Université Claude Bernard Lyon 1

Rapport de Projet

Optimisation de l'analyse de données de ChIP-seq pour des facteurs se liant à des régions répétées du génome comme les ADNt

BOYER Thomas, GALVIS-LASCROUX Johanna, LOHMANN Eugénie
et N'GUYEN Emilien

17 janvier 2019

MAÎTRES D'OUVRAGE : Vincent VANOOSTHUYSE
et Laurent MODOLO
TUTEUR PÉDAGOGIQUE : Philippe VEBER

Table des matières

Résumé

Rapport de Projet.
Ce travail a été mené en collaboration avec le LMBC *: Laboratoire de Biologie et Modélisation de la Cellule.*

1 Introduction

Concernant le positionnement de la condensine dans le génome, des études telles que celle de Sutani et al., 2015[2] et Iwasaki et al.,2014,[1] montrent, chez *S pombe*, qu'il existe un enrichissement de cette protéine à proximité des régions fortement transcrites (donc généralement autour des régions transcrites par les ARN polymerases I, II et III).

S'intéressant à la distribution de la condensine chez S.pombe, L'équipe de l'ENS a essayé de reproduire les résultats obtenus par les auteurs cités, en analysant ses données.

La présence de la condensine à proximité des gènes transcrits par les ARN polymerases I et II a bien été confirmé, mais pour les séquences transcrites par l'ARN pol III, ils trouvent que les pics observées sont aspecifiques (présence de pics dans les controles négatifs comme dans les Input). La raison de cette aspécificité peut être expliquée par le fait que les gènes transcrits par l'ARN polIII sont des séquences répétées représentées par les gènes codant pour les ARNt : les ADNt.

2 Objectif

L'objectif de ce projet est la réanalyse de données ChIP-seq en prenant en compte la nature répétée des ARNt. Pour ce faire, nous avons développé un pipeline dédié à l'analyse des séquences répétées d'ARNt.
Notre approche consiste, dans un premier temps, à définir des consensus de familles de séquences d'ADN génomique codant pour les ARNt, avec ou sans régions flanquantes, par alignement multiple et étude phylogénique. Certains gènes codant pour les ARNt sont très proches les uns des autres (l'ADN espaceur est très petit). De plus, les régions flanquantes autour des ARNt sont parfois identiques entre plusieurs ARNt. La problématique est donc d'établir un programme permettant d'incrémenter les régions flanquantes des ARNt en comparant leurs séquences pour s'assurer que ces régions flanquantes sont bien différentes ou pour déterminer sur quelle longueur elles sont identiques tout en faisant attention à ne pas atteindre la séquence de l'ARNt précédent/suivant.
Dans un second temps, deux protocoles d'analyse sur les données publiques de ChIP-Seq ont été construits : une standard pour la identification et correction rapide de gros artefacts, et l'autre, constitué par des pipelines optimisés dédiés à l'analyse des lectures comprenant ces gènes ADNt. Nous avons produit, selon différentes scénarios possibles, 3 pipelines fonctionnels, tous offrant l'option de suppression des duplicats.
Sachant que dans les premières analyses, les pics observés aux ADNt sont centrés sur la séquence même de l'ADNt, l'option de supprimer les séquences des ADNt du génome de référence -en laissant les régions flanquantes- a été considéré en tant que stratégie pour empêcher la obtention de pics aspécifiques. Aucun artefact de ce type n'ayant été détecté avec les 3 différentes pipelines (même sans utiliser la stratégie de manipulation du génome), nous confirmons que les résultats observés auparavant sont bien des reliquats d'analyse bioinformatique.

3 Déroulement du projet

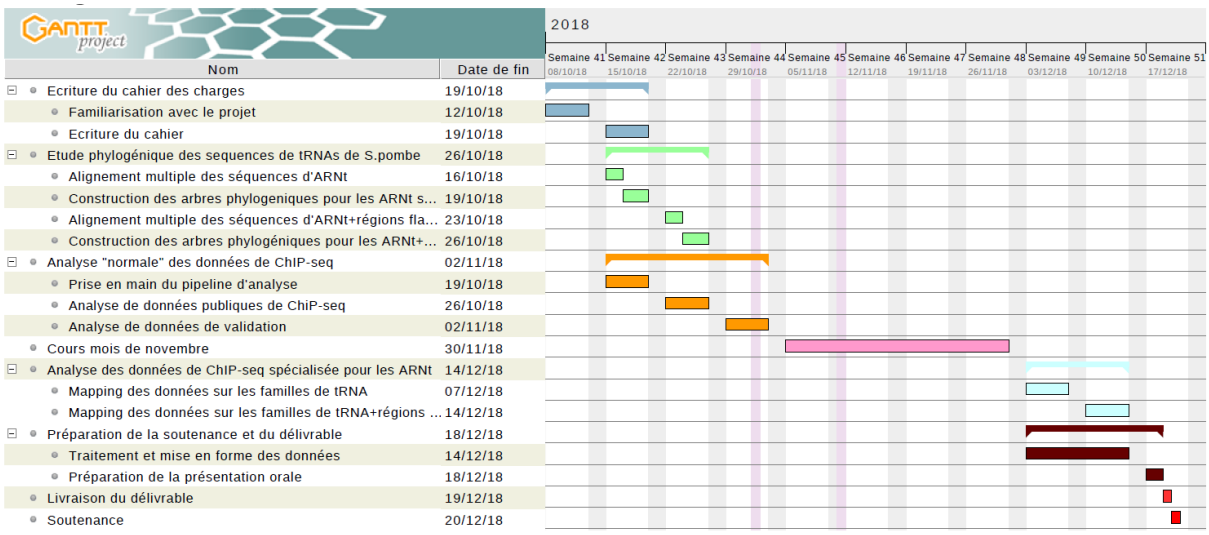

FIGURE 1 – Répartition des tâches dans le temps imparti du Projet.

3.1 Analyse Phylogénétique des séquences répétées correspondant aux gènes ADNt

Les gènes codant pour les ARNt (ADNt) sont des séquences répétées qui sont réparties sur les 3 chromosomes et l'ADN mitochondrial de S. Pombe. Au sein de notre contexte d'analyse, les données issues du ChIP-seq incluent uniquement des ADNt nucléaires, pour un total de 171 gènes. Suivant les recommandations de nos maîtres d'ouvrage, nous avons commencé par définir les familles de ces gènes et identifier des régions qui pourraient potentiellement aider à construire un protocole de *mapping* plus spécifique pour les régions ciblées, tel que les régions flanquantes. Les phylogénies ont été construites et les séquences flanquantes ont été analysées en trouvant les résultats suivants :

3.1.1 Familles des gènes ADNt nucléaires

L'alignement multiple des 171 gènes (représentant 20 familles) avec T-Coffee v.11.00.8cbe486 (algorithme MSA et construction des arbres par la méthode Neighbour-Joining) a permis d'identifier 3 groupes principales de gènes gardant des distances de 1,64, 0,98 et 0,92 par rapport au noeud supérieur, respectivement :

Groupe 1 : Trp, Cys, Thr, Met, Arg, Ser, Leu, Ile, Asn, Lys, Tyr, Asp, Phe
Groupe 2 : Gln
Groupe 3 : Gly, Pro, Val, His, Glu, Ala .

Cependant, sur l'arbre de nos clusters, quelques gènes ne se regroupent pas avec sa famille, ce sont par exemple le cas de SPATRNAILE.02, SPATRNALEU.03 et SPATRNAARG.01, tous se localisant en *cis* sur le chromosome I.
Les séquences consensus établies pour les gènes ADNt (gènes "purs" soit, sans régions flanquantes) sont disponibles dans le fichier 'consensusADNtpropres.csv', mettant en évidence la haute identité des séquences intra-familles (100 %), ce qui augmente le risque d'artefacts pendant l'analyse bioinformatique (mal-alignements, faux *peaks*). Par conséquent, sous l'hypothèse que les régions flan-

quantes peuvent être à certain degré polymorphiques pour être exploitables dans notre pipeline, ces régions ont été aussi étudiées.

3.1.2 Analyse des régions flanquantes

La distance minimale possible parmi deux gènes a été calculé trouvant deux clusters de 2 gènes chacun sur le chromosome I avec une distance inter-gènes de 175 nucléotides. Ainsi la longueur des régions flanquantes (en amont et en aval) pour notre analyse a été fixé en 85 nucléotides.

Les analyses de cluster("phylogénie") ont mis en évidence des groupes, avec une hiérarchie moins claire que dans la "phylogénie" des séquences "purement" codantes, mais permettant encore d'identifier les mêmes familles des gènes codant pour les ARNt.

Les alignements ont permis d'identifier des séquences se regroupant par leur similarité : 18 en amont et 16 en aval, lesquelles montrent une identité intra-familles de 100 % pour ces groupes (Tables 1 et 2). Il est important de noter que la région en amont inclut la *TATAbox* tandis que la région en aval contient la séquence *terminator* riche en Timine (Figure 2).

Gènes (Upstream region)	**Séquence Consensus**
>Consensus-ALA04/ALA05	>AAGCTTCAAAAAATATTAATATTGAGTCTAAAATCAAGTTATTAATGTA TATATTATAATATTCTTACGCTAAAATAAACAATTA
>Consensus-ALA08/ALA09/ALA10/ALA11/ALA12-II	>GGTAGTATCAAAGTGCGTTGGAAAAGAATAATAGCAAATAGCAAAGA AAGTAATAATCGACTAACATAGAGTAAAAATCAACCAT
>Consensus-ARG12/ARG13-III	>TGGGTGCTCTCCACCATTGCTAATTTAAAATTTTTGCCAATAACAGATTT GTACAAATAGTTTTGTTTGAATGCATGTATTAACA
>Consensus-ASP06/ASP07-III	>AAATATATTAAAGATAGTTATTTAATAGAGGTTTAAATATAGTTAGTCTT ATAATTAGGTTAGTTGATGGAGAAAATCAAACAAA
>Consensus-GLU03/GLU04-I	>TTCAAAATAAAAAGTAGCTTATAAAAAAATTGTTTATTTAATTAAATATA CGTTAAATACATATAAACATTGTGTGAACTAAGCA
>Consensus-GLU06/GLU07-II	>TCTCANATATCTTATTTAAAGCAAAATTTAAAATTGATTTGCTAATAGCT CTATTATCTCTAATTTATTAGGGAAGTCATCGAAA
>Consensus-GLY07/GLY08-II	>TAATTATACGCATTTTTTGACAATGCATATTGTAGAATTATCAACTATTT TATATAATAATCAATTTTTGAGATTATATCATTAA
>Consensus-ILE03/ILE04-I	>AGAGTGGTGTCCATCGAAAATATAAACAACAGTTTATTTTAAATTTGTAC ATAAATATGTTAAACCAAAAAAGATTAACAAAAAT
>Consensus-ILE06/ILE07-II	>TATTGCGAAAAAATAAATATTTTATTAATACATCTTTGTGACATCAGAAG TTATTATAATTTATTTTACTCGTGCCATCAATAAT
>Consensus-LEU06/LEU07-II	>CATTTAAATTACTCTTAAAACTTCTTTAACAATTCTTTTTTATTTTATGA TCAAATATTTAATTAAGTAAACTAACTCAACAACA
>Consensus-LEU12/LEU13-III	>ATCAGAATAATAGGTGAGCTTAAATCAAATTATTTTTTGCGAAGTAAG AATAATATGCTTATATATGACACACGAATAACAACAT
>Consensus-LYS07/LYS08-II	>ACTCAATCGTAGCTTATCAACTGCCCCATAATTTTACTTTTTCCTAAA TATATTGATAAGGTATAATTTTAAACTTAACAACGCG
>Consensus-LYS10/LYS11-III	>TTGACTTCTTTCTATGACGTAATATGTGTCCTAAAAAAAGGTAATAAA TATAAATTGTTCTATGTATACTCTCAGCTAAAACCCG
>Consensus-THR08/THR09-III	>GAAGTGCAGACTTATGACCATTGATCATGTTTCCAGGATTATGCTTTTAA ATATAGATAACGTTAATTTACGAAGACAAACAGCA
>Consensus-VAL06/VAL07-II	>TGTATTCGTAACTTCAGACAATTTGACGTTAAGGTTCATTTATAAACCTA TATTACAAATAAAGCTAATAATAAAACCAACATCT
>Consensus-VAL09/10-III	>GCAGTAGACGGTTCACTTCTTTTAATAAATTTTCTAATTTTACACAACT TGAATTAATTTGCATAGATTTTTTATTACTCAACGA

TABLE 1 – Séquences consensus des régions flanquantes en amont

Gènes (Downstream region)	**Séquence Consensus**
>Consensus-ALA04/ALA05-I	>AATTTTTTATCAATTTTTTTGAAATCTCTAAATTCATGAGAATTAATGTGTTTGCCATCTTACAATTCAATAAATACATTCTTTA
>Consensus-ALA08/ALA09/ALA10/ALA11-II	>CTTTTTGTTTATTTTTTTTTTGAAATTTACTGATGTATAGTGTACTGTGAGATTCATTAAAAAATAAAAAACATGGTCACAGCCG
>Consensus-ARG12/ARG13/III	>ATCTTTTTTGCACTAAATCAAAATTTTTGGTATTTTGGTTAAGTAAAGTCACAAAACAAAAACATGTGGAAAAATCGTAAAAGAA
>Consensus-ASN06/ASN07-III	>GATTCTTTTACGATTTTTCCACATGTTTTTGTTTTGTGACTTTACTTAACCAAAATACCAAAAATTTTGATTTAGTGCAAAAAAG
>Consensus-GLU03/GLU04-I	>AAGTTTTATTAATTTATTTTTAACGTTTGTCTTTATATAAAGTAATCTCTTTTTTTTTTTTTTTTTAAGAAAAAATATTTTATT
>Consensus-GLU06/GLU07-II	>TTTTTTTTAAAGGACTTTTATTTTTATCATCGCTAATACTATATCGTGTAGATTAAATTAAATGGAGCGGATTTTTAACAATGGT
>Consensus-GLY07/GLY08-II	>TTGATTTTTTATCAATTATGCTAGTTGCTTGAGAATTATTTTACTAAACATTCATGAGGGACAATGCTTTTACTTTTGTAAACCA
>Consensus-ILE03/ILE04-I	>AATCTTTTTTCTTTTATACTTTTTTATTAAAACACGAATATTGCATATGAAAATAATATTCAATTCATAAAATAAATACTTGCTT
>Consensus-ILE05/ILE06/ILE07/ILE08-II	>TGTTTTTTATTTTTTAATGAATCTCACAGTACACTATACATCAGTAAATTTCAAAAAAAAAAATAAACAAAAAGTGGACAAGCCAG
>Consensus-LEU06/LEU07-II	>CCATTTTTGTTAATTTATGACTGTACAAATTTACAATGAACTTTTCTAAAGTTTTCAAACATAATCCTTCGCTATCATTTTAGGG
>Consensus-LEU12/LEU13-III	>CCATTTTTACTATTTTATTGTACATGTTTTTACATACAAGAGGTACTTTTATTTATTACGTCTCCGCTTTATCTGTCCTTTACTT
>Consensus-LYS07/LYS08-II	>GTTAATTTTTCTTTTTATTGTTGAAAAAATCATTAAATGCTAGTGTATGAAGATTTTGAAAACTGTTTAAAAATATATATACACC
>Consensus-LYS10/LYS11-III	>TTTAGTAATCCTAGTATTGGGATTTAAATTATTTTTTATTCTGATTGAATAGTTCTAAAACCTAACGCTTTTTAACAATTTTTAT
>Consensus-THR08/THR09-III	>ATTTTTTTCATTTTTAATTTATTATACATATCATAATTATAAATGTAATAAATTATGATAAAAAGATGATCGTGCCAAGACTCGA
>Consensus-TYR02/TYR03-II	>AAATTTTTTATATCTTAAAAATATACTTTGTACTAGAGTCCATTGTTTAAAAACCCGCTGTCTTACCCTAAAATGATAGCGAAGG
>Consensus-VAL05/VAL06/VAL07/VAL08-II	>ATCGTTTTTAAGTTTTTAATTTTTTACTTTTAAAAAAGAAATTATATCTGCTTCATTAATTAACGATTCTTTTGGAATGAAATCA
>Consensus-VAL09/VAL10-III	>TCTTTTTATCATAATTTATTACATTTATAATTATGATATGTATAATAAATTAAAAATGAAAAAAATTGCTCCAGCAGTGACTTGA

TABLE 2 – Séquences consensus des régions flanquantes en aval

Étant donné la présence de ces éléments répétés et les hautes identités intra-familles montrant la présence de 'blocs' de séquences identiques (dans toute sa longueur y compris les régions flanquantes) Cela nous mène à considérer la possibilité d'erreurs d'assemblage : en effet, même les méthodes les plus puissantes, tel que les graphes de De Bruijn, rencontrent des difficultés pour calculer l'emplacement correct des séquences répétées dans les génomes eucaryotes.

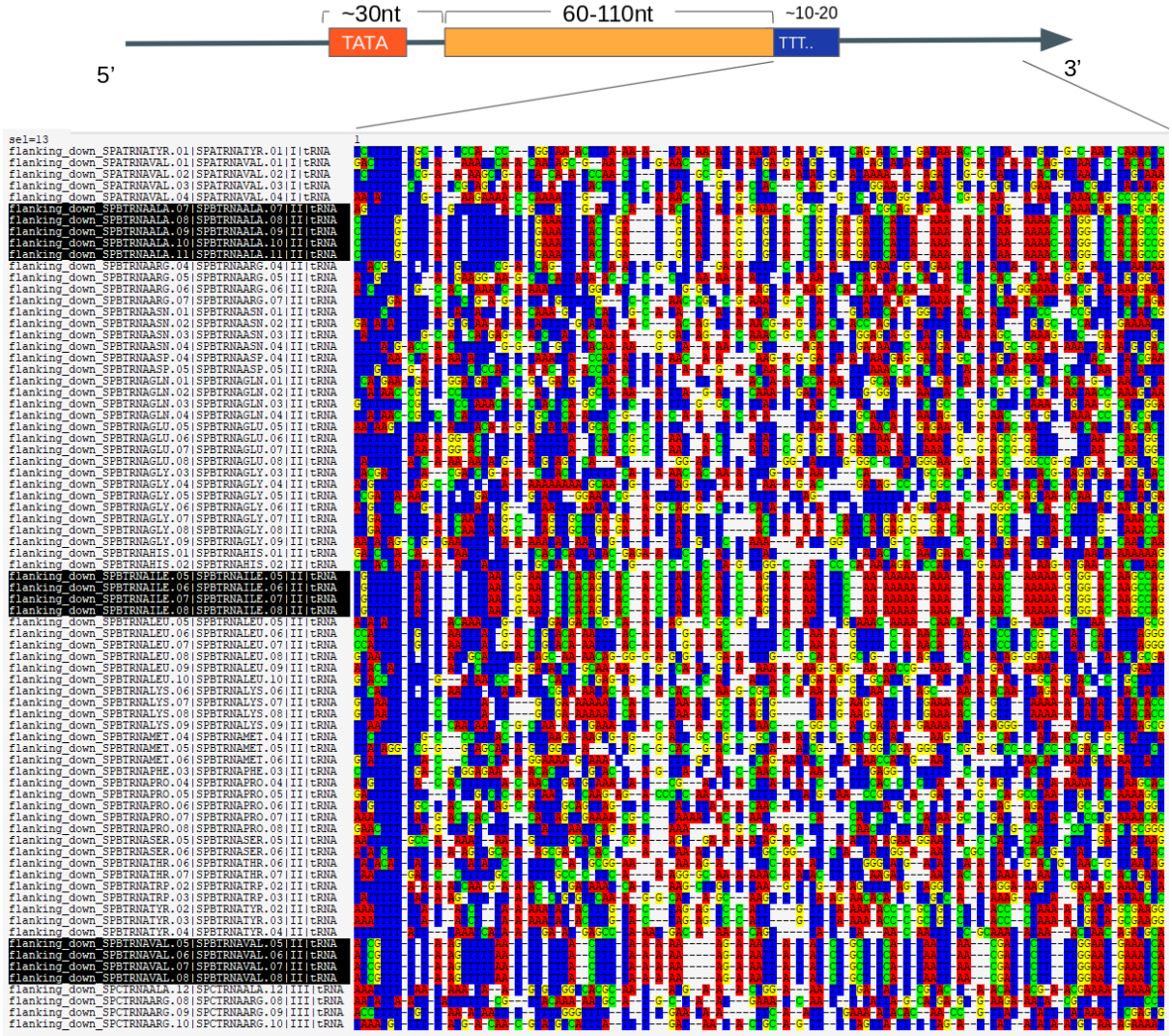

FIGURE 2 – Alignement des régions flanquantes *downstream*. En bleue la séquence *terminator*. Les 'blocs' des familles Ala,Ile et Val signalés en noir

3.2 Analyse standard des lectures et Assurance de la qualité

Deux protocoles ont été établis pour certifier la qualité du travail : une analyse standard initiale, et ensuite, l'analyse dédiée aux séquences répétées en considérant les caractéristiques particulières des données en entrée et la suppression de duplicats.

L'analyse standard construit avec le gestionnaire de pipeline Nextflow utilise Bowtie2 à l'étape du mapping : étant donnée la longueur des reads (35 nucléotides les reads de Iwasaki[1], 50 nucléotides les reads de Sutani[2]) et l'objectif de ce travail, un aligneur tel que BWA n'est pas indiqué. Les reads initialement analysés -suivant le planning prévu au début de ce projet- sont ceux du travail de Iwasaki car le contrôle positif ARNPolIII est disponible sur ce jeux de données et la technologie utilisé est Illumina, facilitant alors une analyse initiale rapide avec les paramètres par défaut Illumina pour ChIP-seq qui nous sont plus familiés.

Cependant, les Input pour ARNPolIII ne sont pas disponibles sur SRA Run Selector ni sur le papier même. Les input correspondent aux fragments d'ADN récupérés après élution en présence de billes sans anticorps, permettant de mesurer le bruit du fond d'une expérience ChIP-seq et ils sont par conséquent indispensables pour garantir un contrôle de qualité optimal de toute analyse bioinforma-

tique. Pour garantir la qualité du travail demandé, nous avons continué la phase de développement des pipelines optimisés en utilisant les lectures du travail de Sutani.

type	code d'accès SRA	séquence ciblé	référence
échantillon	SRR1557175	Cut14 (sous-unité SMC2 de la condensine) Mitosis 1	Sutani et al.
Input	SRR1557176	Cut14 Input Mitosis 1	
échantillon	SRR1557178	Cut14 Mitosis 2	
Input	SRR1559300	Cut14 Input Mitosis 2	
contrôle -	SRR1559301	no-tag	
contrôle -	SRR1564296	no-tag Input	
contrôle -	SRR1564297	NLS-GFP-PK_ChIP_mitosis	
contrôle -	SRR1564298	NLS-GFP-PK_ChIP_input_mitosis	
contrôle +	SRR1541127	Rpc25 (sous-unité rpc8 ARNPol III)	Iwasaki et al.
contrôle -	SRR1541129	no-tag Myc pour Rpc25_Brf1	

TABLE 3 – Lectures *single-end* sélectionnées pour le développement des pipelines.
Les lectures du travail de *Iwasaki et al.*[1] ont servi pour l'analyse standard. Celles du travail de *Sutani et al.*[2] ont été choisies pour l'optimisation, étant les six premiers issues du séquençage ABI-SOLiD.

3.3 Pipelines Optimisés pour l'analyse des séquences répétés issus de ChIP-Seq

3.3.1 Documentation

Travaillant pour le LBMC, nous avons suivi les bonne pratiques en bioinformatique de ce laboratoire. Les bonnes pratique sont l'ensemble des règles régissant l'organisation du projet dans le but de faciliter la collaboration avec d'autres programmeurs. Le projet rendu sur GitLab comporte l'index suivant :

```
  project_name/
bin/
data/
doc/
results/
src/
tests/
CITATION
CONTRIBUTING
README
LICENCE
todo.txt
```

Le fichier `README` inclut les instructions pour dérouler une analyse selon les différentes scénarios possibles de l'utilisateur.

Les analyses phylogénétiques, les .py pour la manipulation des fichiers et les tables des séquences consensus se trouvent dans le répertoire `doc` ainsi que la documentation du code source.

Les résultats sauvegardés dans `results` peuvent être régénérés depuis `data`, `bin` et `src`.

3.3.2 Paramètres

Le schéma générale des trois pipelines est illustré dans la Figure 3. Les pipelines résultantes prennent en compte la possibilité que les données brutes soient issus de technologies différentes (Illumina, ABI-Solid), avec en option la possibilité de suppression de duplicats.

Étant donnée que la taille des lectures et petite (50 nt), l'algorithme Bowtie1 est le plus adapté pour l'étape d'alignement. En effet, les conditions d'analyse par le doctorant qui nous a précédés ont été reprises pour garantir la reproductibilité. Le troisième algorithme implémente Bowtie2 afin de faciliter la comparaison et la détection des potentiels biais liés à l'aligneur.
La description complète des paramètres utilisées dans chaque un des trois pipelines se trouve dans le fichier : '/doc/DOCU_FILE_ChIP-seq.md'

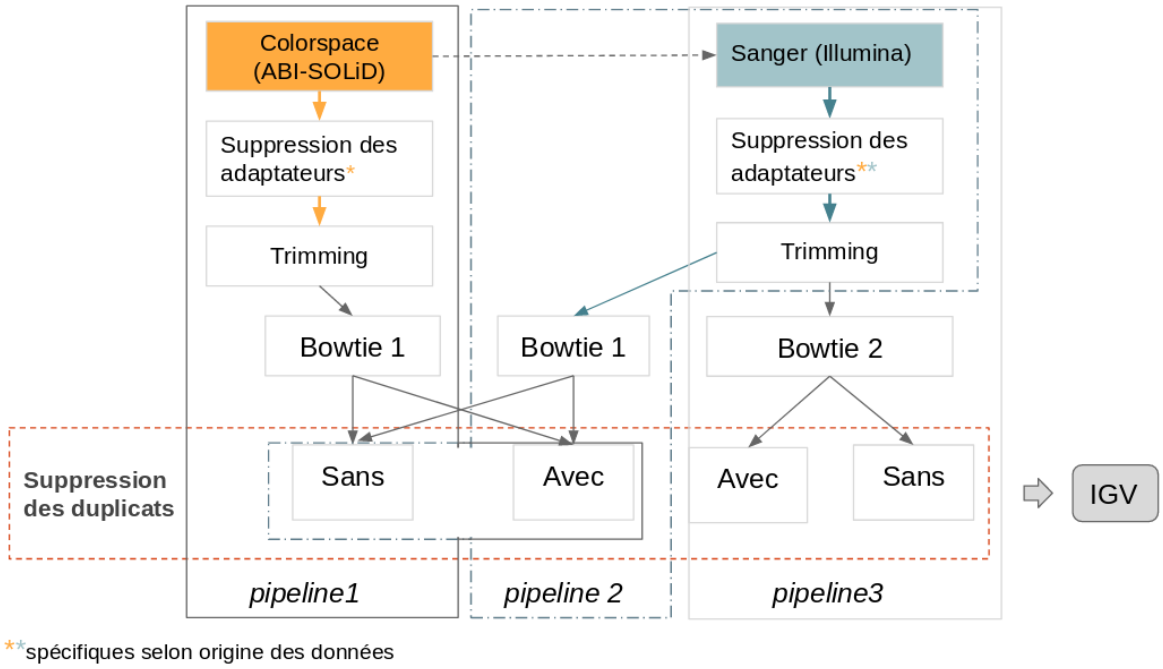

FIGURE 3 – Schéma des pipelines construits

3.4 Résultats déroulement des tests

Pour chaque pipeline, l'échantillon est en bleu, le contrôle négative (input) en vert, et en bas, la localisation des tRNAs. A gauche, les résultats avec duplicats de PCR et à droite, les résultats sans duplicats de PCR.

De manière générale, dans les résultats obtenus avec chaque pipeline, nous pouvons observer des "vrais" pics (pas de pics dans l'input) qui ne sont pas des ARNt ; notre pipeline ne semble donc pas générer de pics aspécifiques. En parallèle, on peut remarquer qu'il n'y a aucun pics sur les ARNt. Nous pouvons aussi observer la présence de nombreux réplicats soit 40 % après alignements. (Figure 4)

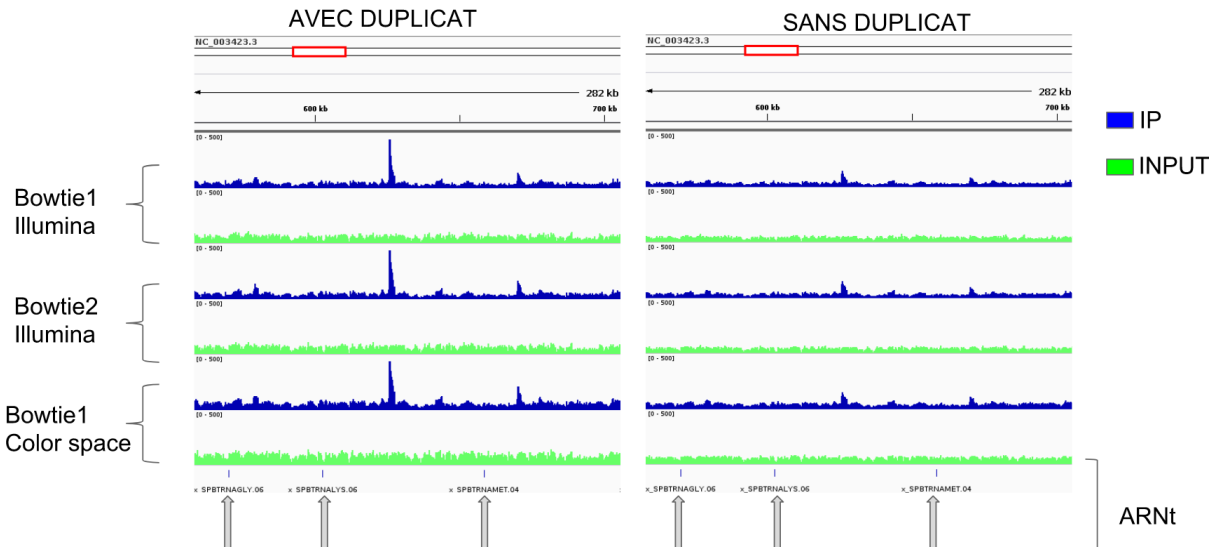

FIGURE 4 – Visualisation des résultats avec IGV

3.5 Conclusion et Perspectives

Par leurs caractéristiques, les séquences ADNt sont exploitables pour la construction des pipelines plus spécifiques, même face aux limitations et potentiels erreurs d'assemblage. Nous avons déjà réussi à réduire considérablement les occurrences d'artefacts et a démontrer que le positionnement de la condensine ne génère pas un enrichissement des lectures s'alignant sur les ADNt.

Reste à envisager la construction d'un métagène, il existe par exemple des outils avancés sous R tel que le package NMF (Non-negative Matrix Factorization) https ://cran.r-project.org/web/packages/NMF/vignettes vignette.pdf, qui exigent d'une expertise de la part du bio-informaticien.

A Annexe

page blast de *Schizosaccharomyces pombe* la levure de référence
https://www.ncbi.nlm.nih.gov/genome/14

base de données de *Schizosaccharomyces pombe* annotations du génome
https://www.pombase.org/

Accès NCBI SRA run selector https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP045414+&go=go et https://www.ncbi.nlm.nih.gov/Traces/study/?acc=+GSE60273&go=go

Références

[1] Osamu IWASAKI et al. « Interaction between TBP and Condensin Drives the Organization and Faithful Segregation of Mitotic Chromosomes ». In : *Molecular Cell* (2015). ISSN : 10974164. DOI : 10.1016/j.molcel.2015.07.007.

[2] Takashi SUTANI et al. « Condensin targets and reduces unwound DNA structures associated with transcription in mitotic chromosome condensation ». In : *Nature Communications* (2015). ISSN : 20411723. DOI : 10.1038/ncomms8815.