Utilisation des méthodes d'inférence démographique pour comprendre l'histoire de la spéciation chez les complexes d'espèces à fort flux de gènes: comparaison des approches par paires d'espèces et multi-espèces

Nous proposons un stage de M2 dans l'équipe 'Ecologie et Génomique Fonctionnelle' de l'UMR BioGeCo du site INRA de Cestas-Pierroton (https://www6.bordeaux-aquitaine.inra.fr/biogeco) sur le thème de la génomique de la spéciation et des inférences basées sur des modèles.

Informations générales

Dates du stage: 1er semestre 2019 (janvier-juin)

Encadrement: Ludovic Duvaux (INRA Pierroton), Carole Smadja (CNRS Montpellier), Roger Butlin (University of Sheffield), Christophe Plomion (INRA Pierroton)

Laboratoire d'accueil: UMR BioGeCo (UMR 1202), Cestas-Pierroton, France

Contact: Ludovic Duvaux (ludovic.duvaux@inra.fr)
Carole Smadja (carole.smadja@umontpellier.fr)

Informations pratiques

Lieu du stage: UMR BioGeCo, site de Cestas-Pierroton (INRA/Université de Bordeaux). Le site offre un cadre agréable au sein d'un domaine forestier situé entre Bordeaux (20km) & Arcachon (40km). Le campus de recherche est accessible par les transports en commun à parti de la gare TER de Cestas (2 navettes aller et 2 navettes retour par jour).

Rémunération: selon les règles en vigueur au sein de l'INRA.

Contexte général du stage

L'accès à de nombreuses données génomiques de qualité nous permet aujourd'hui d'inférer l'histoire démographique des populations (e.g. estimer les tailles de population, les taux de migration entre populations). Cela nous permet d'évaluer l'impact des évènements historiques sur l'évolution de ces populations, comme par exemple l'impact des cycles glaciaires sur les patrons de diversité génétique. Inférer la démographie des espèces émergeantes (ci-après simplement appelées "espèces") est également indispensable pour comprendre le processus de spéciation. L'établissement de l'isolement reproductif conduisant à la spéciation repose généralement sur l'accumulation de barrières à la reproduction durant une période appelée "zone grise de spéciation", période durant laquelle les espèces ne sont pas encore totalement isolées reproductivement (Roux et al. 2016). L'inférence démographique est idéale pour étudier cette période puisque l'étude de la spéciation a montré le rôle prépondérant (i) du temps de divergence, (ii) de la migration inter-espèce (ou flux de gènes) en plus du rôle de l'adaptation dans l'accumulation des barrières (Abbot et al. 2013). Cette approche permet donc de retracer l'histoire de divergence, en testant l'existence et la temporalité du flux de gènes au cours de la divergence, tout en caractérisant l'évolution des tailles des populations et leur temps de divergence (Roux et al. 2016).

L'inférence de l'histoire démographique des espèces se basent généralement sur des modèles de coalescence qui ne considèrent **qu'une seule paire d'espèces à la fois** (Nielsen & Wakeley 2001). Ces modèles permettent des inférences relativement robustes lorsque seulement deux espèces sont réellement impliquées (Beaumont 2008; Strasburg

& Rieseberg 2010). Cependant, dans le cas de complexes d'espèces où plusieurs espèces présentent un isolement reproductif incomplet, les estimations des temps de divergence et de flux de gènes peuvent être biaisées lorsque l'on considère toutes les paires de manière successive. **Il existe des méthodes multi-espèces** mais celles-ci nécessitent de connaître la phylogénie des espèces *a priori* (e.g. Hey 2010). Ce pré-requis est raisonnable si les espèces ont divergé de manière non simultanée et que la dérive génétique est plus forte que la migration puisque dans ce cas la phylogénie estimée sera proche de la phylogénie réelle (la topologie représentant l'ordre des divergences/spéciations successives). C'est par exemple le cas pour les grands mammifères tels que l'homme où les chevaux. Ce postulat est par contre très incertain pour certains complexes d'espèces où un fort flux de gènes et/ou des divergences récentes peuvent fortement fausser l'estimation des phylogénies. En conclusion, il n'existe pas actuellement de méthode satisfaisante pour inférer conjointement la dynamique de spéciation (i.e. les caractéristiques de la "zone grise de spéciation") et l'ordre des spéciations (i.e. la phylogénie) pour les complexes d'espèces à fort flux de gènes.

Objectifs du stage

Le but de ce stage est de déterminer la meilleure approche afin d'inférer l'histoire de la spéciation chez des complexes d'espèces à fort flux de gènes. Pour cela, le/la stagiaire devra comparer des approches d'inférences par paires d'espèces et multi-espèces en utilisant un cadre statistique dit "approximate Bayesian computation" (ABC) et des jeux de données génomiques déjà disponibles.

Déroulement du stage et formation

Afin de répondre aux objectifs du stage, l'étudiant(e) aura à sa disposition un pipeline ABC déjà implémenté et un jeu de données de re-séquençage de milliers d'exons déjà obtenu pour différents biotypes de pucerons. Un jeu de données génomiques d'espèces du chêne blanc sera analysé en parallèle par les encadrants afin de permettre une comparaison à la fin du stage. Afin de limiter le nombre de combinaisons à tester et le fléau de la dimension, le nombre de biotypes/espèces à analyser sera limité à 4 au sein de chaque complexe. Pour les analyses par paires, l'ensemble des combinaisons seront testées (soit 6 par complexe). Pour les analyses multi-espèces, plusieurs solutions seront comparées: (i) l'inférence multi-espèce directe se basant sur une phylogénie *a priori*, (ii) l'utilisation d'une phylogénie informée par les estimations de divergence lors des analyses par paires, (iii) le test des différentes topologies directement par l'ABC, (iv) l'ajout séquentiel de biotypes/espèces.

Quel que soit l'approche considérée, paire d'espèces ou multi-espèces, les inférences suivront 5 étapes majeures (Csilléry et al. 2010, 2015). (i) La première étape est la mise en place des modèles démographiques correspondant aux hypothèses évolutives à tester. Ces modèles, qui concernent l'architecture de l'isolement reproductif, ont déjà été implémenté pour les analyses par paires en suivant Roux et al. (2016). Ils seront adaptés pour les analyses multi-espèces. (ii) L'étudiant(e) testera ensuite par validations croisées la puissance et la robustesse de notre cadre ABC pour discriminer ces modèles. (iii) Les modèles seront ensuite comparés pour retenir celui qui explique le mieux les données. (iv) Le/la stagiaire testera si le modèle retenu produit des données réellement proches des données observées en effectuant deux étapes appelées "goodness-of-fit" et "posterior

predictive check". (v) Enfin, les valeurs de paramètres du modèle retenu pourront être estimées. A la fin du stage l'étudiant(e) comparera et discutera les résultats obtenus chez les deux modèles.

Le/la stagiaire sera tout au long du stage encadré(e) et formé(e) aux méthodes de bioinformatique et de génétique des populations. Ce stage permettra d'acquérir une expérience importante en inférences évolutives basées sur des approches de statistiques Bayésiennes (ABC), et dans leur application aux domaines de la spéciation et de la génomique des populations. Il permettra également à l'étudiant(e) d'être introduit aux données issues des technologies de séquençage à haut débit, actuellement à la pointe de la recherche en biologie.

Profil recherché

Le profil est ouvert à des étudiants provenant (i) de master en biologie évolutive avec un intérêt pour les approches de génomique des populations, la bioinformatique/biologie computationnelle et/ou la modélisation et les biostatistiques, ou (ii) de master de bioinformatique/biostatistiques avec un intérêt pour la biologie évolutive. Des connaissances dans au moins un langage informatique (python, R, dans une moindre mesure C et C++) et dans l'utilisation de cluster de calculs accessibles par des terminaux de commande seraient un plus.

Mots-clé

Inférence démographique, spéciation, flux de gènes, génétique des populations, modèles de coalescence, approximate Bayesian computation.

Références

Abbott R, Albach D, Ansell S, Arntzen JW, Baird SJE, Bierne N, Boughman J, Brelsford A, Buerkle C a, Buggs R, et al. 2013. Hybridization and speciation. J. Evol. Biol. 26:229–246.

Beaumont M. 2008. Joint determination of topology, divergence time, and immigration in population trees. In: Matsumura S, Forster P, Renfrew C, editors. Simulation, Genetics, and Human Prehistory. McDonald I. Cambridge Univ Press.

Csilléry K, Blum MGB, Gaggiotti OE, François O. 2010. Approximate Bayesian Computation (ABC) in practice. Trends Ecol. Evol. 25:410–418.

Csilléry K, Lemaire L, Français O, MGB Blum. 2015. Approximate Bayesian Computation (ABC) in R: A Vignette.

Hey J. 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol.

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158:885–896.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N. 2016. Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence.Moritz C, editor. PLOS Biol. 14:e2000234.

Strasburg JL, Rieseberg LH. 2010. How robust are "isolation with migration" analyses to violations of the im model? A simulation study. Mol. Biol. Evol. 27:297–310.

Références pertinentes des encadrants et du laboratoire d'accueil

Duvaux L, Belkhir K, Boulesteix M, Boursot P. 2011. Isolation and gene flow: Inferring the speciation history of European house mice. Mol. Ecol. 20:5248–5264.

Duvaux L, Geissmann Q, Gharbi K, Zhou J-J, Ferrari J, **Smadja CM**, **Butlin RK**. 2015. Dynamics of Copy Number Variation in Host Races of the Pea Aphid. Mol. Biol. Evol. 32:63–80.

Eyres I, **Duvaux L**, Gharbi K, Tucker R, Hopkins D, Simon J-C, Ferrari J, **Smadja CM**, **Butlin RK**. 2017. Targeted re-sequencing confirms the importance of chemosensory genes in aphid host race differentiation. Mol. Ecol. 26:43–58.

Gossmann TI, Shanmugasundram A, Börno S, **Duvaux L**, Lemaire C, Kuhl H, Klages S, Roberts LD, Schade S, Gostner JM, et al. 2018. The Response to Past Climate Perturbations Explains Extremely Low Genetic Diversity in the Genome of an Abundant Ice-Age Remnant, the Alpine Marmot. SSRN Electron. J.

Leroy T, Rougemont Q, Dupouey J-L, Bodenes C, Lalanne C, Belser C, Labadie K, Le Provost G, Aury J-M, Kremer A, et al [including **Plomion C**]. 2018. Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. bioRxiv 33.

Leroy T, Roux C, Villate L, Bodénès C, Romiguier J, Paiva JAP, Dossat C, Aury JM, **Plomion C**, Kremer A. 2017. Extensive recent secondary contacts between four European white oak species. New Phytol. 214:865–878.

Plomion C, Aury J, Amselem J, Leroy T, Murat F, Duplessis S, Faye S, Francillonne N, Labadie K, Le Provost G, et al. 2018. Oak genome reveals facets of long lifespan. Nat. Plants.

Zhou Y, **Duvaux L**, Ren G, Zhang L, Savolainen O, Liu J. 2017. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. Heredity (Edinb). 118:211–220.