Epigenomic enrichment analysis using Bioconductor

EuroBioc 2019 – Brussels

Dario Righelli – PhD
Istituto per le Applicazioni del Calcolo «M. Picone» – CNR - Napoli
d.righelli@na.iac.cnr.it || dario.righelli@gmail.com

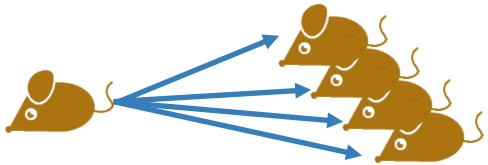
drighelli

What's the aim?

Compare methods and provide guidelines on epigenomic data analysis

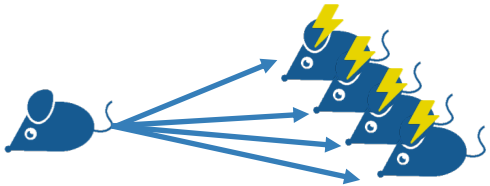
ATAC-seq dataset

Before Fear Induction Condition (E0)



4 biological replicates

After Fear Induction Condition (E1)



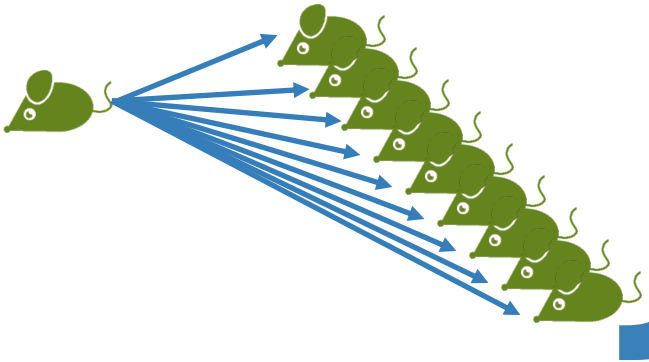
4 biological replicates

Catching differences in open chromatine regions

Yijing Su et al. 2017 - Nature Neuroscience - Neuronal activity modifies the chromatin accessibility landscape in the adult brain

ChIP-seq dataset (NULL dataset)

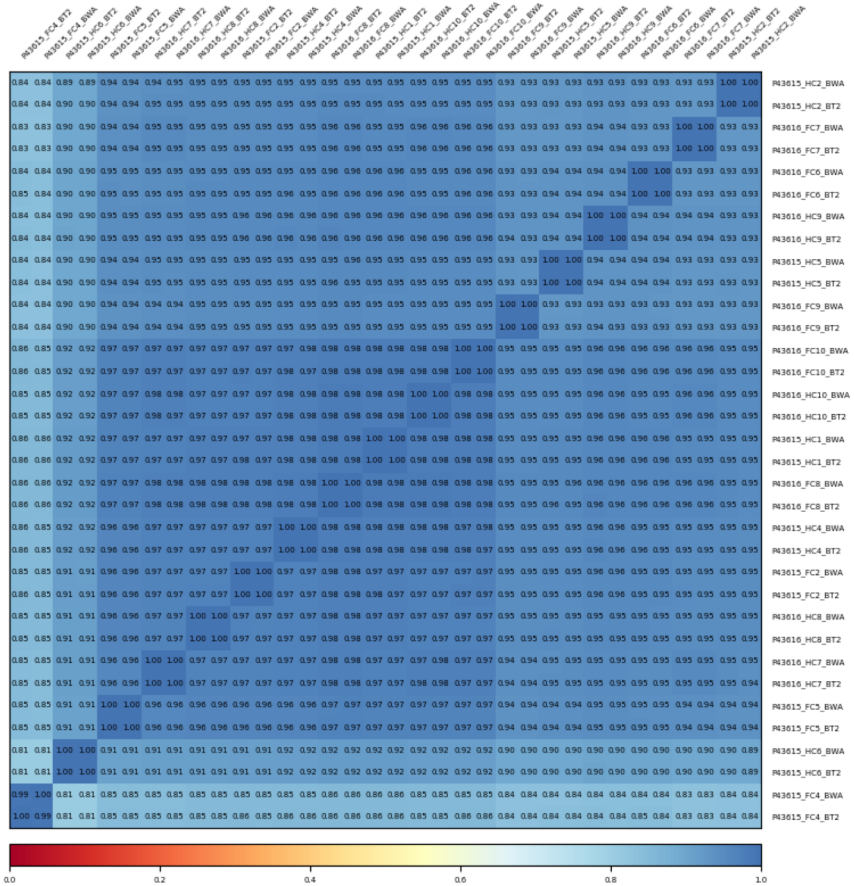
Home Cage Controls - Histon 3, Lysine 9 Acetilation (H3K9ac)



9 biological replicates

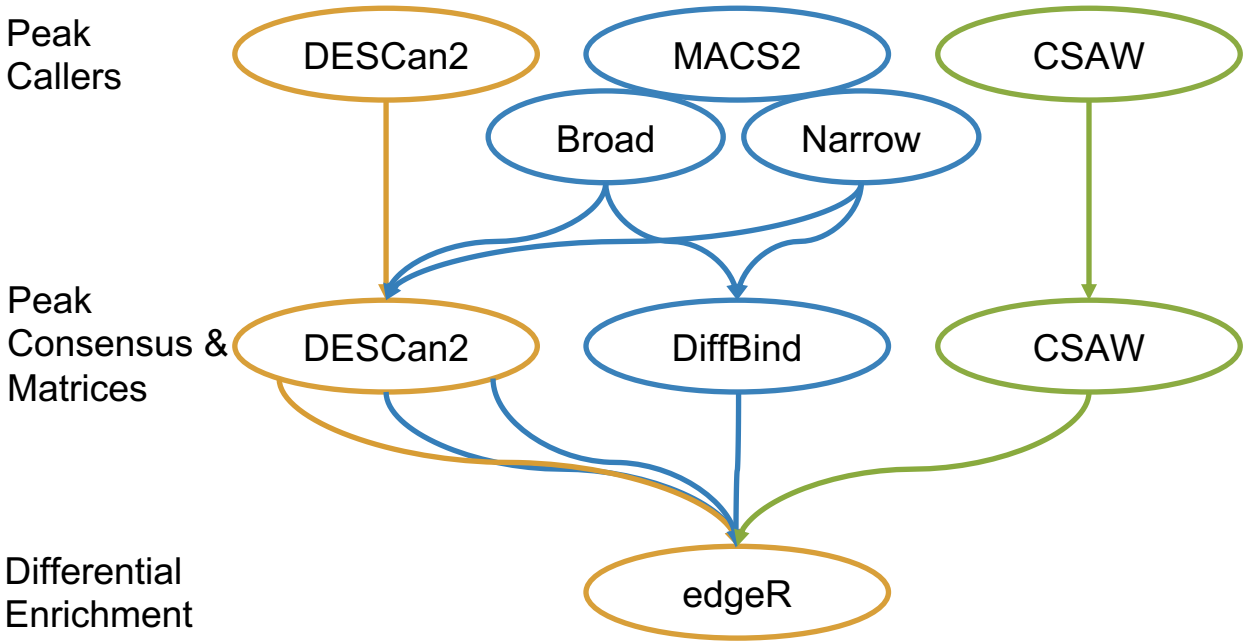
How many random differences are we able to catch inside a control dataset?

BWA and Bowtie2 perform the same



- Most used aligners for epigenomics data
- Correlation computed on ChIP-seq data coverages
- used DeepTools plotCorrelation tool
- Computed correlations on the coverages of the same samples on BWA and Bowtie2 bams have value of 1.

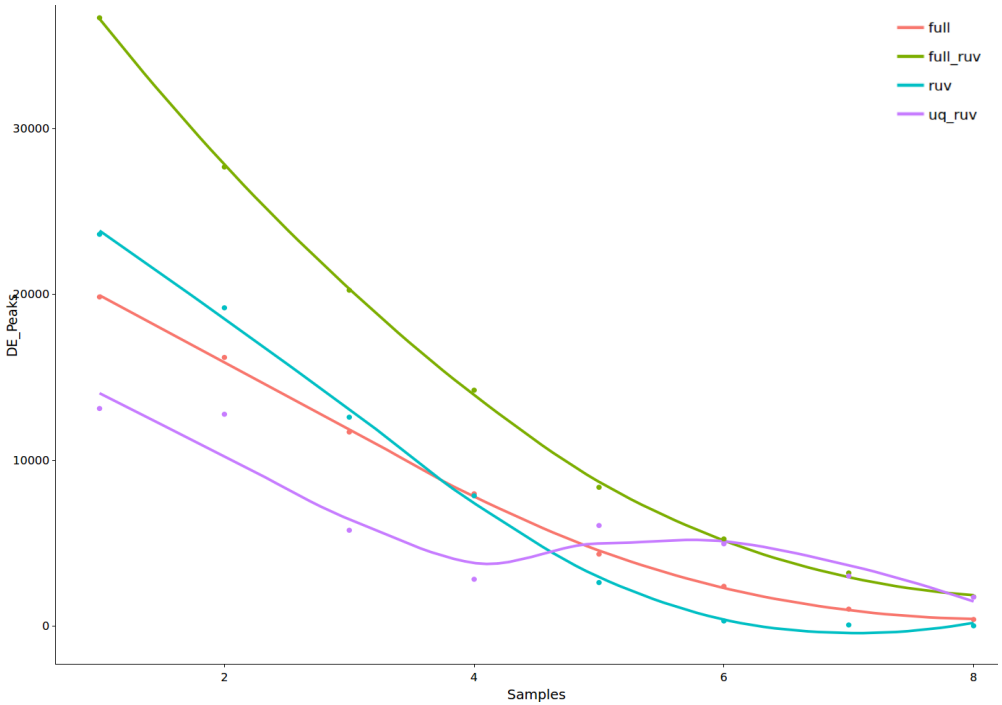
A Bioconductor Approach



- MACS2 (No Bioconductor)
  - Most used peak caller
  - Broad and Narrow peaks option
- DEScan2
  - Has a peak detector in R
  - Peak resolution -> bin size
  - Can work with external peaks
- DiffBind
  - No peak detection
  - Fast on matrix construction
  - Uses external peaks
- CSAW
  - Starts from BAM files
  - Computes matrix of bins x samples
- edgeR
  - Widely used method
  - Very flexible in usage

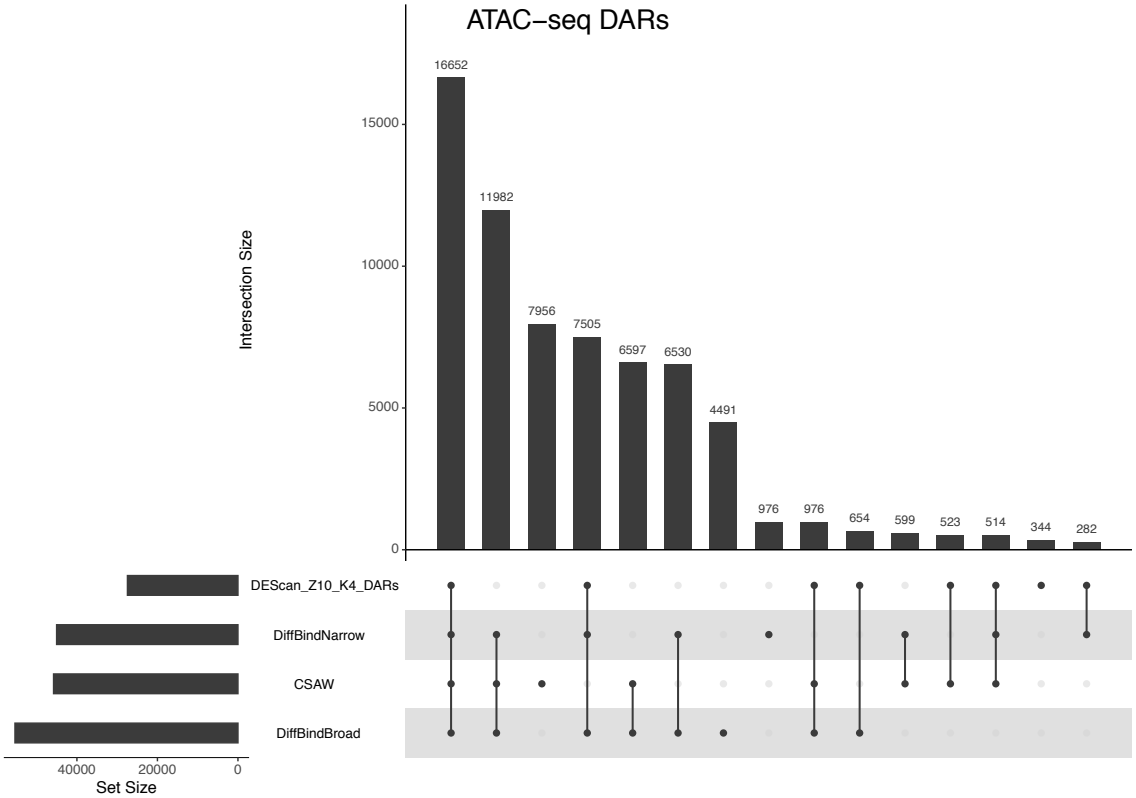
Counts Normalization Affects Differentially Accessible Regions (DARs)

ATAC-seq dataset



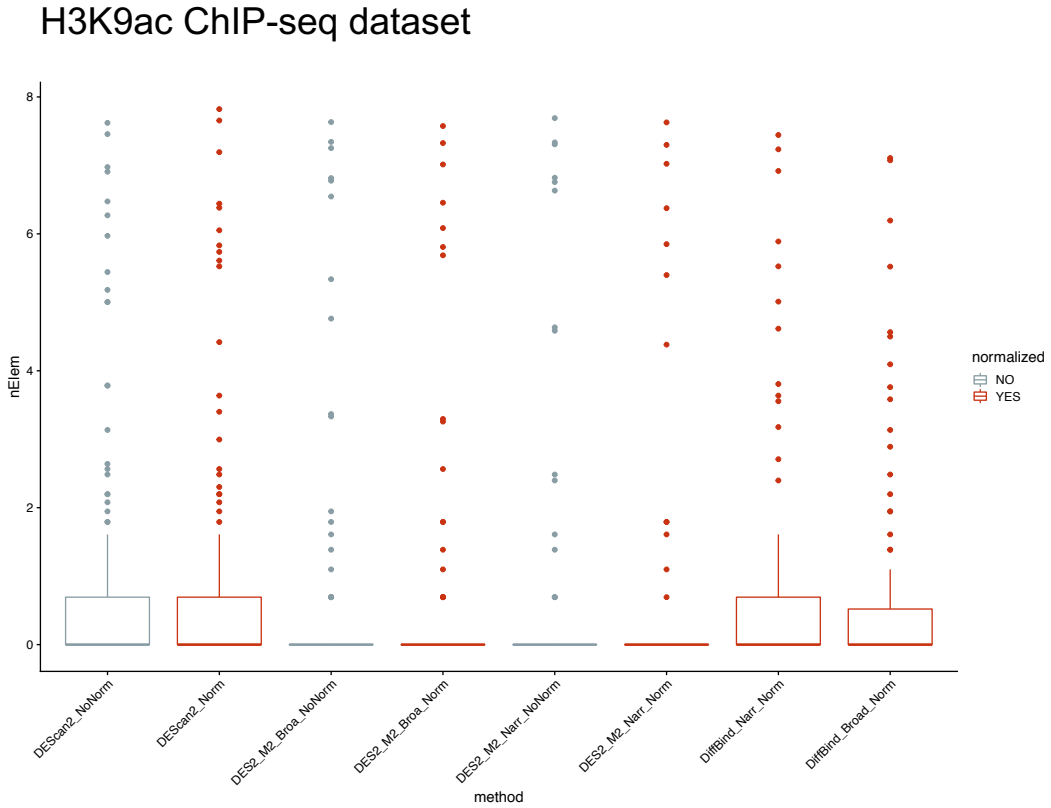
- Pay attention to the normalization process
- One tryes to apply a classic RNA-Seq normalization
- The process does not always give the same results
- Maybe some more specific normalization is required for this kind of data

Comparing DARs across methods



- All the methods have the biggest overlap on the detected peaks
- CSAW and DiffBind show a big amount of not-overlapping regions
- DEScan2 shows the lowest number of not-overlapping regions
- The big amount of not-overlapping regions by CSAW and DiffBind suggests a possible high-level of false positive regions detected.
- Ad-hoc designed **UpsetPlot** on **GRanges**
  - Based on **findOverlaps** method

Peaks contrasts on NULL dataset show no results



- Compared performances on a null dataset of ChIP-seq H3K9ac samples
- Performed 126 permutations of samples
  - Samples are randomly divided in two groups
  - All the possible permutations on 9 samples (126)
- All the methods find mostly 0 Differential Enriched Peaks on the random conditions.
- Sometimes some differences have been found
- With and without normalization

What's Next?

On-going and future works

Some comparisons are still needed

- Compare CSAW on ChIP-seq
- Compare normalization methods with all epigenomics methods
- Explore in-silico biological functions of results
- Testing ATAC—seq Single Cell dataset

Acknowledgements

- Dr. Claudia Angelini – Istituto per le Applicazioni del Calcolo-CNR
- Dr. Davide Risso – Univeristy of Padua
- Dr. Lucia Peixoto – Elon S. Floyd College of Medicine, Washington State University
- Dr. Timothy Triche Jr. - Van Andel Research Institute
- Dr. Ben Johnson - Van Andel Research Institute
- Thank you for your Attention!

Napoli R/Bioconductor Meetup

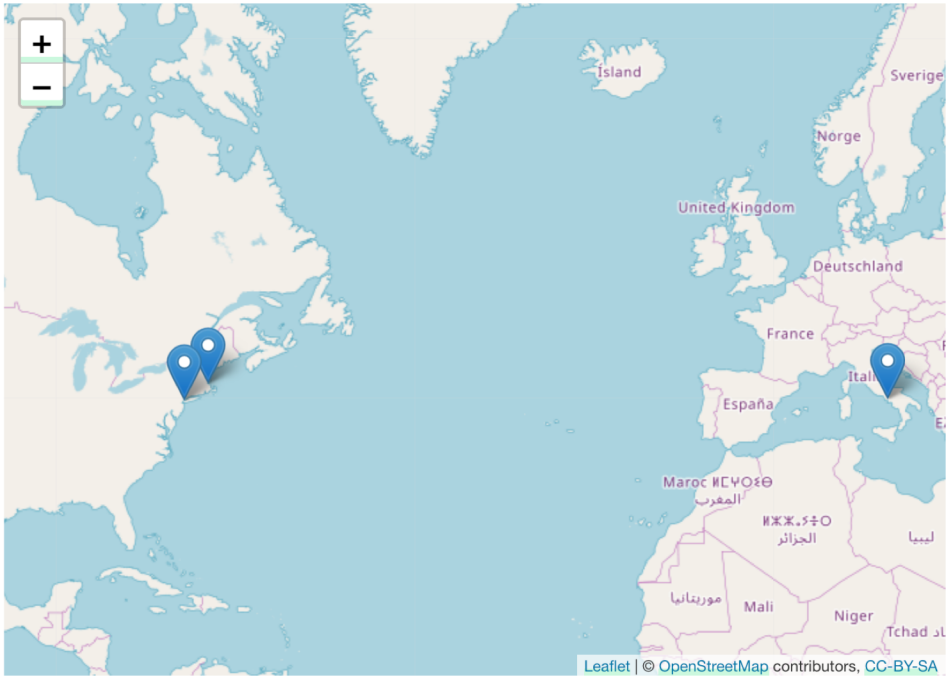
https://www.facebook.com/pg/NapoliRBiocMeetup

http://lists.moo.gs/mailman/listinfo/biocmeetup.naples
napoli.r.bioc@gmail.com

- Since Nov 2018
- R Consortium Array Group
- At least 25 people any event with a good turn-over of attendees
- Eight meetups until now
  - R Package Creation
  - scRNA-seq Analysis
  - Differentially Methylated Regions Analysis
  - Microscope Image Processing
  - Chromosomal Copy Number Changes Detection
  - Bulk RNA-seq Differential Expression
  - Hi-C data analysis using HiCeekR
  - Metagenomics analysis workflow

Napoli R/Bioconductor Meetup

- Part of a wider idea
- Third city in the World
  - Boston (USA)
  - New York (USA)
  - Napoli (IT)
- Useful to
  - share ideas and workflows
  - create new collaborations
  - extend bioinfo community

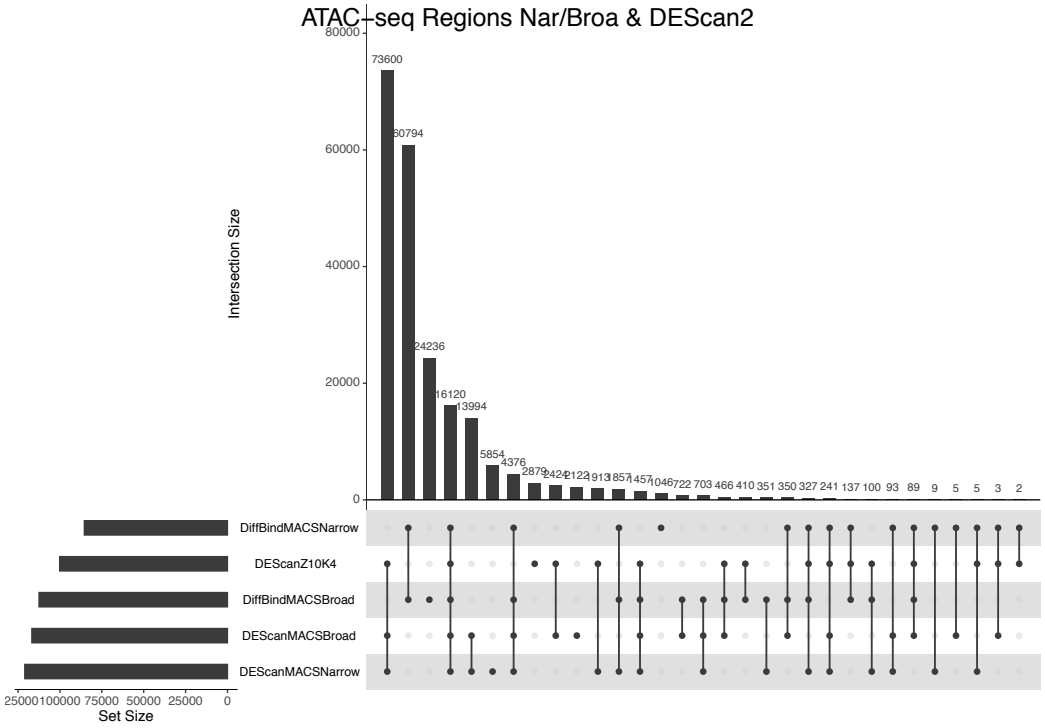


Is there a best Aligner?

Bowtie2 vs BWA

Comparing DARs across methods (2)

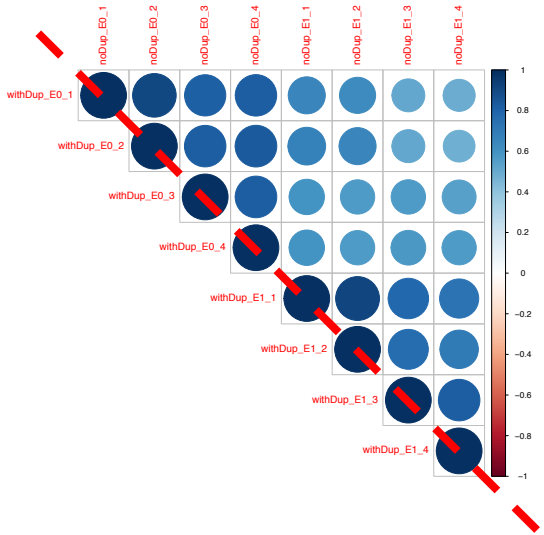
ATAC-seq dataset



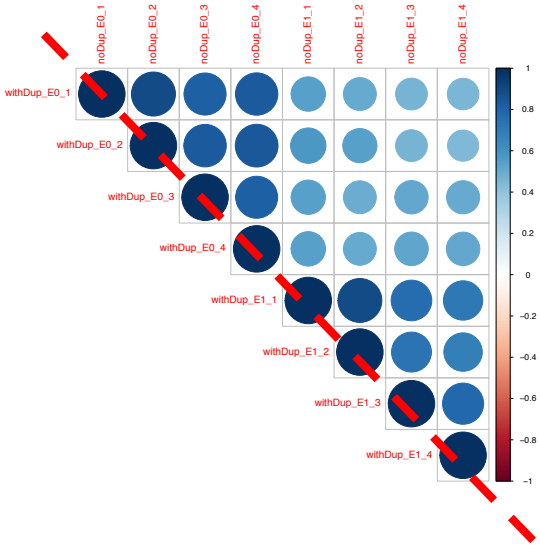
- Ad-hoc designed UpsetPlot on Granges
  - Based on findOverlaps method
- Results description

Duplicates Removal doesn't impact peak detection

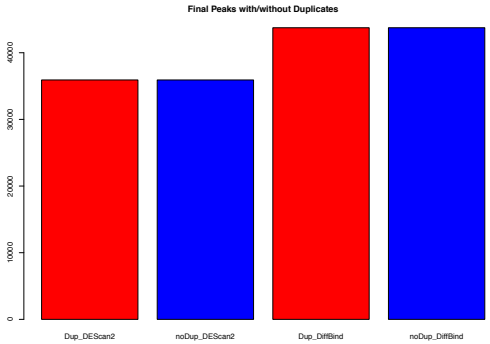
DEScan2



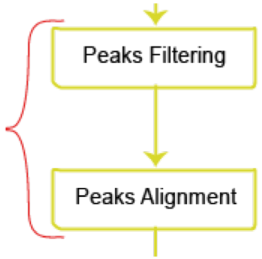
DiffBind

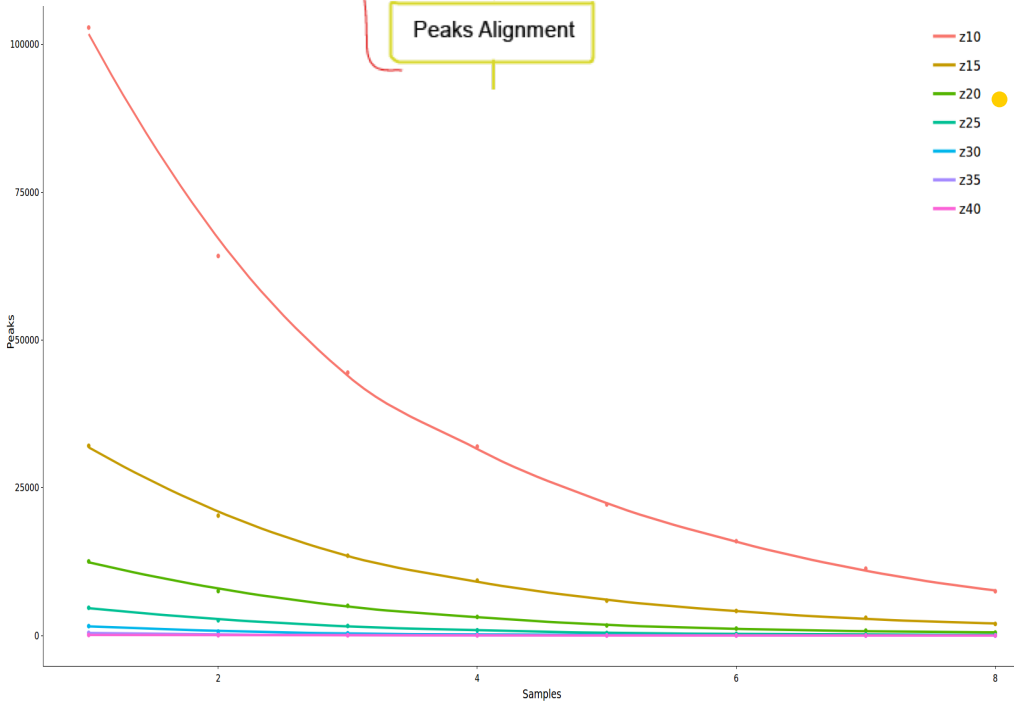


- **Diagonal Correlations** on counts matrices show that there is **no big differences** between duplicates and no-duplicates samples
- rmDup with samtools
- DEScan2 counts matrices
- DiffBind counts matrices



DEScan2 – Differential Enriched Scan 2





- Filter out the peaks with a score lower than a user-defined threshold
- Aligns the peaks over user-defined number of samples
- Different thresholds produce different trends in number of final peaks detected