

## **Extraction de variants et calcul de diversité génétique à partir de graphes de pangénomes : application aux chênes blancs européens**

### **Dates**

**Durée de stage : 6 mois – niveau master 2.**

Date de début : à partir de janvier 2027

***Ce stage s'inscrit dans un projet de recherche qui donnera lieu à une thèse financée de 2027 à 2030, offrant ainsi une possibilité de recrutement doctoral dans sa continuité.***

### **Indemnisation**

Selon les règles en vigueur à INRAE (actuellement 4,50 €/h, soit environ 4 000 € pour 6 mois à temps plein).

### **Localisation**

Unité de recherche d'accueil : [UMR 1202 BIOGECO](#)

Adresse : Cestas Pierroton

### **Contacts**

Ludovic Duvaux, IR à BIOGECO (INRAE Nouvelle-Aquitaine-Bordeaux)

Stéphane De Mita, CRCN à PHIM (INRAE Occitanie-Montpellier)

[Ludovic.duvaux@inrae.fr](mailto:Ludovic.duvaux@inrae.fr) ; [stephane.de-mita@inrae.fr](mailto:stephane.de-mita@inrae.fr)

### **Procédure de candidature**

Envoyer SVP un CV et une lettre de motivation aux courriels précédents **avant le 24 septembre 2026.**

### **Description du stage**

#### **Contexte :**

La diversité génétique constitue une information essentielle pour comprendre l'histoire évolutive des populations et les processus — tels que les changements démographiques, la sélection naturelle, la mutation et la recombinaison — qui ont façonné le génome des espèces. Cette diversité, ainsi que sa variation et sa distribution entre individus et populations, est communément caractérisée à l'aide de statistiques descriptives. Celles-ci permettent à leur tour d'estimer différents paramètres de génétique des populations, tels que le taux de mutation ( $\theta$ ) ou le taux de recombinaison ( $\rho$ ), afin de donner une interprétation biologique et évolutive à la diversité observée. Le calcul de ces statistiques et estimateurs reposent néanmoins encore largement sur les polymorphismes nucléotidiques simples (single nucleotide

polymorphism, ou SNP), alors qu'une part importante de la diversité est constituée de variants structuraux (SV), tels que les insertions, délétions, duplications et réarrangements. Les graphes de variation pangénomiques (Pangenome Variation Graphs, ou PVG) permettent de représenter simultanément plusieurs génomes et de mieux intégrer cette diversité, notamment les SV, tout en limitant les biais associés à une référence linéaire unique [1]. Ils soulèvent cependant de nouveaux problèmes méthodologiques : comment définir et extraire des variants à partir d'une structure de graphe, comment les génotyper et comment adapter les statistiques de génétique des populations à des variants pouvant être imbriqués ou se chevaucher ? Dans ce contexte, notre projet vise à développer des méthodes permettant d'exploiter directement l'information contenue dans les graphes et à l'appliquer au cas d'étude des chênes blancs européens et à d'autres jeux de données.

### Objectifs du stage

Le premier objectif sera d'extraire les variants génétiques à partir d'un PVG. Deux approches complémentaires seront étudiées : (i) une approche fondée sur *VG deconstruct* [2], qui nécessite de définir une séquence de référence et permet d'exporter les variations sous forme de VCF, et (ii) *GraDiv call* [3], qui identifie les variants représentés dans le graphe sans imposer de référence linéaire. Le stage consistera à comparer ces approches et à étudier leur capacité à retrouver les variants et leurs génotypes dans différentes configurations de graphes.

Le second objectif sera d'adapter les statistiques de diversité afin de prendre explicitement en compte les variants structuraux. Les génotypes obtenus à partir du graphe seront utilisés avec *GraDiv compute* pour calculer différentes statistiques, puis les méthodes et leur implémentation seront développées pour tenir compte des propriétés particulières des variations représentées dans un graphe. L'étudiant participera ainsi directement au développement et à l'implémentation de nouvelles fonctionnalités de *GraDiv*, avec l'objectif de conserver au mieux l'information biologique portée par le graphe.

Enfin, les méthodes développées seront évaluées par simulation à l'aide de *MSpangenome* [4], qui permet de générer des graphes de pangénomes à partir de modèles coalescents et de scénarios démographiques explicites. L'étudiant contribuera à la conception et à l'implémentation des évolutions de *MSpangenome* nécessaires au projet, puis comparera les résultats obtenus aux paramètres connus des simulations afin d'identifier d'éventuels biais. Les analyses pourront porter sur les effets de  $\theta$ ,  $\rho$ , de la taille et de l'histoire démographique des populations, ainsi que sur la fréquence et la nature des différents variants. Pour les régions où l'extraction explicite des variants reste difficile, une approche fondée sur les spectres de k-mers pourra également être explorée comme alternative ou complément.

### Méthodologie et compétences requises et développées

Le stage se situe à l'interface entre génétique des populations, la bioinformatique et le développement logiciel. Il s'adresse à un étudiant de Master 2 ayant un intérêt pour la génétique

des populations et la programmation, et souhaitant développer ses compétences dans l'implémentation de méthodes d'analyse de données génomiques. Le travail impliquera notamment la manipulation de graphes et de grands jeux de données, l'analyse statistique et le développement en Python et/ou C. Une première expérience en programmation et/ou en analyse de données génomiques est requise.

Le développement sera conduit dans une démarche reproductible, avec utilisation de Git, ainsi que d'environnements de calcul adaptés aux analyses de génomique. Le projet donnera également l'occasion de travailler avec des outils récents de pangénomique et de développement méthodologique, et d'interagir avec des communautés travaillant sur les graphes de pangénomes et la génétique des populations.

### Environnement scientifique et encadrement

Le stage sera réalisé au sein de l'UMR BIOGECO, sous l'encadrement direct de Ludovic Duvaux et en étroit co-encadrement avec Stéphane De Mita de l'UMR PHIM (Montpellier). Il aura lieu dans l'équipe E4E, sur le site de Cestas-Pierroton. L'équipe E4E étudie l'évolution dans les écosystèmes forestiers et rassemble des compétences en génétique des populations, génomique et écologie évolutive, mobilisables pour ce projet. Les réunions d'équipe permettront au stagiaire de présenter régulièrement son travail et de se familiariser avec différents sujets de recherche.

Lieu du stage : UMR BIOGECO, site de Cestas-Pierroton (INRAE/Université de Bordeaux). Le site est situé dans un domaine forestier entre Bordeaux (20 km) et Arcachon (40 km) et est accessible en transports en commun depuis la gare TER de Cestas. Le stage s'inscrit dans un environnement scientifique connecté à plusieurs projets et réseaux de recherche en pangénomiques (e.g. Get-A-Pan, PanQueSt, PopGraph).

### Références

- [1] Bocs, S. et al. (2026). Adopting graph-based pangenomics for agronomy and biodiversity studies: current resources and challenges. *Peer Community Journal*, 6, e86. doi:10.24072/pcjournal.781.
- [2] Garrison, E. et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36, 875–879. doi:10.1038/nbt.4227.
- [3] Imbert, M. et al. (2025). GraDiv: a tool to compute diversity statistics from a pangenome graph. PHIM/IRD. <https://forge.ird.fr/phim/gradiv/-/tree/main>
- [4] Piat, L. et al. (2026). Simulating population pangenomes under coalescent demographic models with MSpangenome. *bioRxiv*. doi:10.64898/2026.06.29.735168.