Offre de stage : Développement et optimisation d'un outil de simulation en C++ de données génomiques en populations spatialisées.

- Localisations : INRA–CBGP, Montpellier & Université de Montpellier-IMAG.
- Durée : ~ 5 à 6 mois. Indemnités de stage : 554,40 Euros/mois.

Le développement rapide des techniques de séquençage, ainsi que des moyens de calculs informatiques a révolutionné la génétique/génomique des populations ces 20 dernières années. Cependant, du fait de la difficulté à obtenir des données génomiques individuelles et géo-référencées pour de gros échantillons et de la lourdeur des modèles spatialement explicites, l'aspect spatial a souvent été délaissé. On assiste aujourd'hui à deux changement majeurs : (1) les génomes individuels atteignent un prix raisonnable, et (2) les méthodes d'estimations basées sur la simulation de type ABC (Approximate Bayesian Computations) ont gagné un facteur 10 à 100 en terme d'efficacité grâce à l'utilisation des algorithmes de forêts aléatoires (Pudlo et al. 2015). De ce fait, il est maintenant envisageable de faire de l'estimation de paramètres démographiques (dispersion, densité et tailles de populations) et historiques (dater des changements démographiques) à partir de données génomiques sous des modèles démo-génétiques spatialisés de plus en plus réalistes. Améliorer ces techniques d'inférences et les modèles sous-jacents permettra de répondre à des questions essentielles pour mieux comprendre la répartition et l'évolution de la diversité génétique des populations dans le temps et l'espace.

L'objectif du stage est d'implémenter un nouveau simulateur de données génomiques basé sur des algorithmes de coalescences pouvant considérer des modèles spatialisés, dans le but de l'utiliser pour faire de l'inférence démo-génétique. Les techniques modernes d'inférence, par ABC entre autres, nécessitant des algorithmes efficaces, autant en terme de vitesse d'exécution des calculs que de l'espace mémoire nécessaire, le choix des méthodes de stockage et d'indexation des arbres de coalescence et des génomes simulés sera donc crucial pour permettre de simuler de gros jeux de données très rapidement (e.g. Kelleher et al. 2016). Le code développé sera validé par comparaison avec des résultats analytiques et de simulations issues d'anciens algorithmes peu efficaces (e.g. IBDSim). L'ensemble du travail sera valorisé par la rédaction d'une publication scientifique.

Le projet visera le développement d'un logiciel autonome, open source, collaboratif (Git) et si possible en intégration continue. Il sera organisé et prévu pour s'orienter vers une programmation dite « moderne » en C++, utilisant de manière extensive les nouveautés du standard (C++11/14 voire 17), de manière à produire un code lisible, concis, optimisé et immédiatement réutilisable. Une période de 1 à 2 mois en début de stage pourra en partie être dédiée à l'apprentissage de l'état de l'art de la programmation en C++ moderne. Le sujet comporte aussi bien de l'algorithmique, de l'architecture logiciel que du développement C++, le tout teinté d'optimisation et de parallélisation. Si la durée du stage le permet, une encapsulation dans un package R sera réalisée avec sa documentation (en anglais) ainsi qu'une extension ABC qui pourra être ajoutée pour permettre des inférences sur données réelles sous R.

Du fait de son implication complète mais encadrée dans toutes les phases du développement et de l'expérience acquise en C++ moderne orienté calcul, nous estimons qu'à l'issu du stage le stagiaire pourra se positionner sur le créneau très recherché d'ingénieur calcul dans l'industrie, la R&D ou dans

le domaine académique.

Langages demandés : C++, un peu de R. Un ordinateur performant ainsi que l'accès à des clusters de calculs sera fourni pour la durée du stage.

Encadrants :

- Francois David Collin, Univ-montpellier IMAG : Francois-David.Collin@umontpellier.fr
- Raphaël Leblois, INRA-CBGP, Montpellier: raphael.leblois@inra.fr
- Alexandre Dehne-Garcia, INRA-CBGP : alexandre.dehne-garcia@inra.fr

Pudlo, P., Marin, J.-M., Estoup, A., Cornuet, J.-M., Gautier, M. and Robert, C. P. (2015) Reliable ABC model choice via random forests. Bioinformatics 32 (6): 859-866.

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol 12(5): e1004842.