



CENTRE UNIVERSITAIRE DE
FORMATION CONTINUE
ET PAR ALTERNANCE



RÉGION
NORMANDIE



UFR Sciences
et Techniques



Normandie Université



Université de Rouen
Master 2.1 BioInforMatique mi-parcours - Année 2015-2016

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

Stéphanie Levon

IGBMC - Plateforme GenomEast
Encadrante : Stéphanie Le Gras, IE



UNIVERSITÉ DE STRASBOURG



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

Introduction
et objectif

Cahier des
charges

Architecture
du pipeline

Analyse du
projet Ikaros

Conclusion

Perspectives
Planification

- 1 Introduction et objectif
- 2 Cahier des charges
- 3 Architecture du pipeline
- 4 Analyse du projet Ikaros
- 5 Conclusion
- 6 Perspectives
- 7 Planification



ChIP-seq : applications en biologie

Introduction
et objectif

Cahier des
charges

Architecture
du pipeline

Analyse du
projet Ikaros

Conclusion

Perspectives

Planification

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

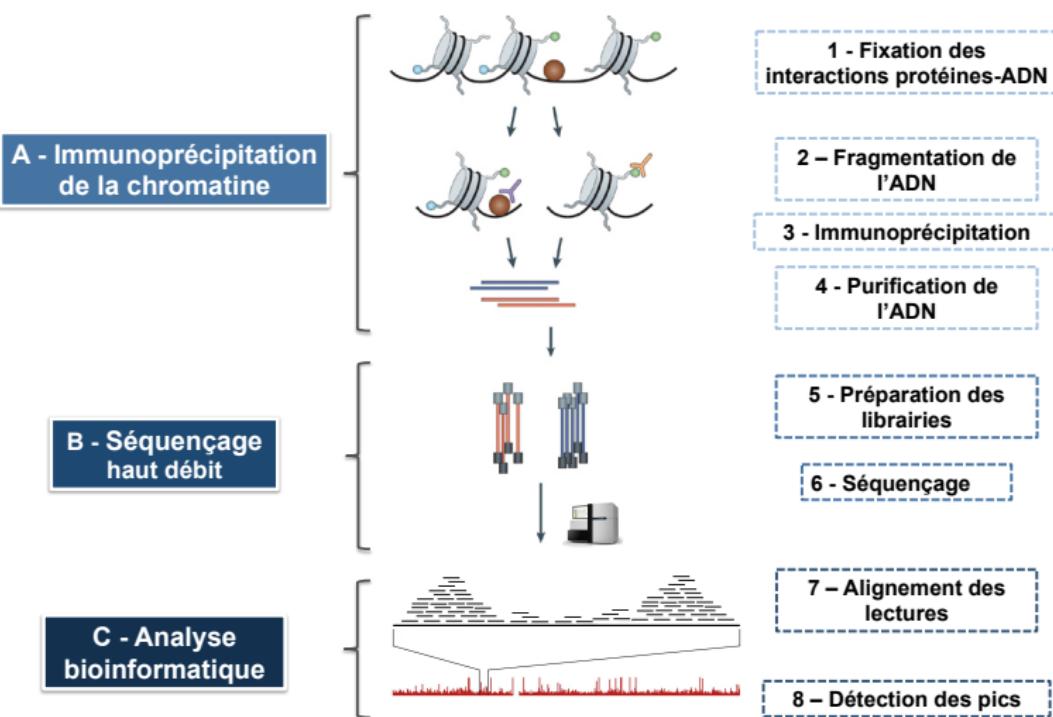
- Déetecter les positions d'ancrage des protéines de liaison à l'ADN
- Déetecter les positions des histones avec ou sans modifications post-traductionnelles

→ Mécanismes de régulation de la transcription



Le ChIP-seq, immunoprécipitation puis séquençage

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



Objectifs de ma mission

Introduction
et objectif

Cahier des
charges

Architecture
du pipeline

Analyse du
projet Ikaros

Conclusion

Perspectives

Planification

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



Conditions à respecter pour l'implémentation

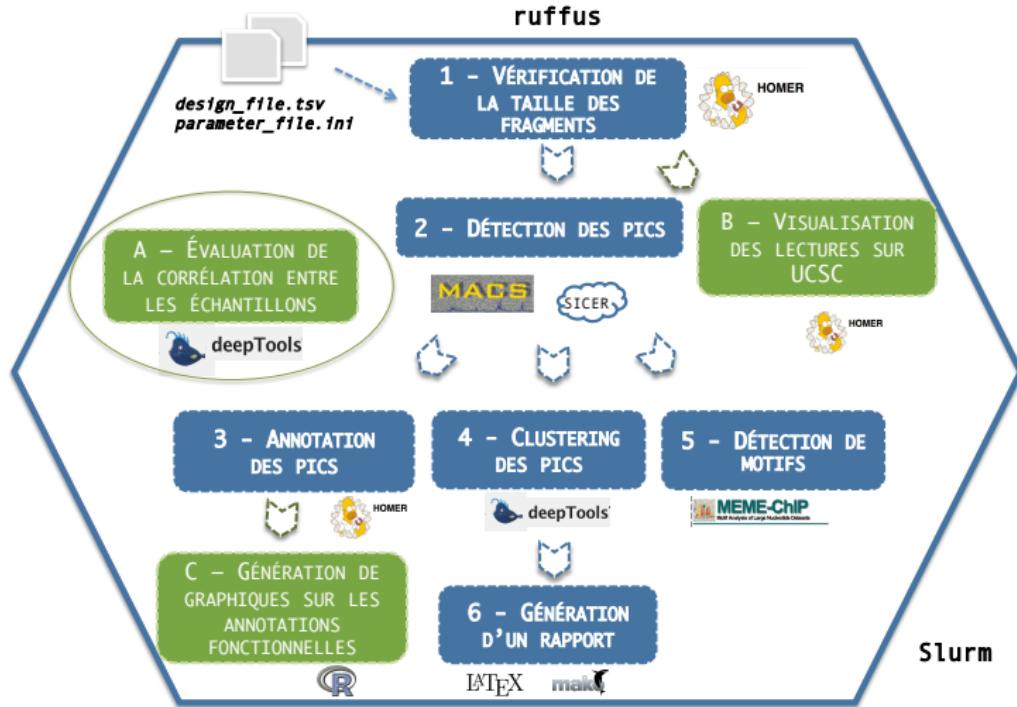
Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification

- Codé en Python avec des scripts R
Respect des bonnes pratiques de programmation
- Construit avec un outil d'aide à la création de pipeline
- Versionné
- Compatible avec le gestionnaire de file de la plateforme Slurm
- Configuré à l'aide de deux fichiers de paramètres
- Implémente les outils déjà utilisés par les bioinformaticiens de la plateforme



Étapes implémentées

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



Fichiers de configuration

Design de l'expérience

Introduction et objectif

Cahier des charges

Architecture du pipeline

Analyse du projet Ikaros

Conclusion

Perspectives

Planification

design_file.tsv

- Quels sont les échantillons à analyser ?
- Quel contrôle pour quelle immunoprécipitation ?
- Paramètres ?

Identifiant de l'échantillon	Nom de l'échantillon	Type cellulaire	Réplicat	Fichier de lectures bam/bed	Identifiant du contrôle	Nom du contrôle	Fichier de lectures bam/bed contrôle	Outil de détection des pics
TLRZ16	J0_IKAROS	CD4	1	TLRZ16.sort.bed.gz	TLRZ22	J0_Input	TLRZ22.sort.bed.gz	MACS
TLRZ18	J0_H3K4m��1	CD4	1	TLRZ18.sort.bed.gz	TLRZ22	J0_Input	TLRZ22.sort.bed.gz	SICER
TLRZ20	J0_H3K27ac	CD4	1	TLRZ20.sort.bed.gz	TLRZ22	J0_Input	TLRZ22.sort.bed.gz	SICER



Fichiers de configuration

Paramètres

Introduction et objectif

Cahier des charges

Architecture du pipeline

Analyse du projet Ikaros

Conclusion

Perspectives

Planification

parameter_file.ini

- Version et chemin des outils
- Numéro/génomé de référence du projet
- Paramètres de chaque outil

```
[SCHEDULER]
scheduler = SLURM

[TOOLS_VERSION]
python_version = 2.7.5
homer_version = 4.4
MACS_version = 1.4.2
SICER_version = 1.1
MEME_version = 4.10
deepTools_version = 2.2.4

[PROJECT INFORMATIONS]
project_number = S15253
working_dir = /ifs/illumina/levon/2/ChIP-seq/ruffus/S15253/160720

[GENOME_INFORMATIONS]
organism = Mouse
genome_assembly = mm9

[ASSOCIATED_FILES]
design_file = /ifs/illumina/levon/2/ChIP-seq/ruffus/S15253/160720/design_file

[MEME_PARAM]
p_meme-nmotifs = 20
p_meme-minw = 8
p_meme-maxw = 25
```

Informations générales

Informations relatives au projet

Paramètres pour l'outil MEME-CHIP

Présentation du projet Ikaros - Dr. Philippe Kastner

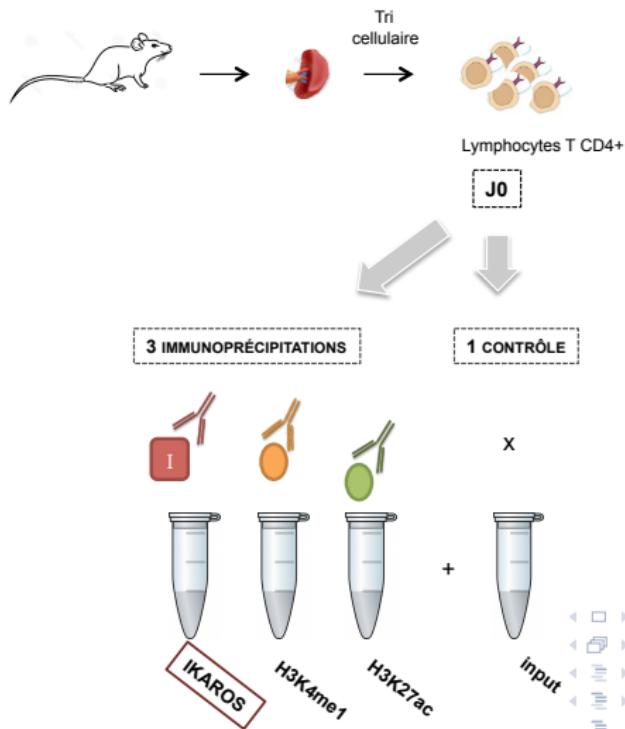
Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification

Hématopoïèse et leucémogenèse chez la souris

Facteur de transcription Ikaros

Exprimé dans les LT CD4+

- Quelles sont les régions en interaction avec la protéine Ikaros ?
- Quels sont les *enhancers* actifs ?



Évaluation de la taille des fragments d'ADN

Homer

Introduction et objectif

Cahier des charges

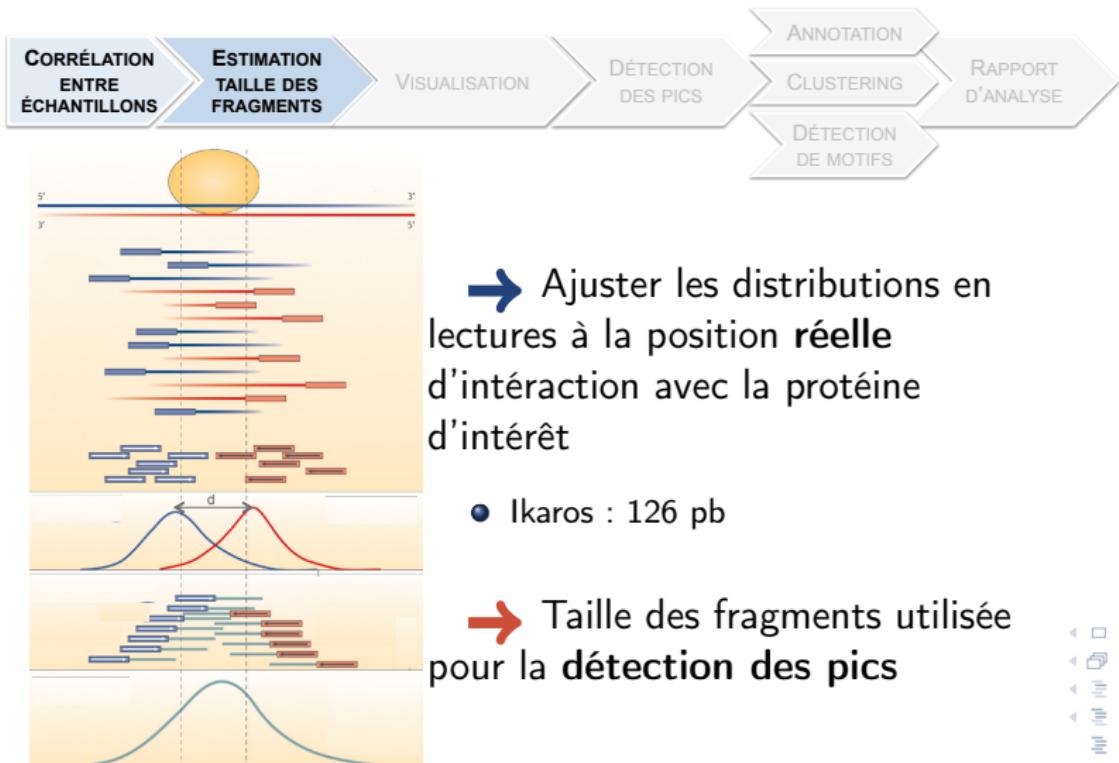
Architecture du pipeline

Analyse du projet Ikaros

Conclusion

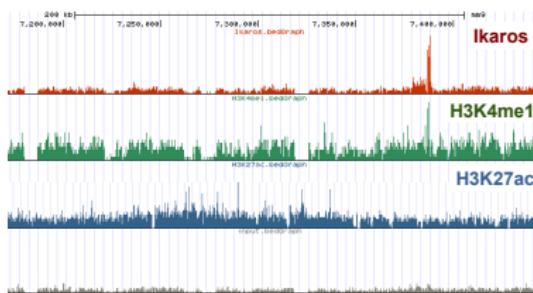
Perspectives

Planification



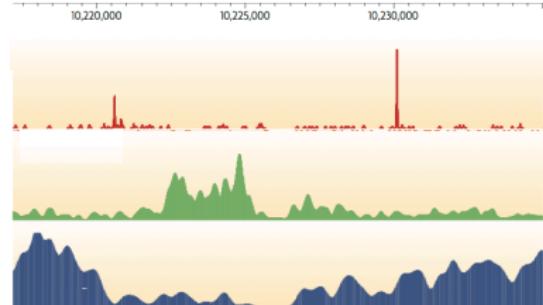
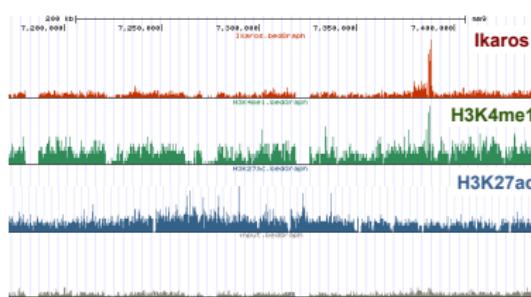
Visualisation des lectures via UCSC

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



Visualisation des lectures via UCSC

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



→ 2 profils de pics différents



Détection des pics

MACS et SICER

Introduction
et objectif

Cahier des
charges

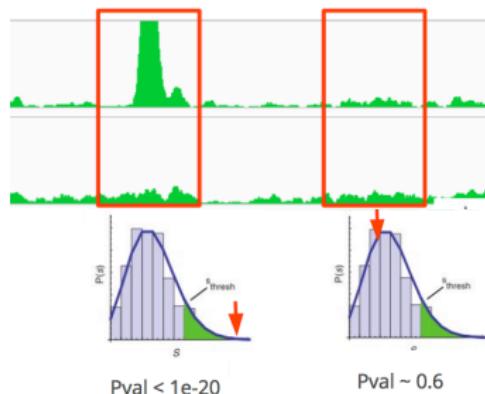
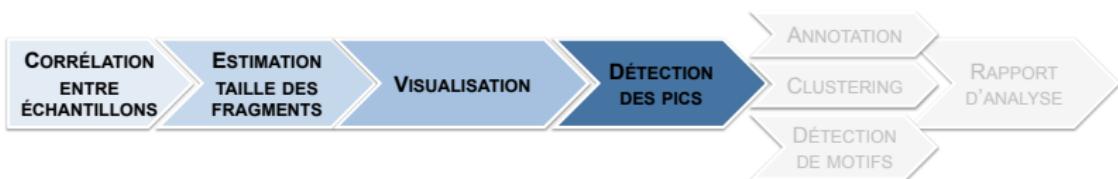
Architecture
du pipeline

Analyse du
projet Ikaros

Conclusion

Perspectives

Planification



- Partitionnement du génome
- Évaluation de l'enrichissement du signal
- Calcul d'une p-value et comparaison avec la valeur seuil de significativité d'un pic
- Correction des tests multiples (FDR)

Annotations des pics

Homer

Introduction et objectif

Cahier des charges

Architecture du pipeline

Analyse du projet Ikaros

Conclusion

Perspectives

Planification



→ HOMER recherche le site d'initiation de la transcription le plus proche

- Génération d'un histogramme

→ HOMER recherche l'annotation fonctionnelle de la région occupée par le centre du pic : 5'UTR, TSS, 3'UTR, TTS, exons, introns, régions intergéniques, mi/nc/sno/rRNA, pseudogènes

- Génération d'un diagramme en secteur



Annotations des pics

Homer

Introduction et objectif

Cahier des charges

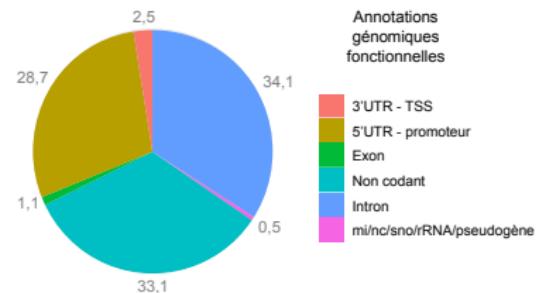
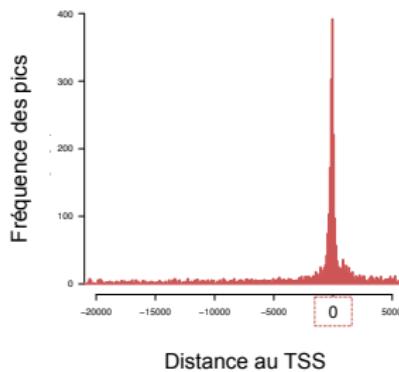
Architecture du pipeline

Analyse du projet Ikaros

Conclusion

Perspectives

Planification



→ Permet d'observer la distribution des fréquences des distances "pics - TSS"

→ Permet d'observer la répartition fonctionnelle des pics détectés

Clustering des régions génomiques

deepTools

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



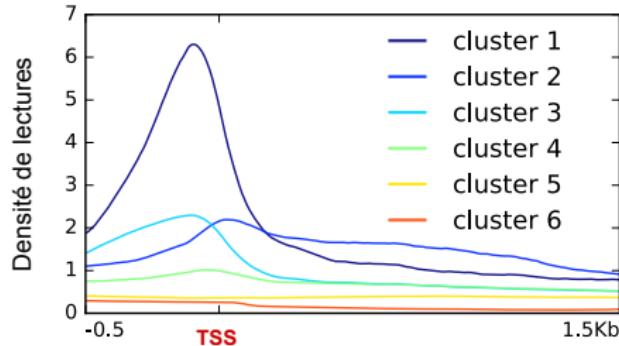
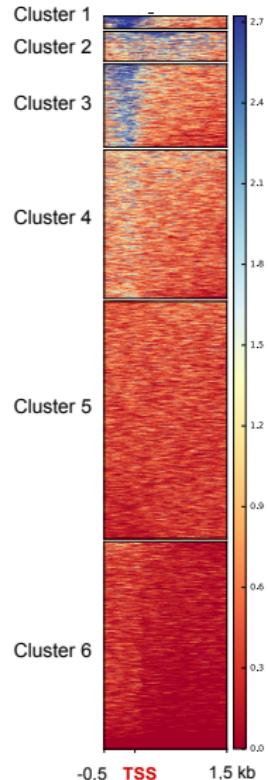
But : Avoir une vision globale des densités en lectures par rapport à une liste de régions d'intérêts



Clustering des régions génomiques

deepTools

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



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

Détection de motifs MEME-ChIP

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



→ Découverte de séquences nucléiques retrouvées fréquemment au sein des pics

Motif	Programme	E-value	Motifs similaires	Distribution
	DREME	2,9.10 ⁻²⁰⁶	FEV ELF5 FLI1	Motif non centré

- MEME et DREME : découverte de motifs
- Tomtom : recherche de motifs similaires en bases de données
- CentriMo : distribution des motifs au sein des pics
- SpaMo : recherche de motifs co-localisés ou *clusters* de motifs



Génération d'un rapport automatique

Librairie Mako, knitr, Latex

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



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



Conclusion

Introduction et objectif

Cahier des charges

Architecture du pipeline

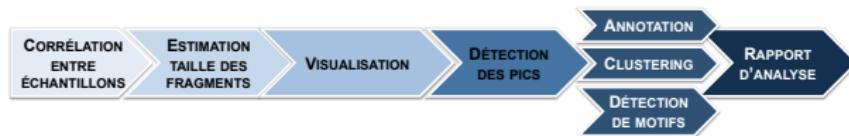
Analyse du projet Ikaros

Conclusion

Perspectives

Planification

- ChIP-seq : *Design* expérimental, étapes ou outil d'analyses
- Apprentissage d'un nouveau langage - Python
- Utilisation d'un gestionnaire de file - Slurm



- Standardisation et automatisation de l'analyse de données de ChIP-seq
- Validation du pipeline par comparaison avec les résultats obtenus par Tao Ye



Perspectives

Introduction
et objectif

Cahier des
charges

Architecture
du pipeline

Analyse du
projet Ikaros

Conclusion

Perspectives

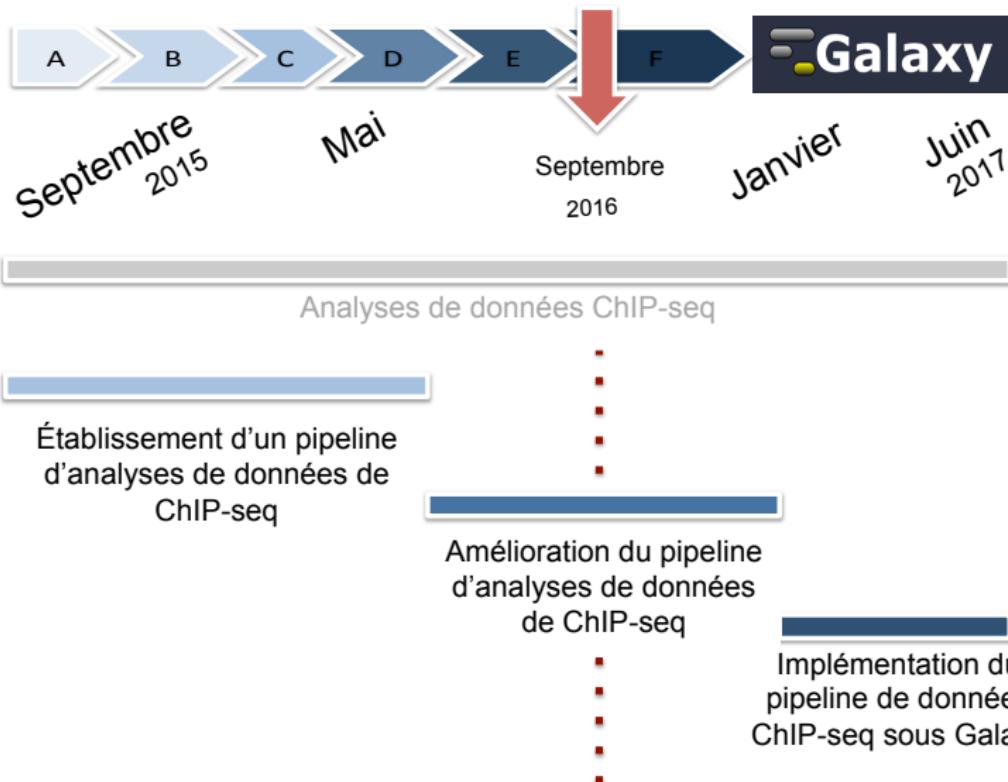
Planification

- Gestion des différents plans expérimentaux
 - Gestion des réplicats
 - Analyse différentielle
- Implémentation sous Galaxy



Planification

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification



Remerciements

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification

Christelle Thibaut-Carpentier
Céline Keime
Stéphanie Le Gras
Matthieu Jung
Tao Ye
Constance Vagne
Philippe Kastner



ainsi que l'ensemble des membres de la plateforme GenomEast

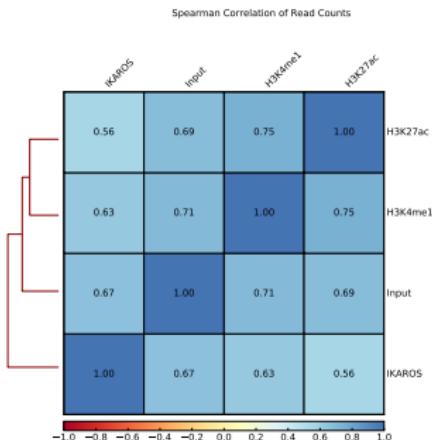
Merci de votre attention

Évaluation de la corrélation entre échantillon deepTools

- Introduction et objectif
- Cahier des charges
- Architecture du pipeline
- Analyse du projet Ikaros
- Conclusion
- Perspectives
- Planification



→ Évaluer la reproductibilité d'une expérience



Visualisation des lectures via UCSC

Fichiers générés avec HOMER

- Introduction et objectif
- Cahier des charges
- Architecture du pipeline
- Analyse du projet Ikaros
- Conclusion
- Perspectives
- Planification



Chargement des bedGraphs



Création d'un trackhub à partir des bigWig



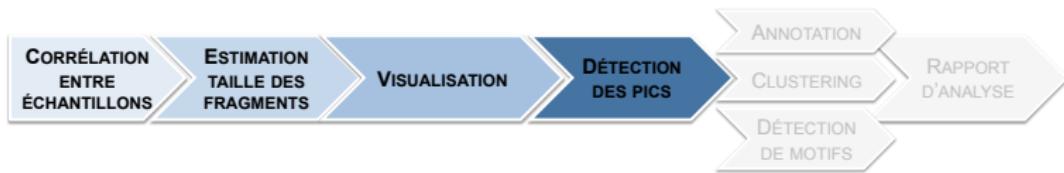
→ Même résultats de visualisation



Détection des pics

MACS

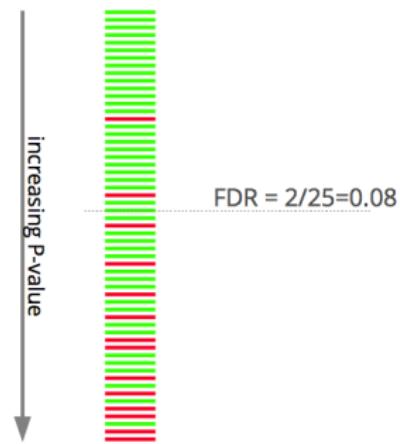
- Introduction et objectif
- Cahier des charges
- Architecture du pipeline
- Analyse du projet Ikaros
- Conclusion
- Perspectives
- Planification



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

Introduction et objectif
Cahier des charges
Architecture du pipeline
Analyse du projet Ikaros
Conclusion
Perspectives
Planification

