Proposition de stage pour le Master 2 Bio-informatique moléculaire: méthodes et analyses

Vers un génome de référence du chevreuil à partir de séquencage long-read nanopore

Contexte

Le chevreuil (*Capreolus capreolus*) est l'un des mammifères artiodactyles les plus étudiés au monde (2621 articles dans Web of Sciences en date du 7 juillet 2023) après le cerf élaphe (*Cervus elaphus*), le daim (*Dama dama*) et le cerf de virginie (*Odocoileus virginianus*). Si ces trois espèces ont toutes un génome de référence de qualité publié, il n'en est rien pour le chevreuil limitant d'autant les études de génomique populationnelle chez cette espèce (seulement 6 articles ces 5 dernières années). Le LBBE conduit depuis plusieurs décennies un projet d'étude à long terme dans différentes populations de chevreuils et ce programme a permis des avancées majeures en écologie évolutive et dynamique des populations (1–3).

A Lyon au LBBE, des projets de génomique se développent par exemple avec l'utilisation du séquençage ddRAdSeq pour notamment travailler sur les causes génomiques de l'hétérogénéité individuelle de la durée de vie. Afin de permettre une analyse correcte en regard d'un génome de référence de ces données ddRadSeq (analyse de type GWAS et rôle fonctionnel des marqueurs SNPs discriminants), il sera nécessaire de disposer d'un génome de référence du chevreuil. Le sujet de stage proposé portera donc sur l'assemblage et l'annotation du génome du chevreuil par un travail bio-informatique massif à partir de données brutes obtenues par un séquençage nanopore long-reads au DTAMB de Lyon.

Mission et déroulé du stage

L'objectif du stage est de réaliser le meilleur assemblage possible compte-tenu des données produites, du temps et des ressources informatiques disponibles (cluster de calcul du LBBE-PRABI). A cette fin la personne recrutée devra réaliser et comparer les assemblages obtenus avec différents assembleurs tels que Flye, Shasta, NextDenovo ou Canu sur la base de statistiques classiques d'évaluation de leur qualité (N50/NG50, nombre de contigs, BUSCO(4) et QV (5) score au sein d'un autre). A ce stade et en fonction de la qualité du meilleur assemblage obtenu une étape d'amélioration de la

contiguité pourra être envisagée par alignement sur les génomes des espèces de Cervidae ou d'autres artiodactyles bien assemblés en utilisant un pipeline comme Ragout2 (6).

Dans un deuxième temps une annotation des séquences répétées avec RepeatModeler2 et RepeatMasker puis des gènes avec un pipeline comme Braker (7) Maker (8) sera entreprise.

En cas de disponibilité en cours de stage des données ddRadSEq et si le travail sur le génome est terminé, le stagiaire pourra être amené à analyser ces données dans le cadre d'une étude GWAS de la durée de vie chez les chevreuils utiliser des pipelines comme Stacks (9,10).

Le stage se déroulera au laboratoire de Biométrie et Biologie Evolutive. Il sera co-encadré par Matthieu Boulesteix et Tomas Carrasco (Département Coévolution multi-échelles) et Sébastien Devillard (Département Ecologie Evolutive).

Contacts et candidature :

Merci de joindre CV + lettre de motivation + relevés de notes de Master 1 à matthieu.boulesteix@univ-lyon1.fr et sebastien.devillard@univ-lyon1.fr .

References

1. Plard F, Gaillard JM, Coulson T, Hewison AJM, Delorme D, Warnant C, et al. Mismatch between birth date and vegetation phenology slows the demography of roe deer. PLoS Biol. 2014 Apr;12(4):e1001828.

2. Garratt M, Lemaître JF, Douhard M, Bonenfant C, Capron G, Warnant C, et al. High juvenile mortality is associated with sex-specific adult survival and lifespan in wild roe deer. Curr Biol. 2015 Mar 16;25(6):759–63.

3. Gaillard JM, Hewison AJM, Klein F, Plard F, Douhard M, Davison R, et al. How does climate change influence demographic processes of widespread species? Lessons from the comparative analysis of contrasted populations of roe deer. Ecol Lett. 2013 May;16 Suppl 1:48–57.

4. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 2019;1962:227–45.

5. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality and phasing assessment for genome assemblies. scholar.archive.org [Internet]. Available from: https://scholar.archive.org/work/fnhdfjlo2zdgllifllwg5gclka/access/wayback/https://www.biorxiv.org/content/10.1101/2020.03.15.992941v1.full.pdf

6. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome assembly of large and complex genomes using multiple references. Genome Res. 2018 Nov;28(11):1720–32.

7. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. p. 65–95.

8. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan;18(1):188–96.

9. Rochette NC, Catchen JM. Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc. 2017 Dec;12(12):2640–59.

10. Arantes LS, Caccavo JA, Sullivan JK, Sparmann S, Mbedi S, Höner OP, et al. Scaling-up RADseq methods for large datasets of non-invasive samples: Lessons for library construction and data preprocessing. Mol Ecol Resour [Internet]. 2023 Aug 30; Available from: http://dx.doi.org/10.1111/1755-0998.13859