 6 x 6
##     Sum  Mean   Var peak                          CV peaksize
##   <dbl> <dbl> <dbl> <chr>                      <dbl>    <dbl>
## 1   499 0.770 1.91  chrY:90811185-90813008      1.79     1823
## 2   430 0.664 1.47  chrX:169993165-169995246    1.82     2081
## 3   412 0.636 1.00  chrY:90809271-90810752      1.58     1481
## 4   343 0.529 4.16  chr16:34823189-34826742     3.85     3553
## 5   337 0.520 0.949 chrM:14587-16222            1.87     1635
## 6   311 0.480 1.04  chrY:90737710-90738964      2.12     1254
```

About half of reads are assigned to a peak

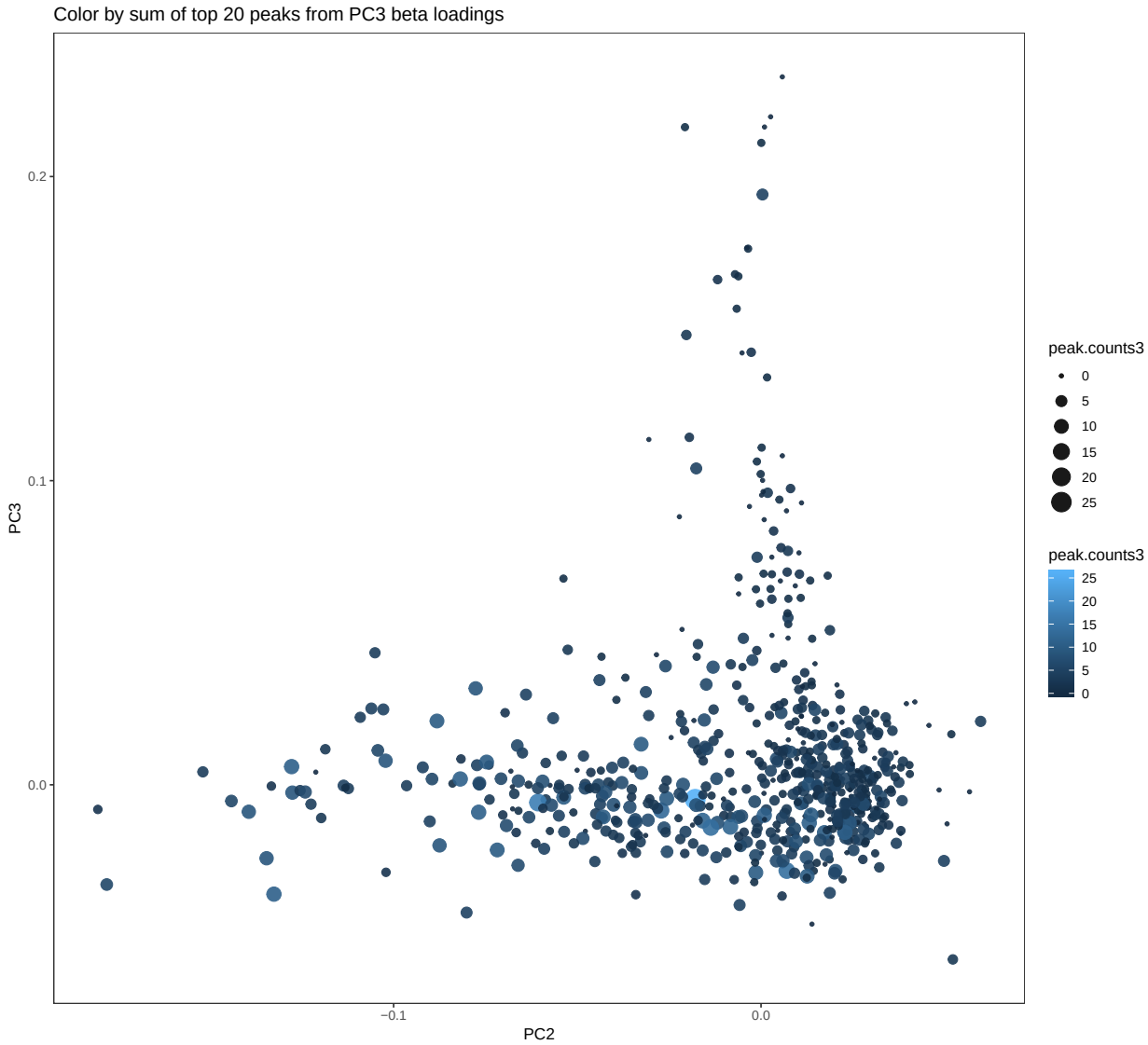
Peak counting results in 3292574 of 6837325 reads across cells to be considered into the count matrix

```
##                                Assigned                  Unassigned_Unmapped
##                                 3292574                                    0
##                  Unassigned_MappingQuality                   Unassigned_Chimera
##                                       0                                    0
##                  Unassigned_FragmentLength                  Unassigned_Duplicate
##                                       0                                    0
##                    Unassigned_MultiMapping                  Unassigned_Secondary
##                                       0                                    0
##                        Unassigned_NonSplit                 Unassigned_NoFeatures
##                                       0                                 353992
## Unassigned_Overlapping_Length                   Unassigned_Ambiguity
##                                       0                                  4831
```

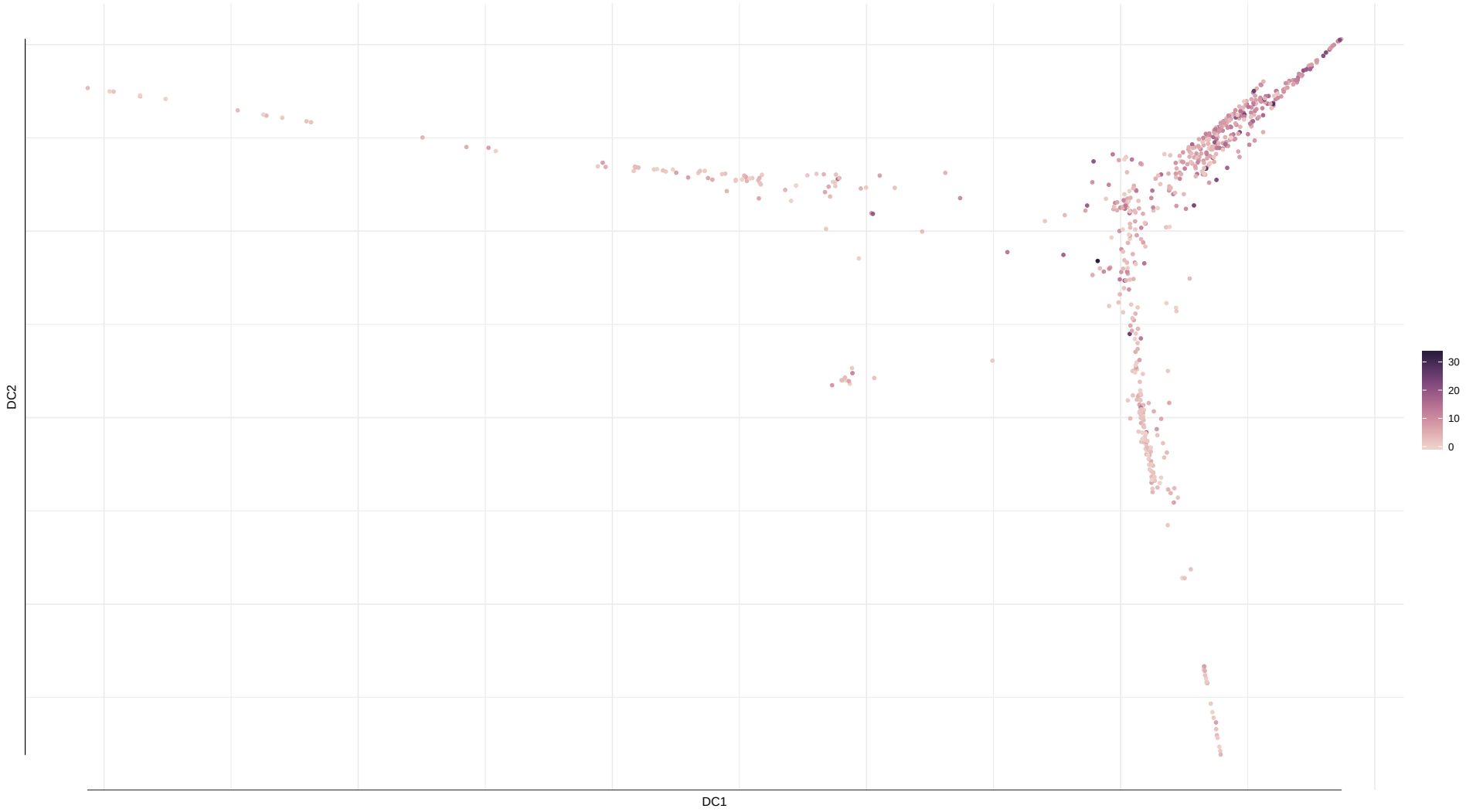
LDA model (K=10) shows cell-to-cell differences in H3K4me1 in murine bone marrow



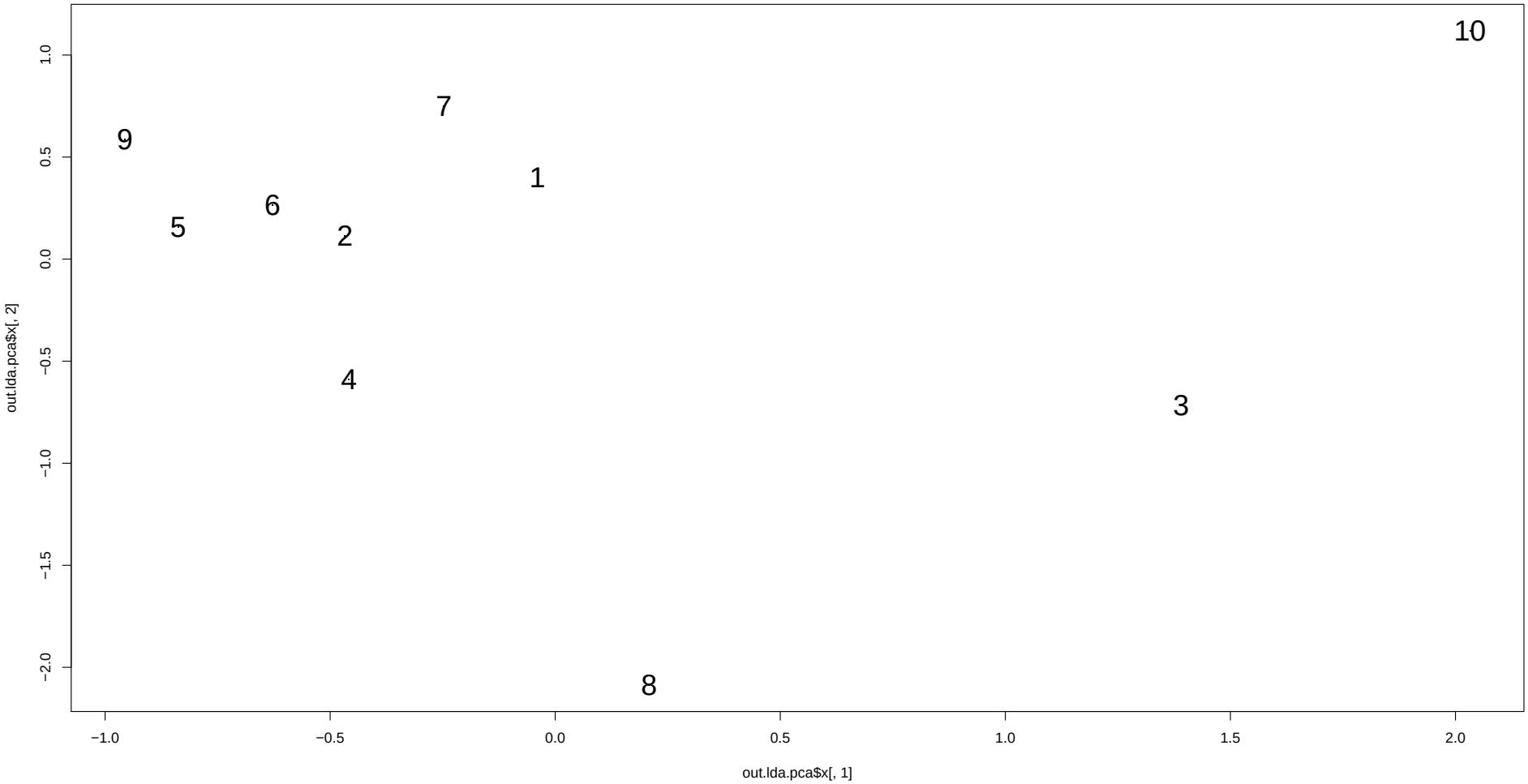
Higher PCs also have structure



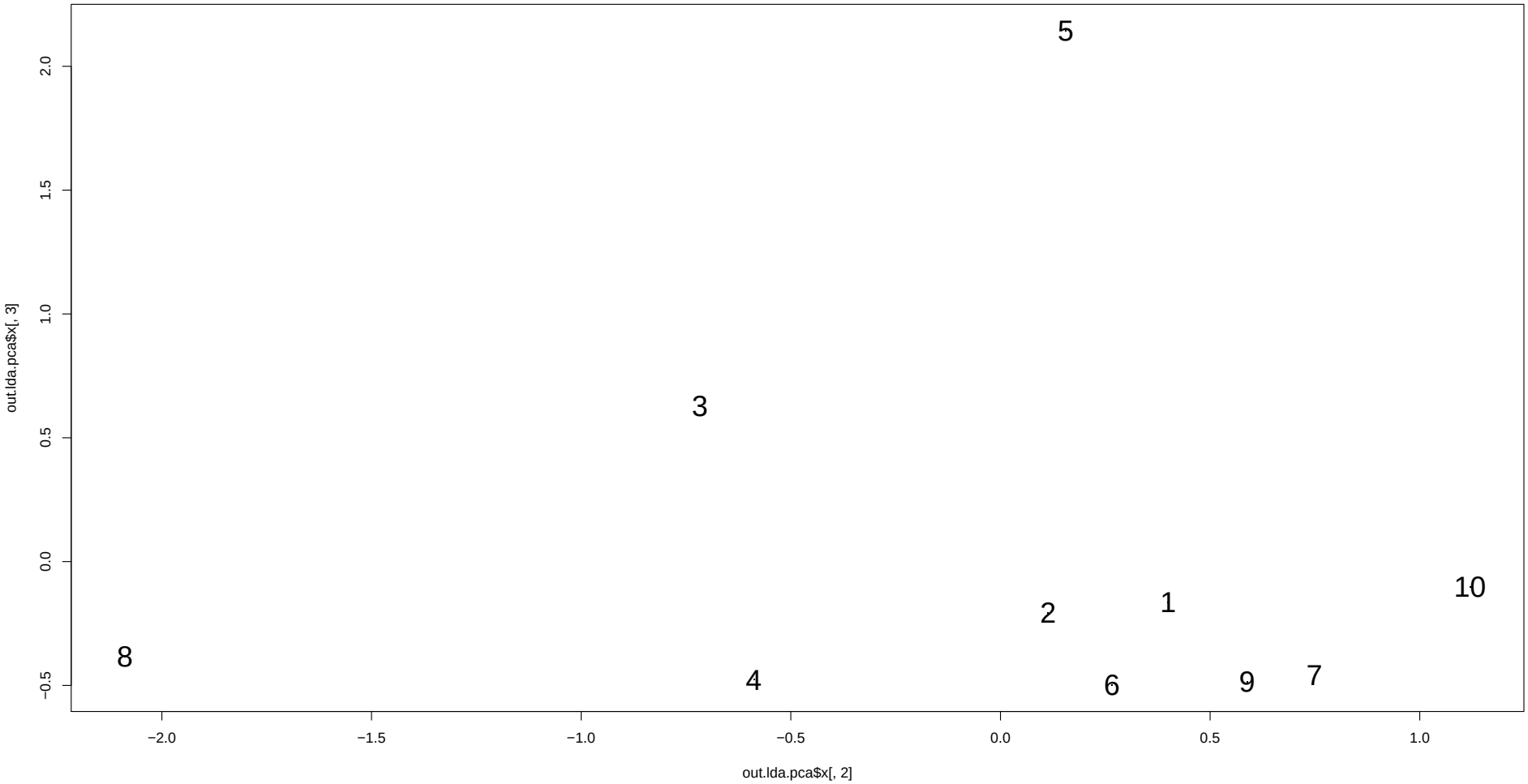
Can do non-linear reductions, and color by clustered peaks



The K=10 clusters are correlated (PC1, PC2)

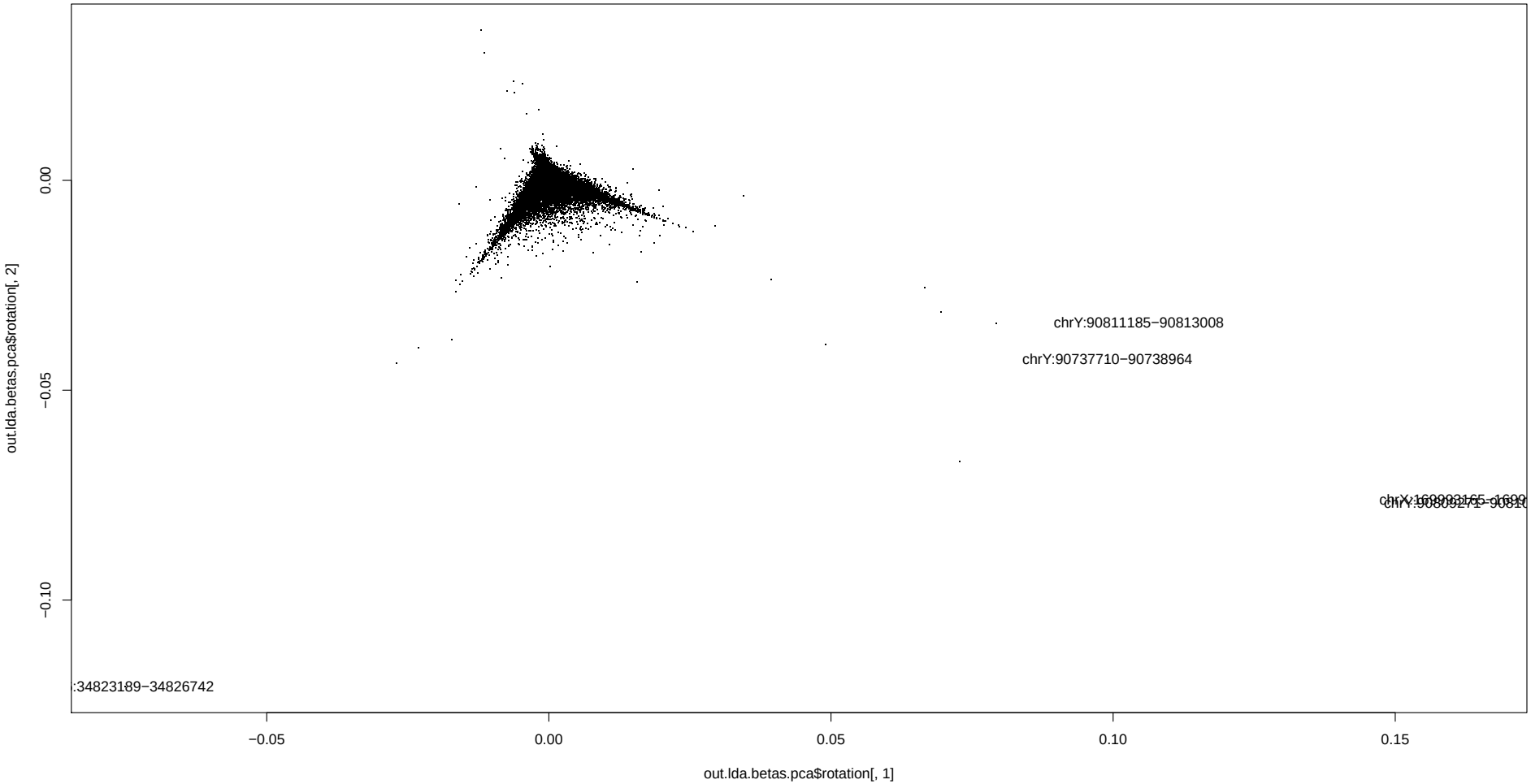


The K=10 clusters are correlated (PC2, PC3)



PCA on beta matrix finds clusters of peaks

The chrY peak is located in gene body of Erdr1



Annotating peaks: 95k intronic, 30k promoter, 20k exon

```
##
##      Distal      Intron    Promoter        Exon          3' Down
##      139933       95244       33432       21174        8296
##          5'
##        1942
```

Annotate peaks, 31% intron, 11% promoter, 7% exon

```
##
##      Distal      Intron   Promoter        Exon          3' Down
##      0.4600      0.3100      0.1100      0.0700      0.0270
##          5'
##      0.0064
```

Conclusions

Preliminary results

- Peak sizes range from 1 to 10kb (mean 1.7 kb), allowing dynamic windows to be captured in the genome.
- Refining the peak-calling (blacklist, mean cutoff > 1) likely improves the output of LDA.
- LDA models the discrete count data to simultaneously find cell clusters and peak clusters (NLP people usually only analyze 'peak/word' clusters, here we also look at 'cell/document' clusters).
- Maybe relevant genes I came across while looking at data: Erdr1, Rock1, Gp49a, Kdm6b

Next steps

- Using the soft clustering more effectively or tuning cutoffs to binarize downstream analyses.
- Consistently clustering different cell-types across histone marks.
- Generating a 3D matrix of cell-type, regions, histone-marks to find relationships across histone marks.
- Filter out distal peaks by a certain distance and ask how LDA changes?