Modélisation et inférence de l'évolution du génome accessoire d'*E. coli*

Les bactéries sont caractérisées par un grand taux de mutation et une grande taille de population. Elles présentent ainsi souvent une extraordinaire diversité génétique. De façon remarquable, la composition en gènes de ces espèces est également extrêmement diverse. Par exemple, un individu de l'espèce *Escherichia coli*, une bactérie commensale de l'intestin, compte environ 2000 gènes que l'on retrouve chez quasiment tous les individus : ces 2000 gènes forment le génome *core*. Un individu portera également environ 2000 autres gènes facultatifs, qui ne se retrouvent que dans une petite proportion des individus. Ces gènes forment le génome accessoire. L'ensemble de ce génome accessoire—le *pangénome*—compte des dizaines de milliers de gènes, dont beaucoup n'ont pas de fonction connue, et dont beaucoup ne sont trouvés que dans un individu (Kallonen et al. 2017; Touchon et al. 2009). Les gènes accessoires sont gagnés par transfert horizontal depuis d'autres individus d'*E. coli*, ou plus rarement, d'autres espèces.

La grande taille de population des micro-organismes ainsi que leur grande variabilité génétique doit impliquer que la sélection naturelle opère efficacement. Par exemple, des résistances aux traitements antimicrobiens peuvent évoluer à des échelles de temps de quelques années ou quelques dizaines d'années. Néanmoins, la structure de population pourrait limiter l'efficacité de la sélection dans certains cas. **Le but de ce projet de thèse est de développer de nouvelles méthodes mathématiques et statistiques pour détecter la sélection chez les micro-organismes et comprendre leur diversification, et d'appliquer ces méthodes à des génomes d'*E. coli* (Touchon et al. 2009; Massot et al. 2016).**

Nous nous intéresserons plus particulièrement à la sélection qui s'exerce sur le génome accessoire des bactéries. La modélisation de la dynamique des transferts de gènes accessoires chez les bactéries présente trois défis qui ne sont pas relevés par les méthodes classiques (Shapiro 2017). (i) Les gènes accessoires sont gagnés à un taux proportionnel à la taille du pool de donneurs potentiels. Ce pool de donneur ne compte pas l'intégralité des individus d'une espèce, car des souches distantes s'échangent moins de matériel génétique (Fraser, Hanage, and Spratt 2007). Des transferts de gènes entre espèces peuvent également se produire, bien que plus rarement. La dynamique de gain et de perte d'un gène accessoire est donc très différente de la dynamique de la substitution d'un nucléotide. (ii) Le taux de transfert horizontal est variable entre gènes, ce qui interdit les hypothèses simplificatrices classiques (0 transfert horizontal ou transfert horizontal très fréquent). (iii) Les populations bactériennes sont typiquement structurées par les hôtes qu'elles habitent, et présentent donc une dynamique en métapopulation avec extinction et recolonisation continues. La sélection s'exerce donc à l'échelle intra-hôte (adaptation à l'environnement de l'hôte, au système immunitaire, au traitement, etc.) et inter-hôte (adaptation pour une meilleure transmission et colonisation).

En dépit de sa complexité, ce système présente plusieurs avantages : des données massives de diversité (des milliers de gènes accessoires), de grandes tailles de population (ce qui peut simplifier les analyses mathématiques), la possibilité de détecter l'évolution dans des jeux de données collectés sur quelques années ou décennies (e.g. Kallonen et al. 2017). Enfin, un gène accessoire comporte des milliers de paires de bases, et donc en général un nombre suffisant de mutations pour documenter par une approche comparative l'histoire évolutive de ce gène.

Ce projet entend éclairer d'importantes questions biologiques : quels sont les processus responsables de la composition et de la diversité des génomes bactériens ? Comment l'adaptation et la diversification des bactéries dépendent-elles de la dérive génétique, des transferts horizontaux de gènes et de la sélection ? Les nouvelles méthodes mathématiques et statistiques développées permettront de répondre à ces questions en exploitant toutes les informations qu'offrent les centaines de génomes disponibles et les milliers de gènes qu'ils

portent. Comprendre ces dynamiques est aussi important du point de vue de la santé publique, car la plupart des gènes qui confèrent la virulence (capacité à infecter l'hôte) ou une résistance aux traitements antibiotiques sont des gènes accessoires.

Le projet comporte plusieurs volets interdépendants : conception et étude mathématique de modèles originaux de diversification des génomes, développement de méthodes d'inférence basée sur ces modèles, analyse de jeux de données de génomes d'*E. coli.* La doctorante ou le doctorant devra donc avoir une bonne maîtrise des outils classiques de modélisation probabiliste et bien sûr un intérêt profond pour l'étude des processus d'évolution moléculaire.

Tâche 1 : Dynamique et génétique des populations bactériennes
Une première étape de la thèse consiste à développer des modèles génériques de dynamique et de génétique des populations d'*E. coli* servant de brique de base pour les tâches suivantes.
La modélisation de la dynamique de ces populations devra prendre en compte leur forte structuration (confinement) à l'intérieur de leurs hôtes, qui forment ce que l'on appelle une métapopulation connectée, ainsi que le mode de structuration spatiale de cette métapopulation (transmissions au plus proche voisin) et le procédé d'échantillonnage des données.
La modélisation de la génétique de ces populations devra prendre en compte, en plus de la dynamique clonale décrite précédemment, les transferts de gènes par conjugaison entre souches co-existant dans un même hôte, ainsi que l'interaction entre les diverses échelles de sélection : intra-hôte (croissance, compétition), inter-hôte (transmission) et inter-espèces (diversification).
Des asymptotiques de grande population, de grande métapopulation et de faible taux d'échantillonnage permettront d'obtenir des modèles universels (on parle en probabilités de principe d'invariance) et pauvres en paramètres.

Tâche 2 : Dynamique et diversité neutre des génomes bactériens
Le but de la deuxième tâche est d'utiliser les modèles obtenus à la Tâche 1 pour faire des prédictions mathématiques sur la distribution de présence-absence des gènes accessoires dans un échantillon de bactéries et sur la diversité génétique observée à chacun de ces gènes en l'absence de sélection (modèles dits neutres).
Pour un seul gène accessoire, cette question généralise un problème classique en génétique des populations, celui de caractériser la diversité génétique d'un échantillon (par exemple) de cellules à un gène donné, c'est-à-dire le nombre de mutations observées sur ce gène et leurs fréquences respectives – cette question se résout classiquement à l'aide d'une représentation mathématique de la généalogie des cellules appelé coalescent. La difficulté est ici de modéliser le couplage (rendu complexe par les transferts horizontaux) entre le coalescent des bactéries séquencées et le coalescent du gène accessoire.
On cherchera à établir un cadre probabiliste permettant d'inférer si les origines des gènes accessoires par transfert sont multiples et plus ou moins distantes dans la phylogénie.

Tâche 3 : Développement de nouvelles méthodes de coalescent pour inférer la sélection et l'épistasie
La troisième tâche aura pour but de modéliser conjointement la dynamique des gènes du génome accessoire pour tirer parti des données massives dont nous disposons : la distribution de leurs effets sélectifs, leurs interactions épistatiques (on parle d'épistasie pour désigner la non-indépendance des effets sélectifs à différents gènes).
Pour contourner les écueils classiques de l'analyse statistique de grands jeux de données (fléau de la dimension, sur-paramétrisation, tests multiples), nous utiliserons une méthode alternative de modélisation inspirée d'un modèle original dû à Amaury Lambert et ses co-auteurs dans le contexte de la macro-évolution (Marin et al. 2019), se basant sur l'observation que le coalescent d'un gène sous sélection a des nœuds peu profonds et que deux gènes co-adaptés ont des coalescents similaires. Une manière simple de traduire cette idée mathématiquement est de supposer qu'en suivant les lignées de gènes dans le sens rétrospectif du temps, les lignées

porteuses d'un même allèle sous sélection coalescent ensemble à un taux plus élevé (attraction homologue) et que les lignées de gènes en épistasie positive sont attirées vers les mêmes hôtes (attraction non homologue). Il s'agira de confronter les prédictions de ces modèles et de les calibrer à celles faites par les modèles des étapes précédentes : parce que cette classe de modèles « backward » alternatifs proposent une manière parcimonieuse de coupler tous les gènes du génome accessoire, il sera possible de les utiliser pour exploiter pleinement la richesse des données génomiques disponibles.

Tâche 4 : Application de ces nouvelles méthodes aux données génomiques

La Tâche 4 s'appuiera sur les modèles et méthodes des tâches précédentes pour analyser un jeu de données généré récemment dans l'unité IAME d'Olivier Tenaillon. Il comprend 500 génomes d'*E. coli* échantillonnés dans la région parisienne, isolés à partir d'échantillons de selles de volontaires sains. Ces échantillons ont été prélevés de 1980 à 2010. Ce jeu de données est exceptionnel par son ampleur temporelle (et sera complété par des génomes de 2020) et par le fait qu'il représente la population commensale naturelle d'*E. coli*, et non des bactéries prélevées dans des infections (qui sont un échantillon biaisé de la population). Nous appliquerons à ce jeu de données les méthodes de vraisemblance naturellement associées aux modèles produits au cours des tâches 2 et 3.

La première étape de la Tâche 4 consistera à reconstruire le génome core et le génome accessoire. L'examen de la distribution de présence-absence des gènes accessoires et de leur diversité génétique permettra, par contraste avec le modèle neutre de la Tâche 2, de déduire la présence de sélection s'exerçant sur les gènes accessoires et le mode prédominant de la sélection (négative ou positive). La deuxième étape consistera à inférer l'arbre phylogénétique du génome *core*, qui représente l'histoire évolutive de l'espèce d'*E. coli*, ainsi que les arbres phylogénétiques individuels pour chaque gène du génome accessoire suffisamment représenté dans le jeu de données. Ces arbres phylogénétiques permettront d'inférer le nombre d'origines indépendantes de chaque gène (par transfert depuis une espèce plus ou moins distante) et de dater ces évènements. Enfin, grâce aux méthodes développées dans la Tâche 3, il sera possible, en analysant conjointement les arbres phylogénétiques des différents gènes accessoires, d'estimer la pression de sélection qui s'exerce sur chaque gène accessoire et d'inférer les ensembles de gènes qui interagissent épistatiquement. Une attention particulière sera portée aux gènes de virulence ou conférant une résistance à un antibiotique

Fraser, Christophe, William P Hanage, and Brian G Spratt. 2007. "Recombination and the Nature of Bacterial Speciation." *Science* 315 (5811): 476–480.

Kallonen, Teemu, Hayley J. Brodrick, Simon R. Harris, Jukka Corander, Nicholas M. Brown, Veronique Martin, Sharon J. Peacock, and Julian Parkhill. 2017. "Systematic Longitudinal Survey of Invasive Escherichia Coli in England Demonstrates a Stable Population Structure Only Transiently Disturbed by the Emergence of ST131." *Genome Research* 27 (8): 1437–1449.

Marin, Julie, Guillaume Achaz, Anton Crombach, and Amaury Lambert. 2019. "The Genomic View of Diversification." *BioRxiv*, May, 413427. In revision for *J Evol Biol*

Massot, Méril, Anne Sophie Daubié, Olivier Clermont, Françoise Jauréguy, Camille Couffignal, Ghizlane Dahbi, Azucena Mora, et al. 2016. "Phylogenetic, Virulence and Antibiotic Resistance Characteristics of Commensal Strain Populations of Escherichia Coli from Community Subjects in the Paris Area in 2010 and Evolution over 30 Years." *Microbiology (United Kingdom)* 162 (4): 642–650. https://doi.org/10.1099/mic.0.000242.

Shapiro, B. Jesse. 2017. "The Population Genetics of Pangenomes." *Nature Microbiology* 2 (12): 1574–1574. https://doi.org/10.1038/s41564-017-0066-6.

Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, and Odile Bouvet. 2009. "Organised Genome Dynamics in the Escherichia Coli Species Results in Highly Diverse Adaptive Paths." *PLoS Genetics* 5 (1): e1000344.