14/10/2022

Offre de stage de Master 2 bioinformatique 2023 (6 mois)
CIRAD, UMR AGAP, équipe « Génome et sélection des pérennes »

Contacts : David Cros, david.cros@cirad.fr ; Xavier Argout, xavier.argout@cirad.fr

Contexte :

Les *topologically associating domains* (TADS) correspondent à un niveau d'organisation 3D du génome à une échelle en dessous de la mégabase et représentent des zones privilégiées d'interaction de la chromatine. Ils sont essentiellement stables au sein des espèces et jouent un rôle clé dans la régulation des gènes, avec des niveaux d'expression qui varient en fonction des tissus, des conditions environnementales et des stades de développement (Long et al., 2020 ; Mota-Gómez et Lupiáñez, 2019 ; Ouyang et al., 2020 ; Szabo et al., 2019 ; Wang et al., 2020). Les TADs sont un front de recherche chez les végétaux, avec un nombre restreint d'études récentes, qui portent essentiellement sur des espèces modèles, comme Arabidopsis (Feng et al., 2014), le riz (Dong et al., 2020 ; Liu et al., 2017) et le maïs (Dong et al., 2020). L'identification des profils de TADs se fait par séquençage Hi-C (*high-throughput chromosome conformation capture* (Lieberman-Aiden et al., 2009). Le développement de cette technique rend aujourd'hui possible ce type d'étude chez les plantes cultivées non-modèles.

Objectif :

L'objectif de ce stage est de :

1) comparer des outils bioinformatiques permettant de définir des profils de TADs et d'identifier les plus performants en termes de répétabilité des résultats, de robustesse face aux variations des paramètres techniques d'appel des TADs, de temps de calcul, de flexibilité d'emploi, etc.
2) caractériser les TADs et les comparer entre espèces (nombre, taille, séquences caractéristiques aux limites de TADs, type d'annotation intra- vs inter-TADs, etc.)

Le travail sera réalisé sur quatre espèces : cacaoyer, palmier à huile, eucalyptus globulus, eucalypus gunnii et hévéa. Ce stage se déroulera dans la poursuite de travaux réalisés récemment par 2 stagiaires dans notre équipe.

L'Annexe ci-dessous présente la procédure suivie pour l'obtention des données Hi-C préalablement au stage, ainsi que des détails concernant les étapes d'analyse des données prévues lors du stage (à partir de l'appel des TADs).

Annexe :

La procédure de séquençage Hi-C est résumée sur la Figure 1. Le formaldéhyde fixe les régions chromosomiques en contact, en créant des liaisons covalentes entre les segments de chromatine spatialement proches (étape de *crosslinking*). L'ADN est ensuite digéré par des enzymes de restriction pour ne garder que les régions de contact. Les extrémités des régions en contact sont marquées à la biotine puis liguées entre elles. L'ADN est fragmenté puis les segments biotinylés sont ligués à des amorces de séquençage. Ces fragments finaux, comprenant deux segments d'ADN physiquement proches dans le noyau au moment de la fixation, séparés par un site de restriction modifié et encadrés par des amorces de séquençage, sont nommés di-tags (Wingett et al 2015). Ils sont séquencés pour obtenir des reads paired-ends. Pour la détection de TADs haute résolution, la profondeur minimale de séquençage des librairies Arima-HiC (kit compatible avec les tissus des plantes) est de 600 millions de paires de reads pour un génome de 3 Gbp. Ceci peut s'obtenir avec deux librairies préparées avec deux répliques biologiques et séquencées chacune à 300 millions de paires de reads, ce qui permet d'évaluer la

reproductibilité de la méthode entre répliques et de combiner ensuite les deux répliques pour une analyse TAD haute-résolution.

Le traitement des données brutes de séquençage Hi-C requiert plusieurs étapes. Tout d'abord, il y a une étape de mapping, où les reads sont positionnés sur un génome de référence. La seconde étape consiste à filtrer les données sur la base de leur qualité, càd à supprimer les artefacts (di-tags comprenant deux fragments d'ADN adjacents ou un seul fragment, di-tags non mappés, etc) pour ne garder que les di-tags valides. Ceci permet d'obtenir une carte de contact sur le génome entier, ou décompte d'interaction, correspondant à une matrice de dimension N x N, avec N le nombre de bins (régions linéaires du génome) et les cellules de cette matrice contenant le nombre de fois que les deux bins correspondants étaient physiquement proches (nombre de reads) (MacKay et Kusalik, 2020). La troisième étape est la normalisation de la carte, destinée à corriger les biais inhérents au Hi-C qui peuvent affecter les nombres de reads, sous l'effet de divers paramètres techniques, tels l'efficacité de la PCR, la densité des sites de restriction, la longueur des fragments de restriction, des biais de mapping, etc (Liu et al., 2017 ; Lyu et al., 2019 ; Zufferey et al., 2018). Les cartes de contact normalisées du génome entier sont ensuite utilisées pour identifier (appeler) les TADs tout le long du génome. De nombreuses méthodes d'appel existent (voir Dali et Blanchette, 2017 ; Forcato et al., 2017 ; Liu et al., 2022 ; Sefer, 2022 ; Zufferey et al., 2018 pour des comparaisons). L'appel des TADs peut être affecté par des aspects techniques, comme la résolution à laquelle sont recherchés les TADs, la méthode de normalisation et la profondeur de séquençage. Les TADs identifiés peuvent aussi être filtrés sur la base de la qualité de leur appel (par exemple concordance entre répliques, profondeur des reads, etc (voir Golicz et al., 2020). La recherche de motifs ADN associés aux limites/frontières de TADs peut se faire avec les programmes MEME, Ame et TRAP (voir des exemples dans Liu et al., 2017 ; Ramírez et al., 2018).

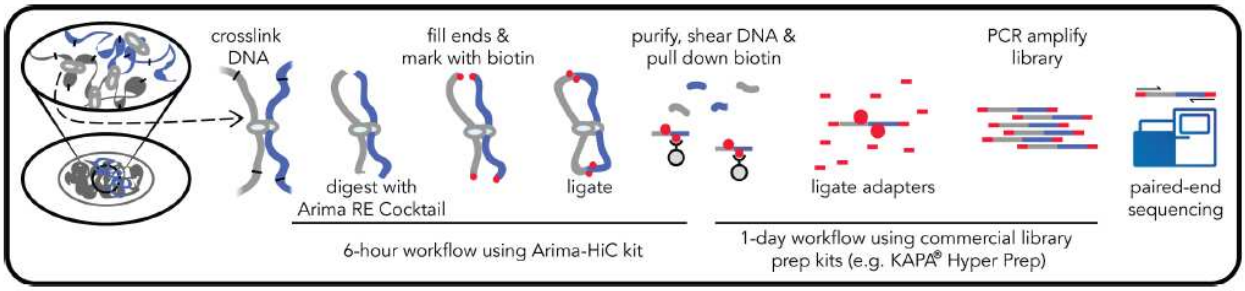

Figure 1 Résumé du protocole du séquençage Hi-C (source : Arima-HiC Kit User Guide for Plant Tissues, 2019)

Références :

Cameron C. J., Dostie J., Blanchette M. HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. *Genome Biology*. **2020**. Vol. 21, n°1, p. 11. https://doi.org/10.1186/s13059-019-1913-y

Dali R., Blanchette M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research*. **2017**. Vol. 45, n°6, p. 2994-3005. https://doi.org/10.1093/nar/gkx145

Dali R., Bourque G., Blanchette M. RobusTAD: A Tool for Robust Annotation of Topologically Associating Domain Boundaries. *bioRxiv*. **2018**. p. 293175. https://doi.org/10.1101/293175

Di Filippo L., Righelli D., Gagliardi M., Matarazzo M. R., Angelini C. HiCeekR: A Novel Shiny App for Hi-C Data Analysis. *Frontiers in Genetics*. **2019**. Vol. 10, p. 1079. https://doi.org/10.3389/fgene.2019.01079

Dong P., Tu X., Li H., Zhang J., Grierson D., Li P., Zhong S. Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains. *Journal of Integrative Plant Biology*. **2020**. Vol. 62, n°2, p. 201-217.

Feng S., Cokus S. J., Schubert V., Zhai J., Pellegrini M., Jacobsen S. E. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. *Molecular cell*. **2014**. Vol. 55, n°5, p. 694-707.

Forcato M., Nicoletti C., Pal K., Livi C. M., Ferrari F., Bicciato S. Comparison of computational methods for Hi-C data analysis. *Nature methods*. **2017**. Vol. 14, n°7, p. 679.

Golicz A. A., Bhalla P. L., Edwards D., Singh M. B. Rice 3D chromatin structure correlates with sequence variation and meiotic recombination rate. *Commun Biol*. **2020**. Vol. 3, n°1, p. 235. https://doi.org/10.1038/s42003-020-0932-2

Kumar V., Leclerc S., Taniguchi Y. BHi-Cect: a top-down algorithm for identifying the multi-scale hierarchical structure of chromosomes. *Nucleic Acids Research*. **2020**. Vol. 48, n°5, p. e26-e26. https://doi.org/10.1093/nar/gkaa004

Lazaris C., Kelly S., Ntziachristos P., Aifantis I., Tsirigos A. HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. *BMC Genomics*. **2017**. Vol. 18, n°1, p. 22. https://doi.org/10.1186/s12864-016-3387-6

Lieberman-Aiden E., Van Berkum N. L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B. R., Sabo P. J., Dorschner M. O., Sandstrom R., Bernstein B., Bender M. A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L. A., Lander E. S., Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **2009**. Vol. 326, n°5950, p. 289-293. https://doi.org/10.1126/science.1181369

Liu C., Cheng Y.-J., Wang J.-W., Weigel D. Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis. *Nature Plants*. **2017**. Vol. 3, n°9, p. 742-748. https://doi.org/10.1038/s41477-017-0005-9
Liu K., Li H., Li Y., Wang Jun, Wang Jianxin. A comparison of topologically associating domain callers based on Hi-C data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **2022**. p. 1-1. https://doi.org/10.1109/TCBB.2022.3147805
Long H. S., Powell G., Greenaway S., Mallon A.-M., Lindgren C. M., Simon M. M. Making sense of the linear genome, gene function and TADs. *bioRxiv*. **2020**. p. 2020.09.28.316786. https://doi.org/10.1101/2020.09.28.316786
Lyu H., Liu E., Wu Z. Comparison of normalization methods for Hi-C data. *BioTechniques*. **2019**. Vol. 68, n°2, p. 56-64. https://doi.org/10.2144/btn-2019-0105
MacKay K., Kusalik A. StoHi-C: Using t-Distributed Stochastic Neighbor Embedding (t-SNE) to predict 3D genome structure from Hi-C Data. *bioRxiv*. **2020**. p. 2020.01.28.923615. https://doi.org/10.1101/2020.01.28.923615
Mota-Gómez I., Lupiáñez D. G. A (3D-Nuclear) Space Odyssey: Making Sense of Hi-C Maps. *Genes (Basel)*. **2019**. Vol. 10, n°6, p. 415. https://doi.org/10.3390/genes10060415
Ouyang W., Cao Z., Xiong D., Li G., Li X. Decoding the plant genome: from epigenome to 3D organization. *Journal of Genetics and Genomics*. **2020**.https://doi.org/https://doi.org/10.1016/j.jgg.2020.06.007
Ramírez F., Bhardwaj V., Arrigoni L., Lam K. C., Grüning B. A., Villaveces J., Habermann B., Akhtar A., Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun*. **2018**. Vol. 9, n°1, p. 189-189. https://doi.org/10.1038/s41467-017-02525-w
Sefer E. A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinformatics*. **2022**. Vol. 23, n°1, p. 127. https://doi.org/10.1186/s12859-022-04674-2
Servant N., Varoquaux N., Lajoie B. R., Viara E., Chen C.-J., Vert J.-P., Heard E., Dekker J., Barillot E. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*. **2015**. Vol. 16, n°1, p. 259. https://doi.org/10.1186/s13059-015-0831-x
Soler-Vila P., Cuscó P., Farabella I., Di Stefano M., Marti-Renom M. A. Hierarchical chromatin organization detected by TADpole. *Nucleic Acids Research*. **2020**. Vol. 48, n°7, p. e39-e39. https://doi.org/10.1093/nar/gkaa087
Szabo Q., Bantignies F., Cavalli G. Principles of genome folding into topologically associating domains. *Sci Adv*. **2019**. Vol. 5, n°4, p. eaaw1668. https://doi.org/10.1126/sciadv.aaw1668
Wang G., Meng Q., Xia B., Zhang S., Lv J., Zhao D., Li Y., Wang X., Zhang L., Cooke J. P., Cao Q., Chen K. TADsplimer reveals splits and mergers of topologically associating domains for epigenetic regulation of transcription. *Genome Biology*. **2020**. Vol. 21, n°1, p. 84. https://doi.org/10.1186/s13059-020-01992-7
Wingett S., Ewels P., Furlan-Magaril M., Nagano T., Schoenfelder S., Fraser P., Andrews S. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*. **2015**. Vol. 4,.
Wolff J., Bhardwaj V., Nothjunge S., Richard G., Renschler G., Gilsbach R., Manke T., Backofen R., Ramírez F., Grüning B. A. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Research*. **2018**. Vol. 46, n°W1, p. W11-W16. https://doi.org/10.1093/nar/gky504
Zufferey M., Tavernari D., Oricchio E., Ciriello G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*. **2018**. Vol. 19, n°1, p. 217. https://doi.org/10.1186/s13059-018-1596-9