Rouen University
Master 2.2 BioInforMatique - 2016-2017

Development of an automated pipeline for ChIP-seq data analysis

Stéphanie Levon

IGBMC - GenomEast platform

Supervisor: Stéphanie Le Gras

Institute of Genetics and Molecular and Cellular Biology
GenomEast platform

- One of the main European research centres in biomedical research

- High throughput technologies to survey genomes, their expression and their regulation
- DNA-seq, RNA-seq, smallRNA-seq, ChIP-seq

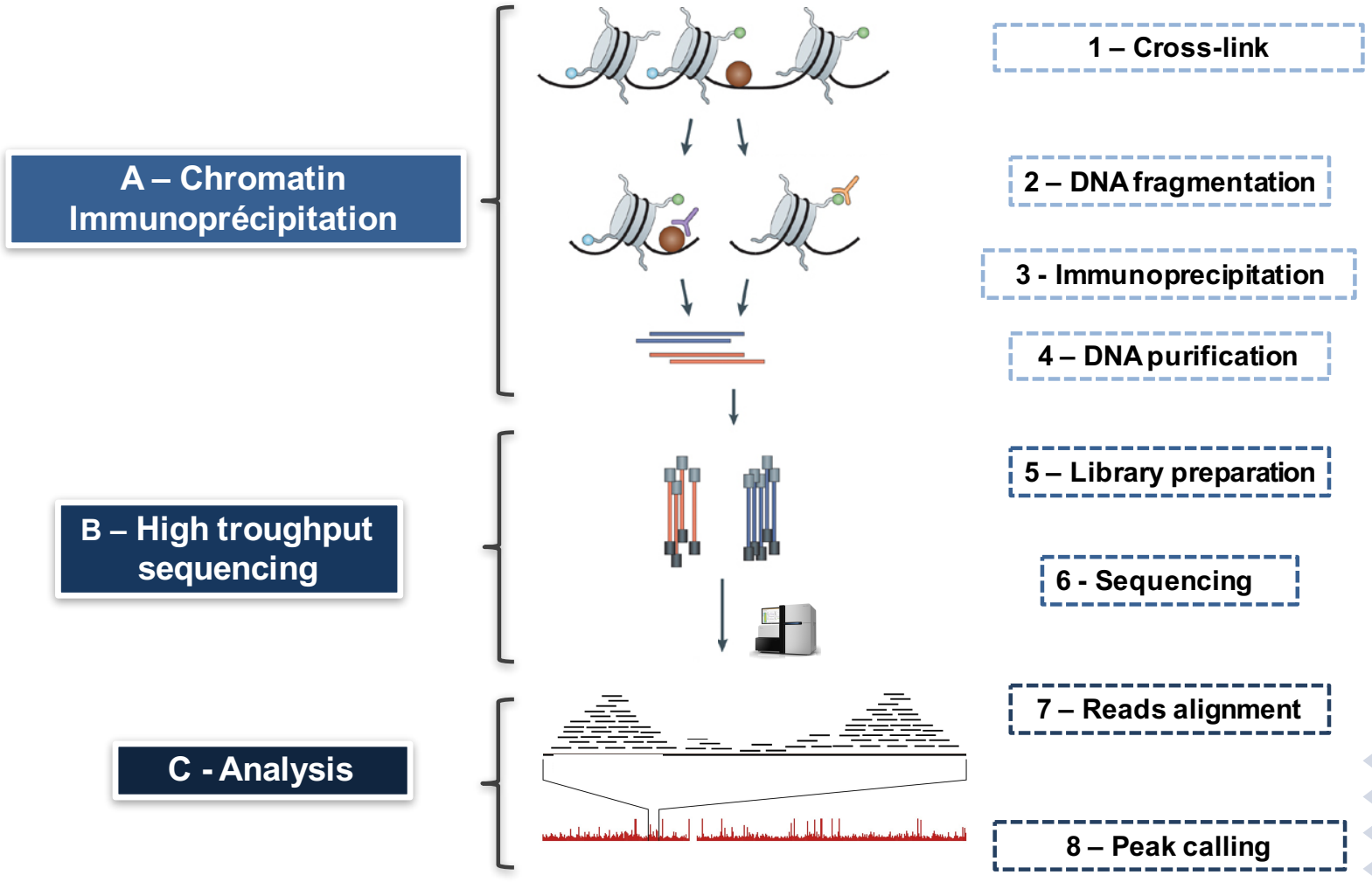
ChIP-seq applications in biology

Enable large scale interaction study between DNA and a specific protein

- Detecting the anchor positions of DNA-binding proteins
- Detecting histone positions with or without post-translational modifications

→ Transcription regulation mechanisms

ChIP-seq: immunoprecipitation followed by sequencing



Goals of my professionnal mission

Development of ChIP-seq analysis pipeline

- Programming languages: coded in Python with R scripts
- Good programming practice:
  - host and versioned by Git
  - verified with Pylint

- Flexibility and parallelization:
  - built with a tool dedicated to pipeline construction
  - interacting with job scheduler of the platform: Slurm

- Flexibility: two configuration files

Configuration files
Experimental design and parameters

design_file.tsv

Sample ID	Name	Cell type	Replicate	Reads file	Control ID	Control name	Control reads file	Peak caller	Motif research
LRCT3	PPARg1	SD48	A	LRCT3.sort.bed.gz	LRCT1	Input1	LRCT1.sort.bed.gz	MACS2	yes
LRCT4	RXRa1	SD48	B	LRCT4.sort.bed.gz	LRCT1	Input1	LRCT1.sort.bed.gz	MACS2	yes
LRCT7	PPARg2	SD48	A	LRCT7.sort.bed.gz	LRCT5	Input2	LRCT5.sort.bed.gz	MACS2	yes
LRCT8	RXRa2	SD48	B	LRCT8.sort.bed.gz	LRCT5	Input2	LRCT5.sort.bed.gz	MACS2	yes

parameter_file.ini

Project information

```
[PROJECT_INFORMATIONS]
project_number =
working_dir =
genome_assembly =
rawdata =
thread_number =
```

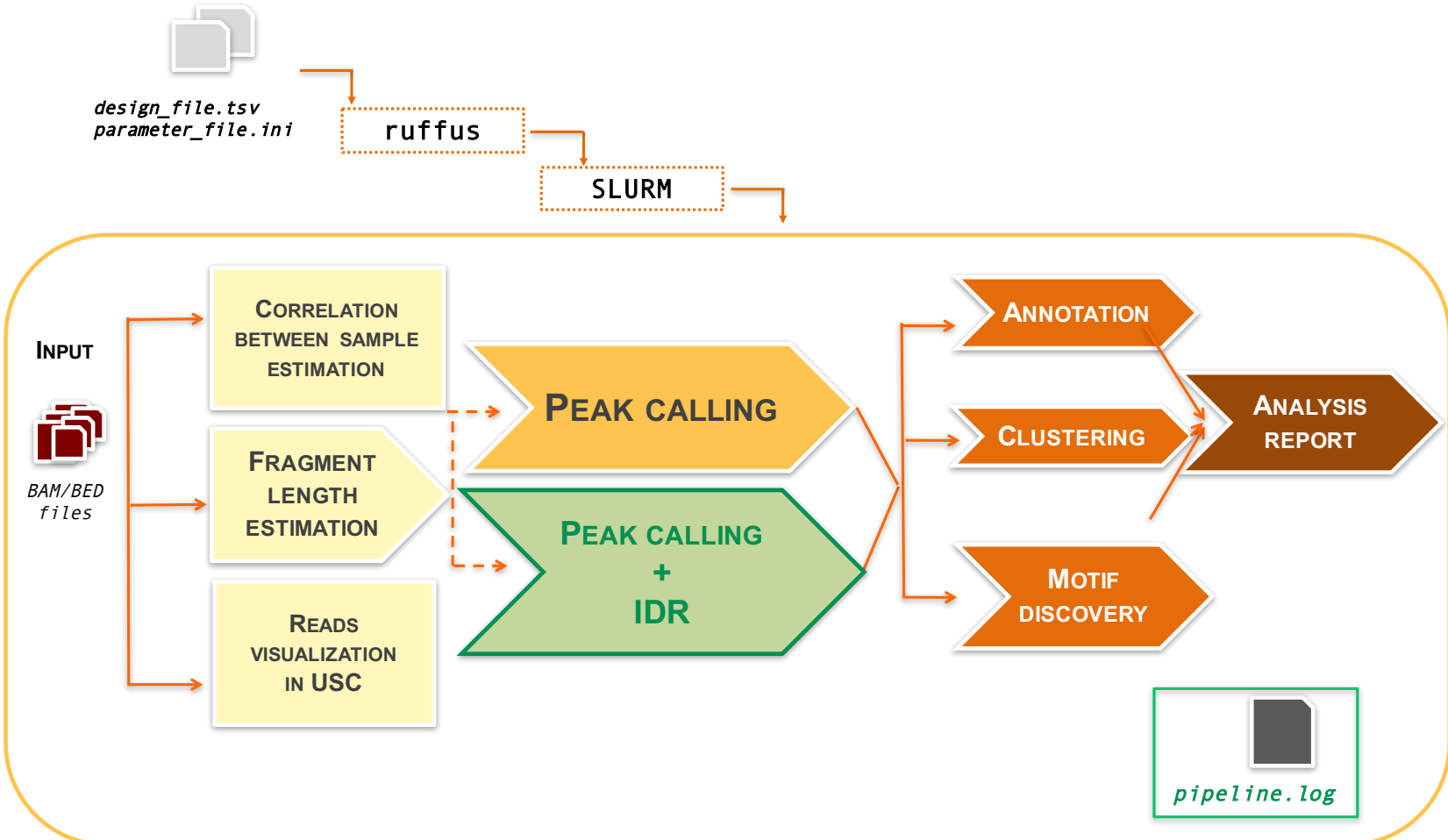
General information

```
[TOOLS_VERSION]
python_version = 2.7.5
homer_version = 4.4
MACS_version = 1.4.2
```

Parameter for each tool

```
[MEME_PARAM]
p_meme-nmotifs = 20
```

The ChIP-seq pipeline analysis



Technical file: the log file

- Enable to trace the analysis

1
```
05/19/2017 02:14:24 PM ##############
```

2
```
Fragment length estimation done for sample RXRa1 - Time spent: 00h 03m 10s
```

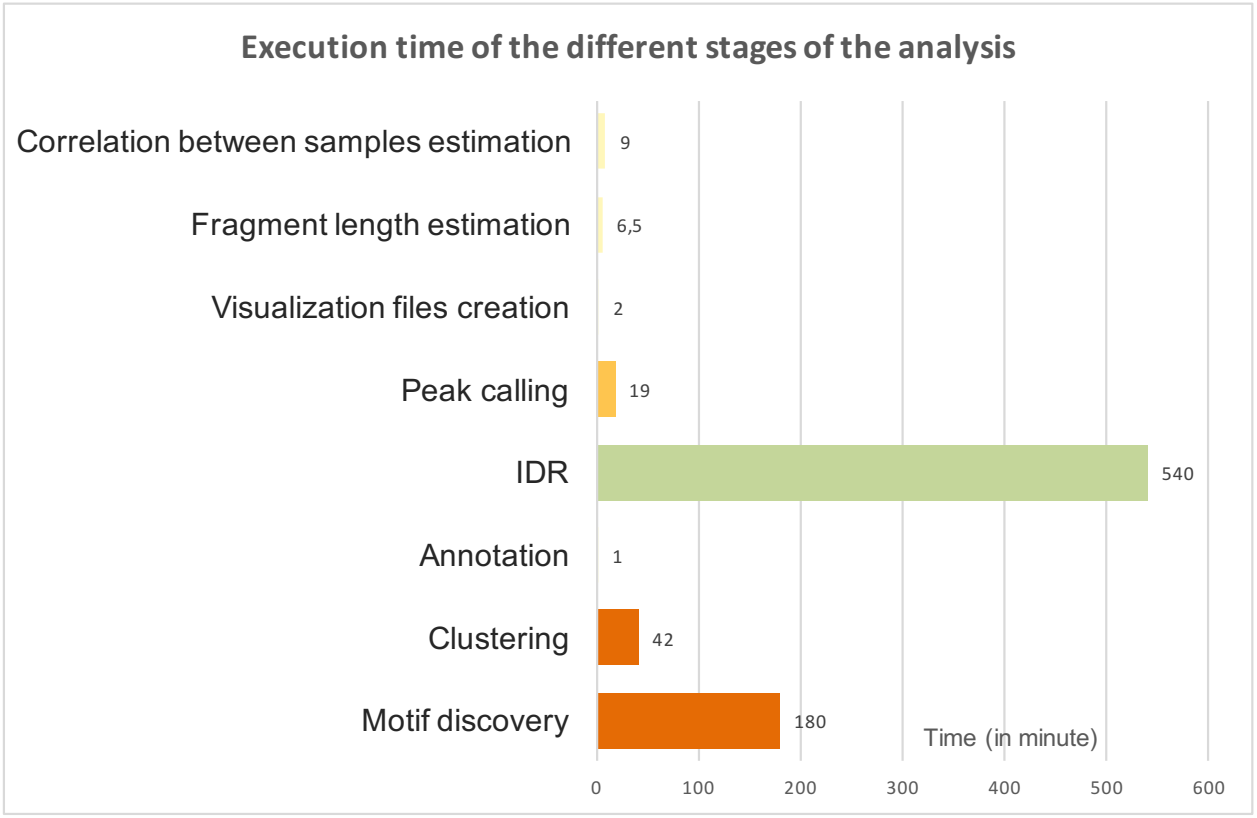
3
```
Command line: makeTagDirectory RXRa1.TagDirectory -tbp 1 RX.sort.bed.gz >> RXRa1.log 2>&1
```

4
```
-------------
Will parse file:RXRa

Creating directory: Input1.TagDirectory and removing existing *.tags.tsv
Treating RX.sort.bed.gz as a GNU zip file
Reading alignment file RX.sort.bed.gz
Guessing that your alignment file is BED format     Optimizing tag files...

Estimated genome size = 39014476
Estimated average read density = 0.025632 per bp
Total Tags = 1000000.0
Total Positions = 882348
Average tag length = 50.0
Median tags per position = 1 (ideal: 1)
Average tags per position = 1.133
Restricting tags per bp...
Fragment Length Estimate: 203
Peak Width Estimate: 254
Autocorrelation quality control metrics:
         Same strand fold enrichment: 1.1
         Diff strand fold enrichment: 1.1
         Same / Diff fold enrichment: 1.0

Guessing sample is ChIP-Seq
```

- Enable to survey the technical performances

Technical performance

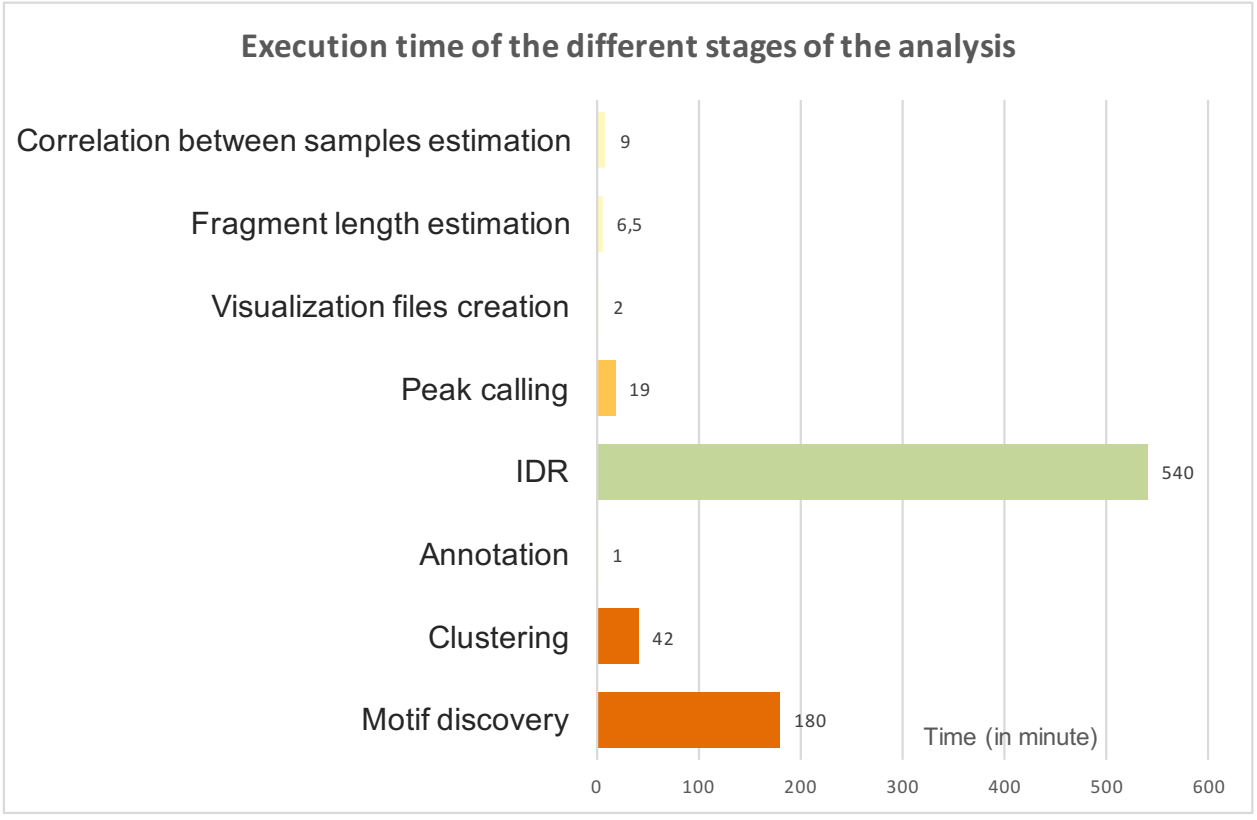
- Project with 4 samples - 45M reads/sample



- Total: 13h08min

Technical performance

- Project with 4 samples - 45M reads/sample



- Total: 13h08min

→ The most time-consuming steps: IDR and motif research

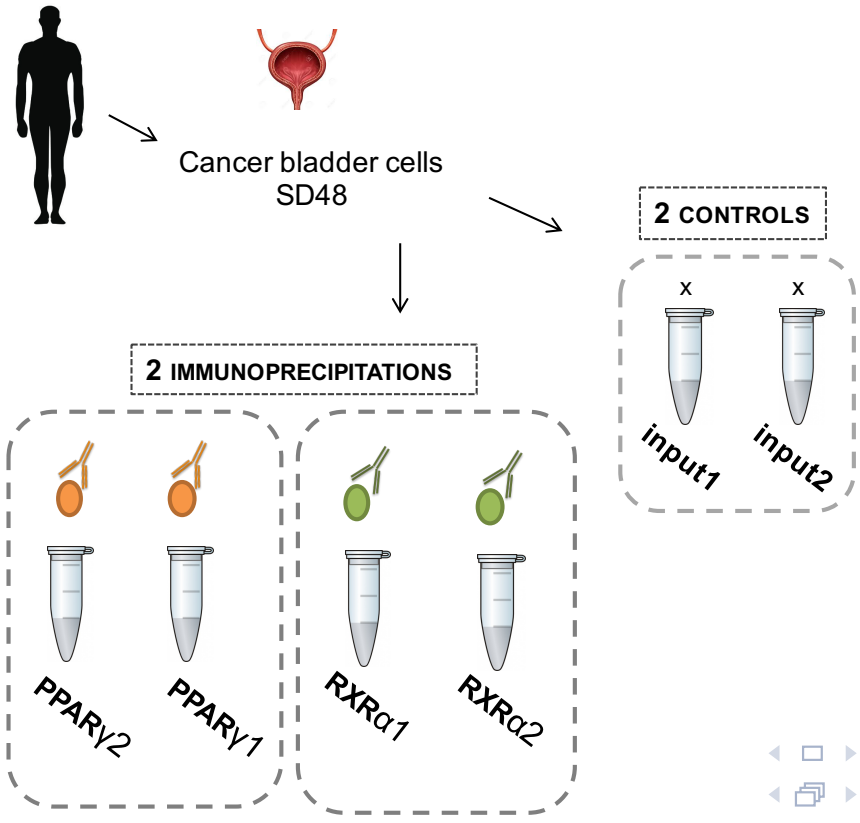
Introducing the project PPARγ-RXRα
Laure Coutos-Thevenot

PPARγ et RXRα
two nuclear receptors

Involved in adipogenesis, anti-inflammatory processes and also in the differentiation of the bladder epithelium

oncogenic role of PPARγ ?

- Identify potentially direct targets of PPARγ et RXRα



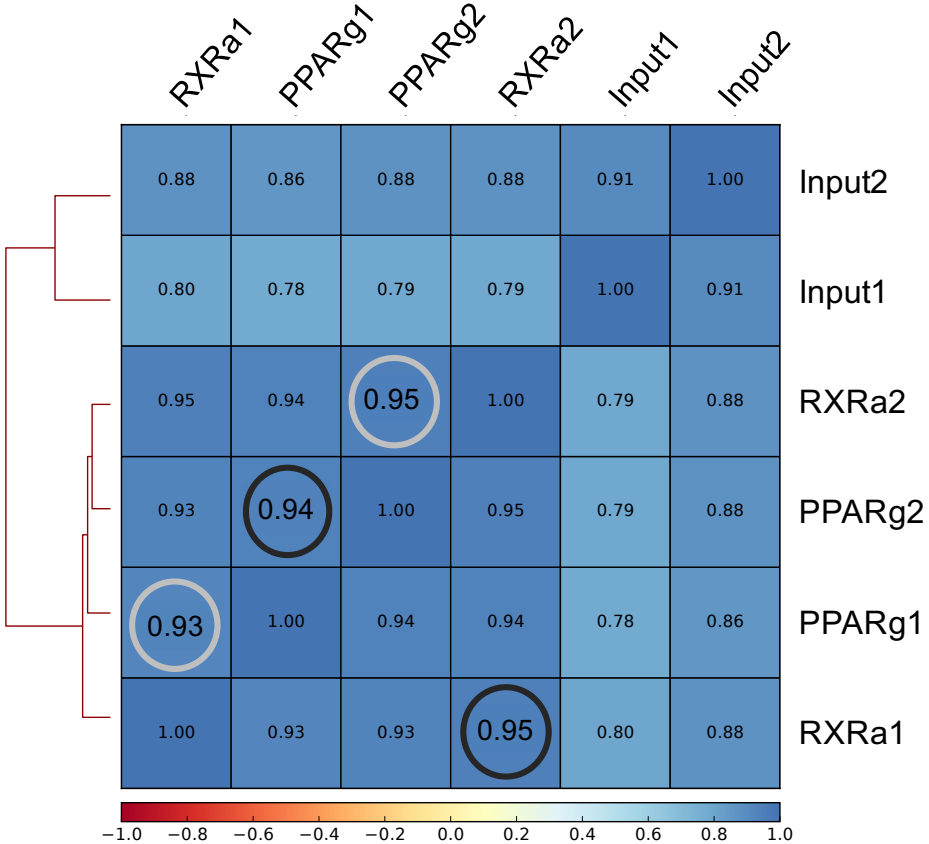
Analysis and visualization of sample correlations
deepTools

→ To assess the experiment reproducibility

- Genome divided in genomic regions
- Computing the overall similarity between two or more files
- Correlation coefficients using Spearman method

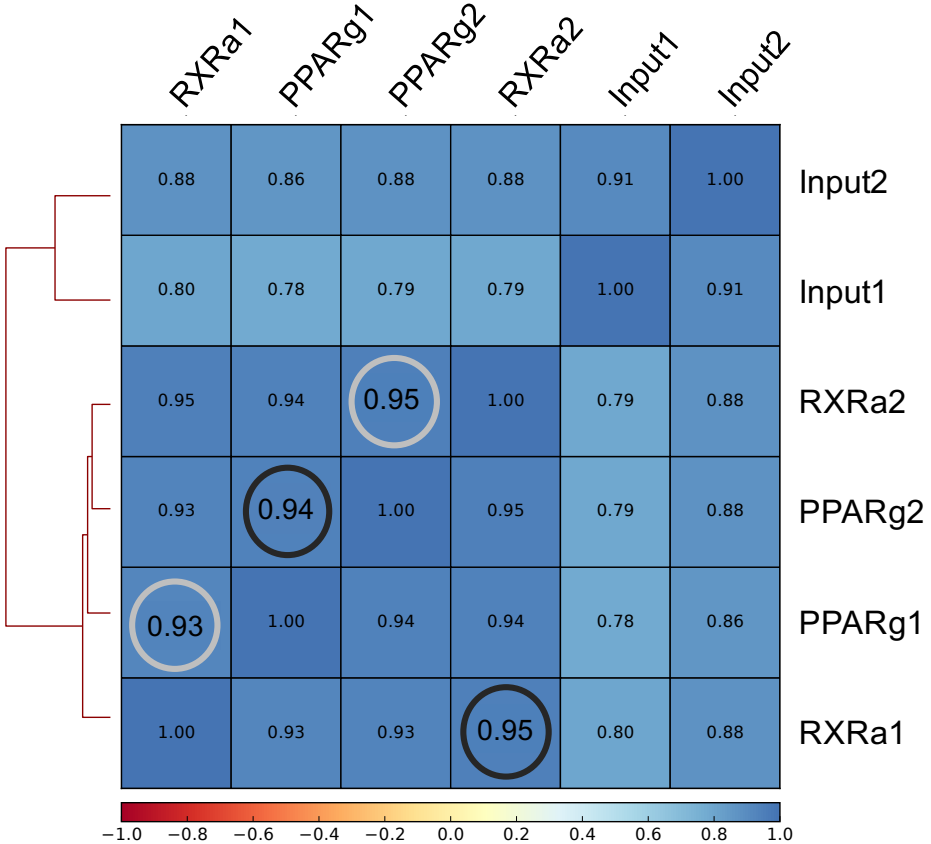
Analysis and visualization of sample correlations
deepTools

→ To assess the experiment reproducibility



Analysis and visualization of sample correlations
deepTools

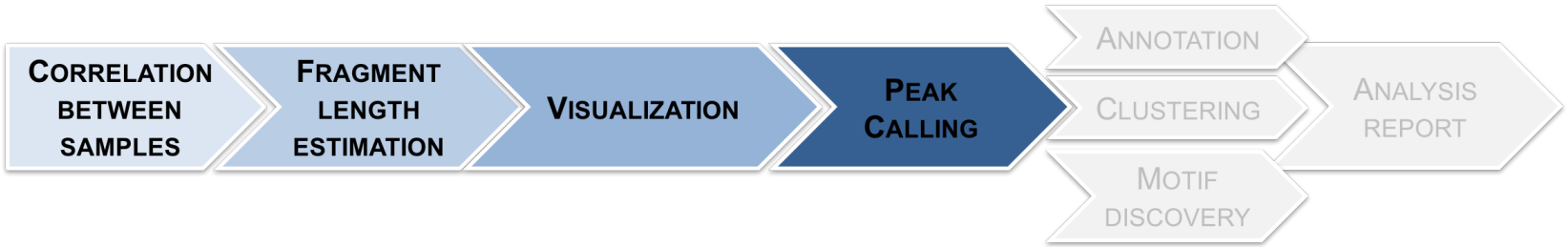
→ To assess the experiment reproducibility

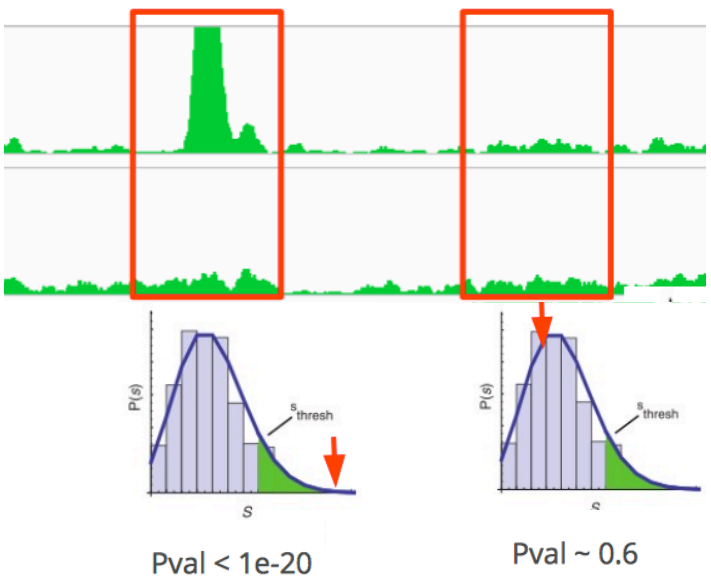


→ Controls clustered

→ Samples clustered

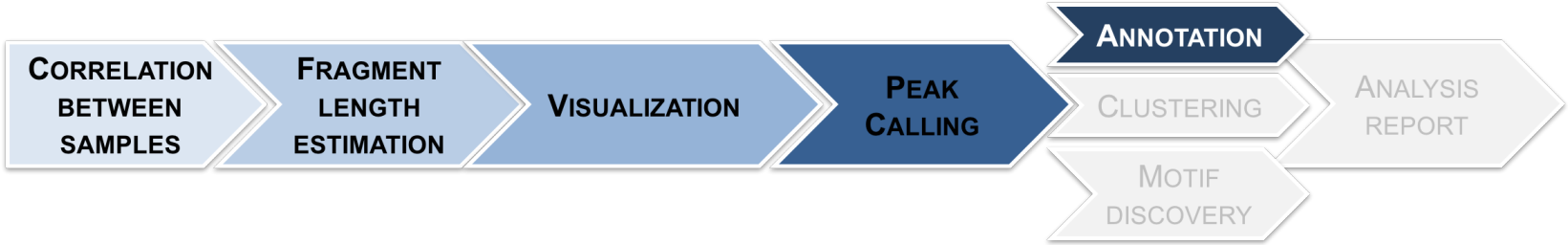
Peak calling
MACS1 et SICER





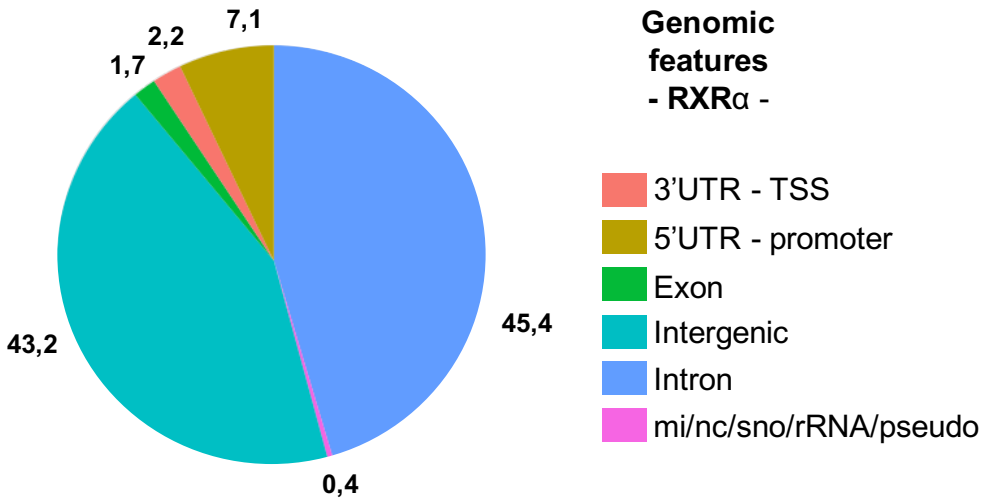
- Dividing the genome into windows
- Evaluation of signal enrichment
- Computing p-value and comparaison with significance threshold
- Multiple tests correction (FDR)

Annotations des pics
Homer

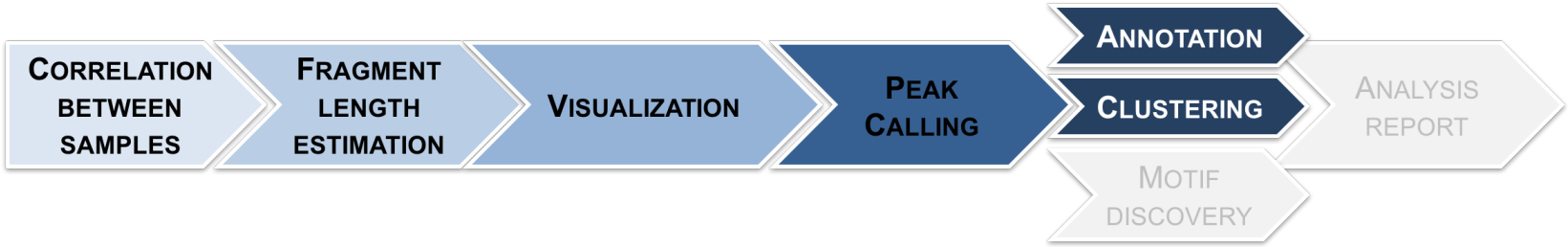


→ Seeking of the functional annotation of the peak center genomic region associated

→ Functional distribution of detected peaks

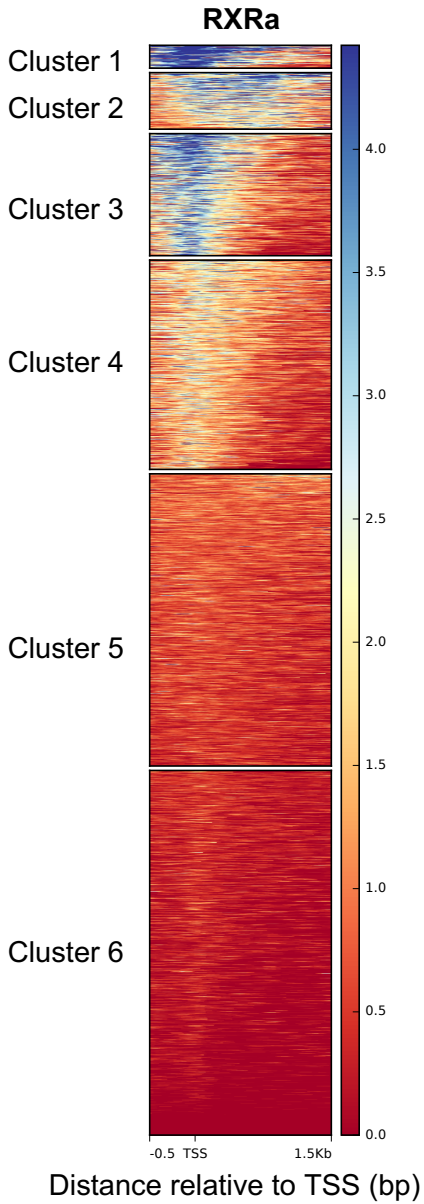


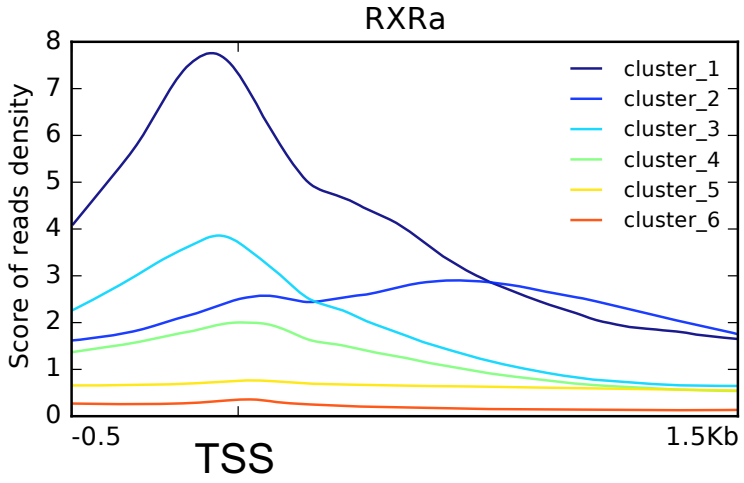
Clustering of genomic regions
deepTools



Goal: having a global vision of reads density surrounding the whole set of TSS of the reference genome

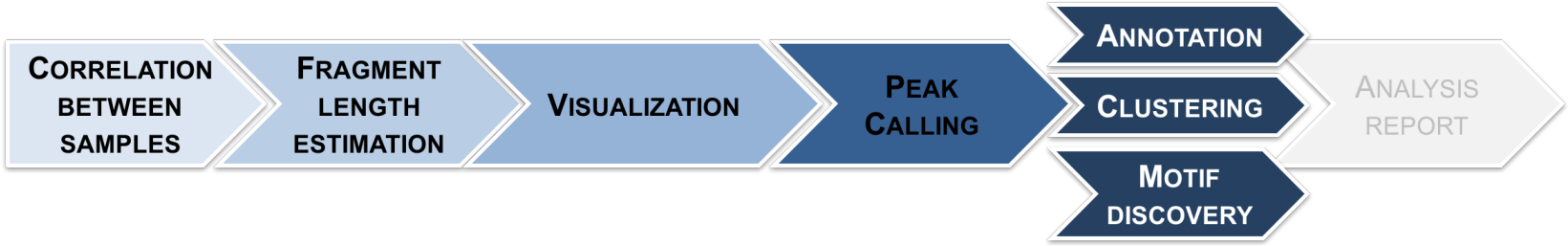
Clustering of genomic regions
deepTools





- *Clusters* 1 and 3: strong enrichment above TSS
- *Clusters* 2 and 4: low enrichment
- *Clusters* 5 and 6: no enrichment

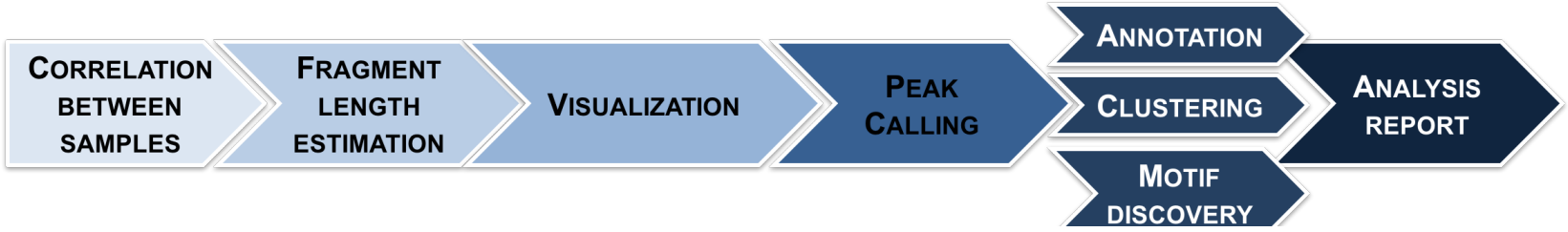
Motif discovery
MEME-ChIP



→ Nucleic sequences overrepresented within the peaks

Outil	E-value	Motif similaire	Distribution
DREME	3,2e-486	ESRRB ESRRA	

Analysis report
Library: Mako, knitr, Latex



1 General information

1.1 Alignment

Alignment was done to the reference **Human** genome (hg38). Table 1 presents statistics of analyzed data.

Sample ID	Sample Name	Uniquely mapped	Unique positions
LRCT1	Input1	42,544,294	37,156,277
LRCT3	PPARg1	45,658,613	33,573,353
LRCT4	RXRa1	52,967,077	40,661,386
LRCT5	Input2	62,533,964	48,350,351
LRCT7	PPARg2	78,386,590	51,165,476
LRCT8	RXRa2	74,406,514	57,136,352

Table 1: **Statistics on input data.** The 'Uniquely mapped' column represents the number of reads uniquely aligned onto the reference genome. The 'Unique positions' column contains the number of different positions in the genome to which uniquely aligned reads are.

1.2 Data visualization

We recommand the use of the UCSC genome browser [1] (`https://genome.ucsc.edu`) to visualize your ChIP-seq data. You can visualize them either :

- **by using a UCSC track hub we prepared** (see Figure 10, section 4 Methods for more details on the usage)

  `ftp://test_trackHub:stephanie@ngs.igbmc.fr/analyzeddata/S16071_170331/hub.txt`

  We would like to point out that this link is only for preliminary visualization of the data and will only be available for **30 days** starting from the data delivery date.

- **by uploading the bedGraph files to UCSC** (see section 3 Delivery and Figure 11, section 4 Methods for more details on the usage)

2 Results

2.1 Sample correlation

Sample correlation was performed using deepTools [2]. See Figure 1.

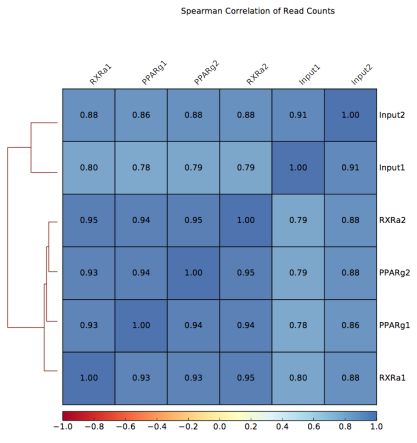


Figure 1: **Sample correlation heatmap using the read enrichment of all 2500pb bins computed across entire genome.** Correlation coefficient are computed with Spearman method. On the clustered heatmap, the colors represent the correlation coefficients and the clusters are joined using the Nearest Point Algorithm (left part of the heatmap).

2.2 Peak calling and irreproducible discovery rate

Peak calling was performed using MACS2. Irreproducible discovery rate is performed when replicates exists. After the peak calling step, a threshold is advised in order to filter significant common peaks. This threshold is set to 0.01 for human and mouse. Table 2 presents the number of significant peaks which overlap between replicate 1 and replicate 2.

Condition	Number of **significant** overlapping peaks
PPARg	36,123
RXRa	41,355

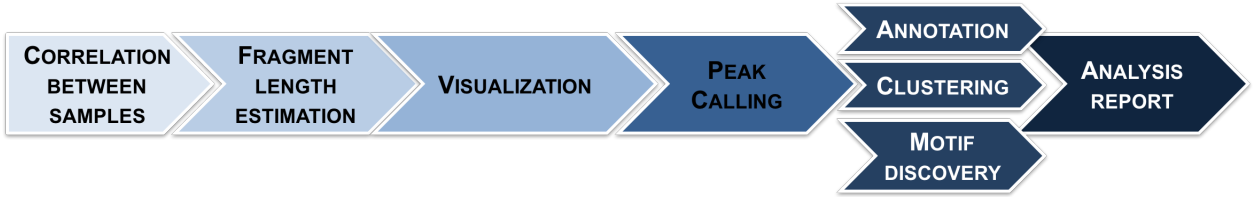
Table 2: **Peaks number overlapping between replicate 1 and replicate 2 after filtering according to IDR threshold.** Significant overlapping peaks are assessed with IDR threshold set to 0.01.

The peaks number can vary from IDR file to annotation file because IDR doesn't merge overlapping peaks while annotation does.

Conclusion and prospects

Conclusion

- ChIP-seq: experimental design, analysis steps and tools
- Standardization and automation of ChIP-seq data analysis
- Compliance with specifications



Prospects

- Improvement of IDR analysis and motif discovery step
- Adding quality control steps
- Differential analysis
- Implementation under Galaxy

Career plans and acknowledgements

Career plans

- Gain experience in the field of NGS analysis and / or pipeline development in the USA

Acknowledgements

Christelle Thibault-Carpentier
Céline Keime
Stéphanie Le Gras
Matthieu Jung
Tao Ye
Damien Plassard
Laure Coutos-Thevenot

and all the members of the GenomEast platform

Thank you for your attention