

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



UFR Sciences
et Techniques



CENTRE UNIVERSITAIRE DE
FORMATION CONTINUE
ET PAR ALTERNANCE



Université de Rouen
Master 2.2 BiolforMatique - Année 2016-2017

Développement d'un pipeline d'analyses de données ChIP-seq

Stéphanie Levon

IGBMC - Plateforme GenomEast

Encadrante : Stéphanie Le Gras, IE



UNIVERSITÉ DE STRASBOURG



Institut de Génétique et de Biologie Moléculaire et Cellulaire

Plateforme GenomEast

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



- Centre de recherche en biomédecine



- Technologies dédiées à l'analyse du génome à grande échelle
- DNA-seq, RNA-seq,
smallRNA-seq, ChIP-seq

ChIP-seq : applications en biologie

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives

Permet l'étude des interactions entre une protéine et de l'ADN
à l'échelle du génome entier

- Déetecter les positions d'ancrage des protéines de liaison à l'ADN
- Déetecter les positions des histones avec ou sans modifications post-traductionnelles

→ Mécanismes de régulation de la transcription



Le ChIP-seq, immunoprécipitation puis séquençage

L'IGBMC

Introduction et objectif

Présentation de l'outil

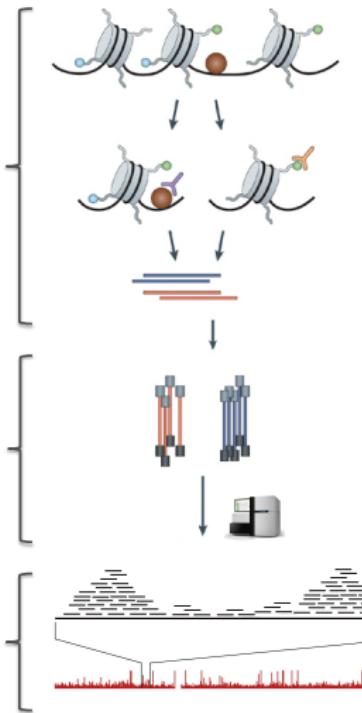
Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives

A - Immunoprécipitation de la chromatine

B - Séquençage haut débit

C - Analyse bioinformatique



1 - Fixation des interactions protéines-ADN

2 - Fragmentation de l'ADN

3 - Immunoprécipitation

4 - Purification de l'ADN

5 - Préparation des librairies

6 - Séquençage

7 - Alignement des lectures

8 - Détection des pics



Objectifs de ma mission

Développement d'un pipeline d'analyse de données ChIP-seq

- Langages : codé en Python avec des scripts R



- Respect des bonnes pratiques de programmation :

- versionné
 - syntaxe vérifiée avec Pylint



GitLab

- Modulable et parallélisable :

- construit avec un outil d'aide à la création de pipeline
 - compatible avec le gestionnaire de file de la plateforme Slurm



Ruffus



- Paramétrisable : configuré à l'aide de deux fichiers de paramètres

Fichiers de configuration

Plan de l'expérience et paramètres

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives

design_file.tsv

Identifiant	Nom	Type cellulaire	Répliquat	Fichier de lectures	Identifiant du contrôle	Nom du contrôle	Fichier de lectures contrôle	Outil de détection des pics	Recherche de motif
LRCT3	PPARG1	SD48	A	LRCT3.sort.bed.gz	LRCT1	Input1	LRCT1.sort.bed.gz	MACS2	yes
LRCT4	RXR α 1	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	RXR α 2	SD48	B	LRCT8.sort.bed.gz	LRCT5	Input2	LRCT5.sort.bed.gz	MACS2	yes

parameter_file.ini

Informations relatives au projet

[PROJECT_INFORMATION]

project_number =
working_dir =
genome_assembly =
rawdata =
thread_number =

Informations générales

[TOOLS_VERSION]

python_version = 2.7.5
homer_version = 4.4
MACS_version = 1.4.2

Paramètres pour l'outil MEME-CHIP

[MEME_PARAM]

p_meme-nmotifs = 20



Description du pipeline d'analyse de données ChIP-seq

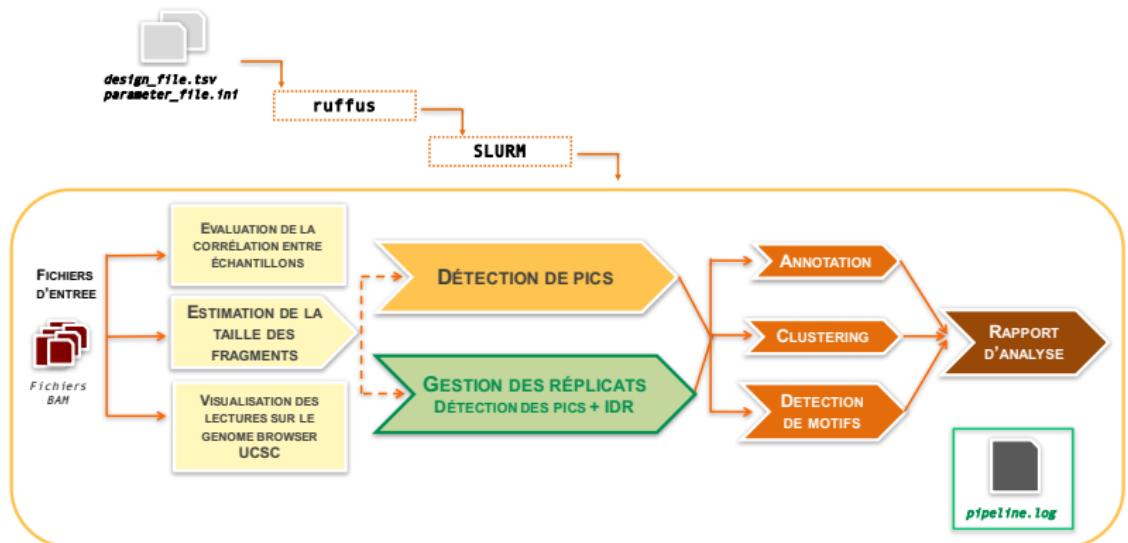
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



Création du fichier de log

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives

- Permet de retracer l'analyse

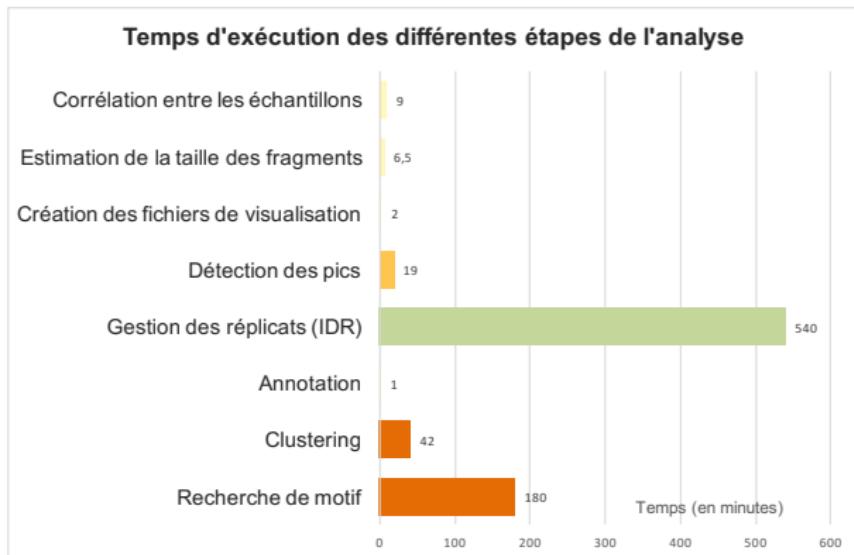
```
1 05/19/2017 02:14:24 PM #####
2 Fragment length estimation done for sample RXRa1 - Time spent: 00h 03m 10s
3 Command line: makeTagDirectory RXRa1.TagDirectory -tbp 1 RX.sort.bed.gz >> RXRa1.log 2>&1
-----
4 Will parse file:RXRa
Creating directory: Input1.TagDirectory and removing existing *.tags.tsv
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
```

- Permet d'avoir accès aux performances techniques

Performances techniques

L'IGBMC
Introduction et objectif
Présentation de l'outil
Analyse du projet PPAR γ -RXR α
Conclusion et perspectives

- Projet de 4 échantillons - 45M lectures/échantillon

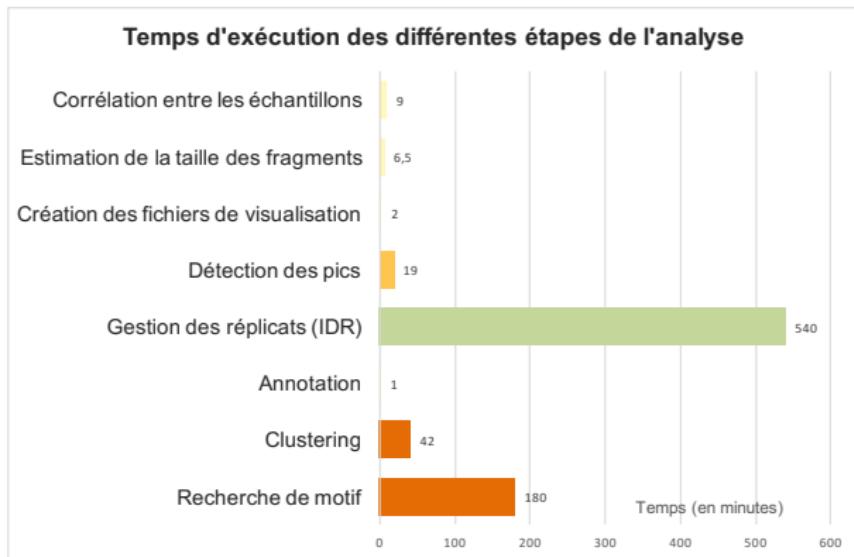


- Temps total : 13h08min

Performances techniques

L'IGBMC
Introduction et objectif
Présentation de l'outil
Analyse du projet PPAR γ -RXR α
Conclusion et perspectives

- Projet de 4 échantillons - 45M lectures/échantillon



- Temps total : 13h08min
- ➔ Étapes les plus chronophages : IDR et détection de motifs

Présentation du projet PPAR γ -RXR α

Laure Coutos-Thevenot

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

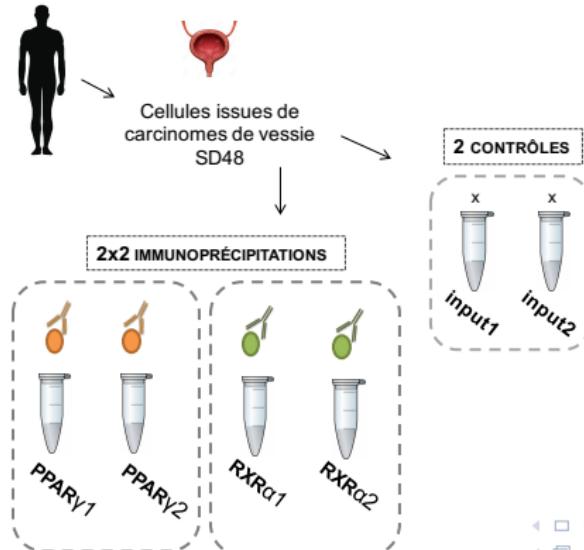
Conclusion
et
perspectives

PPAR γ et RXR α deux récepteurs nucléaires

Rôle dans adipogénèse, les processus
anti-inflammatoire et la
différentiation des cellules
épithéliales de vessie urinaire

Rôle oncogénique de PPAR γ ?

- Quels sont les cibles de PPAR γ et RXR α ?



Évaluation de la corrélation entre échantillons deepTools

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



→ Évaluer la reproductibilité d'une expérience

- Séparation du génome en fenêtres
- Comparaison du nombre de lectures par fenêtre entre deux échantillons
- Calcul d'un coefficient de corrélation par la méthode Spearman



Évaluation de la corrélation entre échantillons deepTools

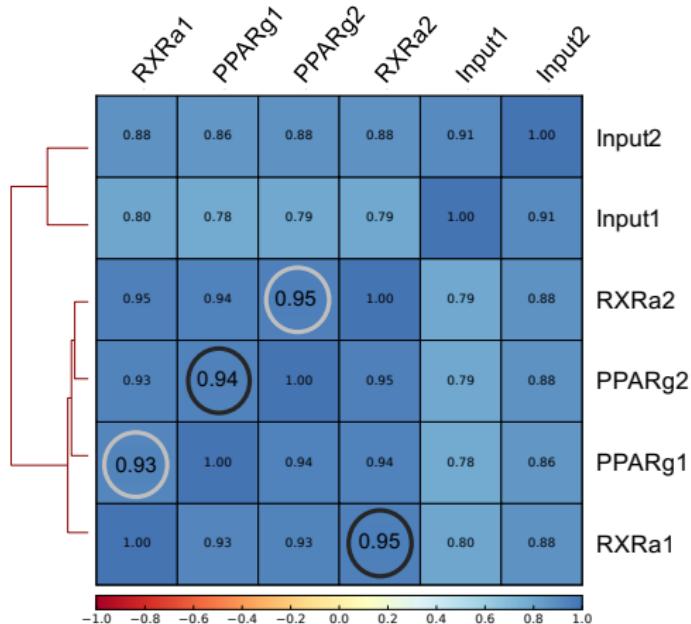
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



Évaluation de la corrélation entre échantillons deepTools

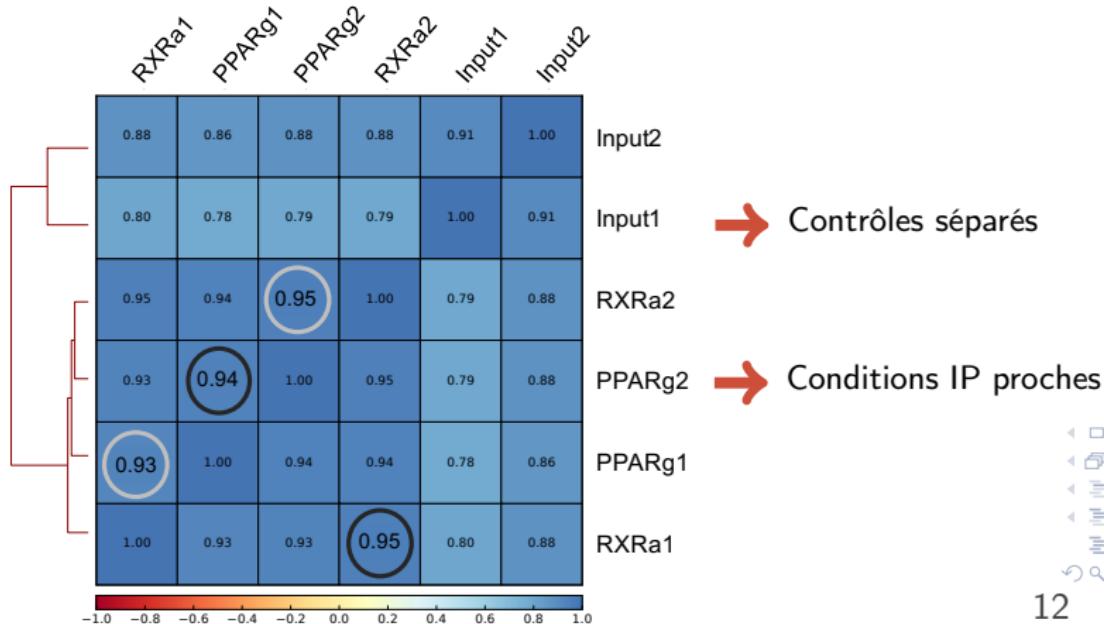
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



Détection des pics

MACS1 et SICER

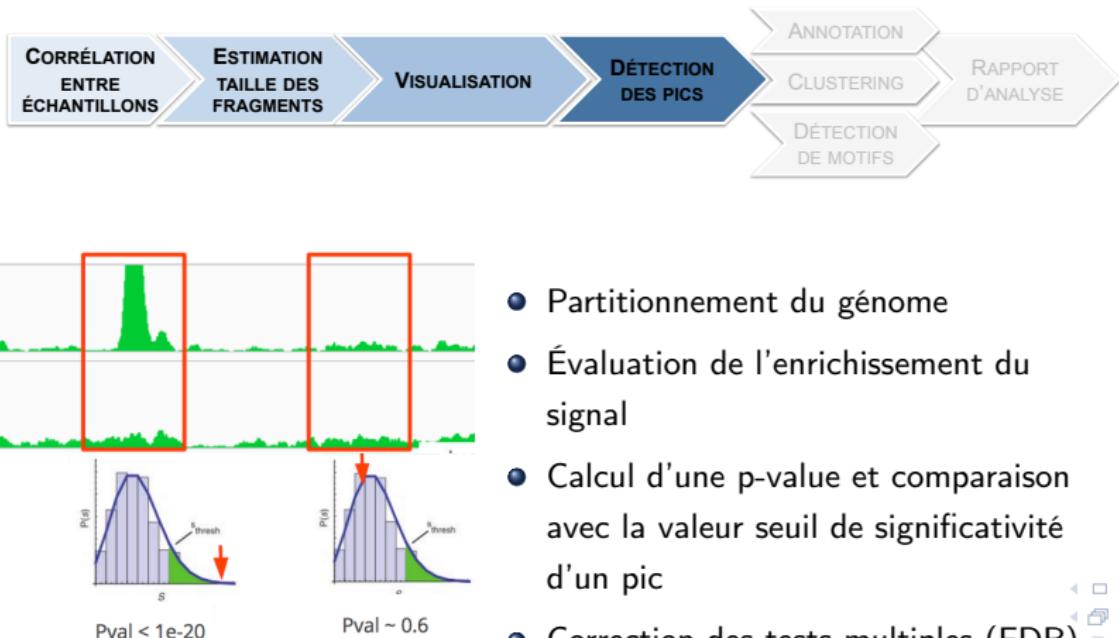
L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



Gestion des réplicats

Méthode IDR : Qunhua Li et Anshul Kundaje

L'IGBMC

Introduction
et objectif

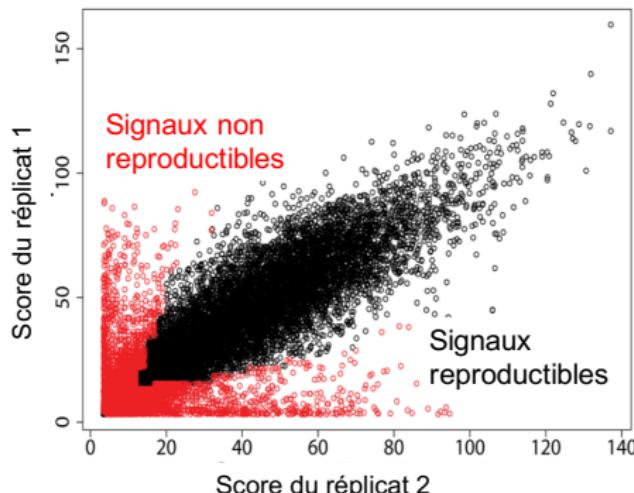
Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



→ Filtrer les pics détectés entre plusieurs réplicats d'une condition pour ne garder que les **pics reproductibles**



Gestion des réplicats

Méthode IDR

L'IGBMC

Introduction et objectif

Présentation de l'outil

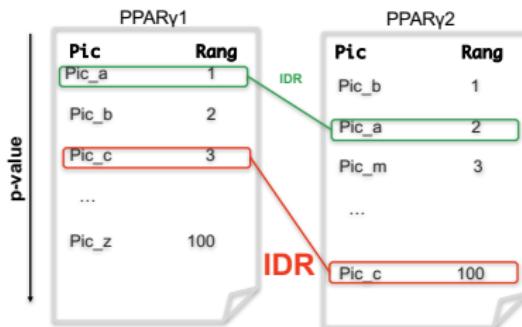
Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



Principe

- Détection de pics avec paramètres peu stringeants (MACS2)
- Ordonnancement des pics par score de significativité (p-value)
- Comparaison des rangs
- Attribution d'un score IDR



Gestion des réplicats

Méthode IDR

L'IGBMC

Introduction et objectif

Présentation de l'outil

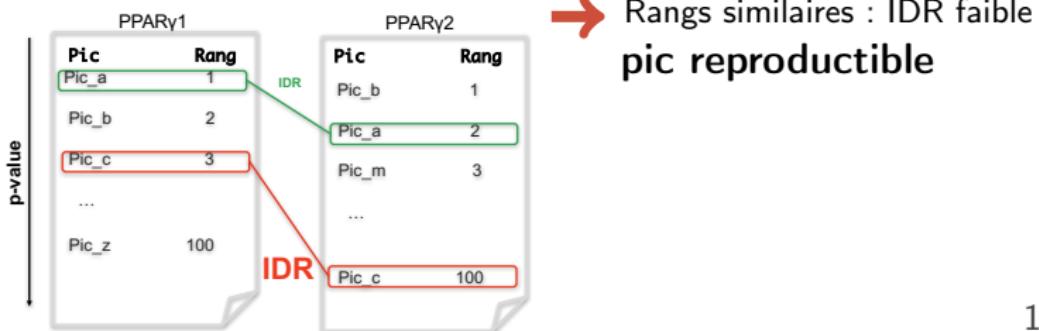
Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



Principe

- Détection de pics avec paramètres peu stringeants (MACS2)
- Ordonnancement des pics par score de significativité (p-value)
- Comparaison des rangs
- Attribution d'un score IDR



Gestion des réplicats

Méthode IDR

L'IGBMC

Introduction et objectif

Présentation de l'outil

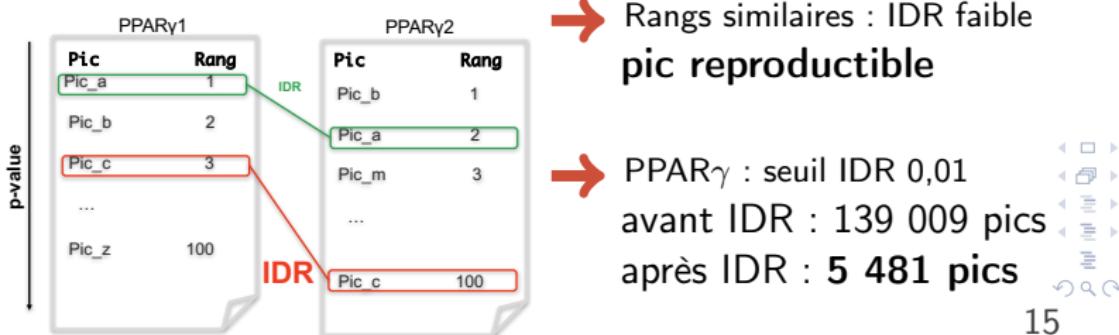
Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



Principe

- Détection de pics avec paramètres peu stringeants (MACS2)
- Ordonnancement des pics par score de significativité (p-value)
- Comparaison des rangs
- Attribution d'un score IDR



Annotation des pics

Homer

L'IGBMC

Introduction et objectif

Présentation de l'outil

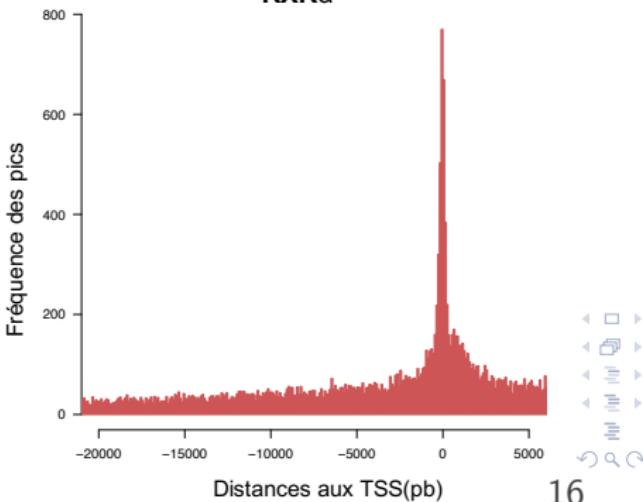
Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



- Recherche le site d'initiation de la transcription le plus proche de chaque pic
- Permet d'observer la distribution des fréquences des distances "pics - TSS"

Distances aux TSS des gènes recensés sur le génome humain
- RXR α -



Annotation des pics

Homer

L'IGBMC

Introduction et objectif

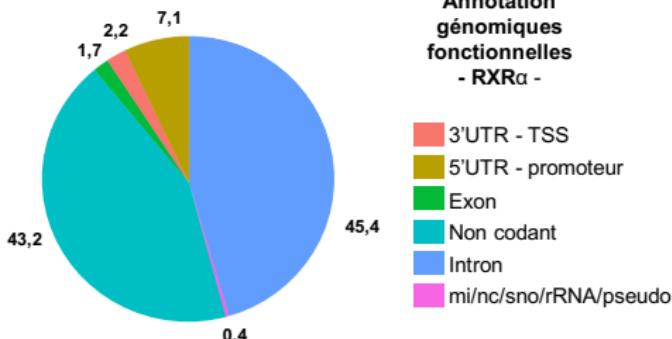
Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



→ Recherche de l'annotation fonctionnelle de la région occupée par le centre des pics
→ Permet d'observer la répartition fonctionnelle des pics détectés



Clustering des régions génomiques

deepTools

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



But : Avoir une vision globale des densités en lectures par rapport à une liste de régions d'intérêts



Clustering des régions génomiques

deepTools

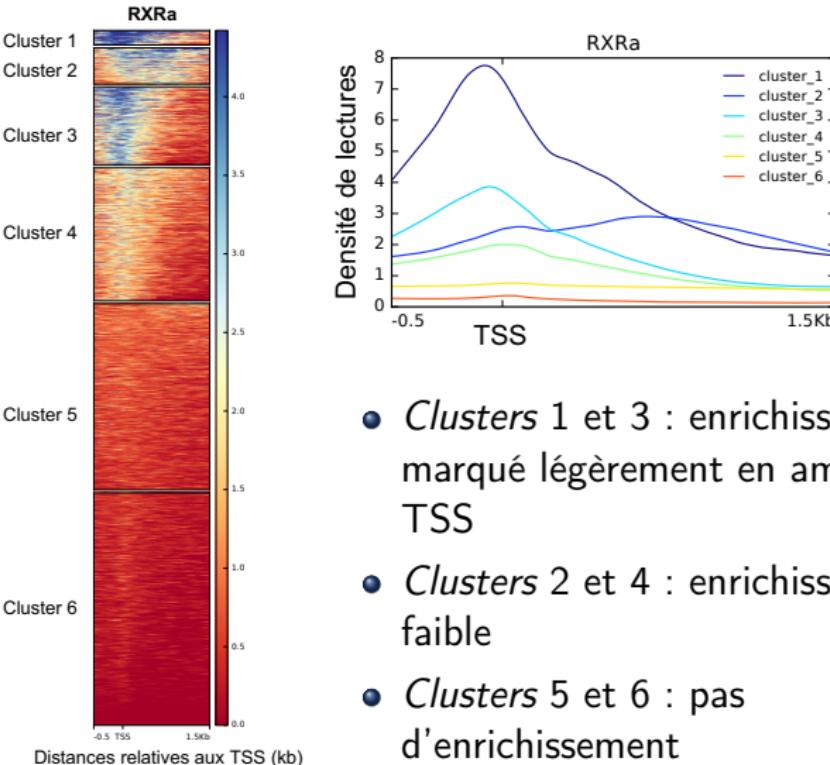
L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



- *Clusters 1 et 3 : enrichissement marqué légèrement en amont du TSS*
- *Clusters 2 et 4 : enrichissement faible*
- *Clusters 5 et 6 : pas d'enrichissement*

Détection de motifs MEME-ChIP

L'IGBMC

Introduction
et objectif

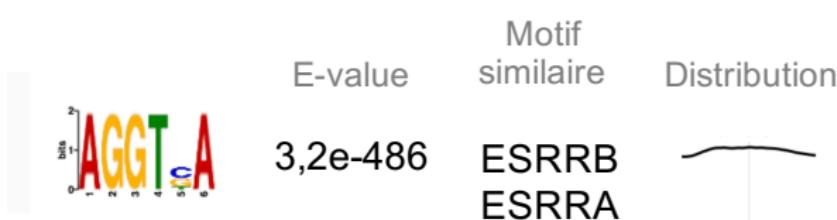
Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives



→ Découverte de séquences nucléiques retrouvées fréquemment au sein des pics de RXR α



Génération d'un rapport automatique

Librairie Mako, knitr, Latex

L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



2 RESULTS

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	87,159,277
LRCT3	PPARG1	45,658,613	33,573,353
LRCT4	RXRA1	52,067,077	40,061,286
LRCT5	Input2	62,533,484	34,533,511
LRCT7	PPARG2	78,389,590	51,105,476
LRCT8	RXRA2	74,400,514	57,136,332

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 recommend 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@igsb.igbmc.fr/analyzedata/B16071_170381/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 hubGraph 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.

2 RESULTS

2.2 Peak calling and irreproducible discovery rate

Spearman Correlation of Read Counts

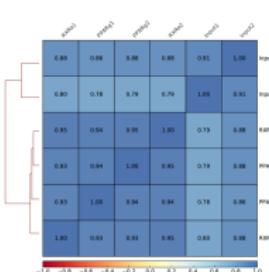


Figure 1: Sample correlation heatmap using the read enrichment of all 2500pb bins computed across entire genome. Correlation coefficient are computed with Spearman method. On the clustered heatmap, the colors represent the correlation coefficients and the clusters are joined using the Nearest Point Algorithm (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 peak number can vary from IDR file to annotation file because IDR doesn't merge overlapping peaks while annotation does.

Conclusion et perspectives

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives

Conclusion

- ChIP-seq : plan expérimental, étapes, outils d'analyses
- Standardisation et automatisation de l'analyse de données de ChIP-seq
- Respect du cahier des charges



Perspectives

- Amélioration des étapes de gestion des réplicats et recherche de motifs
- Ajout d'étapes de contrôles qualités
- Analyse différentielle
- Implémentation sous Galaxy



Remerciements

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

ainsi que l'ensemble des membres de la plateforme GenomEast

Merci de votre attention



Visualisation des lectures via UCSC

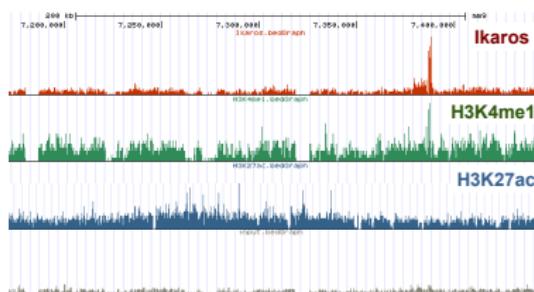
L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
 $PPAR\gamma$ -
 $RXR\alpha$

Conclusion
et
perspectives



Visualisation des lectures via UCSC

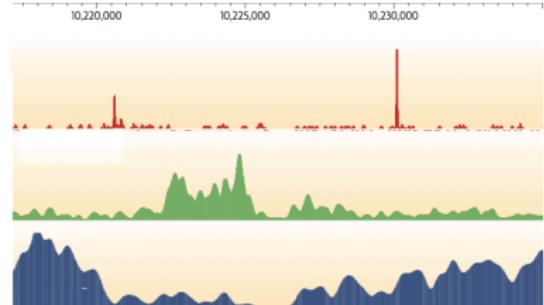
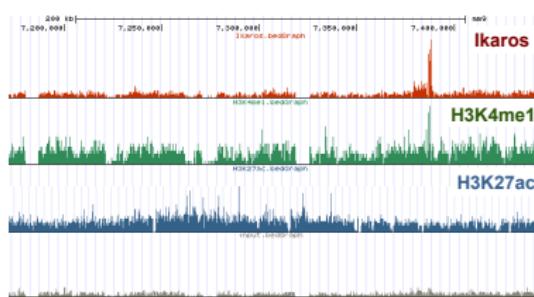
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



→ 2 profils de pics différents



Visualisation des lectures via UCSC

Fichiers générés avec HOMER

L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



Chargement des bedGraphs



Création d'un trackhub à partir des bigWig



→ Même résultats de visualisation



Évaluation de la taille des fragments d'ADN

Homer

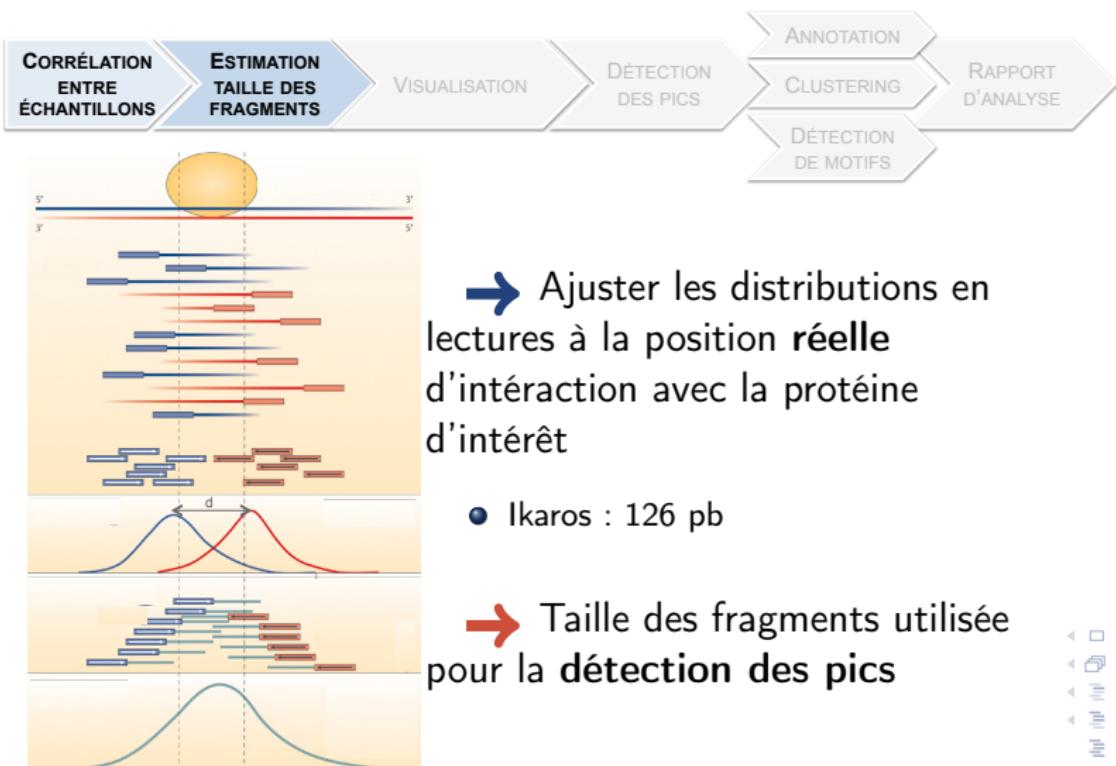
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



Détection des pics

MACS

L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

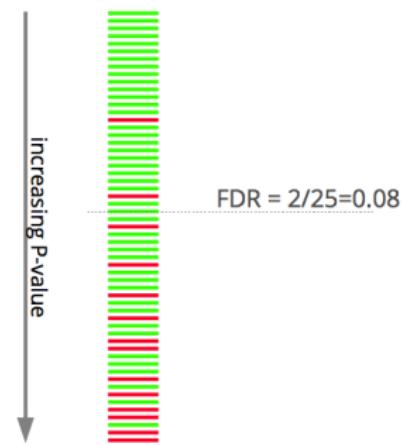
Conclusion et perspectives



Correction des tests multiples - *False Discovery Rate (FDR)*

Échange entre l'IP et le contrôle

$$FDR(p) = \frac{\text{pics negatifs avec } Pval < p}{\text{pics positifs avec } Pval < p}$$



Détection des pics

SICER

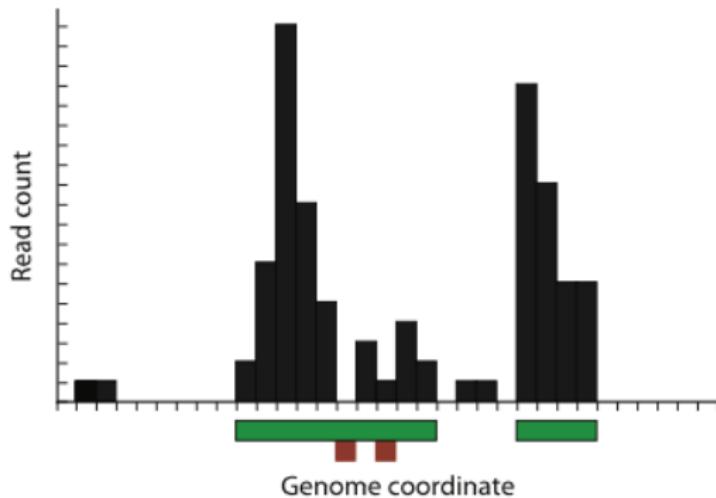
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -RXR α

Conclusion et perspectives



Génération d'un rapport automatique

Librairie Mako, knitr, Latex

L'IGBMC
Introduction et objectif
Présentation de l'outil
Analyse du projet PPAR γ -RXR α
Conclusion et perspectives



Caractéristiques

- Automatiquement généré
- Dynamique

Modèle de rapport

Rnw.mako

mako ↓ python

Rnw

knitr ↓ R

Latex

pdflatex ↓ LATEX

PDF

Rapport d'analyse spécifique d'un projet



Gestion des réplicats

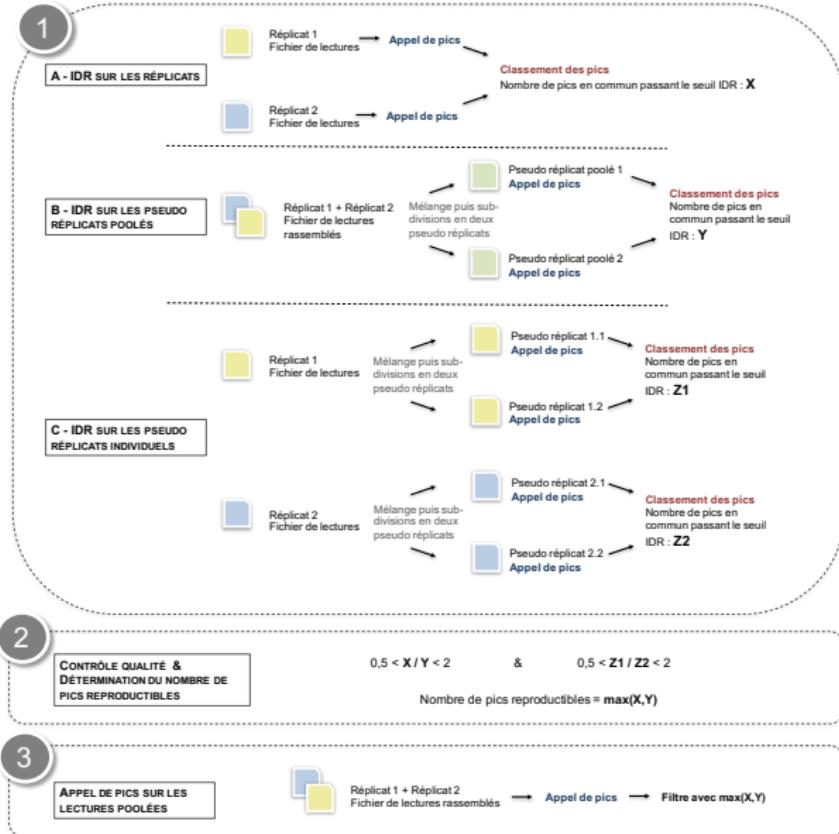
L'IGBMC

Introduction et objectif

Présentation de l'outil

Analyse du projet
PPAR γ -
RXR α

Conclusion et perspectives



Gestion des réplicats

L'IGBMC

Introduction et objectif

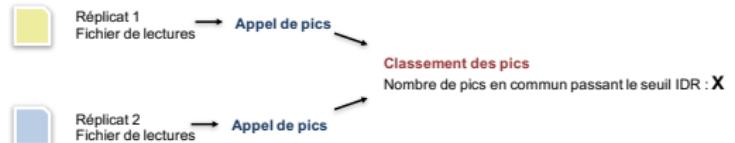
Présentation de l'outil

Analyse du projet
PPAR γ -RXR α

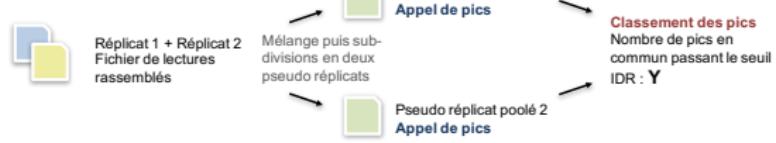
Conclusion et perspectives

1

A - IDR SUR LES RÉPLICATS



B - IDR SUR LES PSEUDO RÉPLICATS POOLÉS



C - IDR SUR LES PSEUDO RÉPLICATS INDIVIDUELS



Gestion des réplicats

L'IGBMC

Introduction
et objectif

Présentation
de l'outil

Analyse du
projet
PPAR γ -
RXR α

Conclusion
et
perspectives

2

CONTROLE QUALITÉ &
DÉTERMINATION DU NOMBRE DE
PICS REPRODUCTIBLES

$0,5 < X / Y < 2$ & $0,5 < Z1 / Z2 < 2$

Nombre de pics reproductibles = $\max(X, Y)$

3

APPEL DE PICS SUR LES
LECTURES POOLÉES



Réplicat 1 + Réplicat 2
Fichier de lectures rassemblés → Appel de pics → Filtre avec $\max(X, Y)$

