Analyse de données Hi-C et de promoteur capture Hi-C

Yorgo EL MOUBAYED, Anthony BAPTISTA et Aitor GONZALEZ.
Aix-Marseille Université (AMU), INSERM, Theories and Approaches of Genomic Complexity (TAGC) UMR1090, Turing Center for Living Systems - Marseille, France.

RÉSUMÉ

La conformation tridimensionnelle du génome est complexe, dynamique et cruciale pour la régulation de l'expression des gènes. Les méthodes de capture de la conformation des chromosomes couplées au séquençage à haut débit (Hi-C, capture Hi-C…) ont révélé comment l'organisation du génome est interconnectée avec l'architecture nucléaire. Ce travail se concentre sur le prétraitement des données de conformation tridimensionnelle du génome des mammifères via un *pipeline* bioinformatique : HiCUP.

The three-dimensional conformation of the genome is complex, dynamic and crucial for the regulation of gene expression. Chromosome conformation capture methods coupled with high-throughput sequencing (Hi-C, capture Hi-C...) have revealed how the organization of the genome is interconnected with the nuclear architecture. This work focuses on the pre-processing of three-dimensional conformation data of the mammalian genome via a bioinformatics pipeline: HiCUP.

INTRODUCTION

Le noyau des cellules humaines abrite 46 chromosomes. Ces chromosomes sont repliés en domaines hiérarchiques à différentes échelles génomiques. Cette hiérarchie permet un empaquetage efficace et l'organisation du génome en compartiments fonctionnels. Les chromosomes occupent des positions distinctes dans le noyau, appelées territoires chromosomiques. Ceux-ci sont divisés en compartiments chromosomiques. Au sein de ces compartiments, il existe des structures de l'ordre du mégabase en interaction. Elles sont appelées *topologically associated domains* (TADs). Les TADs renferment des structures en boucles médiées par le facteur de liaison au CCCTC (CTCF) et les cohésines.

L'organisation tridimensionnelle du génome est une réalité qui a été longtemps négligée. Cependant, elle joue un rôle fonctionnel important dans une série de processus nucléaires: activation/répression de l'expression des gènes, recombinaison, réparation de l'ADN, réplication de l'ADN, sénescence cellulaire...

Le développement des organismes repose sur l'expression différentielle des gènes. Celle-ci est, entre autres, contrôlée par des régions génomiques appelées *enhancers*. Les *enhancers* sont des sites de liaison des facteurs de transcription. Ils peuvent activer l'expression de gènes cibles par une boucle vers les promoteurs **[1]**. Les *enhancers* distants se trouvent à proximité spatiale des promoteurs qu'ils régulent, ce qui est essentiel pour un contrôle spatio-temporel adéquat de l'expression des gènes. Ceci permet de combler des lacunes considérables des distances génomiques en contournant les gènes voisins.

Les interactions promoteurs-*enhancers* sont essentielles pour le contrôle de l'expression des gènes. Elles nécessitent une régulation pour garantir une expression appropriée. Les génomes de l'homme et de la souris abritent chacun environ un million d'*enhancers*. Pour

la grande majorité de ces *enhancers*, les gènes cibles sont inconnus. L'affectation des *enhancers* à leurs gènes cibles reste un défi majeur pour déchiffrer le contrôle de l'expression des gènes chez les mammifères **[2]**.

Les progrès dans la compréhension du repliement au sein des chromosomes ont été limités par le manque d'approches permettant de cartographier les contacts chromosomiques à l'échelle du génome tout en récupérant simultanément des informations spatiales, telles que les distances moléculaires entre différentes régions génomiques. Les études de la conformation tridimensionnelle du génome se limitent à trois grands types d'approches **[3]**:

1. Les méthodes sans ligation : GAM, SPRITE, chIA-Drop.

2. L'imagerie, notamment l'hybridation *in situ* de l'ADN par fluorescence (*DNA-FISH*).

3. Les méthodes basées sur la capture de la conformation des chromosomes par ligation : 3C, 4C, 5C...

La compréhension de l'architecture tridimensionnelle du génome a été révolutionnée par l'introduction du 3C (*chromosomes conformation capture*) **[4]** et de ses variantes. La plus puissante de ces techniques est le Hi-C (*high throughput chromosome conformation capture*) qui est conçue pour identifier l'ensemble des interactions chromosomiques au sein d'une population cellulaire à l'échelle du génome entier. Le protocole expérimental du Hi-C (figure 1-A) est constitué de cinq étapes **[5]**:

1. *Crosslink* : les cellules sont traitées au formaldéhyde pour fixer les liens entre l'ADN et les protéines. Ceci a pour effet de geler l'organisation tridimensionnelle du génome dans le temps.

2. Digestion : l'ADN est digéré par une enzyme de restriction tout en maintenant les liaisons croisées.

3. Biotinylation et ligation : les extrémités 5' des fragments d'ADN qui en résultent sont remplies de nucléotides biotinylés, suivie d'une ligation à l'extrémité émoussée. Une librairie de produits de ligation qui représente les fragments de restriction de l'ADN qui étaient proches les uns des autres dans le noyau au moment de la fixation est ainsi obtenue.

4. *Pulldown* : le marqueur de biotine est utilisé pour effectuer un streptavidine *pulldown* sur les séquences. Ceci permet d'enrichir pour la jonction de ligation.

5. *Séquençage* : les fragments Hi-C purifiés (appelés *di-tags*) qui contiennent les informations tridimensionnelles sont séquencés pour créer la librairie Hi-C.

Les librairies Hi-C sont générées à partir de millions de cellules. Elles sont très complexes, avec 10^{11} de produits de ligature indépendants entre ~4 Kb fragments du génome humain **[6]**. En conséquence, l'identification fiable et reproductible des interactions entre les différents fragments de restriction à partir de données Hi-C n'est pas possible à moins que les librairies Hi-C soient soumises à un séquençage profond. Ceci n'est pas une solution économiquement viable pour la plupart des laboratoires. Pour contourner cette lacune, le promoteur capture Hi-C (figure 1-B-C) a été développé pour enrichir spécifiquement des produits de ligation contenant un promoteur provenant des librairies Hi-C. Une librairie Hi-C est hybridée à un système de capture qui consiste en des sondes d'ARN biotinylé ciblant les extrémités des fragments de restriction de l'ADN. Après hybridation, un streptavidine *pulldown* est effectué pour filtrer les fragments qui se sont hybridés avec les sondes d'ARN. Ceci conduit à un enrichissement pour les fragments d'intérêt. Après une amplification par PCR, la librairie

de promoteur capture Hi-C est prête à être séquencée.

L'analyse des données Hi-C nécessite beaucoup de ressources et de compétences en matière de calcul. Il est essentiel de comprendre le principe du traitement des données Hi-C pour choisir l'outil adéquat et interpréter les résultats. La présente étude se concentre sur le prétraitement des données de conformation tridimensionnelle du génome des mammifères via un *pipeline* bioinformatique : HiCUP (paquet Perl) **[7]**. Ceci constitue le point de départ de l'analyse des données Hi-C.

L'objectif est d'utiliser les données Hi-C et de promoteur capture Hi-C afin de construire des graphes d'interactions génomiques, pour ensuite explorer ces graphes couplés à d'autres ayant des informations moléculaires. Pour donner suite au prétraitement, de nombreux outils existent comme Juicer **[8]**, CHiCAGO (paquet R) **[9]**, FAN-C (paquet Python) **[10]**. Par conséquent, le but est de se familiariser avec ces différentes méthodes avant de les utiliser afin de procéder à l'analyse de données issues d'articles publiées récemment. La compréhension des paramètres influents le nombre d'interactions significatives et par voie de conséquence la construction des réseaux sont au cœur de ce projet.

MATÉRIELS ET MÉTHODES

Prétraitement des données avec HiCUP

Pour parvenir à des conclusions valides concernant les interactions génomiques, il faut que les données Hi-C soient cartographiées de manière non-conventionnelle. Ensuite, les artéfacts du protocole Hi-C doivent être supprimés. Pour répondre à ces exigences, HiCUP (*Hi-C User Pipeline*) fut développé. C'est un *pipeline* bioinformatique (figure 2) pour le prétraitement des données Hi-C qui a peu de dépendances et codé en Perl. HiCUP produit un rapport détaillé de contrôle de qualité dans un format HTML interactif. Ceci permet à l'utilisateur d'évaluer facilement la qualité d'une librairie Hi-C. HiCUP produit également des statistiques récapitulatives à chaque étape du *pipeline*. Ceci permet de contrôler la qualité, d'identifier les problèmes potentiels et d'affiner le protocole expérimental.

Le traitement des données Hi-C contient les étapes suivantes : cartographie, filtration, appariement, *binning*, normalisation, post-traitement et visualisation. HiCUP effectue les étapes de cartographie, filtration et appariement. C'est le point de départ du traitement. Il doit être utilisé en conjonction avec d'autres *pipelines* pour construire une interprétation tridimensionnelle de l'ensemble des données.

Description des *scripts* de HiCUP

HiCUP est constitué de six *scripts* Perl:

HiCUP: un *script* maître qui exécute les *scripts* « HiCUP truncater », « HiCUP mapper », « HiCUP filter » et « HiCUP deduplicator » de manière séquentielle.

HiCUP *truncater*: les paires Hi-C valides sont constituées de deux fragments d'ADN provenant de différentes régions du génome et ligaturés ensemble. En général, le *read forward* correspond à un fragment de ligation et le *read reverse* correspond à l'autre. Cependant, ce n'est pas toujours vrai puisque la jonction de ligation Hi-C peut se trouver dans la région séquencée. Ces *reads* seront probablement retirés du *pipeline* du Hi-C au cours du processus de cartographie. Ceci entraîne la perte de données potentiellement valables. « HiCUP truncater » aide à remédier à cela en identifiant les jonctions de ligation dans les *reads* et en supprimant la séquence en aval du site de reconnaissance de l'enzyme de restriction.

HiCUP *mapper*: les produits de ligation Hi-C valides comprennent deux fragments de restriction de différentes régions du génome ligaturées ensemble. « HiCUP mapper »

permet de cartographier les *di-tags* par rapport à un génome de référence afin de déterminer d'où provient chaque fragment de restriction. Après la cartographie, les *reads forward* et *reverse* sont couplées, c'est-à-dire que deux fichiers d'entrée donnent un fichier de sortie. Ce *script* utilise les logiciels d'alignement de séquences Bowtie **[11]** ou Bowtie2 **[12]** pour effectuer la cartographie.

HiCUP *filter*: la majorité des *reads* générées par « HiCUP mapper » sont très probablement des produits Hi-C valides. Cependant, une minorité ne l'est probablement pas et doit être supprimée. « HiCUP filter » traite les *reads* couplées avec le fichier créé par « HiCUP digester » pour identifier les paires de Hi-C valides et supprimer les artéfacts (figure 3).

HiCUP *deduplicator*: le protocole expérimental Hi-C implique une étape d'amplification PCR pour générer suffisamment de matériel pour le séquençage. Par conséquent, l'ensemble de données générées par « HiCUP filter » peut contenir des copies PCR du même *di-tag*. Ces copies de PCR pourraient entraîner des conclusions incorrectes concernant la conformation génomique. Ces duplicatas doivent être supprimés de l'ensemble de données en ne conservant qu'une seule copie.

HiCUP *digester*: « HiCUP filter » supprime un grand nombre des paires Hi-C invalides. Avant de pouvoir le faire, il a besoin d'un génome de référence digéré, ainsi que des fichiers de séquences appariées. « HiCUP digester » coupe un génome de référence sélectionné avec une ou deux enzymes de restriction spécifiées qui reconnaissent des séquences palindromiques uniques. Le *script* imprime les résultats dans un fichier pour un traitement ultérieur par « HiCUP filter ».

À l'issue du prétraitement, le rapport HTML généré est constitué de quatre sections: tronquage et cartographie, filtration, distribution des tailles des *di-tags* et déduplication.

Installation de HiCUP

HiCUP est écrit en Perl et exécuté en lignes de commande. Pour installer HiCUP, il faut télécharger le fichier hicup_v0.X.Y.tar.gz (https://www.bioinformatics.babraham.ac.uk/projects/hicup/hicup_v0.7.3.tar.gz). Pour assurer une meilleure reproductibilité des analyses dans un système d'exploitation basé sur Unix, un environnement conda a été généré avec les dépendances suivantes: Bowtie2, Perl, R (v.3.1.2) **[13]**, SAMtools (v.0.1.18) **[14]** et gzip.

Exécution du *pipeline* de HiCUP

Afin de maîtriser le *pipeline*, celui-ci a été exécuté avec un jeu de données d'entraînement avec les paramètres par défaut (avec l'assemblage du génome humain GRCh37 comme référence). Les différentes lignes de commandes ont été exécutées individuellement sur un ordinateur en local. La même manipulation a été réalisée sur le cluster de calcul du mésocentre (https://mesocentre.univ-amu.fr/) pour maîtriser l'utilisation des ressources HPC (*High Performance Computing*). Par la suite, le *pipeline* a été exécuté avec un jeu de données issu d'une étude de la conformation tridimensionnelle de lignées cellulaires de lymphoblastoïdes **[15]**. Celui-ci est disponible sur le dépôt public de la base de données GEO (*Gene Expression Omnibus*) **[16]** sous l'identifiant GSM3682164.

Ensuite, toutes les étapes ont été regroupées dans un *workflow* via snakemake **[17]** pour effectuer des analyses de données automatisées, reproductibles et ajustables (figure 4).

Disponibilité des scripts

Les codes générés au cours de la présente étude sont disponibles dans un dépôt public de GitLab via le lien suivant :

https://gitlab.com/Yorgomoubayed/internship-project. Le dépôt contient :

- **Un script snakemake** qui a permis d'automatiser le prétraitement des données avec HiCUP.
- **Un fichier au format YAML** qui a permis de générer l'environnement conda avec toutes les dépendances pour exécuter le *pipeline* de HiCUP.
- **Un script bash** pour le téléchargement des fichiers FastQ du jeu de données GSM3682164.
- **Deux fichiers de configuration** pour exécuter le *pipeline* de HiCUP.
- **Un lien vers le rapport HTML interactif** regroupant les résultats de l'analyse du jeu de données GSM3682164.

RÉSULTATS ET DISCUSSIONS

Prétraitement des données GSM3682164 avec HiCUP

Tronquage et cartographie (figure 5) : la librairie Hi-C traitée contient environ 190 millions de reads. Un certain nombre des *reads* ont été tronqués, bien qu'il s'agisse d'une faible proportion (29.9 %) par rapport au nombre total de *reads* qui ont été traités. Il y a un excès de 80 % des *reads* cartographiés de manière unique sur le génome de référence, ce qui est jugé comme pertinent. Environ 70 % des *reads* ont été couplés avec leurs partenaires, ce qui est également assez pertinent. La cartographie semble donc s'être bien passée.

Filtration (figure 6) : en examinant l'étape de filtration pour éliminer les artéfacts Hi-C typiques, environ 80 % des *di-tags* étaient valides, ce qui est bon pour une librairie Hi-C. Les autres *reads* sont classés comme artéfacts (mauvaise taille, re-ligation…).

Distribution de la taille des *di-tags* (figure 7) : la plupart des *di-tags* se situent dans la fourchette de taille. Cela représente un excès de 150 pb, mais moins de 750 pb. Le graphique est également assez lisse et uniforme dans sa distribution, ce qui suggère que la librairie est complexe. Pour récupérer un nombre plus important de *di-tags*, il est possible de diminuer la taille minimale de sélection jusqu'à environ 100 pb.

Déduplication (figure 8) : la complexité unique de la librairie se confirme en regardant le rapport de déduplication. 93 % des *di-tags* étaient uniques, ce qui est pertinent et suggère qu'il n'y a pas eu trop de cycles d'amplification PCR. Le ratio trans est d'environ 39%, ce qui n'est pas élevé et suggère que la librairie n'est pas bruyante.

Dans l'ensemble, le rapport promet d'obtenir des résultats intéressants pour cette librairie Hi-C à l'issue de l'analyse tridimensionnelle.

CONCLUSIONS ET PERSPECTIVES

L'ensemble du projet avait pour but de se familiariser avec les différentes méthodes d'analyses des données Hi-C et promoteur capture Hi-C avant de les utiliser. La compréhension des paramètres influents le nombre d'interactions significatives et par voie de conséquence la construction des réseaux sont au cœur de ce projet.

Les résultats obtenus avec HiCUP ne sont que le point de départ du traitement. Le *pipeline* n'effectue pas la normalisation et les tests statistiques nécessaires pour interpréter les données Hi-C. Il doit être utilisé en conjonction avec d'autres *pipelines* pour construire une interprétation tridimensionnelle de l'ensemble des données.

Cette interprétation est la clé pour construire des réseaux génomiques pour déterminer les liens entre les maladies partageant des structures génomiques ou moléculaires.

La complexité du traitement des données Hi-C et la combinaison d'outils spécialisés et de formats Hi-C disponibles constituent les prochains défis majeurs du domaine. Des *pipelines* tels que FAN-C ont le potentiel de s'adapter de manière transparente à ces défis, simplifiant considérablement l'analyse des données Hi-C. Son efficacité dans l'interprétation tridimensionnelle des données Hi-C pourrait bien constituer l'objet d'une nouvelle étude.

REMERCIEMENTS

Ce travail a été réalisé dans le cadre d'un projet tutoré (7 semaines) en première année de master de bioinformatique, parcours développement logiciel et analyse des données (DLAD). Le projet se déroulait en télétravail sous la co-supervision de M. Aitor GONZALEZ (enseignant-chercheur) et M. Anthony BAPTISTA (doctorant).

Ce travail a été suivi et évalué par M. Nicolas TERRAPON (enseignant-chercheur) et Mme. Lucie KHAMVONGSA CHARBONNIER (doctorante).

Ce travail a eu accès aux ressources HPC d'Aix-Marseille université financées par le projet Equip@Meso du programme « Investissements d'Avenir » qui est supervisé par l'agence nationale de la recherche.

RÉFÉRENCES

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* **15**, 272–286 (2014).

2. Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S. W. & Fraser, P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. JoVE J. Vis. Exp. e57320 (2018) doi:10.3791/57320.

3. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. Nat. Rev. Genet. 21, 207–226 (2020).

4. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. Science 295, 1306–1311 (2002).

5. Lieberman-Aiden, E. et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009).

6. Aymand, F. et al Genome wide mapping of long range contacts unveils DNA Double Strand Breaks clustering at damaged active genes. Nat Struct Mol Biol **24**, 353–361(2017).

7. Wingett, S. W. et al. HiCUP: pipeline for mapping and processing Hi-C data. F1000Research 4, 1310 (2015).

8. Durand, N. C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 3, 95–98 (2016).

9. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 17, 127 (2016).

10. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: A Feature-rich Framework for the Analysis and Visualisation of C data. bioRxiv 2020.02.03.932517 (2020) doi:10.1101/2020.02.03.932517.

11. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).

12. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

13. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

14. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079 (2009).

15. Gorkin, D. U. et al. Common DNA sequence variation influences 3-dimensional conformation of the human genome. Genome Biol. 20, 255 (2019).

16. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 41, D991–D995 (2013).

17. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522 (2012).

ANNEXES

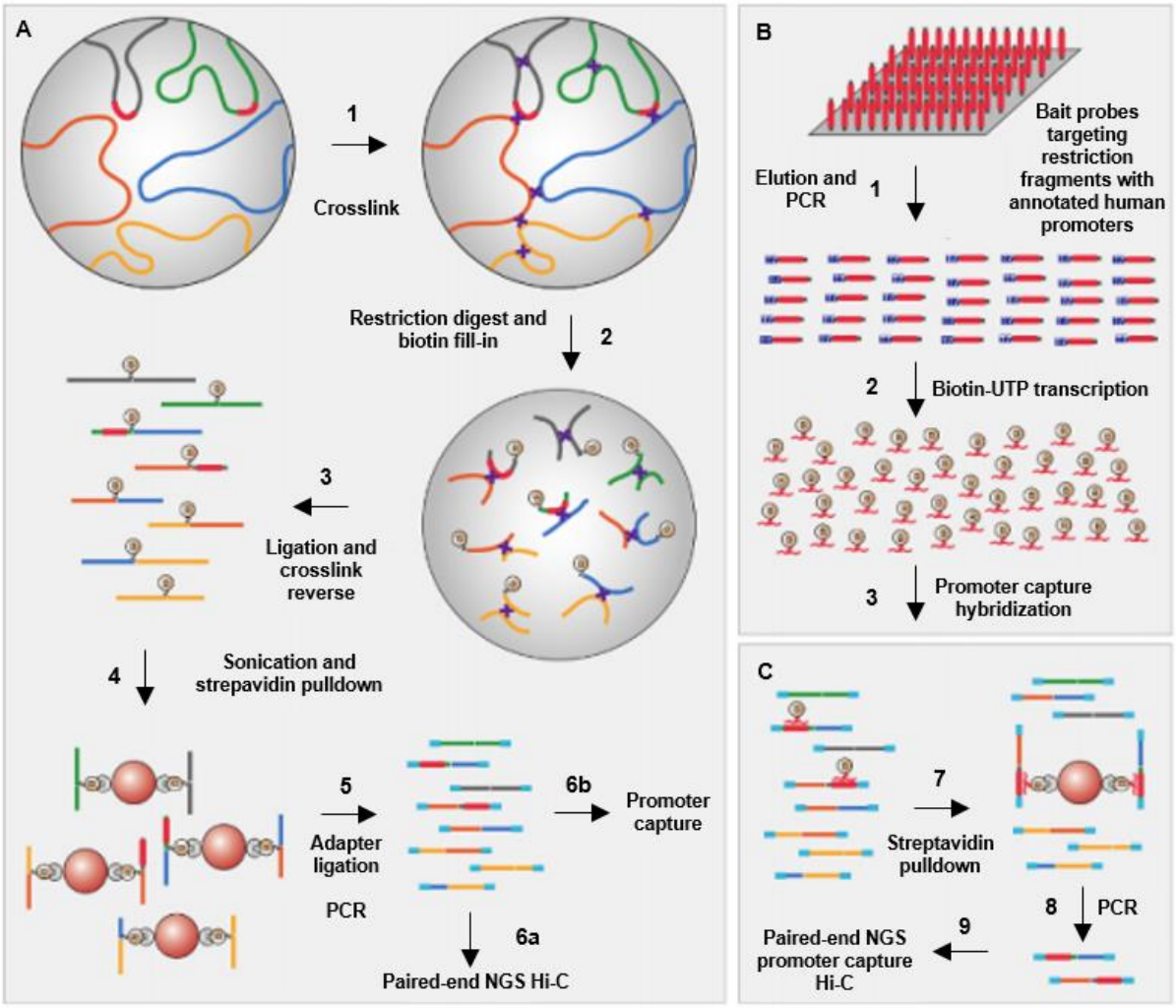


***Figure 1* - Diagramme de flux résumant les protocoles expérimentaux Hi-C et promoteur Capture Hi-C.** *In nucleus* Hi-C (I) suivi de l'hybridation avec les ARN biotinylés (II) qui ciblent les fragments de restriction de tous les promoteurs des gènes de l'espèce étudiée (III).

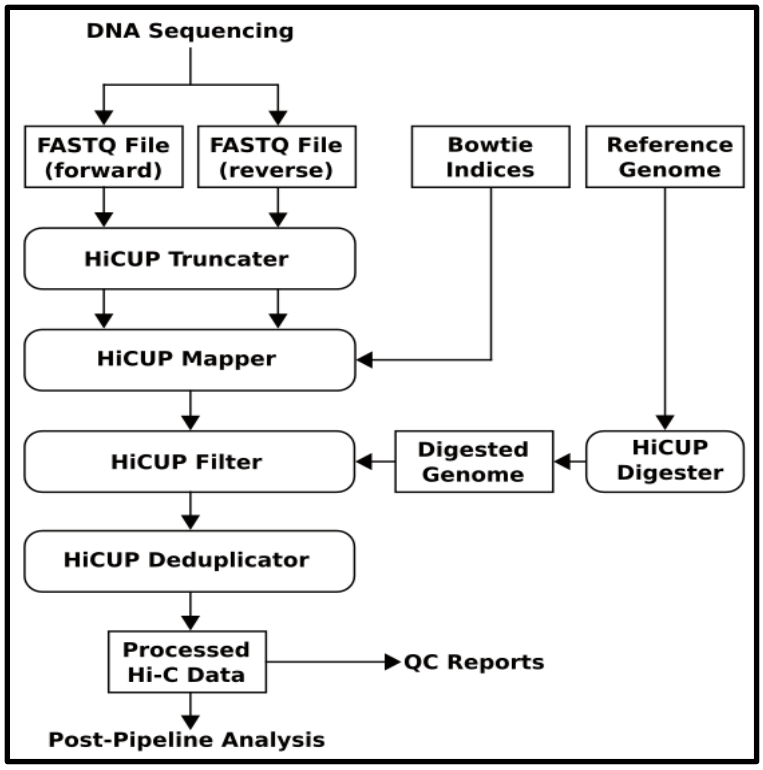


***Figure 2* - Diagramme de flux résumant le *pipeline* de HiCUP**. HiCUP prend les fichiers FASTQ générés par le séquençage de l'ADN et produit des données cartographiées nettoyées accompagnées de rapports de contrôle de qualité. La majeure partie du *pipeline* comprend 4 scripts : truncater, mapper, filter et deduplicator. Ceux-ci sont exécutés à leur tour par le script maître HiCUP qui contrôle le flux de données dans le *pipeline*. HiCUP prend des fichiers FASTQ paired-end avec un génome de référence FASTA et les indices d'alignement associés, puis rapporte les *di-tags* valides au format BAM/SAM.

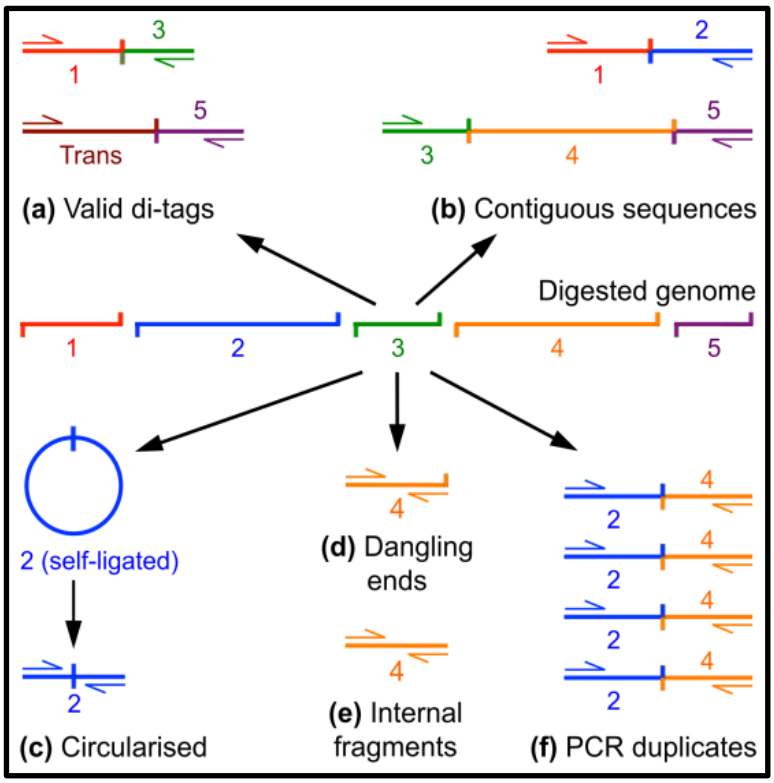


***Figure 3* - Aperçu des artéfacts expérimentaux générés par le protocole expérimental du Hi-C.** Le schéma montre un génome digéré en cinq fragments de restriction. Ces fragments peuvent se liguer les uns aux autres, ou à des fragments dérivés d'un autre chromosome, formant respectivement des *di-tags* cis ou trans valides **(a)**. En revanche, une re-ligature ou une digestion incomplète entraîne la génération de séquences non-valides **(b)**. Un autre artéfact commun se produit lorsque le fragment chimère séquencée correspond à un seul fragment de restriction **(c)**, **(d)** et **(e)**. En outre, la PCR peut générer des copies multiples d'un fragment **(f)**.

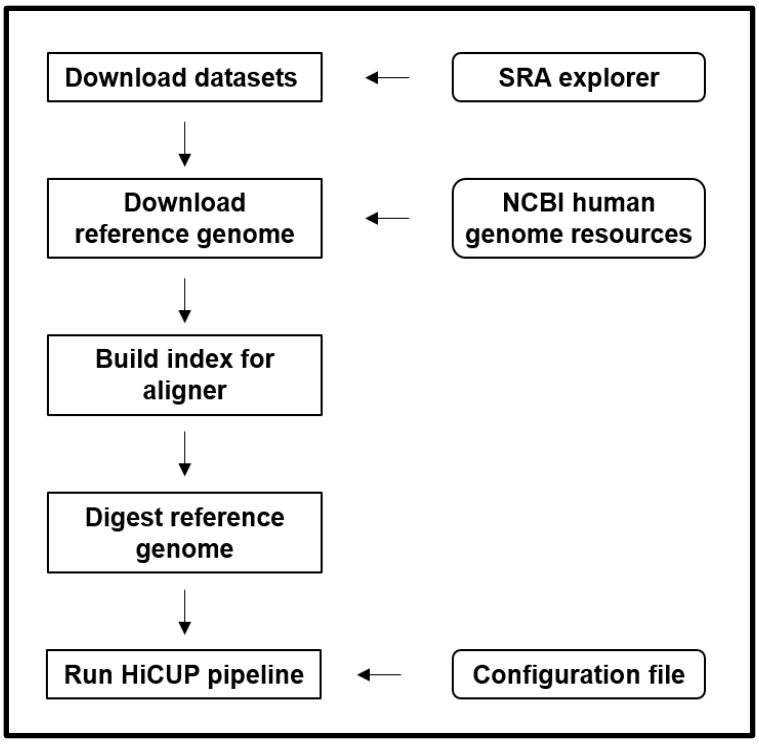
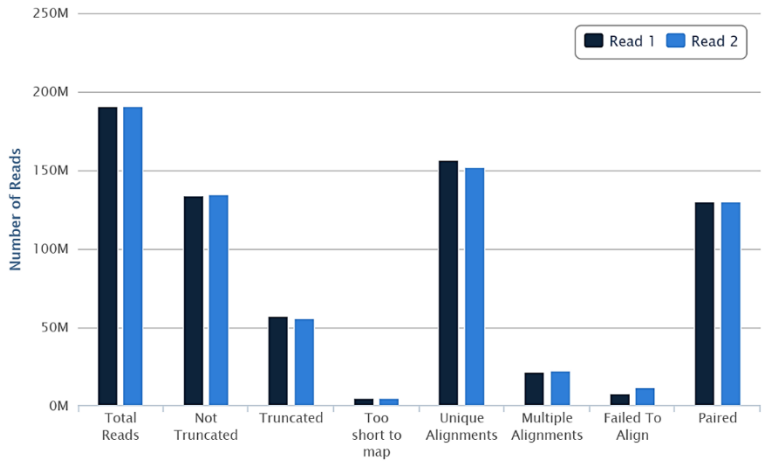


***Figure 4* - Diagramme de flux des étapes complémentaires au *pipeline* de HiCUP implémenté avec snakemake.** Le *pipeline* permet de télécharger les données à analyser via SRA explorer et le génome de référence via le dépôt public de NCBI. À partir du génome de référence, Bowtie2 crée un index et « HiCUP digester » crée un génome de référence digérée *in silico* avec l'enzyme de restriction choisie. Finalement, l'intégralité du *pipeline* de HiCUP est exécutée via un fichier de configuration dans lequel figurent les paramètres.

A

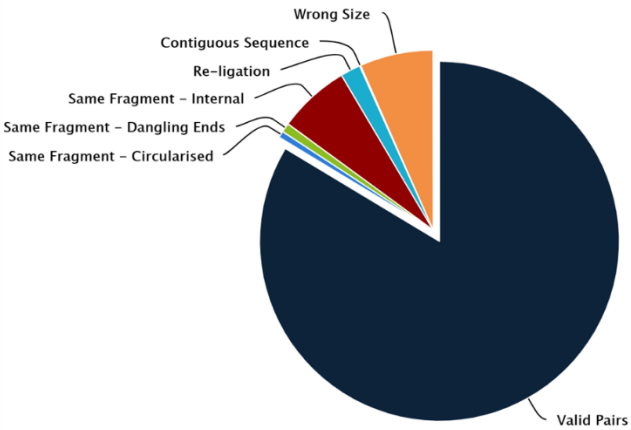


B

	Read 1	Read 2
Total reads	190,478,508	190,478,508
Not truncated	133,560,842	134,626,805
Truncated	56,917,666	55,851,703
Too short to map	4,647,963	4,533,198
Average length of truncated sequence	53.66	53.77
Unique alignments	156,242,477	151,487,635
Multiple alignments	21,453,512	22,546,285
Failed to align	8,134,556	11,911,390
Paired	129,700,060	129,700,060

***Figure 5* - Tronquage et cartographie. A-** Le graphique représente les principales statistiques produites par HiCUP lors du traitement des ensembles de données Hi-C. **B-** Le tableau représente les principales valeurs qui ont permis de générer le graphique dans A.

A



B

	Di-tag count
Valid pairs	108,466,918
Invalid pairs	21,233,142
Same circularised	778,417
Same fragment dangling ends	1,056,037
Same fragment internal	8,373,854
Re-ligation	2,326,053
Contiguous sequence	133,917
Wrong siwe	8,564,864
Total pairs	129,700,060

***Figure 6* - Filtration. A-** Le diagramme en camembert représente un aperçu des proportions qu'occupent les artéfacts techniques issus du protocole expérimental du Hi-C dans les données analysées. **B-** Le tableau représente les principales valeurs qui ont permis de générer le graphique dans A.

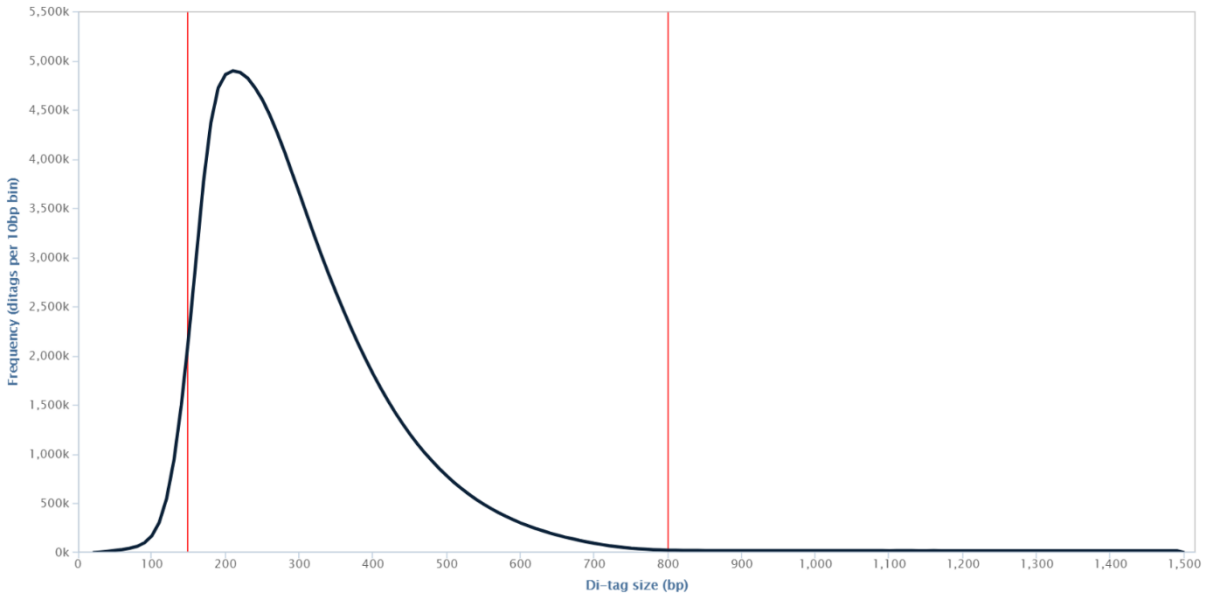
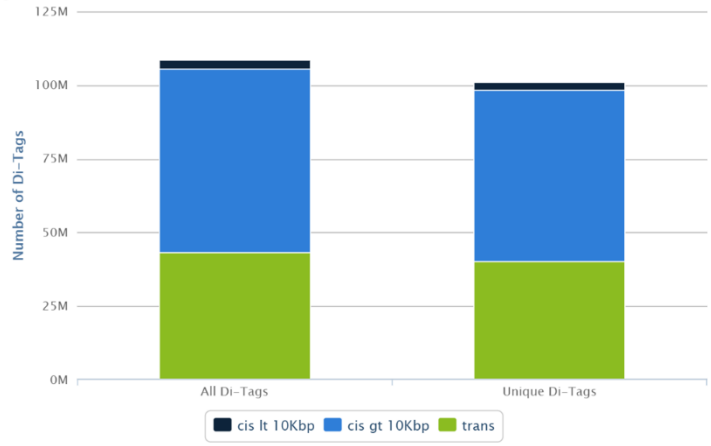


***Figure 7* - Distribution des tailles des *di-tags*.** La courbe représente la distribution des tailles des *di-tags* valides de la librairie Hi-C a l'issue du traitement avec HiCUP pour une gamme de sélection comprise entre 150 pb et 800 pb.

A



B

	All di-tags	Unique di-tags
Read pairs	108,466,918	101,036,425
Cis-close (< 10 Kbp)	3,135,804	2,920,343
Cis-far (> 10 Kbp)	62,406,094	58,115,278
Trans	42,925,020	40,000,804

***Figure 8* - Déduplication.** Le graphique représente le ratio cis/trans des *di-tags* avant (*All di-tags*) et après (*unique di-tags*) élimination des duplicatas de PCR. Un rapport trans/cis élevé est révélateur d'une librairie de mauvaise qualité.