Modèles stochastiques pour l'évolution moléculaire microbienne

Jasmine Gamblin

École Doctorale 227
Sciences de la nature et de l'Homme : évolution et écologie

Thèse de doctorat

Discipline : Biologie des Populations

Modèles stochastiques pour l'évolution moléculaire microbienne

---

Présentée pour obtenir le grade de
Docteure de Sorbonne Université par

Jasmine Gamblin

dirigée par
Amaury Lambert et François Blanquart

Présentée et soutenue publiquement le 28 juin 2024 devant un jury composé de :

Vincent Daubin	Université Lyon I	Rapporteur
Guillaume Martin	Université de Montpellier	Rapporteur
Claude Loverdo	Sorbonne Université	Présidente du jury
Céline Scornavacca	Université de Montpellier	Examinatrice
Marie Touchon	Institut Pasteur	Examinatrice
Amaury Lambert	École Normale Supérieure	Directeur de thèse
François Blanquart	Collège de France	Directeur de thèse

Résumé

Cette thèse se compose de deux parties, dont le point commun est de développer des modèles stochastiques pour étudier l'évolution microbienne.

La première partie s'intéresse aux impacts évolutifs des goulots d'étranglement de population. Le premier chapitre concerne les populations microbiennes subissant des goulots d'étranglement réguliers. Cette dynamique correspond aux expériences d'évolution en laboratoire utilisant des dilutions en série, qui permettent de renouveler le milieu dans lequel croissent les bactéries, mais aussi aux bactéries associées à des hôtes, qui subissent un goulot d'étranglement lors de la colonisation de chaque nouvel hôte. L'effet des dilutions successives sur l'évolution de ces populations n'est pas encore complètement compris. Deux effets ont déjà été mis en avant : le fait qu'atteindre une grande taille de population favorise l'apparition de mutations rares, et la perte de mutations bénéfiques lors des dilutions. Nous avons développé un modèle semi-déterministe afin d'étudier comment ces deux phénomènes se combinent et impactent les chemins évolutifs suivis par la population et sa vitesse d'adaptation, dans un paysage adaptatif minimal comportant une mutation fréquente mais à effet positif faible, et une mutation rare à effet fort. Notre modèle permet aussi de découpler les rôles de deux paramètres importants que sont la taille de population initiale et la taille de population maximale de la souche initiale. Nous trouvons qu'une faible dilution et un cycle bref favorisent l'adaptation par la mutation bénéfique la plus fréquente, tandis qu'une forte dilution et un cycle long favorisent l'adaptation par la mutation la plus avantageuse. Nous calculons aussi le taux de dilution maximisant la vitesse d'évolution.

Le deuxième chapitre est une revue de littérature sur l'effet des goulots d'étranglements sur la capacité d'adaptation des populations, comparant populations microbiennes et animales. Nous suggérons des pistes pour réaliser des expériences microbiennes plus à même d'informer la biologie de la conservation, notamment en utilisant des levures et en incluant de la diversité génétique initiale.

La deuxième partie s'intéresse à des échelles spatiales et temporelles plus larges, en développant un modèle stochastique de la dynamique macroévolutive des gènes dans une espèce bactérienne. De nombreuses espèces bactériennes présentent une diversité de gènes impressionnante : le nombre de gènes présents dans l'espèce est fréquemment bien plus grand que le nombre de gènes présents dans une cellule, ce qui a amené à l'introduction du concept de pangénome. Caractériser les dy-

namiques d'importation et de transfert horizontal de gènes bactériens est crucial pour comprendre l'origine de cette formidable diversité, et est aussi d'une grande importance en santé publique pour étudier la diffusion des gènes de virulence et d'antibiorésistance. Plusieurs modèles ont été formulés pour décrire la dynamique de gènes bactériens au sein de la phylogénie mais demeurent insatisfaisants, car ils prennent en compte soit l'import des gènes dans l'espèce focale (par transfert depuis une autre espèce) soit les transferts horizontaux au sein de l'espèce, mais jamais les deux. Nous avons donc développé un modèle d'évolution de gènes bactériens prenant en compte à la fois le transfert inter- et intra-espèce. Notre modèle comporte trois types de dynamiques : gènes persistants hérités de l'ancêtre commun, gènes privés spécifiques à un clade, et gènes mobiles subissant des transferts fréquents. Nous avons testé ce modèle sur un ensemble de génomes de *Salmonella enterica*, et montré qu'il est capable de reproduire des caractéristiques importantes des pangénomes comme la forme en U de la distribution des fréquences de gènes. Ce modèle permet une classification des gènes en fonction de leur type de dynamique, dont la pertinence biologique est confirmée par l'analyse de la fonction et de la localisation des gènes assignés à chaque type.

Mots-clés : modélisation stochastique ; microbiologie ; goulot d'étranglement ; évolution expérimentale ; macroévolution ; pangénome ; transfert horizontal.

Abstract

This dissertation is composed of two independent parts, sharing the common goal of developing stochastic models to study microbial evolution.

The first part investigates the evolutionary impacts of population bottlenecks. Chapter one deals with microbial populations undergoing periodic bottlenecks. These dynamics correspond in particular to laboratory evolution experiments using serial dilutions to renew the medium in which bacteria grow. These wide variations in population sizes are also found in natural microbial populations, which undergo a bottleneck when transferred to a new host. However, the effects of successive dilutions on the evolution of these populations are not yet fully understood. Two effects have already been brought to light: reaching a large population size favors the appearance of rare mutations, and the loss of beneficial mutations during dilutions. We have developed a semi-deterministic model to study how these two phenomena combine and impact the evolutionary paths followed by the population as well as the adaptation rate on a minimal fitness landscape consisting of two types of beneficial mutations with the empirically supported trade-off between mutation rate and fitness advantage. Our model decouples the effect of two important parameters: the initial population size and the maximum population size of the initial strain. We find that low dilution and short cycles favor adaptation by the most frequent beneficial mutation, while strong dilution and long cycles favor adaptation by the most advantageous mutation. We also calculate the dilution rate maximizing the rate of evolution.

The second chapter is a literature review on the effect of bottlenecks on the adaptive potential of populations, with a comparison between microbial and animal populations. We suggest ways in which microbial experiments could better inform conservation biology, notably by using yeast and including initial standing genetic variation.

The second part looks at larger spatial and temporal scales, developing a stochastic model for macroevolutive gene dynamics in a bacterial species. Indeed, many bacterial species present an impressive gene diversity: the number of genes present in the species is frequently much greater than that carried by one cell, leading to the introduction in 2005 of the "pangenome" concept. Characterizing the dynamics of bacterial gene importations and horizontal transfers is crucial to understand the origins of this formidable diversity. These dynamics are also of great importance

for public health, as they determine the spread of virulence and antibiotic resistance genes within bacterial populations. Over the past 10 years, several models have been formulated to describe the dynamics of bacterial genes along a phylogeny, but they remain unsatisfactory. These models take into account either the importation of genes into the focal species (by transfer from another species) or horizontal transfers within the species, but never both. We have therefore developed an original bacterial gene evolution model to take into account both inter- and intra-species transfer. Our model relies on three types of dynamics: persistent genes inherited from the ancestor, private genes that are clade-specific, and mobile genes undergoing frequent transfers. We have tested this model on a set of *Salmonella enterica* genomes, and shown that it is able to reproduce some important features of pangenomes, such as the U-shape of the gene frequency distribution and the parsimony of their arrangement on the core genome phylogeny. This model is able to classify genes according to their most likely dynamics, and the biological relevance of this classification has been confirmed by analyzing the function and position of genes assigned to each type.

Keywords: stochastic modeling; microbiology; bottlenecks; experimental evolution; macroevolution; pangenome; horizontal gene transfer.

Remerciements

Je remercie en premier lieu les deux rapporteurs de cette thèse, Vincent Daubin et Guillaume Martin. Merci d'avoir pris le temps de lire et de commenter ce manuscrit un peu hétéroclite. À vous deux et à mes examinatrices Claude Loverdo, Céline Scornavacca et Marie Touchon, merci pour l'intérêt que vous avez porté à mon travail et pour les échanges stimulants et chaleureux lors de ma soutenance.

Ensuite, bien sûr, merci infiniment à Amaury et François. Vous m'avez guidée durant cette thèse, et soutenue quand le sommet de la montagne me semblait trop loin. Merci pour votre confiance, pour nos discussions scientifiques fructueuses, pour votre gentillesse, votre inventivité, votre humour et votre sérieux aussi. Je trouve que vous formez un duo qui fonctionne très bien, tant du point de vue scientifique qu'humain. Merci Amaury pour ton côté fin psychologue, ton exigence et le partage de tes préoccupations environnementales. Je ne pensais pas rencontrer un jour un homme de plus de 45 ans aussi woke ! François, merci pour ton optimisme, ta décontraction et ta disponibilité. Bizarrement on ne demande pas leur avis aux doctorants pour décerner une HDR, mais en tous cas je l'aurais validée ;)

Merci aux membres de mon comité de suivi – Bastien Bousseau, Violaine Llaurens et Marie Touchon – pour vous être penchés avec bienveillance et intérêt sur le déroulé de ma thèse. Les réunions avec vous m'ont a chaque fois éclairée scientifiquement et donné un peu plus confiance en mon travail. Un grand merci également à Karine Dubrana que j'ai rencontrée grâce à un programme de mentorat Femmes & Sciences. Nos longues discussions dans divers bars et cafés m'ont aidée à prendre du recul sur ma thèse et à me projeter dans une carrière de chercheuse. Merci pour ces soirées très sympathiques !

Merci à mes collaborateurices, en premier lieu Sylvain Gandon qui a participé à la genèse et à l'encadrement du projet présenté au Chapitre 2. Puis merci à Laure Olazcuaga et Loïc Marrec, ça valait le coup d'aller à une conférence internationale pour rencontrer deux français intéressés par les bottlenecks ^^ C'était très sympa et enrichissant de travailler avec vous sur la revue du Chapitre 3 (oui je sais, c'est pas terminé ...). J'espère que vous verrez bientôt votre talent récompensé par un poste !

Merci à mes parents, Mme. Gabriel, Cédric Villani et M. Bray de m'avoir donné envie de faire des maths, et à Olivier Tenaillon de m'avoir fait découvrir la biologie évolutive quand j'en avais marre des maths fonda.

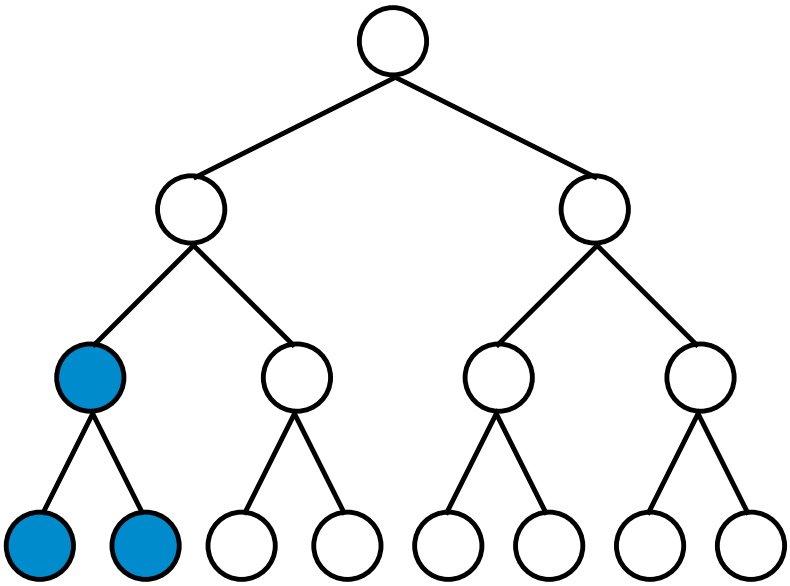
Merci à toustes les membres de cette chouette équipe dans laquelle j'ai eu la chance de passer 4 ans et demi, vous m'avez donné le SMILE chaque jour en arrivant au labo ;) Un merci spécial à Hélix et Félise qui m'ont tout appris de la vie de doctorante, et à Gilou et Thomas avec qui on a vécu presque toute l'aventure ensemble. Merci à Élisa, Augustin, Aurore, Thibaut, Philibert, Mathilde, Nina, Abdel, Adélie, Jean-Jil, Maxime, Pete, Alex, Guillaume (et d'autres !), pour cette ambiance bienveillante et détendue, merci pour les Skribbl du Covid, les SH, les séances d'escalade et les manifs. J'espère vous revoir régulièrement ! Merci aussi aux autres collègues du CIRB et du Collège de France, notamment à l'équipe de gestion du CIRB et au collectif mobilisation-cdf. Merci à Lucile avec qui on s'est suivies de loin.

Enfin, un grand merci à mon entourage qui m'a offert un soutien précieux pendant cette période. À mes amies Castors : Panda, Marie, Fanny et Solène (et les non-franciliennes Julia et Loïse), avec qui on partage de nombreux centres d'intérêt et aventures depuis le collège, et depuis quelques années une certaine radicalisation éco-féministe. À la section escalade, qui a rendu mon passage à l'X assez chill, en particulier Anouk, Hakim mon partenaire de VOS, Mimou et Roché pour les randos MUL, Robin, Alfred, Maxime... À Miléna et ses questions, à Lou la future québécoise, aux amis que j'ai connus en partie par Corentin : Flavien, Félix, Théo, Céline, Clémence. À mes parents, qui nous ont appris avec mon frère à nous émerveiller devant la nature. À mon frère, à toute ma famille : les montpelliérains, les scéens, les perpignanais, et à la famille de Corentin. Je suis très émue d'avoir vu autant d'entre vous à ma soutenance ou en visio, pour partager un peu mon travail de ces dernières années.

Et bien sûr, merci à Corentin. Je ne vais pas lister ici toutes les raisons pour lesquelles la vie est belle avec toi, mais en particulier merci pour ton soutien sans faille durant ma thèse !

Table des matières

Chapitre 1



1

Introduction

Sommaire

Cette introduction présente un tour d'horizon du champ d'étude de la microbiologie évolutive. Cet aperçu ne se veut pas exhaustif, mais expose les motivations de cette thèse et introduit les principaux concepts et outils utilisés dans le reste de ce manuscrit.

1.1 Pourquoi étudier l'évolution microbienne ?

1.1.1 Introduction des concepts

Micro-organismes. Il est généralement considéré que le terme micro-organisme ou microbe désigne les organismes unicellulaires des trois grands domaines de la vie : bactéries, archées, et eucaryotes unicellulaires (incluant des champignons comme les levures, des plantes comme les algues unicellulaires, et les protistes). On inclut aussi dans ce champ d'étude les virus, bien qu'ils ne soient pas toujours considérés comme des êtres vivants à proprement parler en raison de leur incapacité à se reproduire en dehors d'une cellule hôte. Les micro-organismes sont les premières formes de vie a être apparues il y a environ 3,5 milliards d'années, et composent actuellement environ 15% de la biomasse terrestre et océanique (Bar-On, Phillips et Milo 2018).

Mécanismes évolutifs de base. Ces organismes évoluent comme tout le vivant sous l'effet combiné de plusieurs mécanismes. Les mutations – qui sont des modifications aléatoires et accidentelles du matériel génétique, par exemple dues à des erreurs de réplication de l'ADN – apportent de la diversité héritable entre les individus. La sélection naturelle influe sur la fréquence des différents variants dans la population en fonction de l'avantage ou du désavantage qu'ils confèrent à leur porteur en termes de nombre de descendants. La dérive génétique induit des variations aléatoires dans la fréquence des variants, c'est-à-dire qui ne sont pas liées à leur avantage sélectif mais simplement à la stochasticité du processus démographique (naissances, morts) dans une population de taille finie.

Reproduction sexuée et asexuée. Les procaryotes (bactéries et archées) se reproduisent de manière asexuée, c'est-à-dire par division d'un individu en deux individus au génome identique (en dehors des erreurs de réplication). Les eucaryotes unicellulaires peuvent en général alterner entre reproduction sexuée et asexuée en fonction des conditions de vie. La reproduction sexuée implique la formation d'un nouvel individu par la fusion de deux gamètes provenant d'individus de types sexuels différents, le génome de ce nouvel individu étant donc composé à parts égales de matériel génétique provenant des deux parents. Ces deux modes de reproduction impliquent des modes d'évolution différents entre populations sexuées et asexuées, notamment du fait que la reproduction sexuée induit du brassage génétique et donc plus de diversité. Cependant, d'autres mécanismes apportant de la diversité génétique existent chez les procaryotes, comme les transferts horizontaux de gènes entre individus (détaillés Section 1.1.3).

Table 1.1 – Comparaison de trois organismes modèles

Espèce	Temps de génération	Taille du génome	Taux de mutation	Taille de pop. exp. typique	Nb. mut. /site/10j	Nb. génomes sur Genbank
Escherichia coli	20min	5Mb	0.26×10^{-9}	10^9	2×10^2	>260 000
Saccharomyces cerevisae	100min	12Mb	0.33×10^{-9}	10^7	0.5	1618
Drosophila melanogaster	10j	180Mb	4.65×10^{-9}	10^3	5×10^{-6}	118

1.1.2 Un modèle idéal pour étudier les mécanismes évolutifs

Bien que ces organismes microscopiques soient difficiles à observer, ils possèdent des caractéristiques qui rendent leur évolution très facile à quantifier en laboratoire. Cela en fait donc un terrain d'étude extrêmement fertile pour étudier les mécanismes évolutifs (Kawecki et al. 2012 ; McDonald 2019). Les deux caractéristiques qui font des microbes un modèle intéressant sont leur évolution rapide et leur facilité d'utilisation.

Une évolution observable. Le taux de mutation d'un organisme dépend de la fidélité de l'ADN polymérase lors de la réplication de son ADN, ainsi que des mécanismes de correction d'erreurs existants. Les taux de mutation par site nucléotidique par génération sont assez similaires entre les espèces, avec des taux un peu plus élevés chez les organismes pluricellulaires ($\sim 10^{-8}$) que chez les unicellulaires ($\sim 10^{-9}$, toutes les estimations de taux de mutations sur cette page sont tirées de Lynch 2010). Cependant, le temps de génération très court des espèces microbiennes font que leurs taux de mutation par unité de temps sont assez élevés. Le fait que l'on puisse cultiver des populations de grande taille en laboratoire est également un facteur important, car une grande taille de population augmente la probabilité de détecter une mutation d'intérêt. Par exemple chez *Escherichia coli*, le taux de mutation vaut 0.26×10^{-9} par réplication par site nucléotidique et le temps de génération et d'environ 20 min dans des conditions optimales, donc en observant une population de 10^9 individus on obtiendra au bout d'une dizaine de jours une centaine de mutations à un site d'intérêt. Par comparaison, une population de drosophiles aura dans le même temps effectué une seule génération, et on pourra espérer avec un millier d'individus de l'ordre de 10^{-6} mutations sur un site donné (Table 1.1). Les virus présentent des taux de mutation encore plus élevés, notamment les virus à ARN dont la polymérase n'a pas la capacité de réparer ses erreurs.

Praticité d'utilisation. L'autre avantage des micro-organismes est leur compacité : leur petite taille (de l'ordre du micron pour les procaryotes, de la dizaine de microns pour les eucaryotes) permet de cultiver des populations de très grande taille dans un espace restreint. Par exemple, on peut facilement obtenir une concen-

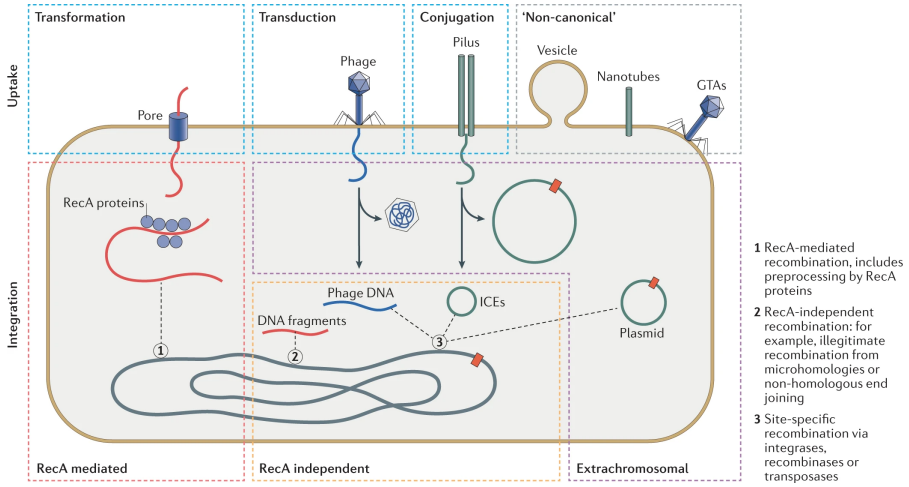


Figure 1.1 – Aperçu des principaux mécanismes de transfert et d'intégration d'ADN chez les bactéries (illustration tirée de B. J. Arnold, Huang et Hanage 2022 et reproduite avec l'accord de Springer Nature).

tration en *E. coli* de 10^9 ufc/ml (ufc signifie « unité formant colonie » et traduit le fait que la méthode couramment utilisée pour compter les cellules vivantes dans un échantillon est de les cultiver sur gélose après dilutions et de compter le nombre de colonies, Son et Taylor 2021). Les espèces couramment utilisées pour l'évolution expérimentale présentent encore d'autre avantages : faciles à élever, à conserver (congélation, déshydratation). Et il existe des méthodes permettant de quantifier facilement certaines de leurs propriétés : estimation du nombre d'individu par densité optique (Beal et al. 2020), cytométrie en flux pour compter et trier les cellules selon certains paramètres morphologiques (taille, quantité d'ADN, présence d'antigènes spécifiques, etc.).

D'autre part, la compacité des génomes microbiens (Table 1.1) rend leur séquençage rapide et bon marché. Il est donc possible de séquencer des dizaines d'individus pour les besoins d'une expérience, et d'avoir accès à de gros jeux de données pour étudier l'évolution de certaines espèces. Dans la base de données en ligne Genbank maintenue par le NCBI (Centre national pour l'information biotechnologique américain), il y a actuellement plus de 260 000 génomes d'*Escherichia coli* disponibles. Durant l'épidémie de SARS-Cov-2, plusieurs millions de coronavirus ont pu être séquencés et leur évolution analysée en direct par des plateformes telles que Nextstrain (Hodcroft et al. 2021).

1.1.3 Processus évolutifs chez les micro-organismes

Malgré cette apparente simplicité d'utilisation en laboratoire, l'évolution des micro-organismes – notamment en milieu naturel – comporte en fait des mécanismes assez complexes que l'on est encore bien loin de comprendre entièrement. J'en passe quelques-uns en revue dans cette section.

Transferts horizontaux de gènes. Chez les procaryotes, l'échange de matériel génétique entre individus est un phénomène courant, et assez surprenant pour qui est habitué à la génétique des eucaryotes pluricellulaires (chez qui les gènes se transmettent uniquement aux descendants). Il en existe trois types principaux :

1. La conjugaison est l'échange direct de matériel génétique entre deux individus à travers un pilus (structure extracellulaire en forme de tube).
2. La transformation est l'intégration par une cellule de matériel génétique issu du milieu environnant, en général des fragments d'ADN de cellules mortes.
3. La transduction est le transfert de matériel génétique médié par un phage (virus de bactérie), dû à une erreur lors de l'excision d'un prophage (génome de phage intégré dans le génome de l'hôte) ou lors de l'empaquetage du génome viral.

Le processus ne s'arrête cependant pas à l'échange, puisqu'il faut également que les fragments d'ADN transférés puisse être exprimés et répliqués dans leur nouvel hôte. Soit le fragment échangé est un plasmide, c'est-à-dire une molécule d'ADN capable de réplication autonome, soit il parvient à s'insérer dans l'ADN chromosomique de l'hôte grâce à la présence d'intégrase, recombinase ou transposase (si c'est un élément génétique mobile), ou alors par un événement de recombinaison homologue ou non. Une recombinaison homologue peut avoir lieu lorsque les séquences flanquant le gène transféré correspondent à des séquences similaires du génome hôte. Cela est donc d'autant plus fréquent que les cellules donneuses et receveuses sont proches génétiquement. En cas de transfert et remplacement d'un gène déjà présent dans le génome hôte, on parle parfois plutôt de conversion génique que de transfert horizontal. Ces mécanismes sont illustrés Figure 1.1.

Chez les eucaryotes unicellulaires, les transferts horizontaux existent mais semblent beaucoup moins fréquents. On estime par exemple qu'environ 1% du répertoire génétique des protistes provient d'événements de transfert (van Etten et Bhattacharya 2020).

Diversité génétique. Ces transferts de gènes sont pour de nombreuses espèces bactériennes vecteurs d'une formidable diversité génétique. En particulier, l'acquisition fréquente de nouveaux gènes provenant d'autres individus de la même espèce ou même d'autres espèces entraîne de grandes différences en termes de répertoire de gènes entre individus d'une même espèce. Cela a mené à la création du terme pangénome pour désigner l'ensemble des gènes présents dans une espèce (Tettelin, Masignani et al. 2005). Chez *E. coli*, l'étude de Touchon, Perrin et al. (2020) portant sur 1300 souches a dénombré plus de 75 000 familles de gènes dans le pangénome, 16 fois plus que le nombre moyen de gènes dans le génome d'un individu (qui est d'environ 4700). Les gènes portés par la majorité des individus, appelé gènes persistants, ne représentent que la moitié d'un génome typique. L'autre moitié correspond à des gènes appelés accessoires, dont le contenu est hautement variable entre individus (Figure 1.2).

Cette grande variabilité entre individus d'une même espèce et la possibilité de transferts de gènes inter-espèce posent la question de la définition d'une espèce bac-

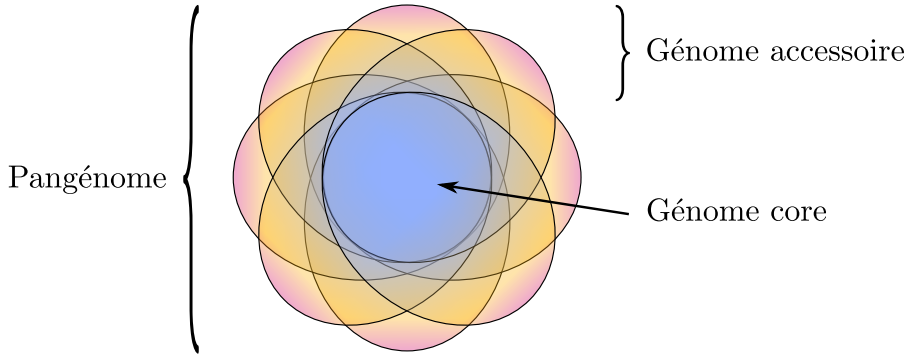


Figure 1.2 – Représentation schématique d'un pangénome bactérien (adaptée de McInerney, McNally et O'Connell 2017). Le génome *core* est l'ensemble des gènes communs à tous les individus de l'échantillon. Les gènes accessoires sont présents chez certains individus seulement. Le pangénome est l'ensemble de tous les gènes présents dans l'échantillon.

térienne (Fraser, Hanage et Spratt 2007). D'un autre côté, la recombinaison entre individus d'une même espèce est au contraire une force qui maintient la cohésion génétique de l'espèce en empêchant certains clones de diverger indéfiniment. Dixit, Pang et Maslov (2017) proposent par exemple un modèle permettant d'estimer le caractère métastable ou divergent d'une population bactérienne en fonction de la fréquence respective des mutations et des événements de recombinaison. Une étude de Sakoparnig, Field et van Nimwegen (2021) suggère d'ailleurs que les arbres phylogénétiques d'espèces bactériennes reflètent la structure de l'espèce en termes de taux de recombinaison (qui recombine avec qui) plutôt que de relations clonales (qui descend de qui). Mais comme la fréquence de recombinaison entre deux souches est supposée dépendre en partie de leur distance génétique, ces deux visions se recoupent probablement. Cette question est discutée à la fin de cette thèse en Section 5.4.2.

Mécanismes de sélection. Une autre question fondamentale est d'identifier les forces évolutives qui régissent la distribution des gènes dans une espèce. Il y a bien sûr de la sélection due aux changements environnementaux. Les gènes conférant des résistances aux antibiotiques par exemple ont augmenté en fréquence en réaction à l'utilisation de ces molécules, avant de se stabiliser à une fréquence intermédiaire. Certains traits peuvent être sélectionnés pour s'adapter à une niche, par exemple à un certain hôte. Chez les bactéries commensales humaines, des gènes de virulence peuvent augmenter en fréquence car ils sont associés à une persistance plus longue dans l'intestin et pourraient donc être sélectionnés (Östblom et al. 2011).

Mais il existe aussi d'autres types de sélection plus complexe, comme suggéré par la revue de Domingo-Sananes et McInerney (2021) qui indique que de multiples interactions sont susceptibles d'influer sur l'évolution des pangénomes bactériens. Ils citent notamment l'épistasie, c'est-à-dire les interactions entre gènes d'un même génome, qui pourrait être un facteur déterminant pour qu'un gène puisse être conservé dans un génome (Whelan, R. J. Hall et McInerney 2021). Les interactions sociales dans les communautés microbiennes jouent aussi un rôle. Par

exemple, il peut être avantageux de perdre un gène permettant de sécréter un métabolite utile (comme les sidérophores qui permettent de piéger le fer) si d'autres le produisent (Ross-Gillespie et al. 2007). On peut ajouter les interactions avec les phages, qui ont mené à l'émergence de divers mécanismes de défense dont le fameux système CRISPR-Cas (Georjon et Bernheim 2023), et pourraient jouer un rôle important dans le maintient d'une grande diversité génétique dans les populations bactériennes (Rodriguez-Valera et al. 2009).

L'importance relative d'une dynamique neutre de gains et pertes de gènes par rapport à ces différents mécanismes de sélection est le sujet d'un débat, que j'aborde en discussion de cette thèse Section 5.4.1.

1.1.4 Applications à la santé humaine

Une autre motivation majeure dans l'étude des micro-organismes est l'enjeu pour la santé humaine.

Pathogènes. Les pathogènes humains sont majoritairement des micro-organismes, et les maladies infectieuses qu'ils provoquent sont une cause importante de décès dans le monde. L'étude Global Burden of Disease de 2019 rapporte 13,7 millions de décès liés à une maladie infectieuse (Ikuta et al. 2022). Il est donc utile d'étudier les micro-organismes pathogènes pour mieux comprendre ces maladies, les prévenir et les traiter. Leur évolution rapide rend aussi nécessaire de comprendre leurs mécanismes évolutifs, pour être capable par exemple de reformuler chaque année le vaccin pour la grippe ou de détecter rapidement un variant plus virulent. En particulier, l'augmentation du nombre de bactéries portant des gènes de résistance aux antibiotiques est très inquiétant. On estime que ce phénomène est lié à plus d'un million de décès par an dans le monde actuellement (Murray et al. 2022).

Microbiote. Cependant, la plupart des micro-organismes ne sont pas pathogènes. Même une bactérie pathogène majeure telle que *Escherichia coli* est en fait la plupart du temps commensale (c'est-à-dire qu'elle tire des bénéfices de son hôte sans lui nuire). Se pencher sur son mode de vie dans l'intestin humain est donc susceptible d'éclairer l'évolution de certains traits favorisant les infections (Tenaillon et al. 2010). Plus généralement, les communautés bactériennes qui composent le microbiote humain sont de plus en plus étudiées à mesure qu'un nombre croissant d'études mettent en évidence des corrélations entre l'état du microbiote et une bonne santé physique et mentale (Valles-Colomer et al. 2019 ; Hou et al. 2022). Les dynamiques évolutives de ces communautés et les liens de causalité avec différentes pathologies sont encore peu connus et sont un champ de recherche en pleine expansion.

Usages thérapeutiques. Les microbes peuvent même être utilisés dans le cadre de thérapies. Je vais en donner cinq exemples, dont les trois premiers appartiennent à la lutte biologique. Un espoir face à la montée du phénomène de résistance aux

antibiotiques est notamment d'utiliser des parasites naturels des bactéries, les bactériophages ou phages. L'avantage par rapport aux substances antibiotiques est que les phages peuvent évoluer pour contourner les systèmes de défense des bactéries, dans une perpétuelle course aux armements évolutive (Froissart et Brives 2021). Un autre parasite, cette fois du moustique, est utilisé pour contrôler les tailles de populations de moustiques *Aedes aegypti* et *Aedes albopictus* et limiter ainsi les cas d'infection par le virus de la dengue dans les populations humaines (virus de la famille des *Flaviviridae*, dont ces moustiques sont vecteurs). Ces bactéries du genre *Wolbachia* imposent en effet un coût sélectif à leurs porteurs, et réduisent leur probabilité d'être vecteurs de ce virus (Utarini Adi et al. 2021). Les modèles de diffusion de ces bactéries dans une population de moustiques permettent d'éclairer les stratégies d'introduction optimales (Dorigatti et al. 2018). Enfin, la transplantation fécale d'un donneur sain permet de lutter contre les infections à *Clostridium difficile* (Cammarota, Ianiro et Gasbarrini 2014).

D'autre part, certaines bactéries sont utilisées dans le cadre de traitements expérimentaux contre le cancer, pour leur capacité à délivrer des substances anticancéreuses de manière ciblée ou à stimuler la réponse immunitaire au niveau des cellules cancéreuses (Duong et al. 2019). Des bactéries génétiquement modifiées sont aussi utilisées pour la production de métabolites d'intérêt entrant dans la composition de médicaments, tels que les antibiotiques ou l'insuline. L'optimisation des propriétés de ces molécules peut être obtenue par des procédures d'évolution expérimentale dirigée (Chartrain et al. 2000).

J'ai évoqué ici les enjeux de l'étude de l'évolution microbienne pour la santé humaine, mais ces organismes sont présents dans tous les types d'environnements et ont une grande importance dans bien d'autres domaines tels que l'agriculture (bactéries du sol, pathogènes des plantes), ou même simplement pour leur rôle primordial dans le fonctionnement des écosystèmes (fixation du carbone par les cyanobactéries, de l'azote par les mycorhizes, biodégradation).

1.2 Approches quantitatives en microbiologie évolutive

J'ai exposé dans la partie précédente plusieurs motivations pour l'étude de l'évolution microbienne. Il est temps à présent de nous pencher sur les méthodes employées pour cette étude. Je me concentre ici sur les approches quantitatives, qui sont rendues naturelles par la simplicité d'expérimentation et les grandes tailles de population des micro-organismes, et dont certaines sont étudiées ou utilisées dans cette thèse.

1.2.1 Approches expérimentales

Cultiver des micro-organismes en laboratoire est l'une des méthodes les plus fondamentales en microbiologie. Cela permet d'étudier certaines de leurs caractéris-

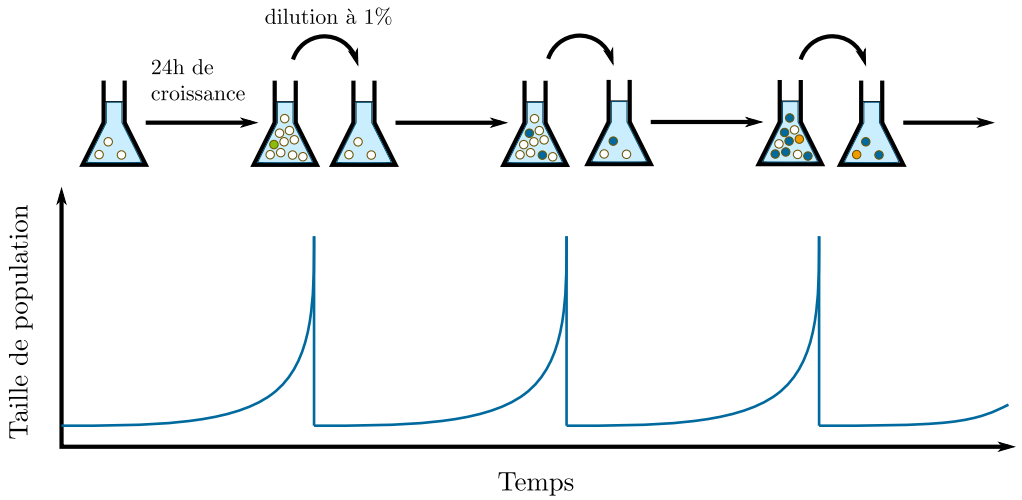


Figure 1.3 – Représentation schématique d'une expérience d'évolution microbienne avec 3 passages en série, montrant la dynamique de la taille de population et l'apparition de mutations (individus colorés).

tiques (taux de croissance, métabolisme, sensibilité aux agents antimicrobiens) ainsi que l'évolution de celles-ci dans un environnement contrôlé, stable ou changeant (addition de substances antibiotiques, modification du milieu nutritif, co-évolution avec des phages).

Passages en série. Une technique couramment utilisée en évolution microbienne expérimentale est celle des passages en série (*serial passage* ou *serial transfer* en anglais). Cette technique consiste à transférer périodiquement une fraction d'une culture microbienne dans un nouveau milieu de culture frais. Le transfert est généralement effectué à la fin de la phase de croissance exponentielle ou au début de la phase stationnaire. Cela permet aux micro-organismes de continuer à se multiplier et à évoluer tout en limitant la taille totale de la population. Ce processus peut être répété de nombreuses fois, parfois sur des milliers de générations, pour observer les changements évolutifs qui se produisent au cours du temps (Figure 1.3).

Exemple : Long-Term Evolution Experiment. Une expérience emblématique utilisant des passages en série est la Long Term Evolution Experiment (LTEE), initiée par l'équipe de Richard Lenski. Elle consiste à cultiver en parallèle 12 populations d'*Escherichia coli* dans des conditions identiques et constantes, en utilisant le glucose comme principale source de carbone, pour étudier leur adaptation sur de nombreuses générations. Cette expérience se démarque en effet par sa durée exceptionnelle : ces populations ont été suivies (presque) sans interruptions depuis 1988, et ont atteint 77 000 générations en 2023. Des échantillons de chaque population sont congelés à intervalles réguliers, permettant de revenir en arrière et de comparer les bactéries actuelles avec leurs ancêtres. Deux objectifs principaux de la LTEE étaient d'étudier la vitesse d'adaptation et la répétabilité de l'évolution.

L'expérience a montré que l'évolution de la valeur sélective moyenne de ces populations est mieux prédite par une fonction puissance que par une fonction hyperbolique (Wiser, Ribeck et Lenski 2013). Cela correspond à une vitesse d'adaptation qui diminue au cours du temps mais ne tend pas vers zéro, suggérant une évolution illimitée. Wiser, Ribeck et Lenski (2013) montrent que cette dynamique est compatible avec un modèle incluant de l'interférence clonale (compétition entre différents clones d'une population portant des mutations bénéfiques différentes) et une épistasie à rendements décroissants (lorsque l'amélioration marginale apportée par une mutation bénéfique diminue avec l'augmentation de la valeur sélective).

Woods et al. (2006) notent qu'un certain degré de parallélisme est observé parmi les 12 réplicats : 4 gènes qui portaient une mutation dans une des 12 populations avaient statistiquement plus de chance d'être muté dans les autres populations qu'un gène tiré au hasard. Cependant, l'importance de la contingence a également été mise en lumière notamment par l'apparition d'une souche capable d'exploiter le citrate (contenu dans le milieu de culture mais normalement non utilisable par *E. coli* en présence d'oxygène) au bout de 31 500 générations (Blount, Borland et Lenski 2008). Des expérimentations supplémentaires ont montré que ce phénomène était probablement dû à l'apparition d'une mutation potentialisante en amont dans cette population, plutôt qu'à un unique événement de mutation extrêmement rare.

Une étude publiée après 60 000 générations de la LTEE met en évidence deux processus importants à l'œuvre dans l'évolution de ces 12 populations (Good et al. 2017). Tout d'abord une forte interférence clonale, qui fait que le destin de nouvelles mutations bénéfiques est majoritairement déterminé par la qualité du fond génétique dans lequel elles apparaissent (plus que par leur avantage sélectif propre). Cela explique par exemple l'apparition de souches hypermutatrices dans la moitié des 12 populations étudiées : les mutations augmentant le taux de mutation sont normalement délétères, mais peuvent se fixer par effet d'auto-stop génétique (*genetic hitchhiking*) en favorisant l'apparition de mutations avantageuses sur la même souche (Sniegowski, Gerrish et Lenski 1997). D'autre part, de nombreux cas de coexistence de long terme entre différentes souches au sein d'une même population ont été identifiés. Ces coexistences persistent pendant trop longtemps pour être uniquement expliquées par l'interférence clonale (qui ralentit l'exclusion des autre souches par rapport à une compétition entre uniquement une souche mutante et une souche résidente), et sont donc probablement dues à des interactions écologiques de type coopération. Cela a été par exemple mis en évidence pour deux phénotypes notés L et S coexistant depuis longtemps dans une des 12 populations : le phénotype L croît plus vite que S sur du glucose, mais sécrète des substances que S est capable de mieux exploiter (Plucain et al. 2014). De manière générale, la LTEE démontre que même l'adaptation à un environnement constant est un processus dynamique et complexe.

Je me suis concentrée ici sur la méthode des passages en série car c'est celle-ci que j'étudie dans le Chapitre 2, cependant il existe aussi des méthodes de culture continue (Gresham et Dunham 2014). C'est le cas du chémostat, qui consiste à

cultiver une population microbienne dans un milieu liquide avec un afflux continu de nutriments et un efflux égal de milieu de culture (contenant nutriments, produits métaboliques et micro-organismes) pour maintenir un volume constant. Cette méthode et ses variantes sont très utilisées en évolution expérimentale, et sont par exemple parfaitement adaptées pour une simulation *in vitro* de microbiote intestinal (Hobson et al. 2022).

1.2.2 Analyses de données génomiques

Un autre angle d'étude des micro-organismes est l'analyse de données génomiques, rendue possible par l'essor des techniques de séquençage de nouvelle génération à la fin des années 2000. Il est depuis plus facile de produire de nouvelles séquences, ou d'utiliser des séquences déjà disponibles dans des bases de données publiques, qu'elles soient collectées lors d'expériences, chez des patients ou dans l'environnement naturel.

Génomique. La génomique est l'étude des gènes et des génomes dans leur globalité et dans leurs diverses dimensions : séquences, produits, fonctions, interactions, évolution... L'analyse d'un ou plusieurs génomes requiert en général les étapes suivantes : séquençage, assemblage et annotation. Les méthodes de séquençage (à haut débit) impliquent en général de fragmenter plusieurs copies du génome en de nombreux segments appelés *reads*, l'étape d'assemblage consiste donc à reconstituer la séquence complète. On utilise pour cela soit un génome de référence s'il en existe un pour cette espèce, soit des méthodes bioinformatiques d'assemblage *de novo* reposant sur des analyses de graphes. Selon la qualité du séquençage et les caractéristiques du génome (présence de séquences répétées), il est parfois compliqué d'obtenir une seule séquence complète. On obtient alors seulement un échafaudage (*scaffold*) du génome, c'est-à-dire un génome en plusieurs fragments. L'étape d'annotation consiste principalement à identifier les parties du génome qui sont traduites en protéines, c'est-à-dire les gènes (souvent les parties non-codantes mais transcrites en ARN sont aussi annotées). Par comparaison avec des séquences de gènes connues, il est parfois possible d'annoter le gène avec sa fonction biologique.

Ces étapes préliminaires peuvent ensuite mener à différents types d'investigation concernant notamment la structure, la fonction et l'évolution des génomes étudiés. La génomique structurale cherche à déterminer la structure des protéines encodées par les gènes. La génomique fonctionnelle étudie la fonction des gènes et de leurs produits (ARN, protéines) ainsi que leurs interactions. Pour cela on peut être amené à étudier *in vitro* ou *in silico* les interactions entre protéines, les réseaux d'interactions de gènes, le niveau d'expression des gènes dans différentes conditions, etc. En génomique comparative, les génomes de différents organismes sont comparés de manière à identifier les éléments conservés ou divergents. Cela passe généralement par un alignement des génomes, qui peut être fait différemment selon l'objectif de l'étude : alignement des séquences codantes uniquement, des séquences entières, cartes ou graphes représentant la conservation (ou non) de l'ordre

des gènes ou synténie. Une fois l'alignement obtenu, un arbre phylogénétique représentant l'histoire évolutive de ces génomes peut être reconstitué, qui peut en retour informer la comparaison des séquences.

Pangénomique. La pangénomique est un type d'analyse génomique qui a pour but d'étudier le répertoire génétique d'une espèce ou d'une population, en général procaryote. Comme les génomes considérés ont un contenu génétique potentiellement très divers, il n'y a pas beaucoup de sens à essayer de les aligner intégralement. Une fois les génomes annotés, l'analyse pangénomique requiert donc une étape supplémentaire : l'identification des familles de gènes présentes dans l'échantillon (réalisée grâce à des méthodes de clustering). On obtient à l'issue de cette étape le répertoire de gènes de l'espèce étudiée, et la distribution de ces gènes dans les génomes échantillonnés, représentée par exemple par une matrice de présence/absence de chaque gène dans chaque génome. On peut alors sélectionner uniquement les gènes présents dans un grand nombre d'individus (gènes dits persistants) pour réaliser un alignement et inférer un arbre phylogénétique. Plusieurs pipelines bioinformatiques sont disponibles pour réaliser ce type d'analyse, tels que PanACoTA (Perrin et Rocha 2021) ou Panaroo (Tonkin-Hill et al. 2020).

Une fois cette analyse effectuée, on peut par exemple tracer les courbes de raréfaction et d'enrichissement pour respectivement le génome *core* et le pangénome, c'est-à-dire le graphe représentant le nombre de gènes observés dans respectivement l'intersection et la réunion des génomes échantillonnés au fur et à mesure que l'on augmente leur nombre. Cela donne une idée du nombre de nouveaux gènes que l'on s'attend à découvrir en séquençant un individu de plus, et donc de la diversité du répertoire génétique de l'espèce. On peut aussi regarder la distribution des fréquences de gènes qui a généralement une forme caractéristique en U, c'est-à-dire montrant une surreprésentation des gènes aux fréquences extrêmes et une sous-représentation aux fréquences intermédiaires (B. J. Arnold, Huang et Hanage 2022). Il est également possible d'inférer la position des événements de gain et perte de gène sur les branches de l'arbre phylogénétique, en choisissant par exemple le scénario le plus parcimonieux en évènements qui explique les données de présence/absence. L'étude Touchon, Hoede et al. (2009) réalise par exemple ces différentes analyses pour le pangénome d'*E. coli*.

Métagénomique. La métagénomique consiste à étudier l'ensemble des génomes présents dans un échantillon environnemental (eau, sol, air, patient), sans passer par une phase de culture en laboratoire. Cette approche permet d'analyser la diversité génétique et le potentiel fonctionnel des communautés microbiennes, sans les biais présents dans les méthodes de culture. En effet, l'essor de cette discipline a montré que la majorité des souches microbiennes présentes dans l'environnement ne sont pas cultivables en laboratoire. Une analyse métagénomique peut inclure soit le séquençage d'un gène unique servant de « code-barre » pour identifier les différentes espèces présentes dans l'échantillon (on parle alors plutôt de métabarcoding), soit un séquençage complet visant à reconstituer les différents génomes ou

au moins faire un inventaire des gènes ou fonctions présentes. Ces analyses présentent de nombreux défis, notamment la représentativité des différentes espèces dans l'échantillon qui peut être influencée par d'autres biais (prélèvement, stockage, méthode de séquençage), ainsi que l'éventuelle étape d'assemblage où il faut pouvoir identifier quels fragments appartiennent à quelle espèce.

1.2.3 Modélisation

La modélisation mathématique est une approche essentielle pour comprendre la dynamique et l'évolution des populations microbiennes. Elle permet la simulation de scénarios biologiques complexes, la prédiction du comportement des systèmes étudiés, ainsi que l'exploitation statistique des données (notamment génomiques) : test d'hypothèse, estimation de paramètres, sélection de modèle.

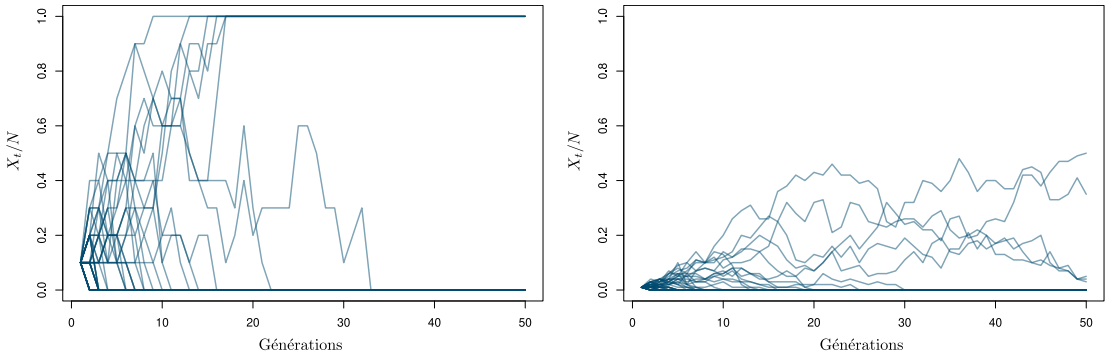
Stochasticité apparente du vivant

Il existe deux grandes approches en modélisation : déterministe ou probabiliste. Je me restreins ici à la seconde, car c'est celle que j'ai principalement employée durant ma thèse. A priori, les systèmes vivants que l'on étudie en biologie évolutive sont régis par des processus déterministes. Cependant ce sont des systèmes extrêmement complexes, constitués de nombreux composants et en général non isolés des influences extérieures. Il en résulte une connaissance incomplète des processus influençant les systèmes vivants, ainsi qu'une impossibilité matérielle de prendre en compte assez d'informations et de mécanismes pour décrire de manière déterministe et exacte leur évolution. La modélisation stochastique prend donc tout son sens pour décrire des événements qui semblent aléatoires à l'échelle populationnelle considérée ici. Deux types de stochasticité apparente dans des processus clés pour l'évolution sont en général pris en compte dans les modèles : la stochasticité démographique et la stochasticité des phénomènes moléculaires.

La stochasticité liée au processus démographique (nombre de descendants par individu, durée de vie, migrations) est d'autant plus importante que les populations considérées sont petites. Bien que les micro-organismes *in vitro* ou *in vivo* soient souvent présents en grande quantité, des sous-populations d'intérêt – par exemple portant une mutation ou un phénotype spécifiques – peuvent être en petits nombres.

Les phénomènes moléculaires les plus couramment modélisés en microbiologie évolutive sont les mutations de l'ADN et les transferts horizontaux de gènes. Ces deux classes d'événements sont assez courantes mais recouvrent une multitude de possibilités (mutation ponctuelle ou structurale, emplacement dans le génome, identité du gène transféré ...) qui prises individuellement sont rares et surtout imprévisibles à l'échelle macroscopique, rendant nécessaire l'emploi de modèles stochastiques.

Par ailleurs quelle que soit la taille de la population ou la rareté des évènements considérés, le cadre probabiliste est intéressant en ce qu'il permet d'étudier la variabilité de ces processus en plus de leur comportement moyen.



(a) $N = 10$ **(b)** $N = 100$

Figure 1.4 – Simulations de trajectoires de nouvelles mutations dans deux populations de taille différente évoluant selon le modèle de Wright-Fisher. Pour chaque figure, 50 trajectoires ont été simulées pendant 50 générations avec une fréquence initiale de $1/N$. (a) $N = 10$: 6 mutations parmi les 50 sont fixées et les autres ont été perdues. (b) $N = 100$: 5 mutations sont encore présentes après 50 générations et les autres ont été perdues.

Utilisation des modèles, avec ou sans données

Simulation : exemple du modèle de Wright-Fisher. Une fois un système modélisé, on peut en dériver des prédictions mathématiques ou simuler son comportement. Cela permet de mieux comprendre les mécanismes à l'œuvre. Je prends ici l'exemple du modèle de Wright-Fisher, modèle phare en génétique des populations pour comprendre l'effet de la dérive génétique sur la fréquence d'un allèle (R. A. Fisher 1930 ; Wright 1931).

Dans ce modèle, on suppose qu'il existe deux allèles différents A et B dans une population de taille constante égale à N. Chaque pas de temps correspond à une génération, et les générations sont donc non-chevauchantes. Au temps $t + 1$, les N nouveaux individus choisissent leur parent au hasard parmi les N individus présents au temps t. La version de base du modèle est sans mutation, donc chaque parent transmet avec fidélité son allèle à ses descendants. Ce modèle est neutre, c'est-à-dire que la sélection naturelle n'est pas prise en compte : les allèles A et B n'ont pas d'effet différent sur le nombre de descendants de leur porteur. On note X_t le nombre d'individus portant l'allèle B dans la population au temps t. Dans ce modèle neutre sans mutation, seule la dérive génétique va influencer la proportion X_t/N d'allèles B dans la population au cours du temps.

Mathématiquement, le fait que chaque individu choisisse son parent de manière uniforme dans la génération précédente se traduit par le fait que X_{t+1} est tiré selon une loi binomiale de paramètres $(N, X_t/N)$:

$$\mathbb{P}(X_{t+1} = j | X_t = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j} \tag{1.1}$$

En particulier, on a $\mathbb{E}(X_{t+1}) = \mathbb{E}(\mathbb{E}(X_{t+1}|X_t)) = \mathbb{E}(X_t)$: l'espérance de X_t ne change pas au cours du temps. On note $p = X_0/N$ la fréquence initiale, qui est

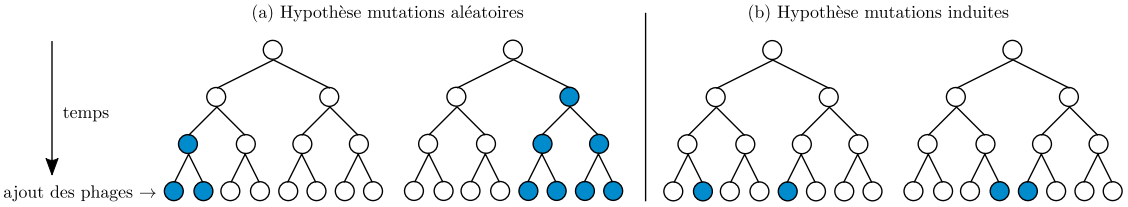


Figure 1.5 – Illustration des deux hypothèses imaginées par Luria et Delbrück pour expliquer la présence de résistance aux phages héritable. (a) Hypothèse des mutations aléatoires : des mutations conférant une résistance apparaissent aléatoirement durant la croissance des bactéries. (b) Hypothèse des mutations induites : les mutations sont induites par l'exposition aux phages. L'hypothèse (a) suppose une variance plus grande du nombre de bactéries résistantes entre deux populations que l'hypothèse (b).

donc l'espérance de la fréquence de l'allèle B dans la population. On peut montrer que p est aussi la probabilité de fixation de l'allèle B. Comme les états $X_t = 0$ et $X_t = N$ sont absorbants (et accessibles depuis les autres états), la probabilité que l'allèle B soit perdu ou fixé tend vers 1 lorsque t tend vers l'infini. On a alors :

$$\begin{aligned} p &= \lim_{t\to\infty} \mathbb{E}\left(\frac{X_t}{N}\right) \\ &= 0 \times \mathbb{P}\left(\frac{X_\infty}{N} = 0\right) + 1 \times \mathbb{P}\left(\frac{X_\infty}{N} = 1\right) \\ &= \mathbb{P}(X_\infty = N) \end{aligned} \tag{1.2}$$

Ce modèle peut donc aider à comprendre certaines caractéristiques de la dérive génétique : c'est une force qui n'affecte pas la fréquence moyenne des allèles et qui peut mener à la fixation d'un allèle avec une probabilité égale à sa fréquence initiale. La Figure 1.4 montre des trajectoires de mutations suivant le modèle de Wright-Fisher.

Sélection de modèle : exemple de l'expérience de Luria-Delbrück. En comparant les prédictions d'un modèle à des données empiriques, on peut estimer si les hypothèses faites par le modèle sont conformes aux observations (test d'hypothèse). En comparant les prédictions de plusieurs modèles à des données observées, on peut déterminer quel est le mécanisme le plus vraisemblable ayant pu conduire à ces observations (sélection de modèle). Tout cela peut être étudié de manière informelle, ou formalisé par des tests statistiques. Un exemple bien connu en microbiologie pour illustrer cette idée est l'expérience de Luria et Delbrück (1943). Dans cette étude, des résultats expérimentaux ont été analysés à la lumière de deux modèles différents pour en déduire un mécanisme important de l'évolution bactérienne.

La question que se sont posée Luria et Delbrück est la suivante : une population bactérienne croît à partir d'un seul individu, puis est mise en présence de phages. On observe que la plupart des bactéries meurent mais certaines résistent à l'infection, et les descendants des ces bactéries survivantes sont aussi résistants. Est-ce que cette résistance héréditaire est le résultat de mutations aléatoires arrivées avant

l'exposition aux phages, ou au contraire d'une résistance induite par la rencontre avec les phages ?

Afin de répondre à cette question, ils élaborent un modèle pour chaque hypothèse (Figure 1.5). Ils supposent que la population de bactéries a une croissance exponentielle à partir d'un individu : $N_0 = 1$, et $N_t = N_0 e^t$ (l'unité de temps choisie est le temps de division moyen d'une bactérie divisé par $\ln 2$). Pour la première hypothèse (mutations aléatoires), des bactéries mutantes apparaissent dans la population à taux aN_t selon un processus de Poisson inhomogène en temps (a est le taux de mutation par individu et par unité de temps). Chaque bactérie mutante forme un clone qui croît ensuite de manière déterministe, à la même vitesse que la souche initiale. Notons M_t le nombre de mutations survenant entre les temps t et $t + \mathrm{d}t$. M_t suit une loi de Poisson de paramètre $aN_t\,\mathrm{d}t$. Au temps t_p où on ajoute les phages, le nombre de bactéries mutantes issues d'une mutation dans l'intervalle de temps $[t, t + \mathrm{d}t]$ vaut $M_t e^{t_p - t}$. Donc si on note R le nombre total de bactéries mutantes au temps t_p, on a avec ce modèle :

$$\begin{aligned} \mathbb{E}(R) &= \int_0^{t_p} \mathbb{E}(M_t e^{t_p - t}) \\ &= \int_0^{t_p} aN_t e^{t_p - t}\,\mathrm{d}t \\ &= \int_0^{t_p} aN_0 e^{t_p}\,\mathrm{d}t \\ &= aN_{t_p} t_p \end{aligned} \tag{1.3}$$

et

$$\begin{aligned} Var(R) &= \int_0^{t_p} Var(M_t e^{t_p - t}) \\ &= \int_0^{t_p} Var(M_t) e^{2(t_p - t)} \\ &= \int_0^{t_p} aN_t e^{2(t_p - t)}\,\mathrm{d}t \\ &= aN_{t_p}(e^{t_p} - 1) \end{aligned} \tag{1.4}$$

La variance est donc supérieure à l'espérance. Alors que pour la seconde hypothèse (mutations induites par la présence de phage), on peut s'attendre à ce que le nombre de bactéries résistantes suive une loi de Poisson de paramètre bN_{t_p} avec b le taux de mutation induite par individu. Dans ce cas, on a $\mathbb{E}(R) = Var(R)$.

Les auteurs ont donc réalisé plusieurs expériences en parallèle de croissance de bactéries puis ajout de phages, et compté à chaque fois le nombre de colonies résistantes. Ils ont trouvé que la variance du nombre de bactéries résistantes était bien supérieure au nombre moyen de ces bactéries, et conclu que l'hypothèse des mutations aléatoires était la plus plausible. Ils ont également étudié d'autres propriétés du modèle avec mutations aléatoires, et notamment proposé un moyen d'estimer le taux de mutation en comptant le nombre de cultures sans bactéries résistantes. Cet article a donné lieu à un prix Nobel, ainsi qu'au développement de toute une littérature mathématique sur des variations du modèle de mutations aléatoires, et

en particulier sur l'étude de la distribution du nombre de mutants à la fin de l'expérience qui est un problème assez compliqué malgré la simplicité du modèle initial (Zheng 1999). Je détaille dans la Discussion de cette thèse comment le modèle développé au Chapitre 2 s'inscrit dans cette littérature (Section 5.1.2).

Inférence. Un modèle peut aussi servir à inférer des propriétés du système étudié (par exemple estimer des paramètres) qui ne sont pas directement observables. Par exemple le taux de croissance d'une souche bactérienne dans un milieu contrôlé est facile à mesurer, mais le taux de transferts horizontaux entre les bactéries présentes dans l'intestin l'est beaucoup moins. La formalisation des mécanismes étudiés sous la forme d'un modèle permet de déterminer les valeurs de paramètres les plus probables au vu de données observables influencées par ces mécanismes, par exemple un échantillon de génomes de ces bactéries. Les principales méthodes d'inférence utilisées en phylogénétique sont présentées en Section 1.3.2.

1.3 Modèles stochastiques pour l'inférence de l'évolution

Dans cette partie, je présente deux modèles stochastiques courants en biologie évolutive et dont des variantes sont utilisées dans cette thèse, les principales classes de méthodes d'inférence ainsi que quelques points techniques concernant la simulation et l'inférence de ces modèles.

1.3.1 Modèles markoviens classiques

Processus de naissance-mort

Le processus de naissance et de mort à taux constant est le modèle stochastique le plus simple permettant de décrire l'évolution du nombre d'individus dans une population. La population est composée au temps t de I_t individus qui indépendamment donnent naissance à taux β et meurent à taux δ (Figure 1.6). Pour un individu donné, donner naissance à taux β signifie mathématiquement qu'au cours d'un intervalle de temps infinitésimal Δt, la probabilité pour cet individu de donner naissance vaut $\beta\Delta t$. De manière équivalente, cela signifie que le temps d'attente avant un événement de naissance est distribué selon une loi exponentielle de paramètre β. Ceci est valable pour un individu. Si la population contient n individus, le temps d'attente avant le prochain événement (naissance ou mort) suit une loi exponentielle de paramètre $n(\beta+\delta)$, la somme des taux des évènements possibles, par indépendance de tous ces évènements. Pour tout entier $n > 0$ on a

$$\begin{aligned}
\mathbb{P}(I_{t+\Delta t} = n+1 \,|\, I_t = n) &= n\beta\Delta t + o(\Delta t) \\
\mathbb{P}(I_{t+\Delta t} = n-1 \,|\, I_t = n) &= n\delta\Delta t + o(\Delta t) \\
\mathbb{P}(I_{t+\Delta t} = n \,|\, I_t = n) &= 1 - n(\beta+\delta)\Delta t + o(\Delta t)
\end{aligned} \tag{1.5}$$

où les termes $o(\Delta t)$ tiennent compte de la possibilité que plusieurs événements se produisent dans l'intervalle Δt. En particulier, les probabilités de transition vers un nouvel état ne dépendent que de l'état présent. $(I_t)_{t\geq 0}$ est une chaîne de Markov à temps continu, dont l'espace d'états est $\mathbb{N}$. La dernière égalité prise pour $n = 0$ traduit le fait que 0 est un état absorbant : $\mathbb{P}(I_{t+\Delta t} = 0 \,|\, I_t = 0) = 1$.

Calcul de quantités d'intérêt. Ce modèle stochastique prenant tout son intérêt lorsque les tailles de populations considérées sont petites, il est par exemple intéressant de pouvoir calculer le nombre de descendants d'un individu unique au bout d'un certain temps $t > 0$, ainsi que la probabilité de survie d'une telle population. Ces résultats ont été établis par exemple par Kendall (1948), en résolvant une équation différentielle impliquant la fonction génératrice du nombre de descendants. Si on note $P_n(t)$ la probabilité qu'un individu ait produit n descendants au bout d'un temps t, alors on peut montrer que pour $\beta \neq \delta$:

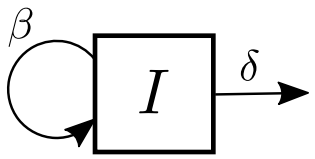


Figure 1.6 – Représentation graphique du modèle de naissance-mort à taux constants.

$$\begin{aligned} P_0(t) &= \frac{\delta(1 - e^{-(\beta-\delta)t})}{\beta - \delta e^{-(\beta-\delta)t}} \\ P_n(t) &= (1 - P_0(t))q(t)(1 - q(t))^{n-1} \text{ pour } n \geq 1, \\ \text{avec } q(t) &= \frac{\beta - \delta}{\beta e^{(\beta-\delta)t} - \delta}. \end{aligned} \tag{1.6}$$

La probabilité que la population soit toujours en vie au bout d'un temps t est $1 - P_0(t)$. On peut observer que c'est une fonction décroissante du temps, qui tend vers 0 si et seulement si $\beta \leq \delta$. Conditionnellement au fait que la population survive, on constate que le nombre d'individus au temps t suit une loi géométrique de paramètre $q(t)$. L'inverse de ce nombre correspond donc à la taille attendue pour la population au temps t conditionnellement à sa survie. Cette taille attendue croît exponentiellement avec le temps lorsque $\beta > \delta$, et décroît jusqu'à la valeur limite $\frac{\delta}{\delta-\beta}$ lorsque $\beta < \delta$. Dans le cas du processus critique ($\beta = \delta$) on peut montrer que $P_0(t) = \frac{\beta t}{1+\beta t}$ et $q(t) = \frac{1}{1+\beta t}$, donc la taille attendue au temps t conditionnellement à la survie jusqu'au temps t est $1 + \beta t$.

Limites et extensions. Ce type de modèle convient bien aux populations asexuées, qui se reproduisent de manière clonale. Pour des populations sexuées, il peut être utile de distinguer les types sexuels et d'avoir un taux de naissance tenant compte de la probabilité d'une rencontre entre deux types différents. D'autre part, la propriété de Markov implique que la probabilité pour un individu de se reproduire ou de mourir est indépendante de son âge, ce qui représente une approximation plus ou moins juste selon la population étudiée.

Il existe de nombreuses extensions possibles à ce modèle : taux variables dans le temps, plusieurs types de populations entre lesquelles il peut y avoir des migrations ou des mutations, etc. Deux variantes sont particulièrement intéressantes pour l'étude des micro-organismes. La première est d'ajouter de la densité-dépendance, en faisant dépendre le taux de naissance et/ou de mort individuel de la taille de population. Cela permet de prendre en compte la limitation des ressources dans l'environnement, en diminuant le taux de naissance et/ou en augmentant le taux de mort à mesure que la taille de population augmente. La question de savoir si la limitation des ressources influence plutôt le taux de naissance ou le taux de mort (ou les deux) n'est pas évidente et dépend du système étudié (Huynh, Scott et P. J. Thomas 2023). La deuxième est une adaptation pour mieux correspondre au cycle de vie des virus, et se nomme le modèle *burst-death* (qui se traduirait par « modèle d'explosion-mort »). En effet, le mode de reproduction des virus à cycle lytique est de se multiplier à l'intérieur d'une cellule jusqu'à son éclatement (ou lyse), libérant alors une grande quantité de particules virales. Le modèle d'explosion-mort garde donc un taux de reproduction et de mort constants par individu, mais lorsqu'un individu parvient à se reproduire (c'est-à-dire, à infecter une cellule) le nombre de descendants produits vaut $B > 1$ (Hubbarde, Wild et Wahl 2007).

Les modèles compartimentaux type SIR (susceptibles-infectés-résistants) sont utilisés pour décrire la dynamique des épidémies et comment l'immunité acquise par la population hôte influence la dynamique épidémique générée par la transmission du pathogène. Ils reposent sur le même type de fonctionnement qu'un processus de naissance-mort multitype, sauf que les changements entre compartiments correspondent à un changement de statut infectieux plutôt qu'à une mutation ou migration. Là encore se pose la question de la validité de l'hypothèse que le temps de séjour dans un compartiment suit une loi exponentielle. Il semblerait par exemple que dans le cas de la pandémie de Covid-19, la distribution du temps passé dans chaque compartiment soit mieux approchée par des lois possédant une densité unimodale comme les lois Gamma ou de Weibull (Linton et al. 2020). Durant cette pandémie, j'ai participé à une publication de l'équipe SMILE proposant une méthode d'analyse de modèles compartimentaux pouvant gérer de nombreux compartiments et des distributions de temps de séjour arbitraires, en structurant la population par âge (au sens du temps écoulé depuis l'infection) et non par type (Foutel-Rodier et al. 2022).

Modèles phylogénétiques d'évolution moléculaire

Ce type de modèles a été initialement introduit par Jukes et Cantor (1969) pour décrire l'évolution d'une séquence ADN le long d'une phylogénie. Les différents sites d'une séquence évoluent indépendamment les uns des autres, chaque site a 4 états possibles (correspondants aux 4 nucléotides A,T,C,G) et les transitions entre états, appelés substitutions, correspondent à des événements de mutation ponctuelle. Ces modèles sont une composante essentielle des méthodes de reconstruction phylogénétique car ils permettent d'estimer quelles sont les topologies d'arbres et les taux d'évolution les plus probables au vu d'un alignement de séquences donné.

Définition du modèle de Jukes-Cantor. Si l'on note S_t l'état d'un site nucléotidique donné à un certain temps t, alors $(S_t)_{t\geq 0}$ est une chaîne de Markov en temps continu, avec un espace d'états fini $\{A, T, C, G\}$. Ce processus est homogène en temps (les taux de substitution sont constants), avec la matrice de taux de substitutions suivante :

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array} \tag{1.7}$$

Dans cette version, les substitutions entre les différents nucléotides ont toutes le même taux. La diagonale de cette matrice n'a pas de signification particulière, ses valeurs sont choisies pour que la somme de chaque ligne soit égale à 0, ce qui permet d'écrire la matrice des probabilités de transitions $P(t) = (\mathbb{P}(S_t = j | S_0 = i))_{i,j\in\{A,T,C,G\}}$ en fonction de Q. Pour un intervalle de temps infinitésimal Δt, on a :

$$P(\Delta t) = I + Q\Delta t \tag{1.8}$$

Plus généralement pour $t \geq 0$, on peut montrer que P vérifie l'équation différentielle matricielle suivante :

$$\frac{d}{\mathrm{d}t}P(t) = QP(t) \tag{1.9}$$

Avec la condition initiale $P(0) = I$ on obtient donc :

$$P(t) = e^{tQ}, \tag{1.10}$$

où pour toute matrice carrée A, l'exponentielle de matrice e^A désigne la somme de la série de terme général $A^n/n!$. Cette expression est utilisée par exemple pour calculer les probabilités de transitions le long des branches d'un arbre phylogénétique. On voit que cela requiert de calculer l'exponentielle de la matrice Q, ce qui est facile lorsque Q est diagonalisable (et c'est par exemple le cas de toutes les matrices symétriques). L'algorithme d'élagage généralement utilisé pour calculer la vraisemblance de ce genre de modèle le long d'une phylogénie est décrit en Section 1.3.3.

Variantes. Ce modèle de d'évolution moléculaire possède de nombreuses variantes : taux de substitutions différents selon les paires de nucléotides considérées, modèles de substitutions par codons, variation des taux entre sites, etc. De manière intéressante, ce type de modèle a également été repris pour modéliser non pas l'évolution moléculaire mais plus généralement l'évolution de traits phénotypiques discrets le long d'une phylogénie (Pagel 1997), sous le nom de modèle Mk. Des modèles similaires ont été utilisés pour décrire l'évolution de la présence (état 1) ou l'absence (état 0) de gènes dans des génomes bactériens (Didelot, Darling et Falush 2008 ; Cohen et Pupko 2010 ; Zamani-Dahaj et al. 2016), sous le nom de modèle FMG (pour Finitely Many Genes). Dans ce cas précis si l'on note g le taux

de gain (transition $0 \longrightarrow 1$) et l le taux de perte (transition $1 \longrightarrow 0$), la matrice des taux de transitions s'écrit :

$$Q = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{pmatrix} -g & g \\ l & -l \end{pmatrix} \end{matrix} \tag{1.11}$$

Cette matrice est diagonalisable et son exponentielle possède une expression analytique :

$$P(t) = e^{tQ} = \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{pmatrix} \frac{l}{g+l} + \frac{g}{g+l}e^{-(g+l)t} & \frac{g}{g+l} - \frac{g}{g+l}e^{-(g+l)t} \\ \frac{l}{g+l} - \frac{l}{g+l}e^{-(g+l)t} & \frac{g}{g+l} + \frac{l}{g+l}e^{-(g+l)t} \end{pmatrix} \end{matrix} \tag{1.12}$$

Cette expression permet notamment d'observer que ce modèle possède la distribution stationnaire suivante : $(\frac{l}{g+l}, \frac{g}{g+l})$. Cela correspond aux proportions respectives de gènes dans l'état 0 et l'état 1 si on laisse un génome évoluer pendant très longtemps selon ce modèle.

1.3.2 Méthodes d'inférence

Comme introduit dans la Section 1.2.3, il est naturel lorsque l'on formule un modèle de vouloir le confronter à des données, afin notamment de comparer différentes hypothèses et d'estimer les valeurs des paramètres. Dans cette section, je passe en revue différentes méthodes d'inférences couramment utilisées en m'appuyant à chaque fois sur l'exemple des modèles d'évolution phylogénétique présentés précédemment.

Maximum de vraisemblance. L'inférence par maximum de vraisemblance a été introduite par R. A. Fisher (1922). Le principe est le suivant : si l'on possède un ensemble de données observées (par exemple, des séquences d'ADN) et que l'on est capable de calculer la probabilité d'observer ces données sous le modèle étudié (par exemple, le modèle de Jukes-Cantor), alors on peut chercher à déterminer la valeur des paramètres du modèle pour lesquelles la probabilité d'observer ces données est maximale. Formellement, on calcule la fonction de vraisemblance :

$$\mathcal{L}(\mathcal{T}, Q_\lambda; D) = \mathbb{P}(D|\mathcal{T}, Q_\lambda) \tag{1.13}$$

où D représente les données, $\mathcal{T}$ l'arbre de parenté et Q_λ la matrice de taux de transitions du modèle de Jukes-Cantor. Ce modèle contient un seul paramètre, λ. Pour trouver la valeur la plus vraisemblable de ce paramètre au vu des données, on maximise la fonction de vraisemblance en λ. Pour cela de nombreuses méthodes d'optimisation existent, adaptées à différents cas de figure : possibilité ou non de calculer un gradient, existence de maxima locaux, etc. Les remarques précédentes concernent le cas où l'on connaît la topologie de l'arbre de parenté entre les individus séquencés. Si ce n'est pas le cas et qu'on aimerait au contraire pouvoir inférer cet arbre à partir des données, alors la topologie de l'arbre devient un paramètre

sur lequel on veut maximiser la vraisemblance. Malheureusement, l'ensemble des topologies possibles est en général beaucoup trop vaste pour l'explorer de manière exhaustive (on peut montrer qu'il existe $\mathcal{O}(e^{n \ln n})$ topologies différentes pour un arbre étiqueté à n feuilles). Les algorithmes d'inférence phylogénétique les plus utilisés (par exemple RaxML de Stamatakis 2014 ou IQ-TREE de Minh et al. 2020) utilisent donc des méthodes heuristiques de marche aléatoire dans l'espace des topologies d'arbres.

Inférence bayésienne. L'inférence bayésienne adopte un point de vue un peu différent, et permet d'inférer la distribution de probabilité des paramètres du modèle prenant en compte à la fois le signal contenu dans les données et la connaissance préalable sur ces paramètres. Cette méthode repose sur la formule de Bayes, qui permet de calculer la distribution postérieure des paramètres – dans notre cas d'inférence phylogénétique, $\mathbb{P}(\mathcal{T}, Q_\lambda|D)$ – à partir de la vraisemblance $\mathbb{P}(D|\mathcal{T}, Q_\lambda)$, de la distribution a priori des paramètres $\mathbb{P}(\mathcal{T}, Q_\lambda)$ et de la vraisemblance marginale des données $\mathbb{P}(D)$:

$$\mathbb{P}(\mathcal{T}, Q_\lambda|D) = \frac{\mathbb{P}(D|\mathcal{T}, Q_\lambda)\mathbb{P}(\mathcal{T}, Q_\lambda)}{\mathbb{P}(D)} \tag{1.14}$$

Ici l'enjeu n'est pas de maximiser la probabilité des données en fonction des paramètres, mais de déterminer la distribution postérieure des paramètres sachant les données. Le calcul analytique de cette distribution étant généralement infaisable (notamment encore à cause du nombre de topologies d'arbres), on utilise des marches aléatoires permettant de l'approximer efficacement. Il existe pour cela une classe de méthodes appelées MCMC (Méthodes de Monte-Carlo par chaînes de Markov), dont l'implémentation la plus utilisée est l'algorithme de Metropolis-Hastings (W. K. Hastings 1970). Cet algorithme permet de construire une marche aléatoire dont la distribution stationnaire est la distribution postérieure des paramètres, le tout sans avoir besoin de calculer la vraisemblance marginale (l'algorithme utilise uniquement des ratios entre les probabilités postérieures, donc les dénominateurs se simplifient).

L'avantage par rapport au maximum de vraisemblance est donc de pouvoir obtenir une distribution de paramètres au lieu d'une seule valeur. Par ailleurs la distribution postérieure donne une interprétation facile de l'incertitude sur la valeur des paramètres, tandis qu'une mesure d'incertitude sur les paramètres liée à la vraisemblance (l'intervalle de confiance) a une interprétation plus délicate. La distribution postérieure permet aussi de mesurer la force du signal contenu dans les données en la comparant avec la distribution a priori. Ceci est valable pour les paramètres réels, mais aussi pour les topologies d'arbres : au lieu d'obtenir uniquement la topologie la plus vraisemblable, on obtient tout un ensemble de topologies probables. Les inconvénients sont un temps de calcul plus élevé et le choix des distributions a priori qui n'est pas toujours évident.

Méthodes basées sur des simulations massives. Il existe d'autres types de méthodes d'inférence pour les cas où l'on est capable de simuler le modèle mais pas

de calculer sa fonction de vraisemblance. C'est le cas des méthodes ABC (Calcul Bayésien Approché), qui permettent d'approcher la distribution des paramètres en faisant de nombreuses simulations pour différentes valeurs de paramètres et en rejetant celles qui produisent des statistiques résumées trop éloignées de celles des données (Tavare et al. 1997). Une prépublication récente propose également une méthode promettant d'accélérer la maximisation de la vraisemblance dans ces cas où elle n'est connue que de manière approchée par des simulations, en fixant l'aléa au début de la procédure pour optimiser une fonction régulière au lieu d'une fonction bruitée (Moinard et al. 2023).

Récemment, des méthodes d'inférence phylogénétique par apprentissage profond on fait leur apparition, dédiées uniquement à l'inférence de paramètres (Lambert, Voznica et Morlon 2023) ou aussi à la reconstruction d'arbres (Z. Wang et al. 2023). Ces méthodes ont l'avantage d'être très rapides pour l'étape d'inférence (plus que les méthodes par maximum de vraisemblance ou bayésiennes), mais nécessitent une étape d'entraînement reposant sur des simulations massives et donc très coûteuse en temps de calcul. Le gain en temps de calcul global dépend donc du nombre d'inférences réalisées par rapport au nombre d'entraînements (pour tester une variante d'un modèle par exemple, il faudra refaire l'étape d'entraînement). Pour le moment, l'outil de reconstruction phylogénétique Fusang permet d'inférer des arbres possédant au plus 40 feuilles (Z. Wang et al. 2023), et n'est donc guère compétitif avec ses homologues basés sur la vraisemblance. Silvestro, Latrille et Salamin (2023) proposent une méthode couplant deux types d'approche : afin d'augmenter la précision de l'inférence phylogénétique lorsque l'hétérogénéité des taux de mutation entre sites suit un modèle complexe, les auteurs infèrent finement le taux de mutation par site par une approche deep-learning, puis utilisent ces estimations dans une reconstruction d'arbre par maximum de vraisemblance. Ils obtiennent ainsi une meilleure précision que par maximum de vraisemblance uniquement.

1.3.3 Défis computationnels rencontrés

Les avantages des populations microbiennes en termes de grandes tailles de population et de quantité de données disponibles peuvent poser des défis dans le traitement computationnel de ces informations. Je me restreint ici aux problématiques rencontrées au cours de ma thèse.

Simulations de grandes populations

Pour le projet présenté dans le Chapitre 2, j'ai simulé l'évolution démographique de populations de grande taille suivant un modèle stochastique de naissance-mort-mutation. J'explique ci-dessous la méthode généralement utilisée pour ce genre de simulations, ainsi que les approximations qui ont été nécessaires en raison des grands nombres utilisés : de 10^6 jusqu'à 10^{25} individus, ce afin de vérifier la convergence vers nos prédictions mathématiques.

Incrémentation du temps. La méthode usuelle pour simuler des chaînes de Markov à temps continu est l'algorithme de Doob, connu en biologie sous le nom d'algorithme de Gillespie (1977) du nom du chimiste qui l'a popularisé. Il repose sur une actualisation du système en trois étapes :

1. Le temps d'attente avant la prochaine transition du système est le minimum des temps d'attente pour chaque type d'événement (ici naissance, mort et mutation). Or ces temps sont indépendants et suivent chacun une loi exponentielle, donc leur minimum suit une loi exponentielle dont le paramètre est la somme des taux des différents événements. C'est dans cette loi qu'il faut tirer le temps d'attente avant le prochain évènement.
2. La probabilité que le prochain événement soit de tel ou tel type est proportionnelle à son taux. Il faut donc tirer le type de l'événement en pondérant chaque type par son taux divisé par la somme des taux, indépendamment du temps d'attente.
3. On incrémente la variable de temps du temps d'attente, et les différentes tailles de populations en fonction de l'événement tiré.

Cette procédure est répétée, par exemple jusqu'à ce que le temps atteigne une certaine date. Un problème couramment rencontré avec cette méthode est que plus la taille de population est grande, plus les taux de transition sont élevés (comme on multiplie les taux individuels de reproduction, mort et mutation par le nombre d'individus). Les incréments de temps deviennent alors de plus en plus petits et la simulation stagne (le temps n'augmente plus). Une solution approchée appelée *τ-leaping* (« saut de τ ») consiste à fixer une valeur minimale pour l'incrément du temps, τ (Gillespie 2001). Si la valeur tirée à l'étape 1 est inférieure à τ, on bascule vers une méthode d'actualisation alternative :

1. On fixe l'incrément de temps à τ.
2. On tire des variables de Poisson pour déterminer le nombre d'événements de chaque type réalisés durant cet intervalle de temps. Par exemple si le taux de mort vaut δ et qu'il y a n individus dans la population, alors on tire une variable suivant une loi $\mathcal{P}ois(n\delta\tau)$ pour déterminer le nombre de morts. Cette simulation serait exacte si le taux de mort global était égal à $n\delta$ durant tout l'intervalle de temps. Comme ce taux change en fait à chaque évènement, c'est une approximation, qui est cependant assez précise lorsque n est grand.
3. On actualise les tailles de population en conséquence. Si certaines tailles sont négatives suite à cette actualisation, on revient au temps précédent et on réitère la procédure en divisant τ par deux.

Approximations de lois. Pour des tailles de populations vraiment grandes, certains taux dépassaient 10^{18} et le simulateur de loi de Poisson de Python ne fonctionnait plus. Dans ce cas, j'ai approximé une loi de Poisson de paramètre λ par une loi normale de paramètres $(\lambda, \sqrt{\lambda})$.

Calcul de vraisemblance phylogénétique

Pour le projet présenté au Chapitre 4, j'ai entre autre implémenté le calcul de la vraisemblance d'un modèle Mk à deux états (0 et 1) le long d'une phylogénie de taille intermédiaire (900 feuilles). Je décris dans cette section l'algorithme d'élagage généralement utilisé pour calculer la vraisemblance sur une phylogénie, puis la solution que j'ai utilisée pour éviter les problèmes de soupassement arithmétique (c'est-à-dire pour manipuler des nombres plus petits que la précision du format numérique utilisé).

Algorithme d'élagage. Notons n le nombre de feuilles de l'arbre et m le nombre de sites de notre alignement. On peut représenter notre alignement par une matrice $S = (s_{i,j})_{1 \leq i \leq n, 1 \leq j \leq m}$. Un calcul naïf de la vraisemblance en sommant sur tous les états possibles des nœuds internes aurait une complexité $\mathcal{O}(m2^n n)$. En effet, pour chaque site j il faudrait calculer

$$\mathbb{P}(S_j|\mathcal{T}, Q) = \sum_{s_{n+1} \in \{0,1\}} \dots \sum_{s_{2n-1} \in \{0,1\}} \pi_{s_{2n-1}} \prod_{l=1}^{2n-2} P_{s_{l_a}, s_{l_b}}(t_l) \tag{1.15}$$

où $s_{n+1}, \dots, s_{2n-1}$ sont les états des nœuds internes pour le site j (la racine étant numérotée $2n-1$), π_X est la probabilité d'avoir l'état X à la racine (on peut prendre par exemple la fréquence d'équilibre de X, ou considérer que tous les états sont équiprobables à la racine), $P_{i,k}(t)$ est la probabilité d'observer une transition $i \to k$ en un temps t, et pour une branche l de l'arbre on a noté l_a et l_b ses extrémités et t_l sa longueur. Ensuite il faut faire le produit des vraisemblances à chaque site pour obtenir la vraisemblance totale (m termes).

Heureusement, l'algorithme d'élagage développé par Felsenstein (1981) permet de réaliser ce calcul avec une complexité linéaire en m et n. Le principe est le suivant : pour un site j donné, on va calculer à chaque nœud k de l'arbre et pour chaque état $X \in \{0, 1\}$ la probabilité $\mathbb{P}(D_k|X)$ d'observer les états aux feuilles descendant de k sachant que le nœud k est dans l'état X (Figure 1.7). On procède comme suit :

1. Pour une feuille f telle que $s_f = Y$, alors $\mathbb{P}(D_k|X) = 1$ si $X = Y$ et 0 sinon.
2. Soit un nœud k pour lequel on a déjà calculé les probabilités des descendants directs g et d, alors pour $X \in \{0, 1\}$ on calcule :
$$\mathbb{P}(D_k|X) = \left(\sum_{Y \in \{0,1\}} P_{X,Y}(t_g) \mathbb{P}(D_g|Y) \right) \times \left(\sum_{Y \in \{0,1\}} P_{X,Y}(t_d) \mathbb{P}(D_d|Y) \right) \tag{1.16}$$
3. Répéter l'opération 2 en progressant des feuilles vers la racine.
4. À la racine, calculer :
$$\mathbb{P}(S_j|\mathcal{T}, Q) = \sum_{X \in \{0,1\}} \mathbb{P}(D_r|X) \pi_X \tag{1.17}$$

Après avoir répété les opérations 1 à 4 pour chaque site, il reste à multiplier la vraisemblance de chaque site pour obtenir le résultat final (ou additionner les log-vraisemblances).

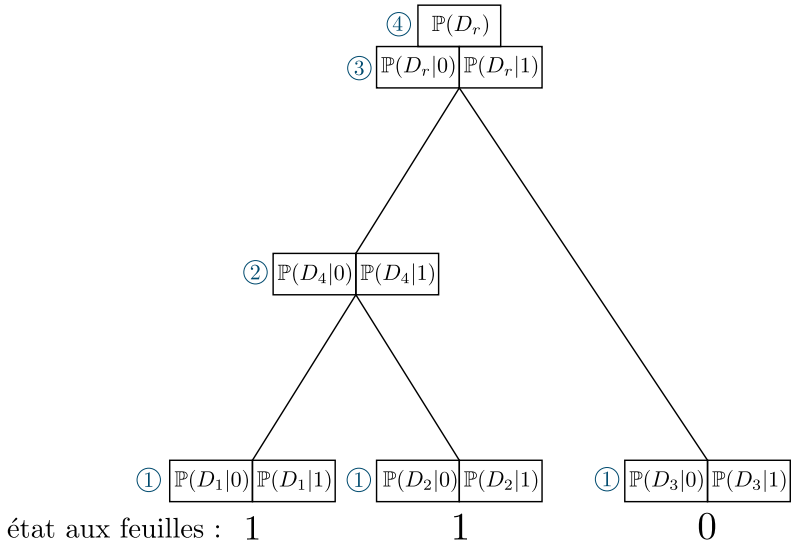


Figure 1.7 – Déroulement de l'algorithme d'élagage de Felsentein pour calculer la vraisemblance des données aux feuilles d'une phylogénie, selon un modèle Mk à deux états (0 et 1). La racine est notée r et les autres nœuds sont numérotés de 1 à 4. Les chiffres entourés correspondent aux étapes de l'algorithme décrites Section 1.3.3.

Soupassement arithmétique. Au fur et à mesure que l'on remonte vers la racine lors du calcul de vraisemblance, les valeurs deviennent de plus en plus petites et peuvent poser des problèmes de soupassement arithmétique, c'est-à-dire d'obtenir des probabilités plus petites que la précision disponible et donc traitées comme étant égales à 0. Pour éviter cela, j'ai toujours travaillé avec la log-vraisemblance au lieu de la vraisemblance. Les produits figurant dans le calcul de la vraisemblance sont facilement remplacés par des sommes de log-vraisemblances. Les sommes dans la vraisemblance nécessitent quant à elles l'utilisation de l'astuce dite « log-somme-exp ». Soient $a = \log(x)$ et $b = \log(y)$. Si on essaye de calculer $\log(x + y) = \log(e^a + e^b)$, il y a un risque de soupassement si x et y sont de très petits nombres. Au lieu de cela, je calcule $\log(x + y) = a + \log(1 + e^{b-a})$. Ici x et y ne sont pas calculés, mais seulement leur rapport $\frac{y}{x} = e^{b-a}$, ce qui est moins susceptible de provoquer un soupassement.

Une autre astuce couramment utilisée lorsque l'on veut plutôt travailler avec la vraisemblance, est de multiplier les probabilités trop faibles par un grand nombre, et de retirer le logarithme de ce nombre autant de fois qu'on l'a ajouté, lors de la dernière étape où l'on additionne les log-vraisemblances de chaque site (expliquée Section 5.2.1 de Schmidt 2010).

1.4 Objectifs de la thèse

Cette thèse se compose de deux parties indépendantes, dont le point commun est d'utiliser des modèles stochastiques pour étudier l'évolution microbienne.

1.4.1 Impacts évolutifs des goulots d'étranglement de population

La première partie porte sur l'étude des populations subissant un ou plusieurs goulots d'étranglement, c'est-à-dire un ou plusieurs événements lors desquels la taille de la population est drastiquement réduite. Les causes de ces réductions drastiques incluent par exemple la dilution d'une population microbienne en laboratoire, la transmission à un nouvel hôte pour un pathogène, le manque de ressources, la perte d'habitat due aux activités humaines ou les événements climatiques extrêmes. Cela concerne donc à la fois les populations *in vitro* et *in vivo*. J'ai cherché à décrire les impacts de tels changements démographiques sur l'évolution de ces populations.

Les goulots d'étranglement peuvent contraindre et canaliser les chemins évolutifs. Le Chapitre 2 présente un travail de modélisation des populations subissant des goulots d'étranglement périodiques. Bien que cette dynamique corresponde notamment aux expériences d'évolution en laboratoire utilisant des dilutions en série (voir Section 1.2.1), l'effet des dilutions successives sur l'évolution de ces populations n'est pas encore complètement compris. Deux effets ont déjà été mis en avant : le fait que la possibilité d'atteindre une grande taille de population favorise l'apparition de mutations rares (A. R. Hall et al. 2010 ; Garoff et al. 2020), et le risque de perte de mutations bénéfiques lors des dilutions (Wahl, Gerrish et Saika-Voivod 2002). Notre objectif était d'étudier comment ces deux phénomènes se combinent et impactent les chemins évolutifs suivis par la population et sa vitesse d'adaptation.
Pour cela, nous avons modélisé une population asexuée qui alterne périodiquement entre des phases de croissance et des dilutions, et évoluant dans un paysage adaptatif minimal où deux types de mutation peuvent survenir, présentant un compromis entre taux de mutation et avantage sélectif. Le modèle que nous avons développé est semi-déterministe : croissance déterministe pour la souche initiale que l'on suppose être présente en grand nombre, et comportement stochastique pour les mutants qui arrivent en petits nombres au début d'une telle expérience. Afin de pouvoir déterminer quels sont les scénarios évolutifs les plus probables, nous avons étudié le comportement limite du modèle lorsque la taille de population est grande et que le taux de mutation est faible. Nous avons par la suite vérifié nos prédictions avec des simulations utilisant des valeurs réalistes de ces paramètres.
Grâce à notre modèle, nous avons déterminé les chemins évolutifs les plus probables en fonction de deux paramètres : l'abondance minimale et l'abondance maximale de la population au cours d'un cycle. Nous avons aussi prédit le nombre de dilutions nécessaires pour voir apparaître des doubles mutants, en fonction des mêmes paramètres démographiques.
Ce travail a été encadré par Amaury Lambert, François Blanquart et Sylvain Gandon. Il a fait l'objet d'une publication dans la revue *Genetics* (Gamblin, Gandon et al. 2023).

Comment les goulots d'étranglement façonnent le potentiel d'adaptation : de la microbiologie à la biologie de la conservation. Le Chapitre 3 présente une revue de littérature sur l'effet des goulots d'étranglements sur la capacité d'adaptation des populations, qu'elles soient microbiennes ou animales. Les populations sauvages subissent fréquemment des changements démographiques qui peuvent déstabiliser leur persistance et, par conséquent, l'équilibre des écosystèmes. En réduisant la diversité génétique, un goulot d'étranglement peut empêcher une population de s'adapter aux changements environnementaux ultérieurs. Avec le changement climatique et la contraction des habitats naturels des populations, il semble essentiel de pouvoir prédire avec précision l'impact des goulots d'étranglement sur le potentiel adaptatif des populations. Dans ce contexte, nous avons cherché à déterminer dans quelle mesure les connaissances acquises grâce à la modélisation et à la microbiologie expérimentale peuvent être appliquées aux populations animales sauvages.
Pour ce faire, nous avons passé en revue les effets des goulots d'étranglement sur le potentiel adaptatif, en les décomposant entre les quatre principaux processus évolutifs (mutation, dérive génétique, sélection naturelle et flux de gènes). Pour chacun d'entre eux, nous abordons les questions suivantes : Que nous apprennent la théorie et les expériences microbiologiques sur l'influence des goulots d'étranglement sur la capacité des populations à s'adapter aux futurs changements environnementaux ? Ces prédictions théoriques peuvent-elles être appliquées aux populations sauvages ? Que manque-t-il pour mieux prédire l'évolution après un goulot d'étranglement ?
Ce travail a été mené en collaboration avec Laure Olazcuaga et Loïc Marrec, et a été déposé sur un serveur de prépublications (Gamblin, Marrec et Olazcuaga 2024).

1.4.2 Évolution des pangénomes bactériens

La deuxième partie s'intéresse à des échelles spatiales et temporelles plus larges, et étudie l'évolution des pangénomes bactériens. Comme abordé dans les Sections 1.1.3 et 1.2.2, de nombreuses espèces bactériennes présentent une diversité de gènes impressionnante : le nombre de gènes présents dans l'espèce peut potentiellement être beaucoup plus grand que celui contenu dans le génome d'un individu, ce qui a mené à l'introduction du concept de pangénome. Or de nombreuses questions demeurent concernant les mécanismes évolutifs qui façonnent ces pangénomes.

Gènes persistants, privés et mobiles : un modèle pour la dynamique des gènes dans les pangénomes bactériens. Le Chapitre 4 présente le développement et l'application d'un modèle de dynamique des gènes dans une espèce bactérienne. Caractériser les dynamiques d'importation et de transferts de gènes bactériens est crucial pour comprendre l'origine de cette formidable diversité. Cette dynamique est aussi d'une grande importance en santé publique pour la diffusion des gènes de virulence et d'antibiorésistance au sein des populations bactériennes. Depuis 10 ans, plusieurs modèles ont été formulés pour décrire la dynamique de gènes bactériens dans une phylogénie (Baumdicker, Hess et Pfaffelhuber 2012 ; Zamani-Dahaj et al. 2016) mais demeurent insatisfaisants. En effet,

ces modèles prennent en compte soit l'import des gènes dans l'espèce focale (par transfert depuis une autre espèce) soit les transferts horizontaux au sein de l'espèce, mais jamais les deux. Nous avons développé un nouveau modèle d'évolution du pangénome bactérien le long de la phylogénie d'une espèce, qui décrit explicitement le moment de l'apparition de chaque gène dans l'espèce et tient compte de trois types génériques de dynamique évolutive des gènes : les gènes « persistants » présents dans le génome ancestral, les gènes « privés » spécifiques à un clade donné et les gènes « mobiles » importés une fois dans le pool génétique et subissant ensuite de fréquents transferts horizontaux. Ce modèle a été nommé PPM pour Persistant-Privé-Mobile. Nous avons développé un algorithme d'inférence par maximum de vraisemblance adapté au modèle PPM et l'avons appliqué à un ensemble de 902 génomes de *Salmonella enterica*. Nous avons montré que ce modèle est capable de reproduire le schéma global de certaines statistiques multivariées telles que le spectre de fréquence des gènes et le diagramme de parcimonie par rapport à la fréquence. De plus, la classification des gènes induite par le modèle PPM nous permet d'étudier la position des gènes accessoires le long du chromosome, ainsi que les fonctions génétiques les plus représentées dans chaque catégorie. Ce travail a été encadré par Amaury Lambert et François Blanquart.

Références du Chapitre 1

Arnold, B. J., I.-T. Huang et W. P. Hanage (2022). Horizontal Gene Transfer and Adaptive Evolution in Bacteria. *Nature Reviews Microbiology* **20**, 206-218.

Bar-On, Y. M., R. Phillips et R. Milo (2018). The Biomass Distribution on Earth. *Proceedings of the National Academy of Sciences* **115**, 6506-6511.

Baumdicker, F., W. R. Hess et P. Pfaffelhuber (2012). The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biology and Evolution* **4**, 443-456.

Beal, J., N. G. Farny, T. Haddock-Angelli, V. Selvarajah, G. S. Baldwin, R. Buckley-Taylor, M. Gershater, D. Kiga, J. Marken, V. Sanchania, A. Sison et C. T. Workman (2020). Robust Estimation of Bacterial Cell Count from Optical Density. *Communications Biology* **3**, 1-29.

Blount, Z. D., C. Z. Borland et R. E. Lenski (2008). Historical Contingency and the Evolution of a Key Innovation in an Experimental Population of Escherichia Coli. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 7899-7906.

Cammarota, G., G. Ianiro et A. Gasbarrini (2014). Fecal Microbiota Transplantation for the Treatment of Clostridium Difficile Infection: A Systematic Review. *Journal of Clinical Gastroenterology* **48**, 693.

Chartrain, M., P. M. Salmon, D. K. Robinson et B. C. Buckland (2000). Metabolic Engineering and Directed Evolution for the Production of Pharmaceuticals. *Current Opinion in Biotechnology* **11**, 209-214.

Cohen, O. et T. Pupko (2010). Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution* **27**, 703-713.

Didelot, X., A. E. Darling et D. Falush (2008). Inferring Genomic Flux in Bacteria. *Genome Research* **19**, 306-317.

Dixit, P. D., T. Y. Pang et S. Maslov (2017). Recombination-Driven Genome Evolution and Stability of Bacterial Species. *Genetics* **207**, 281-295.

Domingo-Sananes, M. R. et J. O. McInerney (2021). Mechanisms That Shape Microbial Pangenomes. *Trends in Microbiology* **29**, 493-503.

Dorigatti, I., C. McCormack, G. Nedjati-Gilani et N. M. Ferguson (2018). Using Wolbachia for Dengue Control: Insights from Modelling. *Trends in Parasitology* **34**, 102-113.

Duong, M. T.-Q., Y. Qin, S.-H. You et J.-J. Min (2019). Bacteria-Cancer Interactions: Bacteria-Based Cancer Therapy. *Experimental & Molecular Medicine* **51**, 1-15.

Felsenstein, J. (1981). Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution* **17**, 368-376.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222**, 309-368.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford : The Clarendon Press. 272 p.

Foutel-Rodier, F., F. Blanquart, P. Courau, P. Czuppon, J.-J. Duchamps, J. Gamblin, É. Kerdoncuff, R. Kulathinal, L. Régnier, L. Vuduc, A. Lambert et E. Schertzer (2022). From Individual-Based Epidemic Models to McKendrick-von Foerster PDEs: A Guide to Modeling and Inferring COVID-19 Dynamics. *Journal of Mathematical Biology* **85**, 43.

Fraser, C., W. P. Hanage et B. G. Spratt (2007). Recombination and the Nature of Bacterial Speciation. *Science* **315**, 476-480.

Froissart, R. et C. Brives (2021). Evolutionary Biology and Development Model of Medicines: A Necessary 'Pas de Deux' for Future Successful Bacteriophage Therapy. *Journal of Evolutionary Biology* **34**, 1855-1866.

Gamblin, J., S. Gandon, F. Blanquart et A. Lambert (2023). Bottlenecks Can Constrain and Channel Evolutionary Paths. *Genetics* **224**. Sous la dir. de K. Jain, iyad001.

Gamblin, J., L. Marrec et L. Olazcuaga (2024). How Bottlenecks Shape Adaptive Potential: From Theory and Microbiology to Conservation Biology. (Visité le 05/02/2024).

Garoff, L., F. Pietsch, D. L. Huseby, T. Lilja, G. Brandis et D. Hughes (2020). Population Bottlenecks Strongly Influence the Evolutionary Trajectory to Fluoroquinolone Resistance in Escherichia Coli. *Molecular Biology and Evolution* **37**, 1637-1646.

Georjon, H. et A. Bernheim (2023). The Highly Diverse Antiphage Defence Systems of Bacteria. *Nature Reviews Microbiology* **21**, 686-700.

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry* **81**, 2340-2361.

— (2001). Approximate Accelerated Stochastic Simulation of Chemically Reacting Systems. *The Journal of Chemical Physics* **115**, 1716-1733.

Good, B. H., M. J. McDonald, J. E. Barrick, R. E. Lenski et M. M. Desai (2017). The Dynamics of Molecular Evolution over 60,000 Generations. *Nature* **551**, 45-50.

Gresham, D. et M. J. Dunham (2014). The Enduring Utility of Continuous Culturing in Experimental Evolution. *Genomics*. Experimental Evolution and the Use of Genomics **104** (6, Part A), 399-405.

Hall, A. R., V. F. Griffiths, R. C. MacLean et N. Colegrave (2010). Mutational Neighbourhood and Mutation Supply Rate Constrain Adaptation in Pseudomonas Aeruginosa. *Proceedings of The Royal Society B: Biological Sciences* **277**, 643-650.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97-109.

Hobson, C. A., L. Vigue, S. Naimi, B. Chassaing, M. Magnan, S. Bonacorsi, B. Gachet, I. El Meouche, A. Birgy et O. Tenaillon (2022). MiniBioReactor Array (MBRA) in Vitro Gut Model: A Reliable System to Study Microbiota-Dependent Response to Antibiotic Treatment. *JAC-Antimicrobial Resistance* **4**, dlac077.

Hodcroft, E. B., M. Zuber, S. Nadeau, T. G. Vaughan, K. H. D. Crawford, C. L. Althaus, M. L. Reichmuth, J. E. Bowen, A. C. Walls, D. Corti, J. D. Bloom, D. Veesler, D. Mateo, A. Hernando, I. Comas, F. González-Candelas, T. Stadler et R. A. Neher (2021). Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020. *Nature* **595**, 707-712.

Hou, K., Z.-X. Wu, X.-Y. Chen, J.-Q. Wang, D. Zhang, C. Xiao, D. Zhu, J. B. Koya, L. Wei, J. Li et Z.-S. Chen (2022). Microbiota in Health and Diseases. *Signal Transduction and Targeted Therapy* **7**, 1-28.

Hubbarde, J. E., G. Wild et L. M. Wahl (2007). Fixation Probabilities When Generation Times Are Variable: The Burst–Death Model. *Genetics* **176**, 1703-1712.

Huynh, L., J. G. Scott et P. J. Thomas (2023). Inferring Density-Dependent Population Dynamics Mechanisms through Rate Disambiguation for Logistic Birth-Death Processes. *Journal of Mathematical Biology* **86**, 50.

Ikuta, K. S. et al. (2022). Global Mortality Associated with 33 Bacterial Pathogens in 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *The Lancet* **400**, 2221-2248.

Jukes, T. H. et C. R. Cantor (1969). CHAPTER 24 - Evolution of Protein Molecules. *Mammalian Protein Metabolism*. Sous la dir. de H. N. Munro. Academic Press, 21-132.

Kawecki, T. J., R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri et M. C. Whitlock (2012). Experimental Evolution. *Trends in ecology & evolution*.

Kendall, D. G. (1948). On the Generalized "Birth-and-Death" Process. *The Annals of Mathematical Statistics* **19**, 1-15.

Lambert, S., J. Voznica et H. Morlon (2023). Deep Learning from Phylogenies for Diversification Analyses. *Systematic Biology* **72**, 1262-1279.

Linton, N. M., T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita et H. Nishiura (2020). Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. *Journal of Clinical Medicine* **9** (2), 538.

Luria, S. E. et M. Delbrück (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491-511.

Lynch, M. (2010). Evolution of the Mutation Rate. *Trends in Genetics* **26**, 345-352.

McDonald, M. J. (2019). Microbial Experimental Evolution – a Proving Ground for Evolutionary Theory and a Tool for Discovery. *EMBO reports* **20**, e46992.

McInerney, J. O., A. McNally et M. J. O'Connell (2017). Why Prokaryotes Have Pangenomes. *Nature microbiology* **2**, 17040.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler et R. Lanfear (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534.

Moinard, S., E. Oudet, D. Piau, E. Coissac et C. Gonindard-Melodelima (2023). The Fixed Landscape Inference MethOd (Flimo): A Versatile Alternative to Approximate Bayesian Computation, Faster by Several Orders of Magnitude. *arXiv.org*.

Murray, C. J. L. et al. (2022). Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis. *The Lancet* **399**, 629-655.

Östblom, A., I. Adlerberth, A. E. Wold et F. L. Nowrouzian (2011). Pathogenicity Island Markers, Virulence Determinants malX and Usp, and the Capacity of Escherichia Coli To Persist in Infants' Commensal Microbiotas. *Applied and Environmental Microbiology* **77**, 2303-2308.

Pagel, M. (1997). Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **255**, 37-45.

Perrin, A. et E. P. C. Rocha (2021). PanACoTA: A Modular Tool for Massive Microbial Comparative Genomics. *NAR Genomics and Bioinformatics* **3**, lqaa106.

Plucain, J., T. Hindré, M. Le Gac, O. Tenaillon, S. Cruveiller, C. Médigue, N. Leiby, W. R. Harcombe, C. J. Marx, R. E. Lenski et D. Schneider (2014). Epistasis and Allele Specificity in the Emergence of a Stable Polymorphism in Escherichia Coli. *Science* **343**, 1366-1369.

Rodriguez-Valera, F., A.-B. Martín-Cuadrado, B. Rodriguez-Brito, L. Pasic, T. F. Thingstad, F. Rohwer et A. Mira (2009). Explaining Microbial Population Genomics through Phage Predation. *Nature Precedings*, 1-1.

Ross-Gillespie, A., A. Gardner, S. A. West et A. S. Griffin (2007). Frequency Dependence and Cooperation: Theory and a Test with Bacteria. *The American Naturalist* **170**, 331-342.

Sakoparnig, T., C. Field et E. van Nimwegen (2021). Whole Genome Phylogenies Reflect the Distributions of Recombination Rates for Many Bacterial Species. *eLife* **10**. Sous la dir. d'A. Nourmohammad et A. M. Walczak, e65366.

Schmidt, B. (2010). *Bioinformatics: High Performance Parallel Computer Architectures*. CRC Press. 372 p.

Silvestro, D., T. Latrille et N. Salamin (2023). Improved Estimation of Molecular Evolution Coupling Stochastic Simulations and Deep Learning. *arXiv.org*.

Sniegowski, P. D., P. J. Gerrish et R. E. Lenski (1997). Evolution of High Mutation Rates in Experimental Populations of E. Coli. *Nature* **387**, 703-705.

Son, M. S. et R. K. Taylor (2021). Growth and Maintenance of Escherichia Coli Laboratory Strains. *Current protocols* **1**, e20.

Stamatakis, A. (2014). RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* **30**, 1312-1313.

Tavare, S., D. J. Balding, R. C. Griffiths et P. Donnelly (1997). Inferring Coalescence Times from DNA Sequence Data. *Genetics* **145**, 505-518.

Tenaillon, O., D. Skurnik, B. Picard et E. Denamur (2010). The Population Genetics of Commensal Escherichia Coli. *Nature Reviews Microbiology* **8**, 207-217.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli et C. M. Fraser (2005). Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome". *Proceedings of the National Academy of Sciences* **102**, 13950-13955.

Tonkin-Hill, G., N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, S. D. Frost, J. Corander, S. D. Bentley et J. Parkhill (2020). Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. *Genome Biology* **21**, 180.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. L. Bouguénec, M. Lescat, S. Mangenot, V. Martinez-Jéhanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Médigue, E. P. C. Rocha et E. Denamur (2009). Organised Genome Dynamics in the Escherichia Coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics* **5**, e1000344.

Touchon, M., A. Perrin, J. A. M. de Sousa, B. Vangchhia, S. Burn, C. L. O'Brien, E. Denamur, D. Gordon et E. P. Rocha (2020). Phylogenetic Background and Habitat Drive the Genetic Diversification of Escherichia Coli. *PLOS Genetics* **16**, e1008866.

Utarini Adi, Indriani Citra, Ahmad Riris A., Tantowijoyo Warsito, Arguni Eggi, Ansari M. Ridwan, Supriyati Endah, Wardana D. Satria, Meitika Yeti, Ernesia Inggrid, Nurhayati Indah, Prabowo Equatori, Andari Bekti, Green Benjamin R., Hodgson Lauren, Cutcher Zoe, Rancès Edwige, Ryan Peter A., O'Neill Scott L., Dufault Suzanne M., Tanamas Stephanie K., Jewell Nicholas P., Anders Katherine L. et Simmons Cameron P. (2021). Efficacy of Wolbachia-Infected Mosquito Deployments for the Control of Dengue. *New England Journal of Medicine* **384**, 2177-2186.

Valles-Colomer, M., G. Falony, Y. Darzi, E. F. Tigchelaar, J. Wang, R. Y. Tito, C. Schiweck, A. Kurilshikov, M. Joossens, C. Wijmenga, S. Claes, L. Van Oudenhove, A. Zhernakova, S. Vieira-Silva et J. Raes (2019). The Neuroactive Potential of the Human Gut Microbiota in Quality of Life and Depression. *Nature Microbiology* **4**, 623-632.

Van Etten, J. et D. Bhattacharya (2020). Horizontal Gene Transfer in Eukaryotes: Not If, but How Much? *Trends in Genetics* **36**, 915-925.

Wahl, L. M., P. J. Gerrish et I. Saika-Voivod (2002). Evaluating the Impact of Population Bottlenecks in Experimental Evolution. *Genetics* **162**, 961-971.

Wang, Z., J. Sun, Y. Gao, Y. Xue, Y. Zhang, K. Li, W. Zhang, C. Zhang, J. Zu et L. Zhang (2023). Fusang: A Framework for Phylogenetic Tree Inference via Deep Learning. *Nucleic Acids Research* **51**, 10909-10923.

Whelan, F. J., R. J. Hall et J. O. McInerney (2021). Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. *Molecular Biology and Evolution* **38**, 3697-3708.

Wiser, M. J., N. Ribeck et R. E. Lenski (2013). Long-Term Dynamics of Adaptation in Asexual Populations. *Science* **342**, 1364-1367.

Woods, R., D. Schneider, C. L. Winkworth, M. A. Riley et R. E. Lenski (2006). Tests of Parallel Molecular Evolution in a Long-Term Experiment with Escherichia Coli. *Proceedings of the National Academy of Sciences* **103**, 9107-9112.

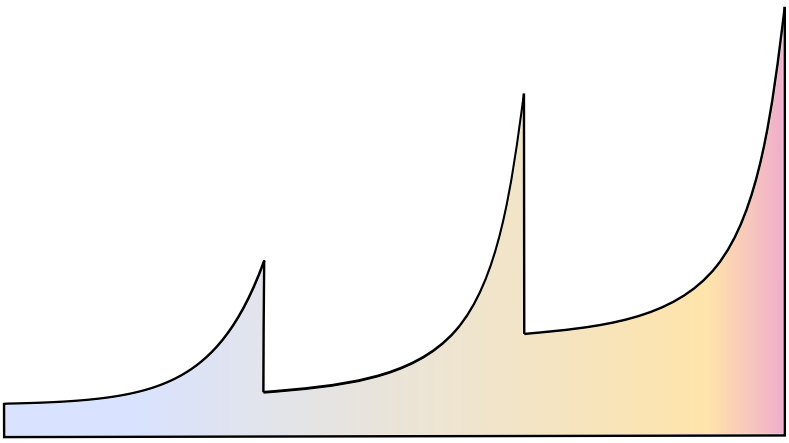
Wright, S. (1931). Evolution in Mendelian Populations. *Genetics* **16**, 97-159.

Zamani-Dahaj, S. A., M. Okasha, J. Kosakowski et P. Higgs (2016). Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Molecular Biology and Evolution* **33**, 1843-1857.

Zheng, Q. (1999). Progress of a Half Century in the Study of the Luria–Delbrück Distribution. *Bellman Prize in Mathematical Biosciences* **162**, 1-32.

Impacts évolutifs des goulots d'étranglement de population

Chapitre 2



2

Bottlenecks can constrain and channel evolutionary paths

Sommaire

2.1 Introduction

Population bottlenecks are sudden, drastic reductions of population size that can arise both *in vivo* and *in vitro*. Pathogen populations experience such bottlenecks during host to host transmission (Abel et al. 2015; Geoghegan, Senior, and Holmes 2016), or they can be induced by resource limitation or seasonality (e.g. the boom and bust dynamics of phytoplankton, Behrenfeld et al. 2017). They also are commonplace in experimental evolution: in serial passaging (or transfer) experiments, a microbial population is periodically subsampled and placed on new medium to grow again (Kawecki et al. 2012). In this way, microbial populations can be followed during several generations (Good et al. 2017) while remaining of a manageable size.

Several studies have investigated how periodic bottlenecks influence the rate of adaptation of such populations. They have mostly focused on the probability of stochastically losing beneficial mutations (Wahl and Gerrish 2001), on the time of arrival of successful mutations (Wahl, Gerrish, and Saika-Voivod 2002), on mutant fixation (A. R. Hall et al. 2010; Garoff et al. 2020; Schenk et al. 2022) and on the predictability of evolution (Szendro et al. 2013; Freitas, Wahl, and Campos 2021) with applications to the study of drug resistance (Huseby et al. 2017; Nicholson and Antal 2019; Mahrt et al. 2021) (see LeClair and Wahl 2018 for a review).

Here we examine theoretically how demography affects not only the rate of adaptation, but also the evolutionary paths followed by bottlenecked populations. In populations with constant size evolving in a regime of strong selection-strong mutation (as is often the case for experimental asexual populations), the distribution of fitness effect of fixed mutations and the rate of adaptation are dictated by the population size, the mutation rate and the shape of the distribution of fitness effects (Desai and D. S. Fisher 2007; Fogle, Nagle, and Desai 2008). In bottlenecked populations, both the supply of beneficial mutations and the probability of establishment change over time, and the accessibility of evolutionary paths can in theory depend on bottleneck size or severity and on cycle duration. Furthermore, adaptation may increase population size, feed back on the mutation supply and enhance the scope for further adaptation. These effects are potentially important for experimental and natural evolution but have not been studied theoretically.

To study how bottlenecks influence the rate and paths of adaptation, we assume a minimal fitness landscape with a trade-off between the rate of appearance of beneficial mutations and their fitness advantage, consistently with the decreasing distribution of fitness effects documented in multiple species (Eyre-Walker and Keightley 2007). We take advantage of large population limit techniques as is standard in population genetics, notably in Luria-Delbrück type fluctuation experiments of microbial populations (Luria and Delbrück 1943; Zheng 1999). We study the timing of emergence and establishment of beneficial mutations to investigate the effect of bottleneck size and cycle duration, or equivalently of the initial and final population sizes of the wild-type, on which paths are accessible to evolution.

2.2 Model

We consider an asexual population adapting to a new environment. We assume two types of beneficial mutations, with a trade-off between fitness and mutation rate. High-rate, weakly beneficial mutations confer a moderate gain in fitness, while low-rate, strongly beneficial mutations confer a large gain in fitness. This setting can be thought of as a coarse discretization of a decreasing (e.g. exponential) distribution of fitness effects (Kassen and Bataillon 2006). These two types of mutations can be thought to target different loci underlying traits linked to adaptation, e.g. resistance to a drug or predator, or the exploitation of a new resource. We thus obtain a simple fitness landscape composed of 4 genotypes: '00' for the wild-type (also denoted WT), '10' for individuals with a weakly beneficial mutation, '01' for individuals with a strongly beneficial mutation, and '11' for individuals carrying both types of mutation. Thus, mutations $00 \longrightarrow 10$ and $01 \longrightarrow 11$ are weakly beneficial, while $00 \longrightarrow 01$ and $10 \longrightarrow 11$ are strongly beneficial. We allow magnitude epistasis but not sign epistasis, thus the growth rates of the four genotypes verify: $r_{11} > r_{01} > r_{10} > r_{00}$ (see Figure 2.1B). We neglect the production of double mutants by recombination between single mutants.

A semi-deterministic model. Population size grows exponentially and is subject to periodic bottlenecks of fixed relative severity, i.e., the fraction of population that survives is constant (as opposed to fixed absolute bottleneck severity, where the number of individuals that survive is constant, see LeClair and Wahl 2018). As a consequence, the WT population goes back to its initial size N_0 at the beginning of each cycle. However, the size of the whole population can increase through successive transfers due to the arrival of beneficial mutants, see Figure 2.1D. We use a semi-deterministic model to describe the dynamics of the population. We assume that N_0 is sufficiently large that the growth of the WT population during one cycle can be described deterministically: $N_t = N_0 e^{r_{00} t}$. In contrast, the population dynamics of mutants, always starting in small numbers of copies, will be described by a stochastic birth-death model. The weakly beneficial and strongly beneficial mutation rates are respectively μ_{high} and μ_{low}, with $\mu_{\text{high}} \gg \mu_{\text{low}}$ (see Figure 2.1A). We denote by $\mu = \mu_{\text{high}} + \mu_{\text{low}}$ the total mutation rate to beneficial mutations. This semi-deterministic setting is similar to the one first used in (Lea and Coulson 1949) to model the Luria-Delbrück experiment, and more recently in (P. Keller and Antal 2014).

Large population, small mutation rate assumption. The mutation rate to a beneficial mutation is typically very small, while the size of microbial populations is usually quite large. Thus we introduce a scaling parameter n that is of the order of $\frac{1}{\mu}$, and we will assume in the following that $n \gg 1$. In the large n limit the probability of most events of interest approaches either 1 or 0, allowing us to determine the most likely scenario in a given parameter setting. To comply with

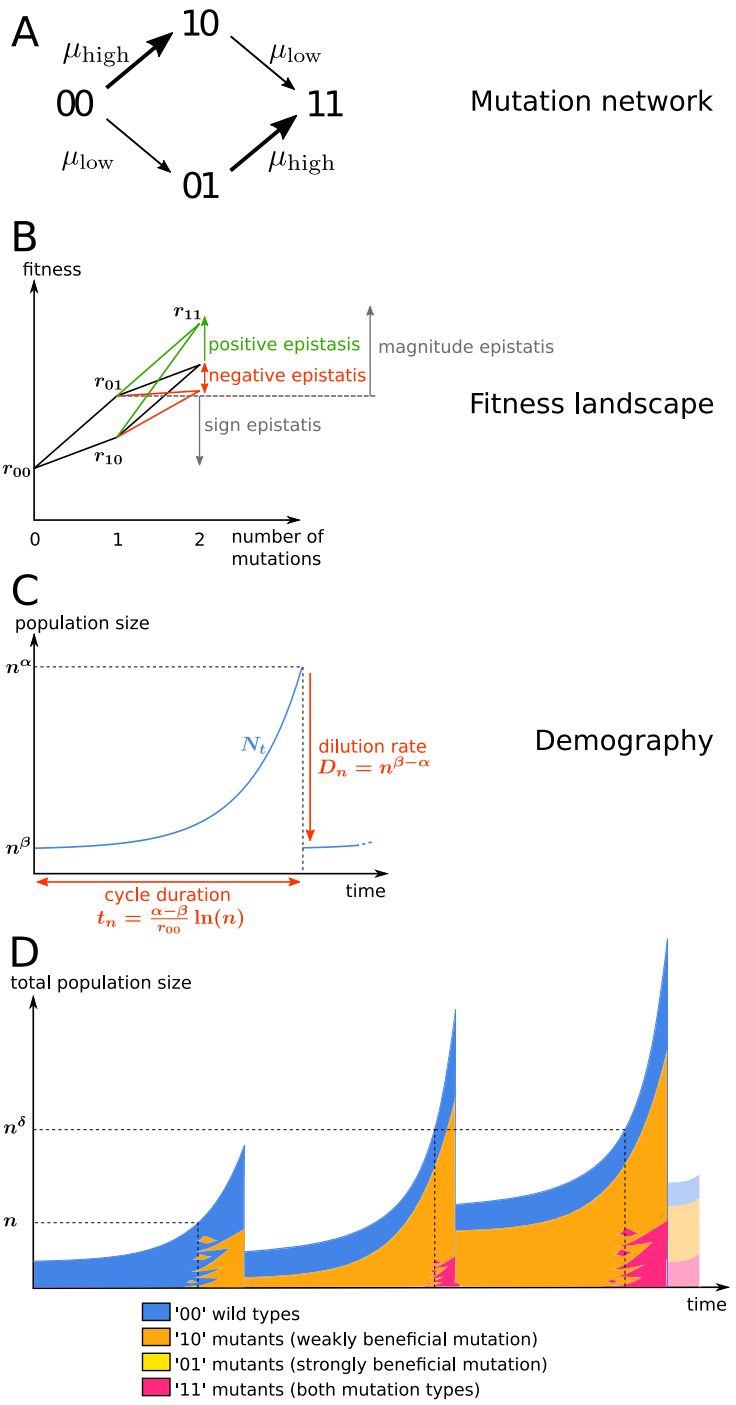


Figure 2.1 – A: Mutation network. B: Fitness landscape, growth rate as a function of the number of mutations. C: Demography during one cycle of growth, WT population size as a function of time. D: Illustration of demography during 3 cycles of growth, total population size as a function of time. Mutants start to appear when the population size is around the inverse of the mutation rate. Mutant subpopulations are composed of multiple independent clones, which can go extinct due to stochasticity (genetic drift, bottleneck). The illustrated scenario corresponds to the blue area in Figure 2.2, where strongly beneficial '01' mutants never establish in the population but weakly beneficial '10' mutants and double mutants do.

Table 2.1 – Main parameters of the model

Notation	Interpretation
r_{00}	growth rate of WT
r_{10}, r_{01}	growth rates of single mutants
r_{11}	growth rate of double mutant
$\mu_{\text{high}} = 1/n$	weakly beneficial mutation rate
$\mu_{\text{low}} = 1/n^{\delta}$	strongly beneficial mutation rate
$\mu = \mu_{\text{high}} + \mu_{\text{low}}$	global beneficial mutation rate
N_t	size of the WT population at time t
$N_0 = n^{\beta}$	initial WT population size
$N_{t_n} = n^{\alpha}$	final WT population size
$t_n = (\alpha - \beta)\ln(n)/r_{00}$	time duration of one cycle
$D_n = n^{\beta-\alpha}$	dilution factor

$\mu_{\text{high}} \gg \mu_{\text{low}}$, we also introduce a parameter $\delta > 1$ such that

$$\mu_{\text{high}} = \frac{1}{n} \quad \text{and} \quad \mu_{\text{low}} = \frac{1}{n^{\delta}}.$$

We suppose that the time t_n between two dilutions is such that during one cycle of the experiment, the WT population grows from a size $N_0 = n^{\beta}$ to n^{α}. As a consequence,

$$t_n = (\alpha - \beta)\frac{\ln(n)}{r_{00}}.$$

We constrain β to be in $(0, 1)$ and α to be greater than 1, in order to have $N_0 \ll n \ll N_{t_n}$. In that way, we ensure that there is no mutant at the beginning of the experiment, and that weakly beneficial mutants will appear during the first cycle with high probability. Because of the large n assumption, there is a sharp transition between a regime where mutations are very unlikely to occur (for $N_t\mu \ll 1$) to a regime where numerous mutations arise (for $N_t\mu \gg 1$). Thus we expect to observe adaptation via multiple-origins soft sweeps in the second regime (Messer and Petrov 2013; Hermisson and Pennings 2017), in agreement with empirical observations in microbes (Nair et al. 2006; Pennings, Kryazhimskiy, and Wakeley 2014; Barroso-Batista et al. 2014). The dilution factor between two cycles must be chosen so that the WT population always starts afresh at the same size $N_0 = n^{\beta}$. Thus the dilution factor is

$$D_n = \frac{1}{n^{\alpha-\beta}}.$$

The notation is summarized in Figure 2.1 and Table 2.1.

2.3 Results

Here we analyze the dynamics of adaptation by characterizing the evolutionary paths followed by evolution, the timing of this process, and how it depends on

demographic parameters. For detailed derivations, see Section 2.5, which also provides insight into the timing of weakly beneficial mutations during the first cycle. When a mutant population arises at cycle k, escapes stochastic extinction and so is present in non-negligible quantity at the end of the growth phase, we will say that this population has **established** during cycle k. If the established mutant population reaches a sufficiently large size that it survives the next bottleneck (and so every subsequent bottleneck), we will say that this subpopulation **survives**. We say that an event A occurs with high probability (w.h.p.) if $P(A) \to 1$ as $n \to +\infty$.

2.3.1 Effect of demography on evolutionary paths

Single mutant '10' establishes w.h.p. at first cycle (assuming $\alpha > 1$). The product of the WT population size at the end of first cycle and of the high mutation rate is $\mu_{\text{high}} N_{t_n} = (1/n)n^{\alpha} \gg 1$. We rigorously showed in Section 2.5.4 that '10' mutants establish w.h.p. during the first cycle and computed an estimate of the number $Z_{10}^{(n)}$ of '10' mutants at the end of this cycle.

Single mutant '01' establishes w.h.p. at first cycle if $\alpha > \delta$ (and if $\alpha < \delta$ w.h.p. never arises). The establishment of mutants '01' depends on the relative values of the final population size and the rate of strongly beneficial mutations, governed by parameters α and δ. If $\alpha > \delta$, then $\mu_{\text{low}} N_{t_n} \gg 1$ and mutants '01' also establish during the first cycle. If on the contrary $\alpha < \delta$, the probability that mutants '01' arise in the first cycle is close to 0. As the WT population has exactly the same size at the end of each cycle, it is unlikely that mutants '01' establish in the course of the experiment. This highlights that the demographic control imposed on the WT population affects the establishment of mutations. In fact, it also affects which evolutionary paths are accessible: in the case where $\alpha < \delta$ the transition $00 \longrightarrow 01$ is not possible (and neither is $01 \longrightarrow 11$).

Significance of demographic parameters α and β. For fixed values of the growth rates (r_{00}, r_{10} and r_{01}) and of δ, we can project on the plane (β, α) areas corresponding to different configurations of evolutionary paths. All path configurations must include the $00 \longrightarrow 10$ transition because '10' establish w.h.p. during the first cycle, thus there are 6 possible configurations. Figure 2.2 shows the areas corresponding to different path configurations for a chosen parameter set. This set was such that these 6 configurations are present, but this is not always the case. Equations for threshold lines (1-5) are derived and explained in Section 2.5.7. These equations are derived for $n \to +\infty$, i.e., for infinitely large populations and vanishing mutation rates. See Figure 2.6 for observed evolutionary paths in simulations with finite values of n. Together, α and β determine the initial and final WT population size at each cycle, the duration of each cycle and the relative severity of each bottleneck. When α increases, the final WT population size, bottleneck severity and cycle duration increase. When β increases, the initial WT population size increases but bottleneck severity and cycle duration decrease.

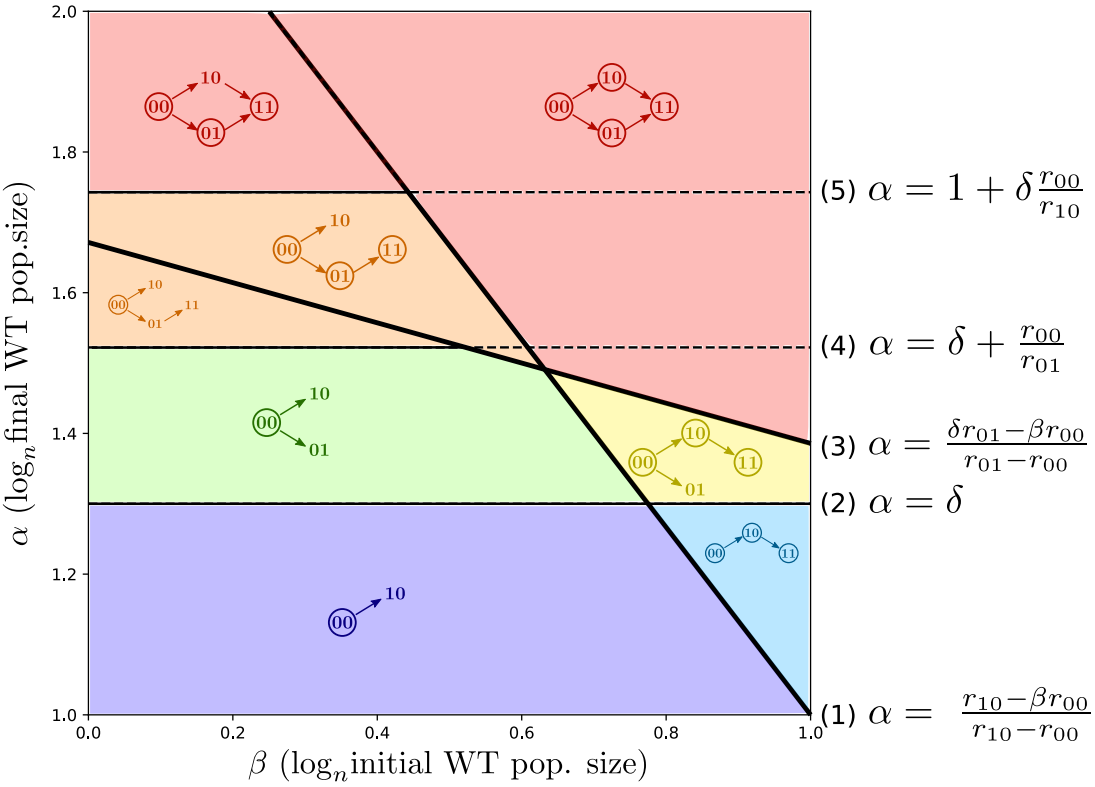


Figure 2.2 – Predicted evolutionary paths, as a function of demographic parameters β and α. These predictions are made for infinitely large populations and vanishing mutation rates. Arrows are observed evolutionary paths. Genotypes that are shown but not circled are establishing but not surviving. Genotypes that are circled are surviving. The six colors (red, orange, yellow, green, blue and purple) correspond to six different predicted path configurations. The dilution ratio is constant along lines of slope 1. ($\delta = 1.3$, $r_{00} = 0.2$, $r_{10} = 0.35$ and $r_{01} = 0.9$)

Single mutants survive if their population size after the first growth phase is larger than the inverse of dilution rate. Once a mutant population is established, it can either disappear because of the bottleneck or survive dilution and pass to the next cycle. For mutants '10', these two outcomes correspond to two regions of parameter space delimited by line (1)

$$\alpha = \frac{r_{10} - \beta r_{00}}{r_{10} - r_{00}}.$$

Indeed, $\alpha > (r_{10} - \beta r_{00})/(r_{10} - r_{00})$ is equivalent to having $Z_{10}^{(n)} D_n \gg 1$ with $Z_{10}^{(n)}$ the number of '10' mutants after one growth phase (see Section 2.5.6), and in that case the probability that at least one mutant '10' survives dilution goes to 1 as $n \to +\infty$. Thus, above line (1) the '10' mutant population survives the bottleneck

w.h.p. and because the dilution factor is constant, it will also survive every subsequent bottleneck. The '10' population will then continue growing, allowing double mutants to arise. Below this line, the '10' mutant population at the end of the first cycle is too small to survive dilution, but re-establishes from the WT w.h.p. at each new cycle of the experiment. For '01' mutants the delimitation between establishment and survival is line (3)

$$\alpha = \frac{\delta r_{01} - \beta r_{00}}{r_{01} - r_{00}}.$$

This line is analogous to line (1) except that mutant '01' appears later (when population size reaches $\sim n^{\delta}$) but grows faster. Thus lines (1) and (3) delineate four different zones that we named the **Southwest**, **Southeast**, **Northwest** and **Northeast** corners. These zones can be further delimited into different colored areas depending on the value of α (Figure 2.2).

Southwest corner: no adaptation. In the Southwest corner, no single mutant population can survive. When $\alpha < \delta$ (purple area under line (2)) only '10' mutants can establish, they never survive but re-establish w.h.p. at each new cycle. Above line (2), in the green area, both '10' and '01' mutants establish at each cycle but w.h.p. never survive. In these purple and green areas, the probability for single mutants to survive (and thus, for double mutants to establish) goes to 0 as $n \to +\infty$ as long as the number of cycles performed is finite. Of course, this is not true anymore for real populations for which n is finite. In that case, we can expect to see weakly beneficial single mutants survive when performing $O(n^{\alpha-\beta-(\alpha-1)r_{10}/r_{00}})$ cycles. When α increases above line (4), the time duration of a growth phase increases and the final population size of '01' mutants becomes large enough for double mutants to arise and establish. These double mutants are lost in dilution w.h.p., except for high values of r_{11} for which they are able to survive even though neither '01' nor '10' survive themselves (case not represented on Figure 2.2).

Southeast corner: adaptation via '10'. In the Southeast corner, only mutants '10' and '11' survive. Depending on the sign of $\alpha - \delta$, we are either in the blue area with no '01' mutant or in the yellow one with '01' establishing repeatedly but not surviving bottlenecks. The example scenario displayed in Figure 2.1D corresponds to the blue area.

Northwest corner: adaptation via '01'. In the Northwest corner, only mutants '01' and '11' survive. The '10' mutant population re-establishes at each cycle. If α increases above line (5), the growth phase lasts sufficiently long for the size reached by the '10' mutant population to also produce double mutants.

Northeast corner: adaptation via both '10' and '01'. In the Northeast corner (in red), all transitions are observed and all mutants will eventually establish and survive if enough cycles are performed.

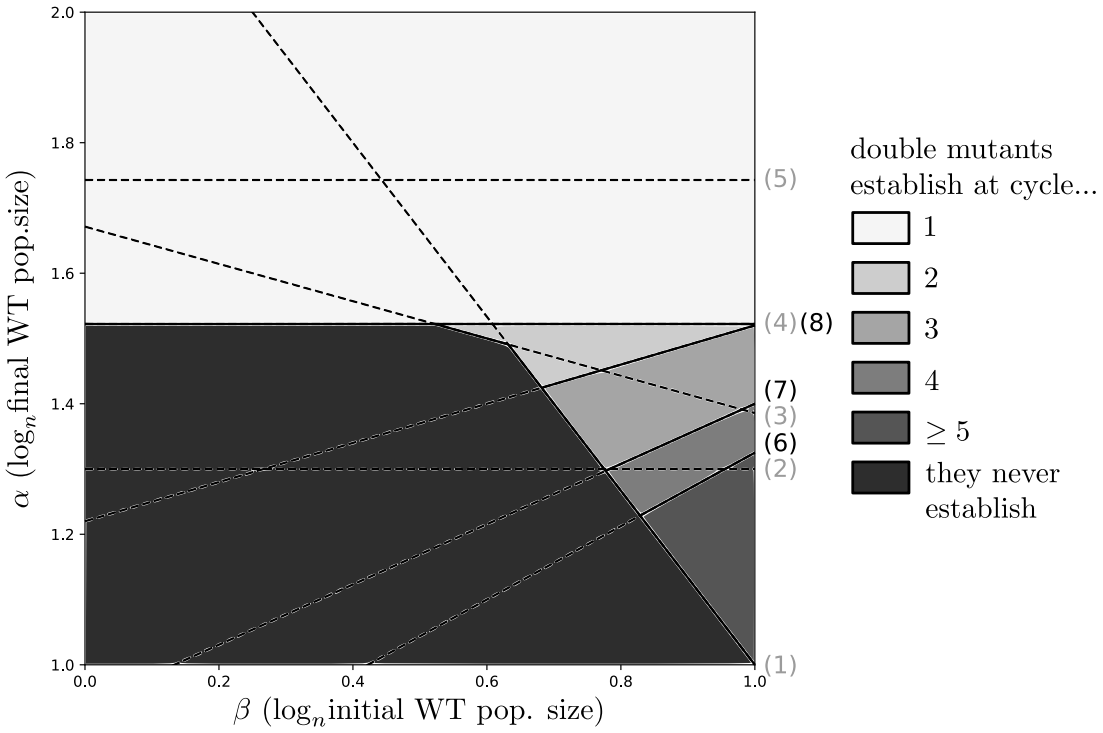


Figure 2.3 – Predicted number of cycles to wait before the establishment of double mutants, as a function of demographic parameters β and α. These predictions are made for infinitely large populations and vanishing mutation rates ($\delta = 1.3$, $r_{00} = 0.2$, $r_{10} = 0.35$ and $r_{01} = 0.9$).

2.3.2 Effect of demography on the timing of establishment of the double mutant

Here we show how demographic parameters affect the timing of adaptation, in particular at which cycle double mutants will establish. In Figure 2.2, lines (4) and (5) determine which value of α is needed for double mutants to establish during the first cycle. Above line (4) double mutants arise during the first cycle from mutant '01', and above line (5) they arise during the first cycle from mutant '10'. Thus, as soon as α is above one of these two lines double mutants establish during the first cycle. Double mutants may also take more time to establish, as we now explain.

The different scenarios of double mutant establishment are illustrated in Figure 2.3, using the same parameter values as in Figure 2.2. Different shades of grey are used to indicate the speed at which the double mutants are predicted to establish. New lines (6-8) delineate these regions (see equations in Section 2.5.8). See Figure 2.9(a) for the observed number of cycles at which the double mutant establishes in simulations with a finite n. It is clear from Figure 2.2 that in the limit where $n \to +\infty$, double mutants cannot establish in a finite number of cycles in the

bottom-left black zone, and that they establish at the first cycle in the top white zone. In the bottom-right grey zones, double mutants are predicted to establish in a number of cycles which is greater than 1 but w.h.p. finite. Lines (6), (7) and (8) indicate at which cycle single mutant populations are large enough for double mutants to establish. Interestingly, in this area the value $\beta = \alpha - (\alpha - 1)r_{10}/r_{00}$ (corresponding to line (1)) maximizes the speed of double mutants establishment.

2.3.3 Simulations with density-dependent division rate

In order to test our theoretical predictions in a more realistic setting, we performed simulations with density-dependent growth rates: intrinsic division and death rates are multiplied by $1 - \frac{N_{tot}}{K}$, with N_{tot} the current, total population size and K the carrying capacity. We consider the case where mutant subpopulations have a small advantage in exploiting the available resources, resulting in a higher carrying capacity. We thus have $K_{\text{WT}} = n^{\alpha}$ for the wild-type and $K_{\text{mut}} = n^{\alpha(1+s)}$ for mutants. This assumption is necessary to observe the blue area from Figure 2.2 where the strongly beneficial mutation can only establish in the '10' background. We further assume that dilution arises when the WT population reaches a proportion p of its carrying capacity. More details about our simulations can be found in Section 2.5.9, and Figure 2.8 provides examples of simulated populations.

Figure 2.4 shows which scenarios were observed after simulating the evolution of a population with 2,000 different pairs of parameters (α, β) and $p = 0.5$. The color of a point corresponds to the evolutionary paths observed during the simulation, with the color coding used in Figure 2.2. Here the WT is able to reach only 50% of its carrying capacity before dilution, thus the dynamics are very similar to our first model without density-dependence, until one of the mutants reaches the stationary phase. On Figure 2.4 we observe almost all the areas predicted by our model, with blurred boundaries due to stochasticity. However, the scenario where adaptation occurs through both weakly beneficial and strongly beneficial single mutants (red area) is less observed than in density-independent predictions because of competition: when both single mutants survive the first bottleneck, the weakly beneficial mutant '10' is driven to extinction by the strongly beneficial mutant '01' and thus is not able to produce a lot of double mutants (see Figure S4(b)).

As we discuss in Section 2.5.9 and Figure 2.7, increasing the value of p or the number of cycles increases the probability that some mutants establish and/or survive in areas where they were not expected to, thus shifting the position of certain areas compared to model predictions.

In conclusion, when doing simulations with finite values of n and density-dependent growth we recover the qualitative effects of α and β on evolutionary pathways. Note, however, that density dependence reduces the occurrence of the fixation of double mutants via two different pathways (i.e. red dots in Figure 2.4).

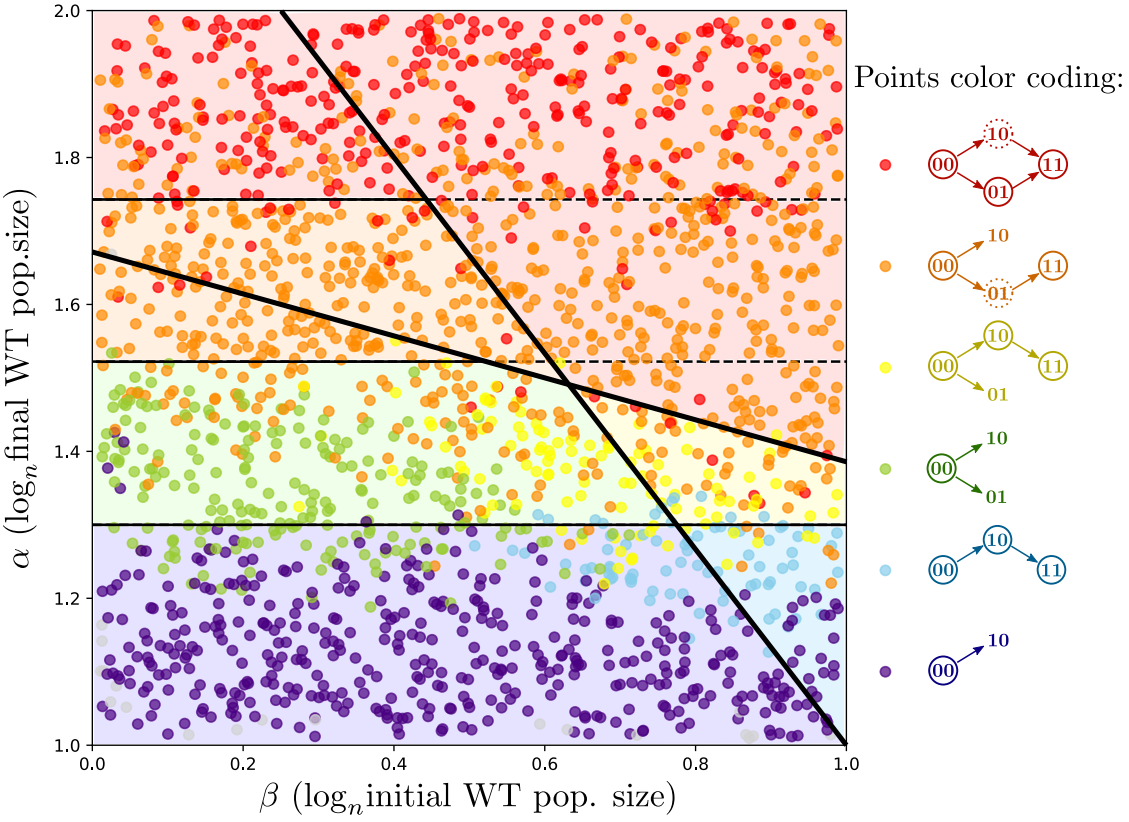


Figure 2.4 – Observed paths in simulations with density-dependent division: each point corresponds to a simulation and is colored according to the observed scenario. Each simulation was run for 7 cycles, with parameters $n = 10^{12}$, $p = 0.5$, $s = 0.1$, $\delta = 1.3$, $r_{00} = 0.2$, $r_{10} = 0.35$, $r_{01} = 0.9$, $r_{11} = 1$ and death rates equal to 0.1. Boundaries and background colors are theoretical predictions for exponential growth and large n from Figure 2.2.

2.4 Discussion

We characterize the trajectory and the speed of adaptation of an asexual, exponentially growing population subject to periodic bottlenecks. We studied adaptation on a minimal fitness landscape where only two classes of mutations are available: high-rate weakly beneficial mutations and low-rate strongly beneficial mutations.

Our main result is that 1) depending on initial and final population sizes, a unique evolutionary path unfolds and that 2) varying these two parameters, all paths can be explored. Establishment of a mutant is possible when the population size is of the order of the inverse of the relevant mutation rate. Surviving the bottleneck is possible when the final population size is of the order of the bottleneck severity. Tuning initial and final population sizes enables us to determine the evolutionary paths that the population will follow. A particularly interesting implication is that evolutionary paths can appear constrained not only because of sign epistasis (Weinreich et al. 2006) and rugged landscapes (Handel and Rozen 2009; Hartl 2014;

de Visser and Krug 2014), but also because of fluctuating demography limiting the mutational input and causing the loss of beneficial mutations. This phenomenon had been observed in several previous experimental studies (Garoff et al. 2020; Mahrt et al. 2021; Schenk et al. 2022), and here we dissected the mechanisms by which demographic parameters can constrain and direct mutational pathways.

2.4.1 Effect of demographic parameters

Our model predicts which evolutionary paths will be observed, in the limit where the population size is large and the mutation rate is small. The predicted outcome depends both on the initial population size (β in logarithmic scale) and the final wild-type population size (α in logarithmic scale), where 1 in logarithmic scale corresponds to the inverse of the higher mutation rate. As α or β decreases, the accessibility of the double mutant decreases until it is no longer possible for the population to acquire both mutations. Indeed, decreasing population size at the end of the growth phase (decreasing α) limits the supply of mutations and increasing bottleneck severity (decreasing β) prevents mutant populations to survive until the next cycle.

Demographic parameters affect the rate of evolution. Increasing the final population size (α) always speeds up adaptation, as a large population size favors the emergence of mutations and also gives more time for the mutant subpopulation to reach a size large enough to survive the bottleneck. Interestingly, increasing the initial population size (β) has more complex effects. If the final population size is above the threshold for the establishment of the double mutant from the strongly beneficial '01' mutant at the first cycle (α above line (4)), then the initial population size has no influence on the outcome. However, when the final population size is smaller than this threshold, establishment of the double mutant is fastest (in terms of number of cycles) for an intermediate initial population size. Indeed, when β is too small the bottleneck is too severe for mutations to survive. On the contrary if it is too large, then the bottleneck is less severe but the growth phase is shorter, leading to an overall effect of slowing down double mutant establishment. This non-monotonous effect of β is similar to that of the dilution ratio in (Wahl, Gerrish, and Saika-Voivod 2002): a small ratio allows few mutations to survive, but a high ratio reduces the duration of a cycle and yields fewer mutations.

2.4.2 Limitations

Our analysis relies on both a large population size and a small mutation rate approximation. However, these approximations correctly predict the outcome even for a finite population size. As shown on Figure 2.6, boundaries between different scenarios are blurred due to stochasticity but still visible. An important limitation of our model is that the population grows exponentially and is only bounded by the periodic bottlenecks, not by resource limitation. Coupled to the fact that the dilution ratio is kept constant throughout the experiment, this unlimited growth allows different clones to coexist indefinitely without interfering, which does not

seem realistic (Miralles et al. 1999; Lang et al. 2013; Maddamsetti, Lenski, and Barrick 2015). If instead the total inoculum size were constant we would observe clonal interference (Campos and Wahl 2009; Campos and Wahl 2010). However, we do observe clonal interference when relaxing the assumption of exponential growth in a set of additional simulations with density-dependence. In this setting, the emergence of double mutants from both single mutants is no longer possible. Nevertheless, these simulations show that the rest of our results hold qualitatively when weakly beneficial mutants can reach a higher population size than wild-types. The mutants can reach a higher final population size than WT when the WT does not reach stationary phase at the end of the cycle, or because beneficial mutations can enable a larger stationary size: for example, in the Long Term Evolution Experiment the evolution of the ability to use citrate causes a ten-fold increase in final optical density, a proxy for population size (Blount, Borland, and Lenski 2008).

2.4.3 Application to experimental design

Can our results be used to guide the design of evolution experiments? If mutation and division rates are known, it is possible to choose the size of the inoculum and of the final population size (or carrying capacity, for our model with density-dependence) to decide which mutants emerge and when. For example, if the goal is to obtain double mutants as quickly as possible from 10 ml of a bacterial population saturating at 10^9 individuals/ml with a beneficial mutation rate of 10^{-7}, then we would recommend based on our model to have a dilution factor of $n^{-(\alpha-1)r_{10}/r_{00}} = (10^{-3})^{r_{10}/r_{00}}$ (corresponding to the optimal value of β mentioned above). More details for computing the optimal dilution factor are available in Section 2.5.10. However all configurations may not be accessible in any given population or species, depending on the mutation rate and distribution of fitness effects. For example when r_{10} and r_{01} are close to r_{00} (weak selection), the lines (1) and (3) on Figure 2.2 are almost vertical. Thus we observe a greater diversity of scenarios when beneficial mutations confer a substantial fitness advantage, i.e., under strong selection.

In a more general setting where we have k beneficial mutations with a similar rate-benefit trade-off, and even without knowing precisely the mutation rate and distribution of fitness effects, what remains true is that increasing the initial and/or final population size will allow the population to access more evolutionary paths. Furthermore, a large initial population size combined with a small final population size (relaxed bottlenecks - short cycles) will favor paths going first through frequent mutations, while a small initial population size with a large final population size (severe bottlenecks - long cycles) will favor paths where the first mutations are rare but strongly beneficial.

The possibility to speed up the rate of adaptation by tuning demographic parameters could also alleviate the problem of bottlenecks in directed evolution (Bloom and F. H. Arnold 2009; Badran and Liu 2015).

2.4.4 Interpretation of experiment outcome

Our results imply that sign epistasis is not necessary for evolution to follow a specific evolutionary path over others. For example, Figure 2.1D illustrates a scenario corresponding to parameter values falling into the blue area of Figure 2.2: the emergence of the strongly beneficial '01' mutant is highly unlikely, but the weakly beneficial '10' mutant establishes in the first cycle and the double mutant establishes during the second cycle. Without prior knowledge on traits and demography, an interpretation of the emergence of '11' mutants exclusively from '10' and never from '01' mutants is sign epistasis: the fitness of the '01' mutant is lower than the WT, but this mutation confers a benefit in the background of the other (weakly beneficial) mutation. However, this interpretation is incorrect here: the strongly beneficial mutation is beneficial in all backgrounds but emerges from '10' and not from '00', simply because '10' mutants reach a higher population size than the WTs during the course of the experiment. The first set of beneficial mutations could thus enable access to other rarer mutations not through epistatic relationship but a larger final population size.

The phenomenon that we highlight here has been evidenced in experimental evolution. For example, Garoff et al. experimentally evolved *E. coli* under ciprofloxacin antibiotic in bottlenecks of varying severity (Garoff et al. 2020). They showed that when the final population size is small, evolving mutations are weakly beneficial but affect mechanisms with large mutational target, for example efflux pump repressors. Rarer and more beneficial mutations only evolve when the final population size is larger. A similar observation has been made by Schenk et al. for β-lactam antibiotic resistance (Schenk et al. 2022).

All in all, these new mathematical results shed light on the factors shaping adaptation in repeatedly bottlenecked populations, showing that all paths can be followed by adaptation depending on demographic controls, and that the repeated appearance of specific evolutionary paths over others does not imply sign epistasis. This work calls for models studying the effect of demographic controls on evolution in more complex fitness landscapes and for inference methods disentangling the role of epistasis and demography in realized evolution experiments.

2.5 Supplementary Information

This section contains supplementary information about our model (with derivation of the equations for threshold lines from Figure 2.2 and 2.3) and about stochastic simulations. Code used to perform simulations of the model and generate Figure 2.4 can be found at `https://github.com/JasmineGamblin/periodic-bottlenecks`.

2.5.1 A semi-deterministic model

As described in the main text, we use a semi-deterministic model in order to study the evolution of an asexual population experiencing periodic bottlenecks.

The results presented in Sections 2.5.2 to 2.5.5 of this Supplementary Information focus on the first growth phase of this population. During this phase, the wild-type population grows exponentially from a size n^β ($\beta \in (0,1)$) at time 0 to a size n^α ($\alpha > 1$) at time t_n. The wild-type population size at time $t \in [0, t_n]$ is $N_t = N_0 e^{r_{00}t}$. r_{00} is the growth rate and can be decomposed as $r_{00} = b_{00}(1-\mu) - d_{00}$, with b_{00} the division rate, μ the beneficial mutation rate and d_{00} the death rate.

We suppose that two categories of beneficial mutations can arise: weakly beneficial mutations with rate $\mu_{\text{high}} = \frac{1}{n}$ and strongly beneficial mutations with rate $\mu_{\text{low}} = \frac{1}{n^\delta}$ ($\delta > 1$). We focus here on weakly beneficial mutations and thus on mutants with genotype '10' (i.e. having a weakly beneficial mutation but no strongly beneficial mutation). Weakly beneficial mutations arise on the wild-type population according to a time-inhomogeneous Poisson process with parameter $\Lambda(t) = b_{00}\mu_{\text{high}}N_t$. Each one of these mutations gives rise to a clone of mutants '10' following a birth-death process with birth rate $b_{10}(1-\mu_{\text{low}})$ and death rate d_{10}.

A few known results on birth-death processes

We denote by Y_t the random variable measuring the size of a clone arising from a weakly beneficial mutation after a growth time of t. Then it can be shown (e.g. in Kendall 1948) that:

$$\mathbb{P}(Y_t = 0) = \frac{d_{10}(1 - e^{-r_{10}t})}{b_{10}(1-\mu_{\text{low}}) - d_{10}e^{-r_{10}t}}$$

and for $k > 0$

$$\mathbb{P}(Y_t = k) = (1 - \mathbb{P}(Y_t = 0)) \left(1 - \frac{b_{10}(1-\mu_{\text{low}})}{d_{10}}\mathbb{P}(Y_t = 0)\right) \left(\frac{b_{10}(1-\mu_{\text{low}})}{d_{10}}\mathbb{P}(Y_t = 0)\right)^{k-1}$$

We denote by $s_{10}(t)$ the probability for this clone to escape stochastic extinction at least until time t. Then:

$$s_{10}(t) = 1 - \mathbb{P}(Y_t = 0) = \frac{r_{10}}{b_{10}(1-\mu_{\text{low}}) - d_{10}e^{-r_{10}t}}.$$

For $k > 0$, we can rewrite $\mathbb{P}(Y_t = k)$ as:

$$\mathbb{P}(Y_t = k) = (1 - \mathbb{P}(Y_t = 0))e^{-r_{10}t}s_{10}(t)\left(1 - e^{-r_{10}t}s_{10}(t)\right)^{k-1}.$$

Thus

$$\mathbb{P}\left(Y_t = k | Y_t > 0\right) = e^{-r_{10}t}s_{10}(t)\left(1 - e^{-r_{10}t}s_{10}(t)\right)^{k-1}.$$

This shows that Y_t conditioned on survival is a geometric random variable with parameter

$$p_{10}(t) = e^{-r_{10}t}s_{10}(t) = \frac{e^{-r_{10}t}r_{10}}{b_{10}(1-\mu_{\text{low}}) - d_{10}e^{-r_{10}t}}.$$

It follows that the expected clone size at time t conditioned on the clone survival until t is

$$\mathbb{E}\left(Y_t | Y_t > 0\right) = \frac{1}{p_{10}(t)}.$$

We denote this quantity as $c_{10}(t)$ in Section 2.5.3.

2.5.2 Waiting time to the first mutation

We denote by T_n the time at which the first weakly beneficial mutation to escape stochastic extinction occurs. We define σ_n as the time at which the wild-type population reaches a size of n during the first growth phase. As the WT population size has an exponential growth $N_t = N_0 e^{r_{00}t}$ with $N_0 = n^\beta$, we have $\sigma_n = \frac{1-\beta}{r_{00}} \ln(n)$. In this section, we show that $T_n - \sigma_n = O(1)$ for $n \to \infty$. If we denote by U_n the random variable equal to $T_n - \sigma_n$ and take $t > -\sigma_n$, then

$$\mathbb{P}(U_n \leq t) = \mathbb{P}\left(T_n \leq t + \sigma_n\right).$$

Weakly beneficial mutations arise according to a non-homogeneous Poisson process of parameter $\Lambda(t) = b_{00}\mu_{\text{high}} n^\beta e^{r_{00}t}$, and the probability that a mutant '10' escapes stochastic extinction is $s_{10}(\infty) = \frac{r_{10}}{b_{10}(1-\mu_{\text{low}})}$. Thus the probability that no such mutant occurs before time $t + \sigma_n$ is

$$\mathbb{P}\left(T_n > t + \sigma_n\right) = \exp\left(-\int_0^{t+\sigma_n} \Lambda(s) \frac{r_{10}}{b_{10}(1-\mu_{\text{low}})}\, \mathrm{d}s\right).$$

Thus

$$\begin{aligned}
\mathbb{P}(U_n \leq t) &= 1 - \exp\left(-\int_0^{t+\frac{1-\beta}{r_{00}}\ln(n)} b_{00}\mu_{\text{high}} n^\beta e^{r_{00}s} \frac{r_{10}}{b_{10}(1-\mu_{\text{low}})}\, \mathrm{d}s\right) \\
&= 1 - \exp\left(\frac{b_{00}\mu_{\text{high}} r_{10}}{b_{10}(1-\mu_{\text{low}})r_{00}} n^\beta \left(1 - e^{r_{00}t + \ln(n)(1-\beta)}\right)\right) \\
&= 1 - \exp\left(\frac{b_{00}\mu_{\text{high}} r_{10}}{b_{10}(1-\mu_{\text{low}})r_{00}} n^\beta\right) \exp\left(-\frac{b_{00}\mu_{\text{high}} r_{10}}{b_{10}(1-\mu_{\text{low}})r_{00}} n e^{r_{00}t}\right).
\end{aligned}$$

Recalling that $\mu_{\text{high}} = \frac{1}{n}$ and $\mu_{\text{low}} = \frac{1}{n^\delta}$ with $\delta > 1$, we obtain

$$\mathbb{P}(U_n \leq t) \underset{n\to\infty}{\to} 1 - \exp\left(-\frac{b_{00}r_{10}}{b_{10}r_{00}} e^{r_{00}t}\right).$$

The limit of the cumulative distribution function of U_n shows that it converges in distribution to a random variable U_∞, where $-U_\infty$ follows a Gumbel distribution with parameters $\left(\frac{1}{r_{00}} \ln\left(\frac{b_{00}r_{10}}{b_{10}r_{00}}\right), \frac{1}{r_{00}}\right)$. Thus we have:

$$T_n - \sigma_n \underset{n\to\infty}{=} O(1).$$

This shows that the first weakly mutation to escape stochastic extinction occurs around the time where the WT population size is of the order of the inverse of the mutation rate.

2.5.3 Timing of weakly beneficial mutations

Here we look more closely at the times when weakly beneficial mutations occur during the first growth phase. We are interested in mutant clones that survive

the first bottleneck, and we seek to investigate how their mutational origins are distributed over time. Mutational events giving birth to a surviving clone are called **surviving** mutations (see main text). The rate at which these mutations are produced at time t during the first growth phase is

$$b_{00}\mu_{\text{high}}N_t \propto e^{r_{00}t}.$$

Thus most mutations arrive at the end of the growth phase. Their distribution is plotted in pink on Figure 2.5. However, early mutations have more descendants. The final expected clone size (conditioned on survival) for a weakly beneficial mutation occurring at time t is $c_{10}(t) = \frac{1}{r_{10}}\left(b_{10}(1-\mu_{\text{low}})e^{r_{10}(t_n-t)} - d_{10}\right)$. The probability that a mutation occurring at time t survives until the end of the growth phase being $s_{10}(t) = \frac{r_{10}}{b_{10}(1-\mu_{\text{low}})-d_{10}e^{-r_{10}(t_n-t)}}$, the rate of weakly beneficial mutations weighted by final expected clone size at time t is

$$b_{00}\mu_{\text{high}}N_t s_{10}(t)c_{10}(t) \propto e^{-(r_{10}-r_{00})t}$$

(see Section 2.5.1 for a derivation of c_{10} and s_{10}). This distribution is plotted in purple on Figure 2.5. The consequence is that clones that manage to survive the bottleneck have more chance to arise from a mutation arising in the middle of the growth phase (see solid blue curve on Figure 2.5). Indeed, if we suppose that the dilution is a binomial sampling, then the rate of surviving weakly beneficial mutations at time t is:

$$b_{00}\mu_{\text{high}}N_t s_{10}(t)\left(1-\left(1-n^{-(\alpha-\beta)}\right)^{c_{10}(t)}\right)$$

This distribution was already studied by Wahl, Gerrish, and Saika-Voivod (2002) in a setting where the selective advantage is small. They found that surviving mutations were likely to occur at all times during a growth phase, with a relatively flat distribution. This is indeed what we can observe on Figure 2.5 when $r_{10} \simeq r_{00}$ (dashed blue curve).

2.5.4 Number of mutants '10' after one growth phase

$Z_{10}^{(n)}$ is the random variable counting the number of mutants with genotype '10' at the end of the first growth phase. In the Appendix 2.5.10, we show that $Z_{10}^{(n)} \underset{n\to+\infty}{\sim} n^{(\alpha-1)\frac{r_{10}}{r_{00}}}Z_{10}$. This shows in particular that mutants '10' **emerge** during the first growth phase, i.e. they reach a large population size ($Z_{10}^{(n)} \gg 1$).

Distribution of Z_{10}

The characteristic function of Z_{10} is

$$\begin{aligned}\phi_{Z_{10}}(t) &= \mathcal{L}_{Z_{10}}(-it)\\ &= \exp\left(-\frac{b_{00}\pi}{b_{10}\sin\left(\frac{r_{00}\pi}{r_{10}}\right)}\left(\frac{b_{10}}{r_{10}}\right)^{\frac{r_{00}}{r_{10}}}|t|^{\frac{r_{00}}{r_{10}}}(-i\,\mathrm{sgn}(t))^{\frac{r_{00}}{r_{10}}}\right)\\ &= \exp\left(-\frac{b_{00}\pi}{b_{10}\sin\left(\frac{r_{00}\pi}{r_{10}}\right)}\left(\frac{b_{10}}{r_{10}}\right)^{\frac{r_{00}}{r_{10}}}\cos\left(\frac{\pi r_{00}}{2r_{10}}\right)|t|^{\frac{r_{00}}{r_{10}}}\left(1-i\,\mathrm{sgn}(t)\tan\left(\frac{\pi r_{00}}{2r_{10}}\right)\right)\right).\end{aligned}$$

This expression tells us that Z_{10} has a one-sided (Lévy) stable distribution supported by $[0,\infty)$ (Bertoin 1996). The stability parameter is $\frac{r_{00}}{r_{10}} \in (0,1)$, which implies that all moments are infinite. Sadly, the probability density function of Z_{10} is not analytically expressible in the general case (Penson and Górska 2010), nor are its median and mode.

2.5.5 Number of mutants '01' after one growth phase

For mutants with genotype 01, there are two different cases depending on the sign of $\alpha-\delta$. Indeed, the probability that no mutant '01' appears during a growth phase is:

$$\begin{aligned}p_{01}^{(n)} &= \exp\left(-\int_0^{t_n} b_{00}\mu_{\text{low}}N_t\,\mathrm{d}t\right)\\ &= \exp\left(-b_{00}n^{\beta-\delta}\int_0^{t_n} e^{r_{00}t}\,\mathrm{d}t\right)\\ &= \exp\left(-b_{00}n^{\beta-\delta}\frac{1}{r_{00}}(n^{\alpha-\beta}-1)\right)\\ &= \exp\left(-\frac{b_{00}}{r_{00}}(n^{\alpha-\delta}-n^{\beta-\delta})\right)\end{aligned}$$

Case $\alpha < \delta$

In this case, $p_{01}^{(n)}$ goes to 1 as $n \to \infty$. It is highly unlikely to see any '01' mutant appear during the course of the experiment.

Case $\alpha > \delta$

In that case, $p_{01}^{(n)}$ goes to 0 as $n \to \infty$. Furthermore, we can do the same computations as with mutants '10' and find that the random variable $Z_{01}^{(n)}$ counting the number of mutants '01' at the end of the first growth phase verifies:

$$Z_{01}^{(n)} \underset{n\to+\infty}{\sim} n^{(\alpha-\delta)\frac{r_{01}}{r_{00}}}Z_{01}$$

with Z_{01} a stable distribution with stability parameter $\frac{r_{00}}{r_{01}}$. This shows that mutants '01' emerge during the first growth phase.

2.5.6 Survival of single mutants

From Section 2.5.4 we know that mutants '10' **establish** with high probability during the first cycle. However this does not guarantee that they will **survive**, that is, reach a sufficiently large size to survive the next bottleneck. Mutants '10' survive if $A = -(\alpha - \beta) + (\alpha - 1)\frac{r_{10}}{r_{00}}$ is positive. Indeed the expected quantity of mutants '10' present at the beginning of the second cycle is $D_n Z_{10}^{(n)}$, with $D_n = n^{-(\alpha-\beta)}$ being the dilution factor. This quantity is equivalent to $n^{-(\alpha-\beta)+(\alpha-1)\frac{r_{10}}{r_{00}}} Z_{10}$ when n is large, thus the population size of '10' mutants starting in the second cycle is large if $A > 0$ and negligible if $A < 0$. If $A > 0$, the extinction probability of '10' mutants is close to zero and their growth is deterministic for the remainder of the experiment, so that surviving one bottleneck entails the survival of all bottlenecks. If $A < 0$, with high probability no '10' mutant survives the dilution as $n \to \infty$. Indeed, the probability that no mutant survives dilution knowing that $Z_{10} = z$ is

$$\left(1 - n^{-(\alpha-\beta)}\right)^{zn^{(\alpha-1)\frac{r_{10}}{r_{00}}}} \underset{n\to\infty}{\sim} \exp\left(-zn^{A}\right).$$

Then we discuss the **establishment** and **survival** of the strongly beneficial '01' mutant in the population. It establishes during the first cycle if $\alpha > \delta$. In this case, the number of '01' mutants at the end of the first cycle $Z_{01}^{(n)}$ is equivalent to $n^{(\alpha-\delta)\frac{r_{01}}{r_{00}}} Z_{01}$ for $n \to \infty$. Similarly to '10' mutants, they survive if $-(\alpha - \beta) + (\alpha - \delta)\frac{r_{01}}{r_{00}}$ is positive.
On the contrary if $\alpha < \delta$, then the probability $p_{01}^{(n)}$ that no '01' mutant is present at time t_n goes to 1 as $n \to \infty$ (see Section 2.5.5). As the WT population has the same growth trajectory in every phase, $p_{01}^{(n)}$ is in fact the probability to have no '01' mutant at the end of any cycle.

2.5.7 Constrained evolutionary paths

Parameters α, β and r_{00} govern the demography of the wild-type population, while parameters δ, r_{10} and r_{01} drive the appearance and growth rates of mutants. We have seen in Section 2.5.6 that their relative values determine which mutations can survive, and a consequence is that they also determine which evolutionary paths are accessible. This is illustrated in Figure 2.2 of main text, where the different areas are delimited by the following threshold lines:

- (1) $\alpha > \frac{r_{10} - \beta r_{00}}{r_{10} - r_{00}}$ condition for the survival of mutants 10 (this is equivalent to $A > 0$);
- (2) $\alpha > \delta$ condition for the establishment of mutants '01' during first cycle;
- (3) $\alpha > \frac{\delta r_{01} - r_{00}}{r_{01} - r_{00}}$ condition for the survival of mutants 01;
- (4) $(\alpha - \delta)\frac{r_{01}}{r_{00}} > 1$ condition for the establishment of double mutants from mutant '01' during first cycle;
- (5) $(\alpha - 1)\frac{r_{10}}{r_{00}} > \delta$ condition for the establishment of double mutants from mutant '10' during first cycle.

Conditions (4) and (5) are explained in the following section.

2.5.8 Establishment of double mutants

Demographic parameters constrain evolutionary paths, but also the timing of evolution. In particular, they determine in which cycle the double mutants '11' will establish. In the case where mutants '10' survive the first bottleneck, we can consider that their growth is deterministic during the following cycles. At the end of the k^{th} cycle they have gone through $k-1$ dilutions and deterministic growth phases, thus their population size is

$$\begin{aligned} & n^{(\alpha-1)\frac{r_{10}}{r_{00}}} Z_{10} \times D_n^{k-1} \times \left(e^{r_{10}t_n}\right)^{k-1} \\ & \underset{n\to\infty}{=} O\left(n^{(\alpha-1)\frac{r_{10}}{r_{00}}+(k-1)\left(\frac{r_{10}}{r_{00}}-1\right)(\alpha-\beta)}\right) \end{aligned}$$

Thus double mutants produced from mutant '10' establish during the first cycle k_1 such that

$$(\alpha-1)\frac{r_{10}}{r_{00}} + (k_1-1)\left(\frac{r_{10}}{r_{00}}-1\right)(\alpha-\beta) > \delta.$$

Similarly, double mutants produced from mutant '01' establish during the first cycle k_2 such that

$$(\alpha-\delta)\frac{r_{01}}{r_{00}} + (k_2-1)\left(\frac{r_{01}}{r_{00}}-1\right)(\alpha-\beta) > 1.$$

As a result, double mutants establish at cycle $\min(k_1, k_2)$. This is illustrated in Figure 2.3 from main text, where additional lines are plotted compared to Figure 2.2:

- (6) $(\alpha-1)\frac{r_{10}}{r_{00}} + \left(\frac{r_{10}}{r_{00}}-1\right)(\alpha-\beta) > \delta$ condition for the establishment of double mutants from mutant '10' during the second cycle;
- (7) $(\alpha-1)\frac{r_{10}}{r_{00}} + 2\left(\frac{r_{10}}{r_{00}}-1\right)(\alpha-\beta) > \delta$ condition for the establishment of double mutants from mutant '10' during cycle 3;
- (8) $(\alpha-1)\frac{r_{10}}{r_{00}} + 3\left(\frac{r_{10}}{r_{00}}-1\right)(\alpha-\beta) > \delta$ condition for the establishment of double mutants from mutant '10' during cycle 4.

Lines representing conditions for the establishment of double mutants from mutant '01' during cycles 2, 3 and 4 are not plotted because they are below lines (6-8) and thus are not relevant for area delimitation.

2.5.9 Simulations

The code that was used to run the following simulations is available on Github at `https://github.com/JasmineGamblin/periodic-bottlenecks`, in the form of a Python script and a Jupyter notebook from which simulations can easily be run.

Exponential growth

We simulated populations evolving according to a fully stochastic model: individuals were grouped by subpopulations carrying the same genotype, each following

a birth-death-mutation process. Double mutants originating from the weakly beneficial single mutant or the strongly beneficial single mutant were grouped separately. We used the principle of the Gillespie (1977) algorithm to simulate the evolution of the system during growth phases, and binomial sampling to perform the dilution step. We resorted to a τ-leaping approximation to speed-up the simulation when population sizes become large (Gillespie 2001).

For each simulation, we recorded at the end of each growth phase which mutant subpopulations were present above a fixed threshold of 50 individuals (chosen so that the probability of extinction from this size is $< 1\%$). This way we could associate a color to the simulation according to the observed evolutionary paths, following the color coding from Figure 2.2 of main text.

We obtained Figures 2.6(a,b,c) by randomly drawing 2,000 pairs of parameters (α, β) and simulating a population for 7 cycles, for different values of n. Populations for $n \geq 10^9$ and $\alpha > 1.4$ were simulated only for 4 cycles to avoid dealing with too large numbers. We can observe that for increasing values of n, the results of the simulations converge to the deterministic behavior that we predicted for $n \to \infty$. Grey points correspond to simulations where no mutant subpopulation was detected above the threshold.

Figure 2.9(a) was obtained by drawing 2,000 pairs of demographic parameters. For each pair, a population was simulated for 7 cycles and the cycle at which the double mutant emerges was recorded.

Density-dependent growth

For simulations with density-dependent growth rates, we multiplied intrinsic birth and death rates by $1 - \frac{N_{tot}}{K}$, with N_{tot} the total population size and K the carrying capacity. We chose $K_{\mathrm{WT}} = n^{\alpha}$ for wild-types and $K_{\mathrm{mut}} = n^{\alpha(1+s)}$ for all mutant subpopulations, with s being a small advantage that mutants may have in exploiting the available resources. We also introduced a parameter p which is the proportion of the carrying capacity that wild-types are able to reach before dilution. Hence, we now have

$$t_n = \frac{1}{r_{00}} \ln \left(\frac{n^{\alpha-\beta} - 1}{\frac{1}{p} - 1} \right)$$

and

$$D_n = \frac{n^{\beta-\alpha}}{p}.$$

Figure 2.8 shows examples of simulated evolution for 11 different combinations of parameters α and β. We chose parameter values $p = 0.5$ and $s = 0.1$, thus wild-types are able to reach 50% of their carrying capacity before dilution. Here the fact that growth is density-dependent introduces competition between the different subpopulations: we observe that wild-types and small-benefit mutants can be driven to extinction by fitter mutants. This effect of competition can be observed on Figures 2.6(d,e,f), where we see fewer red points in comparison to simulations

with exponential growth (Figures 2.6(a,b,c)). This is due to the fact that in most cases the weakly beneficial single mutant is outcompeted by the strongly beneficial one and thus is not able to produce double mutants.

Figure 2.7 shows the effect of two parameters: the proportion p of the carrying capacity reached by wild-types at the end of the growth phase, and the number of cycles performed. When varying p between 0.5 and 0.9, we can verify that the case $p = 0.5$ is very similar to the exponential case as we would expect (because at this point the growth is still nearly exponential). For a higher value of p, we observe that strongly beneficial single mutants have some chance so establish even if $\alpha < \delta$, and that both single mutants may survive dilution even below their survival thresholds. These observations are due to the fact that with density-dependent growth it takes more time for wild-types to reach their final population size compared to the exponential case. Thus the cycle duration is longer, giving more time to mutants to establish and/or survive. When varying the number of cycles performed between 7 and 30, we observe as expected that performing more cycles also increases the probability that mutants establish and/or survive in areas where they are not expected to do so in large n predictions.

Figure 2.9(b) was obtained by drawing 2,000 pairs of demographic parameters. For each pair, a population was simulated for 7 cycles and the cycle at which the double mutant emerges was recorded.

2.5.10 Optimal dilution factor

If for a given population we know the maximum density m, the volume v and the beneficial mutation rate μ, then we can compute the parameters of our model:

$$n = \mu^{-1} \quad \text{and} \quad \alpha = \frac{\ln(mv)}{\ln(n)}.$$

For a rare, strongly beneficial mutation with rate μ_{low}, we predict that:

— If $\delta = -\frac{\ln(\mu_{\text{low}})}{\ln(n)} < \alpha$, this mutation will establish at the first cycle of growth.

— If not, it may still establish in the background of more frequent but less advantageous mutations if they are able to reach a higher population size than the wild-type. In that case, we can maximize the speed of adaptation by choosing the parameter β to be just above line (1) of Figure 2.2 from the main text. To be on that line, β must satisfy the equation $\alpha = (r_{10} - \beta r_{00})/(r_{10} - r_{00})$, which leads to an optimal dilution factor of:

$$n^{-(\alpha-1)r_{10}/r_{00}} = (\mu^{-\frac{\ln(mv)}{\ln(\mu)}-1})^{r_{10}/r_{00}}.$$

This optimal dilution factor can be bounded above by $\mu^{-\frac{\ln(mv)}{\ln(\mu)}-1}$ if the growth rates are not known.

Appendix: Limit of $Z_{10}^{(n)}$ for $n \to +\infty$

Here we show that $Z_{10}^{(n)} \underset{n\to+\infty}{\sim} n^{(\alpha-1)\frac{r_{10}}{r_{00}}} Z_{10}$. We will first introduce some notations, then compute the Laplace transform of $\frac{Z_{10}^{(n)}}{n^q}$ and show that it converges to a non-trivial function for $q = (\alpha-1)\frac{r_{10}}{r_{00}}$ when $n \to \infty$.

Notations

We define the following random variables:

$M_{t_n} \in \mathbb{N}$ counts the number of weakly beneficial mutations arising among the WT population and still present at time t_n. M_{t_n} is distributed as a Poisson random variable of parameter $\Lambda(t_n) = \int_0^{t_n} b_{00}\mu_{\text{high}} N_t s_{10}(t_n - t)\,\mathrm{d}t$.

$T_1, ..., T_{M_{t_n}} \in [0, t_n]$ are the times when these mutations arise. Conditioned on $\{M_{t_n} = m\}$, they are independent and identically distributed with probability density $\frac{b_{00}\mu_{\text{high}} N_t s_{10}(t_n - t)}{\Lambda(t_n)}$.

$Y_t^{(i)} \in \mathbb{N}$ is the size after a growth time of t of the clone arising from the i^{th} weakly beneficial mutation. Conditioned on $\{M_{t_n} = m\}$, the $(Y_t^{(i)})$ with $i \in [1..M_{t_n}]$ are independent and identically distributed. $Y_t^{(i)}$ is a birth-death process starting at time T_i and killed at time t_n, conditioned to survive until t_n. We recalled in Section 2.5.1 that $Y_t^{(i)}$ is a geometric variable with parameter $p_{10}(t) = \frac{e^{-r_{10}t} r_{10}}{b_{10}(1-\mu_{\text{low}}) - d_{10}e^{-r_{10}t}}$.

With that we can express $Z_{10}^{(n)}$ as:

$$Z_{10}^{(n)} = \sum_{i=1}^{M_{t_n}} Y_{t_n - T_i}^{(i)}$$

Laplace transform of $Z_{10}^{(n)}/n^q$

Let us compute the Laplace transform of $\frac{Z_{10}^{(n)}}{n^q}$. For $u > 0$, we have:

$$\begin{aligned}
\mathcal{L}_{Z_{10}^{(n)}/n^q}(u) &= \mathbb{E}\left(e^{-u\frac{Z_{10}^{(n)}}{n^q}}\right) \\
&= \mathbb{E}\left(e^{-\frac{u}{n^q}\sum_{i=1}^{M_{t_n}} Y_{t_n-T_i}^{(i)}}\right) \\
&= \mathbb{E}_{M_{t_n}}\left(\mathbb{E}_{T_i|M_{t_n}}\left(\mathbb{E}_{Y_{t_n-T_i}^{(i)}|T_i}\left(e^{-\frac{u}{n^q}Y_{t_n-T_i}^{(i)}}\right)\right)^{M_{t_n}}\right).
\end{aligned}$$

The Laplace transform of a geometric random variable Y of parameter p is $\mathbb{E}\left(e^{-uY}\right) = \frac{p}{e^u-(1-p)}$, thus

$$\begin{aligned}
\mathcal{L}_{Z_{10}^{(n)}/n^q}(u) &= \mathbb{E}_{M_{t_n}}\left(\mathbb{E}_{T_i|M_{t_n}}\left(\frac{p_{10}(t_n-T_i)}{e^{\frac{u}{n^q}}-1+p_{10}(t_n-T_i)}\right)^{M_{t_n}}\right)\\
&= \mathbb{E}_{M_{t_n}}\left(\left(\int_0^{t_n}\frac{p_{10}(t_n-t)}{e^{\frac{u}{n^q}}-1+p_{10}(t_n-t)}\times\frac{b_{00}\mu_{\text{high}}N_t s_{10}(t_n-t)}{\Lambda(t_n)}\,\mathrm{d}t\right)^{M_{t_n}}\right)\\
&= \sum_{m=0}^{\infty}\left(\int_0^{t_n}\frac{p_{10}(t_n-t)}{e^{\frac{u}{n^q}}-1+p_{10}(t_n-t)}\times\frac{b_{00}\mu_{\text{high}}N_t s_{10}(t_n-t)}{\Lambda(t_n)}\,\mathrm{d}t\right)^{m}\frac{\Lambda(t_n)^m}{m!}e^{-\Lambda(t_n)}\\
&= e^{-\Lambda(t_n)}\sum_{m=0}^{\infty}\left(\int_0^{t_n}\frac{p_{10}(t_n-t)}{e^{\frac{u}{n^q}}-1+p_{10}(t_n-t)}b_{00}\mu_{\text{high}}N_t s_{10}(t_n-t)\,\mathrm{d}t\right)^{m}\frac{1}{m!}\\
&= \exp\left(-\Lambda(t_n)\right)\exp\left(\int_0^{t_n}b_{00}\mu_{\text{high}}N_t s_{10}(t_n-t)\frac{p_{10}(t_n-t)}{e^{\frac{u}{n^q}}-1+p_{10}(t_n-t)}\,\mathrm{d}t\right).
\end{aligned}$$

Then recalling that $\Lambda(t_n) = \int_0^{t_n} b_{00}\mu_{\text{high}}N_t s_{10}(t_n-t)\,\mathrm{d}t$, we have

$$\mathcal{L}_{Z_{10}^{(n)}/n^q}(u) = \exp\left(\int_0^{t_n} b_{00}\mu_{\text{high}}N_t s_{10}(t_n-t)\left(\frac{p_{10}(t_n-t)}{e^{\frac{u}{n^q}}-1+p_{10}(t_n-t)}-1\right)\,\mathrm{d}t\right)$$

and by replacing $p_{10}(t_n-t)$ and $s_{10}(t_n-t)$ by their expression we obtain:

$$\mathcal{L}_{Z_{10}^{(n)}/n^q}(u) =$$
$$\exp\left(\int_0^{t_n} b_{00}n^{\beta-1}e^{r_{00}t}r_{10}\frac{1-e^{\frac{u}{n^q}}}{e^{r_{10}(t-t_n)}(b_{10}(1-\mu_{\text{low}})-d_{10}e^{\frac{u}{n^q}})-b_{10}(1-\mu_{\text{low}})(1-e^{\frac{u}{n^q}})}\,\mathrm{d}t\right).$$

Using the change of variables $y = \frac{e^{r_{00}t}-1}{e^{r_{00}t_n}-1}$, $\mathrm{d}y = \frac{r_{00}e^{r_{00}t}}{e^{r_{00}t_n}-1}\,\mathrm{d}t$ leads to:

$$\begin{aligned}
\mathcal{L}_{Z_{10}^{(n)}/n^q}(u) &= \exp\left(-\frac{b_{00}r_{10}}{r_{00}}n^{\beta-1}(n^{\alpha-\beta}-1)\frac{(e^{\frac{u}{n^q}}-1)}{b_{10}(1-\mu_{\text{low}})-d_{10}e^{\frac{u}{n^q}}}\right.\\
&\qquad\left.\times\int_0^1\frac{\mathrm{d}y}{\frac{((n^{\alpha-\beta}-1)y+1)^{\frac{r_{10}}{r_{00}}}}{n^{\frac{r_{10}}{r_{00}}(\alpha-\beta)}}+\frac{b_{10}(1-\mu_{\text{low}})(e^{\frac{u}{n^q}}-1)}{b_{10}(1-\mu_{\text{low}})-d_{10}e^{\frac{u}{n^q}}}}\right)\\
&= \exp(g(n)I_n)
\end{aligned}$$

with

$$\begin{aligned}
g(n) &= -\frac{b_{00}r_{10}}{r_{00}}n^{\beta-1}(n^{\alpha-\beta}-1)\frac{(e^{\frac{u}{n^q}}-1)}{b_{10}(1-\mu_{\text{low}})-d_{10}e^{\frac{u}{n^q}}}\\
&\underset{n\to+\infty}{\sim}\frac{-ub_{00}}{r_{00}}n^{\alpha-1-q}
\end{aligned}$$

and

$$I_n = \int_0^1\frac{\mathrm{d}y}{\frac{((n^{\alpha-\beta}-1)y+1)^{\frac{r_{10}}{r_{00}}}}{n^{\frac{r_{10}}{r_{00}}(\alpha-\beta)}}+\frac{b_{10}(1-\mu_{\text{low}})(e^{\frac{u}{n^q}}-1)}{b_{10}(1-\mu_{\text{low}})-d_{10}e^{\frac{u}{n^q}}}}.$$

For $n \to \infty$, I_n is equivalent to

$$I_n^* = \int_0^1 \frac{\mathrm{d}y}{y^c + a_n} = \int_0^\infty \frac{e^{-x}}{e^{-cx} + a_n}\,\mathrm{d}x$$

where $c = \frac{r_{10}}{r_{00}}$ and $a_n = \frac{b_{10}(1-\mu_{\text{low}})u}{r_{10}n^q}$. We can compute this integral using the residue theorem (Ahlfors 1979): we integrate the function $h(x) = \frac{e^{-x}}{e^{-cx}+a_n}$ on a rectangle with width R and height $\frac{2\pi}{c}$ with bottom left corner on the origin. If we respectively name I_R, K_R, J_R and L_R the integrals along the four sides of the rectangle, then we have $I_R \underset{R\to\infty}{\to} I_n^*$, $J_R = -e^{-\frac{2i\pi}{c}} I_R$ and $J_R + K_R \underset{R\to\infty}{\to} 0$. As this rectangle contains one simple pole of $h(x)$ (for $x = -\frac{\ln(a_n)}{c} + \frac{i\pi}{c}$), we obtain:

$$\begin{aligned}
I_n^*(1 - e^{-\frac{2i\pi}{c}}) &= 2i\pi \mathrm{Res}\left(\frac{e^{-x}}{e^{-cx}+a_n}, -\frac{\ln(a_n)}{c} + \frac{i\pi}{c}\right) \\
I_n^* &= \frac{2i\pi}{1 - e^{-\frac{2i\pi}{c}}} \times \frac{e^{\frac{\ln(a_n)}{c} - \frac{i\pi}{c}}}{-ce^{-c(-\frac{\ln(a_n)}{c} + \frac{i\pi}{c})}} \\
I_n^* &= \frac{2i\pi}{e^{\frac{i\pi}{c}} - e^{-\frac{i\pi}{c}}} \times \frac{a_n^{\frac{1}{c}}}{ca_n} \\
I_n^* &= \frac{\pi}{c\sin(\frac{\pi}{c})} a_n^{\frac{1}{c}-1}
\end{aligned}$$

It follows that

$$I_n \underset{n\to+\infty}{\sim} \frac{r_{00}\pi}{r_{10}\sin\left(\frac{r_{00}\pi}{r_{10}}\right)} \left(\frac{b_{10}u}{r_{10}n^q}\right)^{\frac{r_{00}}{r_{10}}-1}$$

and

$$g(n)I_n \underset{n\to\infty}{\sim} \frac{-b_{00}\pi}{b_{10}\sin\left(\frac{r_{00}\pi}{r_{10}}\right)} \left(\frac{b_{10}u}{r_{10}}\right)^{\frac{r_{00}}{r_{10}}} n^{\alpha-1-\frac{r_{00}}{r_{10}}q}.$$

Thus $\mathcal{L}_{Z_{10}^{(n)}/n^q}(u)$ has a finite limit when $n \to \infty$ for $q = (\alpha-1)\frac{r_{10}}{r_{00}}$:

$$\mathcal{L}_{Z_{10}^{(n)}/n^{(\alpha-1)\frac{r_{10}}{r_{00}}}}(u) \underset{n\to\infty}{\to} \exp\left(-\frac{b_{00}\pi}{b_{10}\sin\left(\frac{r_{00}\pi}{r_{10}}\right)} \left(\frac{b_{10}u}{r_{10}}\right)^{\frac{r_{00}}{r_{10}}}\right).$$

It follows that

$$Z_{10}^{(n)} \underset{n\to+\infty}{\sim} n^{(\alpha-1)\frac{r_{10}}{r_{00}}} Z_{10}$$

with Z_{10} a random variable independent of n with Laplace transform:

$$\mathcal{L}_{Z_{10}}(u) = \exp\left(-\frac{b_{00}\pi}{b_{10}\sin\left(\frac{r_{00}\pi}{r_{10}}\right)} \left(\frac{b_{10}u}{r_{10}}\right)^{\frac{r_{00}}{r_{10}}}\right).$$

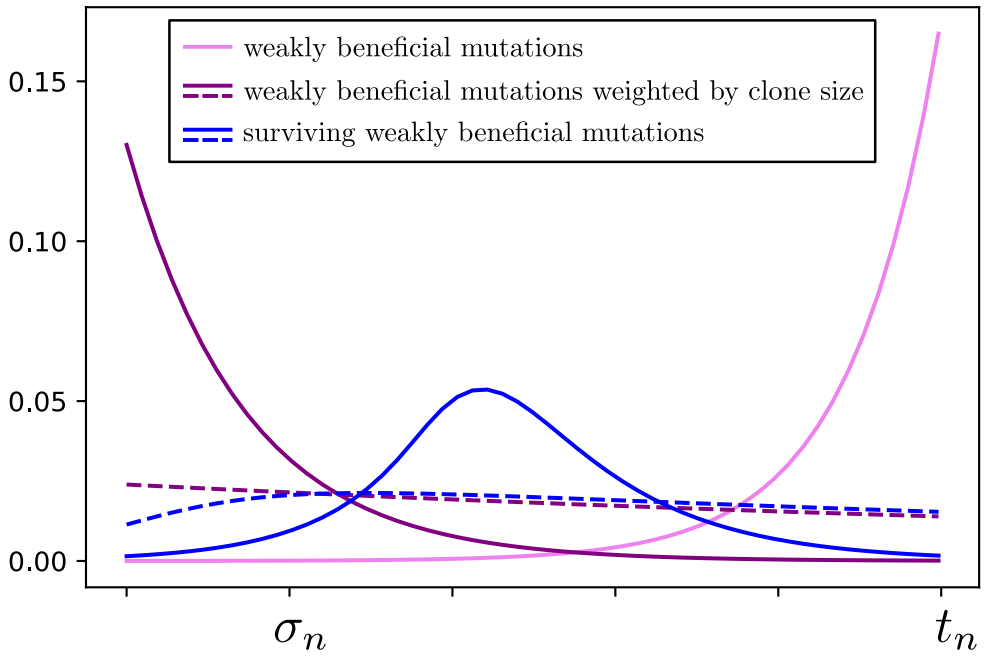
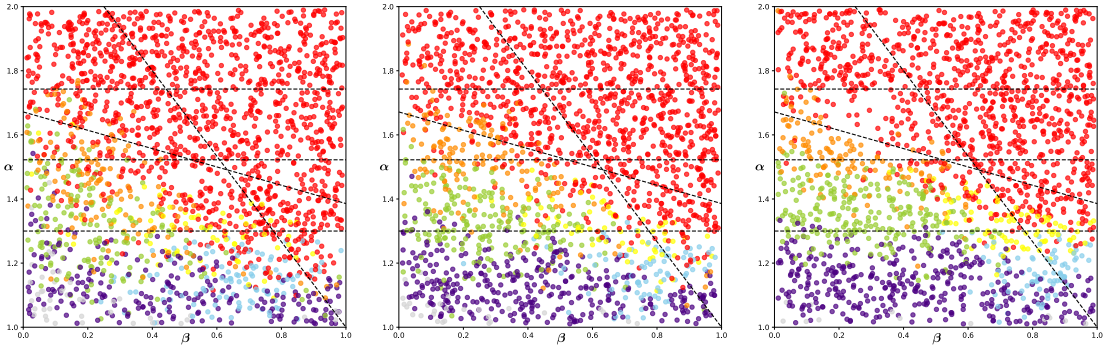
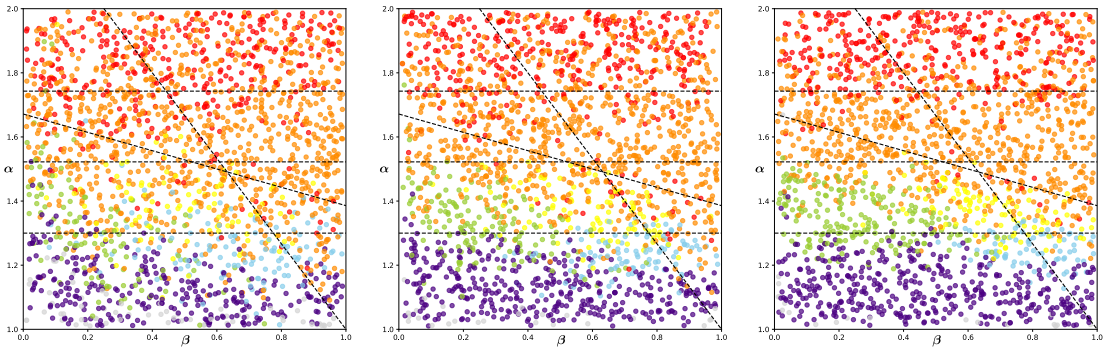


Figure 2.5 – Distribution in time of weakly beneficial mutational origins during first cycle. Pink curve: total weakly beneficial mutation rate, purple curves: total weakly beneficial mutation rate weighted by final expected clone size, and blue curves: rate of emergence of weakly beneficial mutations that survive the first bottleneck. Solid curves are plotted for $r_{10} = 0.3$, dashed ones for $r_{10} = 0.18$ ($r_{00} = 0.17$, $n = 10^8$, $\alpha = 1.4$, $\beta = 0.9$).

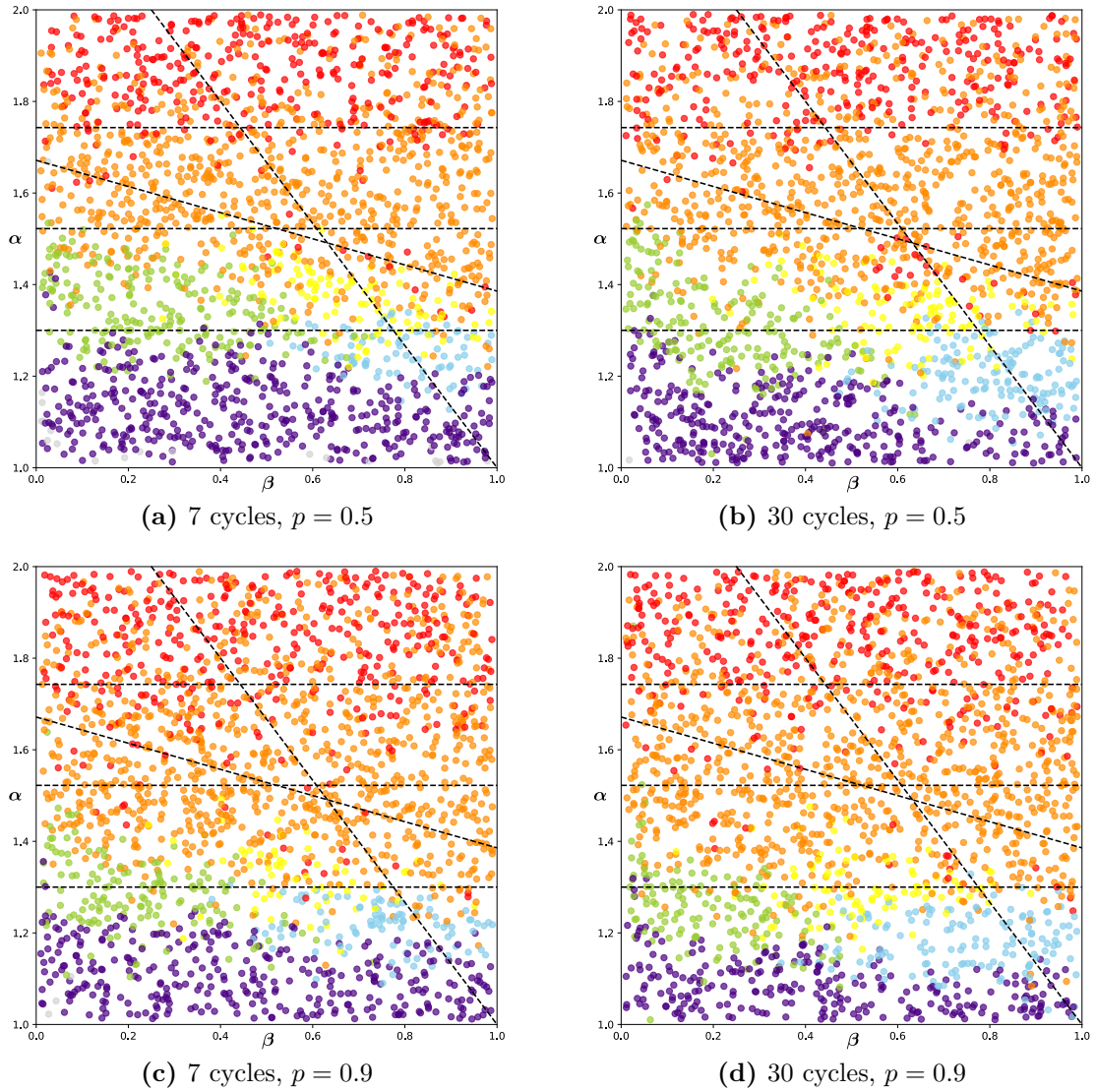


(a) Exponential growth, $n = 10^6$ **(b)** Exponential growth, $n = 10^9$ **(c)** Exponential growth, $n = 10^{12}$



(d) Density-dependent growth, $n = 10^6$ **(e)** Density-dependent growth, $n = 10^9$ **(f)** Density-dependent growth, $n = 10^{12}$

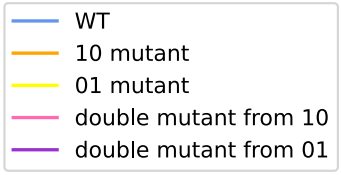
Figure 2.6 – Observed paths in simulations with exponential (a,b,c) and density-dependent (d,e,f) growth: each point corresponds to a simulation and is colored according to the observed scenario. For each subplot, 2,000 simulations were run for 7 cycles with $n = 10^6$ (a,d), $n = 10^9$ (b,e) or $n = 10^{12}$ (c,f). $\delta = 1.3$, $b_{00} = 0.3, b_{10} = 0.45, b_{01} = 1, b_{11} = 1.1$ and $d_{00} = d_{10} = d_{01} = d_{11} = 0.1$. For simulations with density-dependence, $p = 0.5$ and $s = 0.1$.



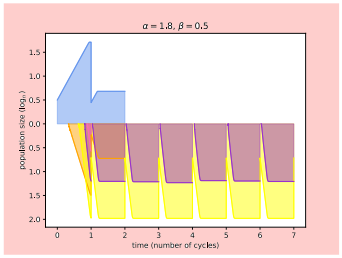
(a) 7 cycles, $p = 0.5$ **(b)** 30 cycles, $p = 0.5$

(c) 7 cycles, $p = 0.9$ **(d)** 30 cycles, $p = 0.9$

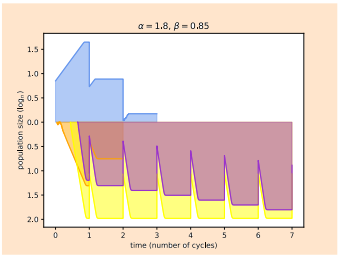
Figure 2.7 – Influence of density-dependence and the number of cycles: observed paths in simulations with density-dependent growth. Each point corresponds to a simulation and is colored according to the observed scenario. Simulations were run with $p = 0.5$ (a,b) or $p = 0.9$ (c,d), for 7 (a,c) or 30 (b,d) cycles. $n = 10^{12}$, $s = 0.1, \delta = 1.3, b_{00} = 0.3, b_{10} = 0.45, b_{01} = 1, b_{11} = 1.1$ and $d_{00} = d_{10} = d_{01} = d_{11} = 0.1$.



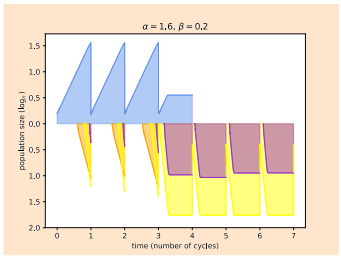
(a) Legend



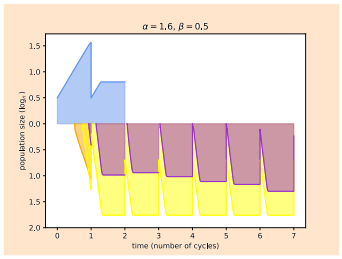
(b) Adaptation mainly from 01



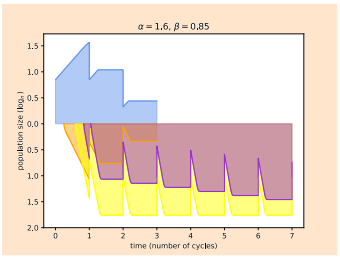
(c) Adaptation from 01



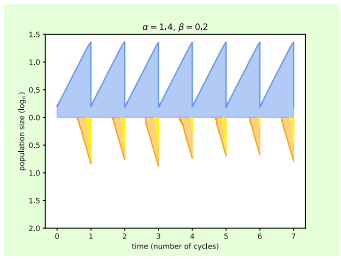
(d) Adaptation from 01



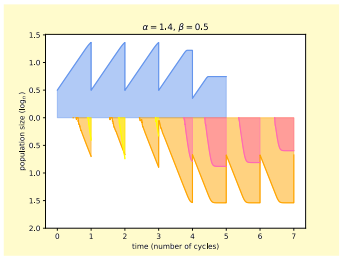
(e) Adaptation from 01



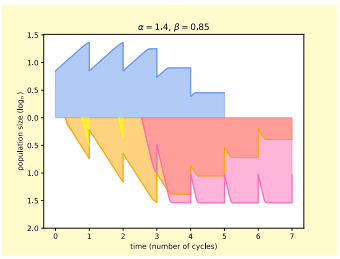
(f) Adaptation from 01



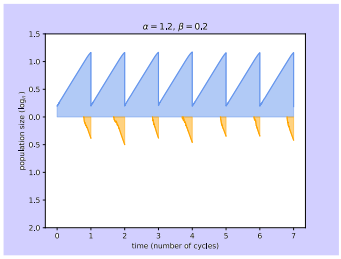
(g) No adaptation



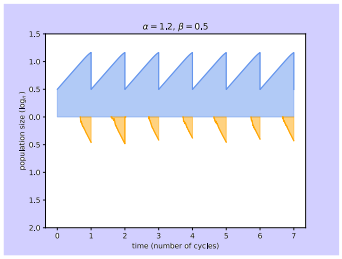
(h) Adaptation from 10



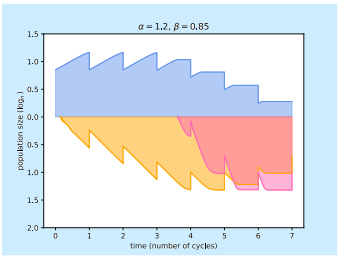
(i) Adaptation from 10



(j) No adaptation

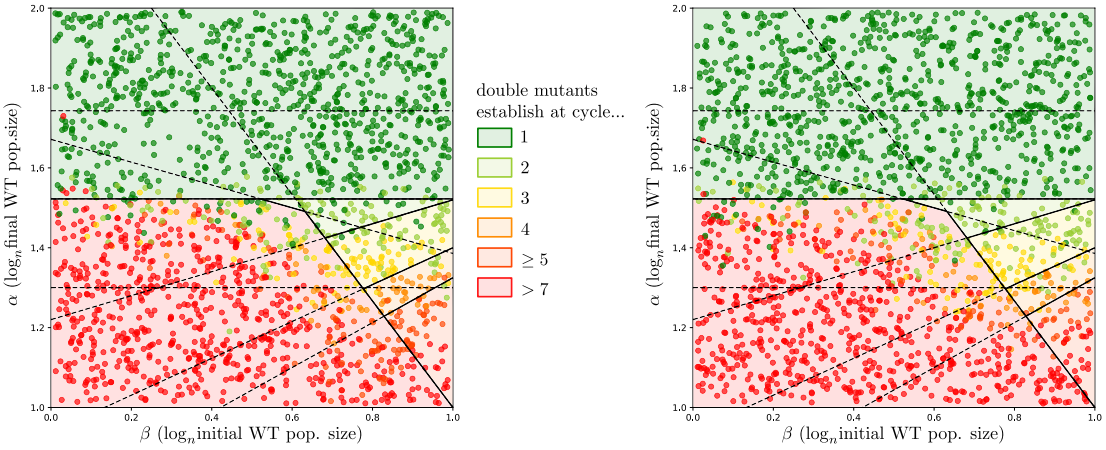


(k) No adaptation



(l) Adaptation from 10

Figure 2.8 – Examples of simulated trajectories for $(\alpha, \beta) \in \{1.2, 1.4, 1.6, 1.8\} \times \{0.2, 0.5, 0.85\}$, with density-dependence. WT population size is above, mutant population sizes are plotted below and superimposed. Population sizes are plotted on $\log_n$ scale. Each simulation was run for 7 cycles, with parameters $n = 10^9$, $p = 0.5$, $s = 0.1$, $\delta = 1.3$, $b_{00} = 0.3$, $b_{10} = 0.45$, $b_{01} = 1$, $b_{11} = 1.1$ and $d_{00} = d_{10} = d_{01} = d_{11} = 0.1$.



(a) Exponential growth **(b)** Density-dependent growth

Figure 2.9 – Observed cycles of double mutant emergence in 2,000 simulations with exponential growth (a) and density-dependent growth (b) for 7 cycles, with $n = 10^{12}$. $\delta = 1.3$, $b_{00} = 0.3, b_{10} = 0.45, b_{01} = 1, b_{11} = 1.1$ and $d_{00} = d_{10} = d_{01} = d_{11} = 0.1$. For simulations with density-dependence, $p = 0.5$ and $s = 0.1$. Point are colored according to the cycle at which double mutants established. Boundaries and background colors are theoretical predictions for exponential growth and large n.

Historique du projet

J'ai commencé à travailler sur ce projet durant mon stage de Master 2 en janvier 2020. Initialement, l'idée était de modéliser une expérience d'évolution de phages qui devait être réalisée dans l'équipe de Sylvain Gandon à Montpellier. Cette expérience avait pour but d'étudier l'adaptation d'un phage à un environnement hétérogène : les phages sont mis en présence d'une population de bactéries dont certaines sont susceptibles, et d'autres sont résistantes aux phages grâce à des systèmes de défense de type CRISPR-Cas ciblant différentes parties du génome viral. Les phages initiaux peuvent donc exploiter uniquement les bactéries susceptibles, tandis que des phages portant une mutation au niveau du site ciblé par une population de bactéries résistantes seront capables de contourner cette résistance et d'exploiter aussi cette population (Chabas et al. 2019). Les phages sont régulièrement transférés dans une nouvelle population bactérienne dont la composition est toujours identique, afin de pouvoir étudier uniquement l'adaptation des phages et non la coévolution phages-bactéries (Figure 2.10). Une partie des phages sont est régulièrement séquencée afin de détecter l'apparition de mutations permettant de contourner les résistances.

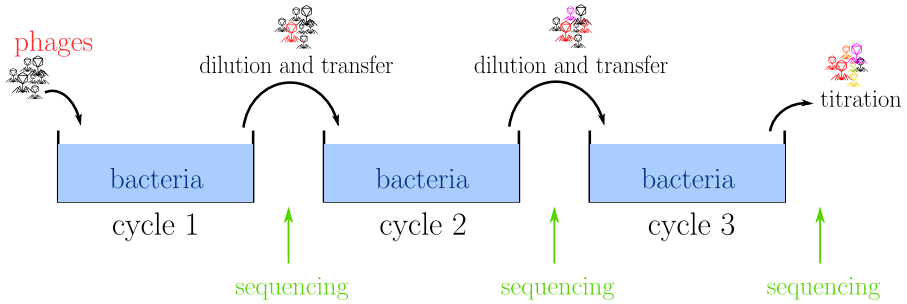


Figure 2.10 – Expérience d'adaptation de phages imaginée au début du projet.

Le but de cette expérience était d'étudier l'ordre d'apparition des mutations dans la population de phages, le coût de ces mutations, ainsi que la vitesse d'adaptation en fonction de la proportion de bactéries résistantes. L'objectif de notre modèle était de prédire l'évolution du système et de faire des recommandation pour le protocole expérimental. Le projet de modèle initial était donc plus complexe que celui présenté dans ce chapitre, avec 6 types de mutations différentes et un taux de croissance dépendant de la fraction de la population bactérienne pouvant être exploitée par chaque variant. Un des enjeux de cette expérience étant de pouvoir observer l'ordre d'apparition des différentes mutations, nous avions déterminé une zone de paramètres favorable pour laquelle les mutations les plus fréquentes arrivent au premier cycle et les mutations plus rares arrivent aux cycles suivants. Cette zone est représentée en blanc et orange sur la Figure 2.11, qui est donc l'ancêtre de la Figure 2.2.

Les confinements successifs ayant eu lieu à cette période 2020-2021 ont retardé la mise en place de ces expériences, et nous avons donc continué le travail de modélisation dans le cadre plus général présenté dans le reste du chapitre. Depuis,

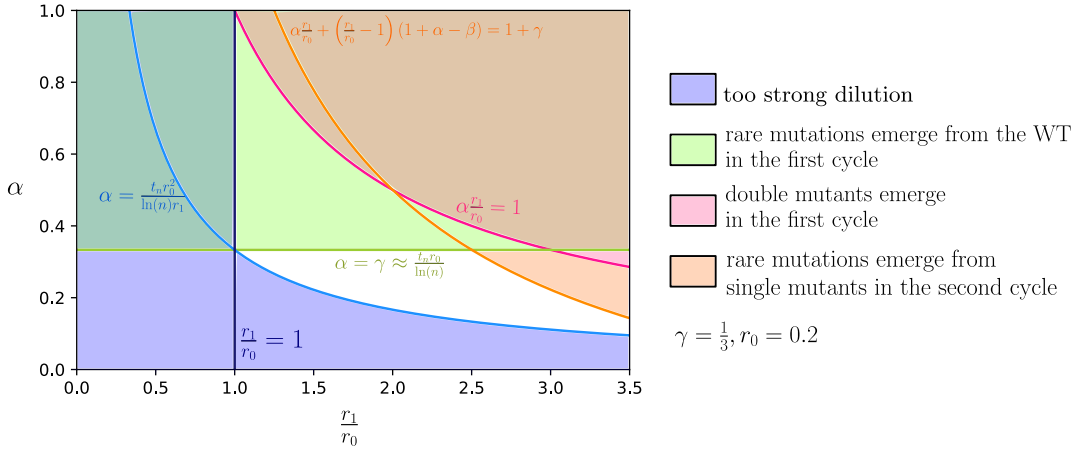


Figure 2.11 – Prédiction des comportements possible de la population de phages au cours des deux premiers cycles, en fonction des paramètres α (qui règle la taille de population maximale, défini à l'époque comme $n^{1+\alpha}$ au lieu de n^{α}) et $\frac{r_1}{r_0}$ (ratio des taux de croissance entre les simples mutants et la population initiale de phages). γ était alors le paramètre réglant la différence de taux entre les mutations rares et fréquentes (noté δ par la suite).

des expériences similaires ont été réalisées (sur la co-évolution phages-bactéries : Guillemet et al. 2022, et sur la vitesse d'adaptation des phages confrontés à des populations bactériennes hétérogènes portant un nombre variable de résistances : Gandon et al. 2024) mais la divergence entre les deux projets ne permet malheureusement plus une comparaison directe entre modèle et expérience.

Références du Chapitre 2

Abel, S., P. S. Z. Wiesch, B. M. Davis et M. K. Waldor (2015). Analysis of Bottlenecks in Experimental Models of Infection. *PLOS Pathogens* **11**.

Ahlfors, L. V. (1979). *Complex Analysis*. T. 3. McGraw-Hill New York.

Badran, A. H. et D. R. Liu (2015). In Vivo Continuous Directed Evolution. *Current Opinion in Chemical Biology* **24**, 1-10.

Barroso-Batista, J., A. Sousa, M. Lourenço, Marta Lourenço, M.-L. Bergman, D. Sobral, J. Demengeot, K. B. Xavier et I. Gordo (2014). The First Steps of Adaptation of Escherichia Coli to the Gut Are Dominated by Soft Sweeps. *PLOS Genetics* **10**.

Behrenfeld, M. J., Y. Hu, R. T. O'Malley, E. Boss, C. A. Hostetler, D. Siegel, D. A. Siegel, J. L. Sarmiento, J. A. Schulien, J. W. Hair, Xiaomei Lu, Xiaomei Lu, X. Lu, S. Rodier et A. J. Scarino (2017). Annual Boom-Bust Cycles of Polar Phytoplankton Biomass Revealed by Space-Based Lidar. *Nature Geoscience* **10**, 118-122.

Bertoin, J. (1996). *LÉVY PROCESSES*. Sous la dir. de S. J. Taylor. Cambridge University Press edition. T. 30. Cambridge.

Bloom, J. D. et F. H. Arnold (2009). In the Light of Directed Evolution: Pathways of Adaptive Protein Evolution. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 9995-10000.

Blount, Z. D., C. Z. Borland et R. E. Lenski (2008). Historical Contingency and the Evolution of a Key Innovation in an Experimental Population of Escherichia Coli. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 7899-7906.

Campos, P. R. A. et L. M. Wahl (2009). The Effects of Population Bottlenecks on Clonal Interference, and the Adaptation Effective Population Size. *Evolution* **63**, 950-958.

Campos, P. R. A. et L. M. Wahl (2010). The Adaptation Rate of Asexuals: Deleterious Mutations, Clonal Interference and Population Bottlenecks. *Evolution* **64**, 1973-1983.

Chabas, H., A. Nicot, S. Meaden, E. R. Westra, D. M. Tremblay, L. Pradier, S. Lion, S. Moineau et S. Gandon (2019). Variability in the Durability of CRISPR-Cas Immunity. *Philosophical Transactions of the Royal Society B* **374**, 20180097.

Desai, M. M. et D. S. Fisher (2007). Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics* **176**, 1759-1798.

De Visser, J. A. G. M. et J. Krug (2014). Empirical Fitness Landscapes and the Predictability of Evolution. *Nature Reviews Genetics* **15**, 480-490.

Eyre-Walker, A. et P. D. Keightley (2007). The Distribution of Fitness Effects of New Mutations. *Nature Reviews Genetics* **8**, 610-618.

Fogle, C. A., J. L. Nagle et M. M. Desai (2008). Clonal Interference, Multiple Mutations, and Adaptation in Large Asexual Populations. *Genetics* **180**, 2163-2173.

Freitas, O., L. M. Wahl et P. R. A. Campos (2021). Robustness and Predictability of Evolution in Bottlenecked Populations. *Physical Review E* **103**, 042415-042415.

Gandon, S., M. Guillemet, F. Gatchitch, A. Nicot, A. C. Renaud, D. M. Tremblay et S. Moineau (2024). Building Pyramids against the Evolutionary Emergence of Pathogens. *Proceedings of the Royal Society B: Biological Sciences* **291**, 20231529.

Garoff, L., F. Pietsch, D. L. Huseby, T. Lilja, G. Brandis et D. Hughes (2020). Population Bottlenecks Strongly Influence the Evolutionary Trajectory to Fluoroquinolone Resistance in Escherichia Coli. *Molecular Biology and Evolution* **37**, 1637-1646.

Geoghegan, J. L., A. M. Senior et E. C. Holmes (2016). Pathogen Population Bottlenecks and Adaptive Landscapes: Overcoming the Barriers to Disease Emergence. *Proceedings of The Royal Society B: Biological Sciences* **283**, 20160727.

Gillespie, D. T. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry* **81**, 2340-2361.

— (2001). Approximate Accelerated Stochastic Simulation of Chemically Reacting Systems. *The Journal of Chemical Physics* **115**, 1716-1733.

Good, B. H., M. J. McDonald, J. E. Barrick, R. E. Lenski et M. M. Desai (2017). The Dynamics of Molecular Evolution over 60,000 Generations. *Nature* **551**, 45-50.

Guillemet, M., H. Chabas, A. Nicot, F. Gatchich, E. Ortega-Abboud, C. Buus, L. Hindhede, G. M. Rousseau, T. Bataillon, S. Moineau et S. Gandon (2022). Competition and Coevolution Drive the Evolution and the Diversification of CRISPR Immunity. *Nature Ecology & Evolution* **6**, 1480-1488.

Hall, A. R., V. F. Griffiths, R. C. MacLean et N. Colegrave (2010). Mutational Neighbourhood and Mutation Supply Rate Constrain Adaptation in Pseudomonas Aeruginosa. *Proceedings of The Royal Society B: Biological Sciences* **277**, 643-650.

Handel, A. et D. E. Rozen (2009). The Impact of Population Size on the Evolution of Asexual Microbes on Smooth versus Rugged Fitness Landscapes. *BMC Evolutionary Biology* **9**, 236-236.

Hartl, D. L. (2014). What Can We Learn from Fitness Landscapes. *Current Opinion in Microbiology* **21**, 51-57.

Hermisson, J. et P. S. Pennings (2017). Soft Sweeps and beyond: Understanding the Patterns and Probabilities of Selection Footprints under Rapid Adaptation. *Methods in Ecology and Evolution* **8**, 700-716.

Huseby, D. L., F. Pietsch, G. Brandis, L. Garoff, A. Tegehall, A. Tegehall et D. Hughes (2017). Mutation Supply and Relative Fitness Shape the Genotypes of Ciprofloxacin-Resistant Escherichia Coli. *Molecular Biology and Evolution* **34**, 1029-1039.

Kassen, R. et T. Bataillon (2006). Distribution of Fitness Effects among Beneficial Mutations before Selection in Experimental Populations of Bacteria. *Nature Genetics* **38**, 484-488.

Kawecki, T. J., R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri et M. C. Whitlock (2012). Experimental Evolution. *Trends in ecology & evolution*.

Keller, P. et T. Antal (2014). Mutant Number Distribution in an Exponentially Growing Population. *Journal of Statistical Mechanics: Theory and Experiment*.

Kendall, D. G. (1948). On the Generalized "Birth-and-Death" Process. *The Annals of Mathematical Statistics* **19**, 1-15.

Lang, G. I., D. P. Rice, M. J. Hickman, E. Sodergren, G. M. Weinstock, D. Botstein et M. M. Desai (2013). Pervasive Genetic Hitchhiking and Clonal Interference in Forty Evolving Yeast Populations. *Nature* **500**, 571-574.

Lea, D. E. et C. A. Coulson (1949). The Distribution of the Numbers of Mutants in Bacterial Populations. *Journal of Genetics* **49**, 264-285.

LeClair, J. S. et L. M. Wahl (2018). The Impact of Population Bottlenecks on Microbial Adaptation. *Journal of Statistical Physics* **172**, 114-125.

Luria, S. E. et M. Delbrück (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491-511.

Maddamsetti, R., R. E. Lenski et J. E. Barrick (2015). Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Experiment with Escherichia Coli. *Genetics* **200**, 619-631.

Mahrt, N., A. Tietze, S. Künzel, S. Franzenburg, C. Barbosa, G. Jansen et H. Schulenburg (2021). Bottleneck Size and Selection Level Reproducibly Impact Evolution of Antibiotic Resistance. *Nature Ecology and Evolution* **5**, 1233-1242.

Messer, P. W. et D. A. Petrov (2013). Population Genomics of Rapid Adaptation by Soft Selective Sweeps. *Trends in Ecology and Evolution* **28**, 659-669.

Miralles, R., P. J. Gerrish, A. Moya et S. F. Elena (1999). Clonal Interference and the Evolution of RNA Viruses. *Science* **285**, 1745-1747.

Nair, S., D. Nash, D. Sudimack, A. Jaidee, M. Barends, A.-C. Uhlemann, S. Krishna, Sanjeev Krishna, S. Krishna, F. Nosten, T. J. C. Anderson et Timothy J. C. Anderson (2006). Recurrent Gene Amplification and Soft Selective Sweeps

during Evolution of Multidrug Resistance in Malaria Parasites. *Molecular Biology and Evolution* **24**, 562-573.

Nicholson, M. D. et T. Antal (2019). Competing Evolutionary Paths in Growing Populations with Applications to Multidrug Resistance. *PLOS Computational Biology* **15**.

Pennings, P. S., S. Kryazhimskiy et J. Wakeley (2014). Loss and Recovery of Genetic Diversity in Adapting Populations of HIV. *PLOS Genetics* **10**.

Penson, K. A. et K. Górska (2010). Exact and Explicit Probability Densities for One-Sided L\'evy Stable Distributions. *Physical Review Letters* **105**, 210604.

Schenk, M. F., M. P. Zwart, S. Hwang, P. Ruelens, E. Severing, J. Krug et J. A. G. M. de Visser (2022). Population Size Mediates the Contribution of High-Rate and Large-Benefit Mutations to Parallel Evolution. *Nature Ecology and Evolution*.

Szendro, I. G., J. Franke, J. A. G. M. de Visser, J. A. G. M. de Visser et J. Krug (2013). Predictability of Evolution Depends Nonmonotonically on Population Size. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 571-576.

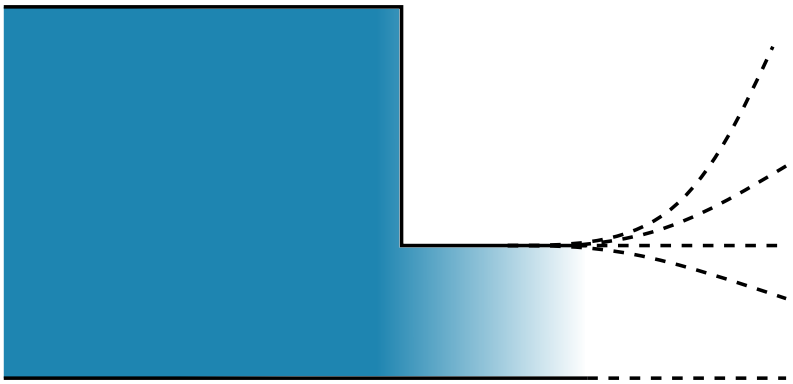
Wahl, L. M. et P. J. Gerrish (2001). The Probability That Beneficial Mutations Are Lost in Populations with Periodic Bottlenecks. *Evolution* **55**, 2606-2610.

Wahl, L. M., P. J. Gerrish et I. Saika-Voivod (2002). Evaluating the Impact of Population Bottlenecks in Experimental Evolution. *Genetics* **162**, 961-971.

Weinreich, D. M., N. F. Delaney, M. A. DePristo et D. L. Hartl (2006). Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312**, 111-114.

Zheng, Q. (1999). Progress of a Half Century in the Study of the Luria–Delbrück Distribution. *Bellman Prize in Mathematical Biosciences* **162**, 1-32.

Chapitre 3



3

How bottlenecks shape adaptive potential: from theory and microbiology to conservation biology

Sommaire

3.1 Introduction

Most natural populations experience bottlenecks that can be caused, for example, by severe climate events, habitat loss, and overhunting (Lande 1988; Frankham, Briscoe, and Ballou 2002). The bottleneck-induced population size reduction increases the extinction risk of populations and, thus, may destabilize ecosystems (Frankham, Lees, et al. 1999). As the current intensive anthropogenic perturbations of Earth's systems are increasing the occurrence of bottlenecks (Barnosky et al. 2011; Ceballos et al. 2015), there is an urgent need to understand how bottlenecks impact the evolutionary fate of at-risk populations. This review examines the evolutionary fate of populations undergoing sudden random decreases in size that are selectively neutral, as opposed to those caused by directional selection. Specifically, this review focuses on bottlenecks involving random reductions in population size rather than selective bottlenecks involving non-random reductions in population size (but see Section 3.4.2).

Forecasting the effects of population bottlenecks is crucial in conservation biology (Frankham, Briscoe, and Ballou 2002). However, our ability to predict the long-term consequences of bottlenecks under natural conditions remains weak. First, predictions in the wild are generally made retrospectively, i.e., after a bottleneck has occurred and the population has gone extinct or survived (Bouzat 2010). Second, predictions are often made case-by-case, preventing their applicability to other systems. To fill this knowledge gap and improve our understanding of bottlenecks, we need to develop a comprehensive overview combining theoretical predictions and empirical evidence. Although connections between fundamental evolutionary biology and wildlife conservation have slowly developed, they are increasingly strengthening, highlighting their importance (Hohenlohe, Funk, and Rajora 2021).

Microbiology is a field that allows for experimentally assessing the impact of bottlenecks on the adaptive potential of microbial populations (LeClair and Wahl 2018). A common experimental evolution technique is subjecting micro-organisms to serial passaging, which involves repeated bottlenecks. In such experiments, a microbial population is inoculated into a medium and grows. Then, the experimenter takes a fraction of this population, inoculates a new medium, and repeats the process several times. This common technique explains why the literature in experimental microbiology has so well documented the impact of bottlenecks on the adaptation of microbial populations (LeClair and Wahl 2018). Micro-organisms, such as bacteria and fungi, can reproduce rapidly and reach large numbers in small spaces, allowing for highly replicated experiments and, therefore, highly accurate predictions (Elena and Lenski 2003). In summary, the simplicity of these experiments makes microbiology an excellent field for testing theoretical predictions.

To the best of our knowledge, no study has yet bridged the gap between what microbiology tells us about the impact of bottlenecks on the adaptive potential of populations and conservation biology. The lack of direct links between both fields likely results from the many differences between microbial populations in controlled laboratory experiments and wild populations in natural ecosystems (see Box 4 in Kawecki et al. 2012):

1. Microbes used in experiments substantially differ from endangered natural species targeted by conservation efforts, mostly diploid and sexual.

2. The demography of microbial populations studied in laboratory conditions also differ from that of wild populations. For example, the size of microbial populations is likely larger than that of wild populations. As a result, the genetic load is likely higher in wild populations than in laboratory populations.

3. Bottlenecks in the wild likely vary in intensity and frequency, whereas they are typically periodic in microbiology experiments.

4. The number of generations before adapting to a new environment differs for microbial and wild populations.

Many other differences exist, such as variations in environmental constraints, and anthropogenic pressures, which affect wild population dynamics but are absent from laboratory conditions. Despite these differences, our review describes how microbiology findings apply to wild populations.

The evolution of wild populations involves multiple evolutionary and ecological processes that act simultaneously. Understanding the influence of each process independently is crucial for a better understanding of the overall effect of a bottleneck during demographic history on future response to selection to a new environment. Yet, empirical studies under laboratory conditions involving models other than micro-organisms have never examined the effects of bottlenecks in anything other than a holistic way (e.g., fish populations of *Heterandria formosa* in Klerks et al. 2019, insect populations of *Tribolium castaneum* in Olazcuaga et al. 2023, *Drosophila melanogaster* in Ørsted, Hoffmann, et al. 2019). Therefore, it is impossible to quantify the contribution of each process. Conversely, theory and microbiology have studied each ecological and evolutionary process independently (e.g., the fraction of beneficial mutations lost due to bottlenecks in Wahl, Gerrish, and Saika-Voivod 2002). Therefore, we discuss in this review the impact on adaptive potential of each of the following evolutionary processes occurring during a bottleneck: mutation, genetic drift, natural selection, and gene flow.

This review aims to enhance our comprehension of how bottlenecks impact adaptive potential to a new environment of a population. This aim is achieved by synthesizing theoretical and microbiological knowledge and applying it to wild populations. For each of the evolutionary processes of interest, we inquire: (i) What do theory and microbiology experiments tell us about the impact of bottlenecks during demographic history on the ability of populations to adapt to a future environmental change? (ii) Do these theoretical predictions apply to wild populations? and (iii) What is needed to better predict the wild population evolution after a bottleneck? Our review aims to increase the effectiveness of conservation efforts by anticipating the evolutionary consequences of demographic changes in wild ecosystems.

3.2 Disentangling the influence of each evolutionary process during a bottleneck on the future adaptive potential of populations

3.2.1 Mutation

Population bottlenecks can impact mutations' appearance, fixation, and frequency, potentially disturbing future adaptation. Indeed, mutations introduce genetic variation on which selection can act, allowing populations to adapt to their environment. A population will adapt to future environmental changes by increasing the frequency of mutations that are beneficial in the new environment. The population can adapt via (i) the appearance of new beneficial mutations, known as an adaptation from *de novo* mutations, and (ii) the pre-existence of beneficial mutations in the population, known as an adaptation from standing genetic variation. The relative importance of these two mechanisms varies depending on the population properties, such as the population size, and some timescales, such as the number of generations between the bottleneck and the environmental change. In the following section, we discuss how these two mechanisms, i.e., the appearance of *de novo* mutations and mutations from standing genetic variation (or pre-existing mutations), can be impacted during a bottleneck.

Impact of bottleneck on *de novo* mutations

Theoretical work has shown that bottlenecks can reduce the mutation supply and the fixation probability of beneficial mutations in populations experiencing them (Wahl, Gerrish, and Saika-Voivod 2002). More precisely, populations adapting mainly from *de novo* mutations have an adaptation rate limited by the mutation supply, which depends on population size and the appearance rate of adaptive beneficial mutations. These theoretical predictions were experimentally confirmed with asexual populations where adaptation depends mainly on *de novo* mutations (see, e.g., de Visser and Rozen 2005). Campos and Wahl (2010) even derived a more complex expression of the adaptation rate of asexuals, taking into account clonal interference. Mechanically, a bottleneck reduces the population size and, thus, limits adaptation.

Additionally, the adaptation rate of populations adapting through *de novo* mutations also depends on the distribution of fitness effects of beneficial mutations, which may also be affected by bottlenecks. Indeed, the fixation probability of all beneficial mutations is predicted to be reduced in a population undergoing bottlenecks (Wahl, Gerrish, and Saika-Voivod 2002). However, mutations are affected differently depending on their rate and the fitness benefit they confer, which are usually negatively correlated. Gamblin, Gandon, et al. (2023) used a stochastic model to show that severe bottlenecks following a long growth phase favor rare beneficial mutations (as shown by Wahl, Gerrish, and Saika-Voivod 2002). In contrast, relaxed bottlenecks following short growth phases favor frequent weakly ben-

eficial ones. A similar effect has been observed in microbial experiments studying antimicrobial resistance. Garoff et al. (2020) and Schenk et al. (2022) found that antimicrobial resistance evolved through weakly beneficial mutations with large mutational targets when the population size prior to the bottleneck was small. When the population size prior to the bottleneck was larger, antimicrobial resistance evolved through rarer and more beneficial mutations.

The theoretical and experimental studies mentioned above mostly deal with asexual populations, which rely more on *de novo* mutations to adapt to environmental changes than sexual populations, the latter being more frequently genetically diverse. Even if most endangered species, which are the focus of conservation efforts, are sexual, these predictions could apply to them. Indeed, recent evidence suggests that some animal species are also limited by mutation supply on recent evolutionary scales (Rousselle et al. 2020).

Impact of bottleneck on pre-existing genetic variation

Populations experiencing a bottleneck are theoretically expected to have reduced genetic diversity (R. Chakraborty and Nei 1982; Lynch and Hill 1986; Nei, Maruyama, and R. Chakraborty 1975; Tajima 1989; Tajima 1996). This reduction can limit their ability to adapt to future changing environments (Frankham, Briscoe, and Ballou 2002; Willi, Tregenza, et al. 2006). This effect is especially important for sexual populations, as their adaptation in a short timescale is mostly driven by standing genetic variation. Indeed, sexual populations have constrained access to beneficial mutations due to Haldane's sieve, which results from selection mainly acting on heterozygotes, thus decreasing the fixation probability of beneficial recessive mutations compared to asexual populations (see Marad, Buskirk, and Lang 2018 for a comparison in yeast). Also, sexual reproduction allows selection to act on individual loci rather than haplotypes, thus making it possible to exploit the standing diversity (see Burke, Liti, and Long 2014 for this observation in yeast). Finally, higher organisms typically have smaller population sizes and lower mutation rates than micro-organisms, resulting in a limited supply of new mutations to rely on for adaptation.

Studies on asexual yeasts have shown that standing genetic variation drives adaptation along with *de novo* mutations (Vázquez-García et al. 2017; Ament-Velásquez et al. 2022). However, this aspect of microbial adaptation has not been extensively studied because experiments involving asexual individuals often start with a clonal population. Thus, there are no results yet from microbial experiments based on standing genetic variation that could apply to wild endangered populations.

Nonetheless, some observations from *Drosophila* experiments show how bottlenecks impact adaptation from standing genetic variation by disrupting allele frequencies. Rare alleles are likely to be lost during the bottleneck, resulting in a reduced allelic diversity (Allendorf 1986; Fuerst and Maruyama 1986). Swindell and Bouzat (2005) performed an empirical test of the drift-mutation model using *Drosophila*. This drift-mutation model aimed to predict the adaptive potential of

a population through genetic variation, which is modeled as an equilibrium between mutations and fixation due to inbreeding (Lynch and Hill 1986; Clayton and Robertson 1955). In particular, this model assumes that the adaptive potential only depends on heterozygosity and not on allelic diversity (Falconer 1960), which are two different aspects of the genetic diversity of a population. During their experiment, Swindell and Bouzat (2005) observed a good agreement between model predictions and empirical observations, except after a severe bottleneck. The authors hypothesized that not considering the loss of allelic diversity during the bottleneck leads to overestimating the adaptive capacity following this event. This result suggests that the loss of heterozygosity and rare alleles are to be accounted for when predicting the effect of a bottleneck on a wild endangered population.

Additive genetic variance from a quantitative genetics perspective

For quantitative traits, the additive component of the genetic variance, which is denoted V_A, is often taken as a proxy for the adaptive potential. The effect of a bottleneck on quantitative genetic variation is more complex to predict (Hoffmann, Sgrò, and Kristensen 2017). In theory, the genetic variance should decrease after a bottleneck as it is proportional to the effective population size (Lynch and Hill 1986). This effect is usually observed in morphological traits (Willi, Van Buskirk, et al. 2007), whereas fitness-associated life-history traits often show an increased genetic variance following a bottleneck (Van Heerwaarden et al. 2008). A possible explanation is that, after a bottleneck, the disruption of allele frequencies could result in a transfer of epistatic and dominance variance to additive variance, especially for life-history traits, which are more influenced by these non-additive effects (Crnokrak and Roff 1995; Roff and Emerson 2006). However, these life-history traits also typically experience high inbreeding depression (DeRose and Roff 1999). As a result, an increase in additive variance for the genes associated with these traits may mitigate the fitness decrease. Still, it will not allow an increase in fitness compared to the pre-bottleneck level.

Overall, Willi, Tregenza, et al. (2006) and Lopez-Fanjul and Villaverde (1989) concluded that genetic variance, and thus the future adaptive response, may increase in a population facing a bottleneck ("bottlenecked population" hereafter). However, this phenomenon is unlikely to result in a full fitness recovery, let alone a fitness increase.

The implication for endangered species management is that computing the additive genetic variance just after a bottleneck event may not reflect long-term adaptive potential but merely short-term adaptation in reaction to this event.

Prospects for filling knowledge gaps

Understanding the effect of bottlenecks simultaneously on pre-existing mutations and *de novo* mutations appearing during a bottleneck is important to predict the impact of bottlenecks on adaptability better. Performing microbial experiments that include initial genetic variation, allowing for better differentiating the effect of

de novo mutations from standing genetic variation after a bottleneck, could improve this understanding. A recent review (Burke 2023) suggested using yeast evolution experiments with standing genetic variation to study eukaryote adaptation. Indeed, yeast can combine short generation time and easy handling in the lab with sexual reproduction. Performing such experiments, particularly with small populations undergoing different types of bottlenecks, would help quantify the importance of standing genetic variation versus *de novo* mutations for eukaryote adaptation.

To improve our knowledge of which quantitative traits have their variance decreased after a bottleneck, we would need experiments to estimate the genetic variance in a wide range of traits (Willi, Tregenza, et al. 2006). Estimating the genetic variance seems essential to predict the overall effect of a bottleneck on the population's adaptive potential. Moreover, in natural populations, the link between genetic diversity and response to selection is not always clear (Pujol et al. 2018). For example, Albatross persists despite losing genetic diversity (Milot et al. 2007). Clarifying the link between genetic variation and response to selection would help understand if genetic diversity can be used to accurately predict the natural populations' potential to adapt after one or more bottlenecks.

3.2.2 Genetic drift

Populations suffering from bottlenecks are particularly affected by genetic drift, which is the change in allele frequencies caused by population size fluctuations rather than by selection, mutation, or migration. Indeed, the strength of this process is inversely proportional to the effective size of the population (Kimura 1955). These population size fluctuations caused by chance likely lead to negative impacts on a population, such as (i) the fixation of deleterious mutations, which decreases the population's fitness; (ii) the reduction in the fixation probability of beneficial mutations, which limits adaptation, and (iii) the increase of alleles at extreme frequencies (i.e., 0 or 1), which reduces genetic variation (Falconer 1960). The latter point has already been covered in Section 3.2.1 about mutation. Taken together, these effects may limit the adaptation of bottlenecked populations to future environmental changes.

Fixation of deleterious mutations

The accumulation of deleterious mutations caused by genetic drift in a population undergoing repeated bottlenecks predicted by theoretical work was highlighted numerous times in microbial experiments (H. Muller 1964). In particular, many experiments used clone-to-clone transfers to maximize the rate and the speed of accumulation (Clarke et al. 1993) (see the review Elena and Lenski (2003) for references on viruses, bacteria, and yeast). For instance, a linear decay of the average fitness of a hypermutator *Escherichia coli* strain subject to repeated single-cell bottlenecks was observed in (Heilbron et al. 2014).

Sexual populations are also theoretically expected to suffer from deleterious mutation accumulation (Lynch, Conery, and Burger 1995). This accumulation was

observed in domesticated species due to bottlenecks and selective sweeps (Marsden et al. 2016; Xie et al. 2018), but also in some wild bottlenecked populations such as the Florida panther (Roelke, Martenson, and O'Brien 1993). However, the deleterious mutation is potentially less common than in asexual microorganisms, where, due to the absence of recombination, the offsprings are expected to bear at least as much mutational load as their ancestors, a process called Muller's ratchet (H. Muller 1964; H. J. Muller 1932). In wild populations that reproduce sexually, recombination can break this process (McDonald, Rice, and Desai 2016). Therefore, the populations being the focus of conservation efforts are probably less affected by this particular bottleneck effect.

Reduced fixation probability of beneficial mutations

Some theoretical studies reproducing microbiology experiments showed that bottlenecks can reduce the fixation probability of a beneficial mutation. For example, Wahl, Gerrish, and Saika-Voivod (2002) found that the fixation probability of a beneficial mutation in a periodically bottlenecked population was reduced by a factor accounting for the probability of losing the mutation during the dilution. As a reminder, the fixation probability of a beneficial mutation in a fixed-size population is approximately twice the selective advantage (Haldane 1927). Heffernan and Wahl (2002) also considered that genetic drift is increased in bottlenecked populations due to a lower size resulting from the bottleneck. This effect reduces the fixation probability by about 25% compared to previous estimates. These theoretical predictions were confirmed by experiments using microorganisms where bottlenecks and genetic drift hindered adaptation. In the case of experimental evolution of antibiotic resistance, Huseby et al. (2017) found a positive correlation between bottleneck size and ciprofloxacin tolerance. In addition, Garoff et al. (2020) highlighted that a low-intensity bottleneck (i.e., a small reduction in population size) leads to higher fluoroquinolone tolerance than a high-intensity bottleneck. This effect is also likely to impact wild populations of endangered species, though the extent of this impact is not clear as their adaptation relies mostly on standing genetic variation.

Prospects for filling knowledge gaps

As bottleneck-induced drift affects both new and existing mutations, estimating its impact on the adaptive potential of bottlenecked populations would require quantifying the respective importance of *de novo* mutations versus standing genetic variation for the adaptation of a given population on a given timescale. This problem was previously mentioned in Section 3.2.1 about mutation. The review (R. D. H. Barrett and Schluter 2008) presents the relative contribution of these two sources of genetic variation in wild populations. This review suggested ways to detect molecular signatures of adaptation from standing genetic variation (R. D. H. Barrett and Schluter 2008; Przeworski, Coop, and Wall 2005).

Another open question deals with the potential beneficial effect of genetic drift on the adaptive potential of populations. In a modeling study, Handel and Rozen (2009) found that small asexual populations could reach higher fitness peaks than large ones on rugged landscapes because drift prevents them from being stuck on a local maximum. The authors concluded that there is an optimal population size to maximize adaptation, depending on the fitness landscape's characteristics and the relative importance of adapting rapidly versus reaching high fitness peaks (Handel and Rozen 2009). Assessing whether these effects are also observed in wild populations would be interesting.

3.2.3 Natural selection

Natural selection will act more or less effectively on the bottlenecked population depending on several factors, such as the severity of the bottleneck, its duration, and the population's genetic diversity before and after the bottleneck. Some general predictions can be drawn from microbiology about the impact of natural selection after a bottleneck, regardless of the characteristics of the bottleneck: (i) Several mechanisms tend to reduce the population fitness, such as genetic load and inbreeding depression, but (ii) natural selection can also purge deleterious alleles in sexual populations. These mechanisms will modify the population's fitness, impacting its ability to adapt to future environmental changes.

Fitness decrease due to bottlenecks

Several factors may explain why a bottlenecked population experiences a fitness decline, even without any environmental change.

As detailed in Section 3.2.2 about genetic drift, bottlenecks are expected to cause an accumulation of deleterious mutations in the population due to genetic drift, leading to a decrease in fitness if the bottlenecks are severe.

It is important to note a major difference between microbial populations and the endangered wild populations of diploid eukaryotes that we are considering in this review: the latter can suffer from inbreeding depression because bottlenecks reduce population sizes (D. Charlesworth and B. Charlesworth 1987; L. F. Keller, Lukas F. Keller, and Waller 2002), leading to more reproductive events between related individuals. This inbreeding depression results in a loss of heterozygosity that can unmask recessive deleterious alleles, ultimately decreasing this population's fitness and adaptive potential (S. C. H. Barrett and Kohn 1991; Ellstrand and Elam 2003). More generally, genetic load (i.e., the actual or potential reduction in population mean fitness due to drift load, inbreeding load, and mutation load) is responsible for a direct decline in population fitness following a bottleneck (Hedrick and Garcia-Dorado 2016; Kirkpatrick and Jarne 2000).

In less well-understood ways, bottlenecks can affect other characteristics of the populations than genetics but still influence their fitness and future ability to adapt. Specifically, a bottleneck can impact the balanced relationship between host and microbiome in eukaryotes. For instance, Ørsted, Yashiro, et al. (2022) showed

that *Drosophila* populations that had undergone bottleneck treatment also lost the diversity and richness of their microbiome. The direct consequence of this loss is a reduction in the fitness of individuals belonging to bottlenecked populations (Ørsted, Yashiro, et al. 2022).

Purge of deleterious alleles by natural selection increases population fitness

Whereas a bottleneck can increase the frequency of deleterious alleles (see Section 3.2.2 about genetic drift), natural selection can purge these deleterious alleles (Kirkpatrick and Jarne 2000; Hedrick and Garcia-Dorado 2016). If a purging process is more efficient during a bottleneck, then going through a bottleneck could be beneficial for the adaptive potential of the population (Bouzat 2010; Bertorelle et al. 2022; Dussex et al. 2023).

Purifying selection can play out in microorganism experiments and yet population evolution, but inbreeding facilitates the purge in diploid eukaryotes (Hedrick and Garcia-Dorado 2016). As mentioned above, inbreeding increases homozygosity and, thus, the unmasking of recessive deleterious alleles. Whereas inbreeding depression decreases population fitness, it is also an opportunity for selection to act on these deleterious alleles and purge them.

The empirical evidence for a purge of deleterious alleles following a bottleneck appears to be mixed (Bouzat 2010). There is some evidence that purge can strongly affect experimental populations (Crnokrak and S. C. H. Barrett 2002) and captive populations (López-Cortegano, Moreno, and García-Dorado 2021; Boakes, J. Wang, and Amos 2006). In experimental yeast populations, Agrawal and Whitlock (2011) used data from the Saccharomyces Genome Deletion Project to estimate fitness and dominance coefficient at about 1000 loci. From this, the authors estimated that the effect of one generation of purging (i.e., deliberate inbreeding) in an already partially inbred population would substantially decrease inbreeding depression.

Evidence of a purge process in wild endangered populations is mostly indirect (Bouzat 2010). However, some direct evidence exists. For example, the deleterious load was significantly lower for the endangered species Iberian lynx (*Lynx pardinus*) than for the widespread Eurasian lynx (*Lynx lynx*) due to a genetic purging process (Kleinman-Ruiz et al. 2022). Other examples of purging in natural environments after a bottleneck exist (e.g., the Alpine ibex: Grossen et al. 2020). As discussed in (Bouzat 2010), the role of purging during a bottleneck and the factors influencing its role in natural populations still need to be discovered.

Prospects for filling knowledge gaps

While allele purging appears to be a key process for understanding the adaptive potential of bottlenecked populations, empirical evidence is still mixed, proving that we do not yet fully understand this mechanism. Therefore, it would be useful

to use diploid eukaryotic microorganisms to test the factors and conditions under which allele purging occurs.

In addition, a study suggested carefully handling the results of selection detection methods when working with bottlenecked populations (Leigh et al. 2021). Indeed, Leigh et al. (2021) found that, due to the high level of genetic drift, methods commonly used to detect selection (e.g., F_{ST} outlier scans and Genome-Environment Association analyses) presented high false positive rates when applied to bottlenecked Alpine Ibex populations. Detecting adaptation is essential for managing endangered populations, so testing other methods' false positive and negative rates and developing new methods to distinguish between drift and selection is necessary.

3.2.4 Gene flow

Human activity causes fragmentation of populations in the wild, leading to spatially structured populations divided into sub-populations (also called demes or islands) of reduced sizes between which individuals may migrate. Thus, population fragmentation results in a bottleneck that risks reducing genetic diversity within sub-populations (i.e., genetic depression), losing adaptive potential, and accumulating deleterious mutations (Keyghobadi 2007; Frankham, Ballou, et al. 2017). However, the migration of individuals between sub-populations can induce gene flow, which represents an opportunity to diversify the gene pool of the sub-populations despite the fragmentation-induced bottleneck. Quantifying the genetic diversity of sub-populations is crucial to assessing the adaptive potential of fragmented populations, particularly in the case of changing environments threatening their persistence. This section reviews how and in what conditions gene flow can restore the adaptive potential of a bottlenecked population. Gene flow can (i) restore lost genetic variation, (ii) mitigate inbreeding depression, (iii) resulting in a decreased probability of extinction and restoration of adaptive potential, and (iv) amplify natural selection depending on the meta-population structure.

Restoration of lost genetic variation

One of the main effects of gene flow in a bottleneck population is the restoration of lost genetic variation (Soulé 1987; Franklin and Frankham 1998). This theoretical expectation is observed experimentally (Swindell and Bouzat 2006) and in natural populations (Jangjoo et al. 2016; Goodman et al. 2001; Chiucchi and Gibbs 2010).

Habitat fragmentation can cause the extinction of bottlenecked populations. Gene flow between sub-populations can mitigate the negative effects of bottlenecks by restoring lost genetic diversity (Ingvarsson 2001). For example, Jangjoo et al. (2016) discovered that connectivity in a meta-population of the alpine butterfly *Parnassius smintheus* preserves genetic diversity before, during, and after a two-generation bottleneck. The Roseate Tern (*Sterna dougallii dougallii*) is an endangered Atlantic seabird population that provides another example of how con-

nectivity and gene flow across populations can help retain genetic diversity despite a severe bottleneck (Dayton and Szczys 2021). Seed dispersal with water facilitates gene flow between bottlenecked populations, mitigating the decrease in allelic diversity caused by a bottleneck (Yu et al. 2020). These examples are not isolated cases and are found in many endangered species. To further elaborate, gene flow between sub-populations undergoing a bottleneck can even erase the genetic variation effects of a bottleneck to the point where no negative genetic effects can be detected (e.g., *Actinidia chinensis* populations: Yu et al. 2020).

Change in genetic load composition

One less studied effect of gene flow on bottleneck populations, which could be predominant in the populations' fate, is its impact on genetic load. Gene flow is theoretically expected to reduce the deleterious effects of inbreeding in bottleneck populations. Gene flow in metapopulations can mitigate inbreeding load by preventing the fixation of deleterious alleles in bottleneck populations (Whitlock 2003). However, gene flow can also reduce the effectiveness of the purge of these deleterious alleles by increasing heterozygosity. With individual-based simulations, a study found that an intermediate rate of gene flow can minimize the mutation load and prevent the extinction of local populations while still allowing some purging of deleterious alleles (Sachdeva, Olusanya, and Barton 2022).

Gene flow mitigate extinction risk

Bottlenecked populations are highly vulnerable to extinction via (i) demographic stochasticity (e.g., random births and deaths), demographic heterogeneity and sampling variation in sex ratios (Melbourne and A. Hastings 2008), and (ii) environmental stochasticity (e.g., catastrophic events: Lande 1988). Specifically, bottlenecked populations can fall into an "extinction vortex", often characterized by a complex interplay between genetic drift, demographic stochasticity, and environmental fluctuations (Soulé 1986). A population bottleneck reduces fitness directly through increased genetic load and indirectly through erosion of genetic variation, leading to population decline, exacerbating the effects of genetic drift, demographic stochasticity, and environmental fluctuations until extinction (Nordstrom et al. 2023). Theoretical models have highlighted a critical level of gene flow that allows a metapopulation to survive over long timescales, even if it is often ultimately driven to extinction (Hanski and Ovaskainen 2003; Gyllenberg and Hanski 1992; Lande et al. 1998).

The process of restoring gene flow in these bottlenecked populations to alleviate genetic load, increase genetic variation, and increase persistence probability is termed genetic rescue (Bell, Fugère, et al. 2019; Tallmon, Luikart, and Waples 2004; Whiteley et al. 2015). Much empirical evidence suggests that recently fragmented populations will likely receive a demographic benefit from gene flow, beyond the addition of immigrant individuals, through genetic rescue (Whiteley et al. 2015; Frankham 2015; Hufbauer et al. 2015; Fitzpatrick and Reid 2019). A recent meta-

analysis revealed the significant and consistent benefits of gene flow for the adaptive potential of endangered species experiencing a fragmentation-induced population bottleneck (Frankham 2015).

When a population faces deteriorating environmental conditions and is doomed to extinction, gene flow may allow its evolutionary rescue. For example, Bell and Gonzalez (2011) set up an experiment in which a yeast metapopulation was subjected to salt-induced environmental stress. This experiment showed that local yeast dispersal and gradual deterioration favored the evolutionary rescue of the metapopulation, which would otherwise die out due to environmental stress. This experimental result later led to the development of theoretical models investigating the probability of evolutionary rescue by including a hitherto overlooked ecological factor: population structure, which may result from fragmentation. Interestingly, these models showed that the probability of evolutionary rescue in an island model, in which demes deteriorate one by one, does not vary monotonically with gene flow rate (Uecker, Otto, and Hermisson 2014). In other words, there is a gene flow rate that optimizes the evolutionary rescue of a population. Gene flow allows genetic variation and the introduction of beneficial mutants necessary for adaptation. However, a too-strong gene flow risks preventing local beneficial mutants from becoming permanently established, hence the need for intermediate gene flow to optimize adaptation (Tomasini and Peischl 2022). Further studies showed that directed gene flow based on habitat choice could favor evolutionary rescue (Czuppon et al. 2021). Habitat choice occurs, for example, when individuals, whether mutant or wild-type, preferentially immigrate to demes whose environment has already changed. Other details of population fragmentation, such as between which sub-populations gene flow is allowed (e.g., island model, stepping-stone model) or its asymmetry, impact the probability of evolutionary rescue (Tomasini and Peischl 2020; Tomasini and Peischl 2022). To our knowledge, the above-mentioned theoretical predictions have not been tested experimentally.

Meta-population structure can amplify or suppress natural selection

The fragmentation of a population induces a bottleneck that divides the population into smaller sub-populations. This bottleneck accentuates genetic drift within sub-populations, but its effect on natural selection is more subtle. Its effect may depend on the meta-population structure resulting from fragmentation and the gene flow pattern.

Many scientific publications address whether population fragmentation and gene flow change the fixation probability of a mutation compared to a non-fragmented population of the same size. Pollak (1966) focused on a population fragmented into a finite number of demes between which individuals can migrate and showed that symmetric migrations lead to the same fixation probability as in a non-subdivided population.

Whitlock (2003) and Whigham, Dick, and Spencer (2008) challenged this independence of the fixation probability from the meta-population structure resulting from fragmentation. Further works showed that the meta-population structure re-

sulting from fragmentation could either amplify or suppress natural selection (i.e., increase or decrease the efficacy of natural selection, respectively) (Lieberman, Hauert, and Nowak 2005; Houchmandzadeh and Vallade 2011). Amplifying natural selection means reducing the fixation probability of deleterious mutations and increasing that of beneficial ones, whereas suppressing natural selection does the opposite. Importantly, the meta-population structure alone is insufficient to assess the impact of a fragmentation-induced bottleneck on natural selection (i.e., amplifier, suppressor, or without effect) as the gene flow pattern needs to be taken into account (Marrec, Lamberti, and Bitbol 2021). Many theoretical studies assessing the impact of fragmentation on evolutionary dynamics focused on the fixation probability, but other important quantities are impacted, such as the adaptation rates (Hindersin and Traulsen 2014). An experiment in which ciprofloxacin-resistant mutants were tracked in a *Pseudomonas aeruginosa* meta-population showed that for low migration rates, natural selection is amplified in a star topology compared to a well-mixed population (P. P. Chakraborty, Nemzer, and Kassen 2023), thus confirming a theoretical prediction made by (Marrec, Lamberti, and Bitbol 2021).

Prospects for filling knowledge gaps

Human activity fragments populations into several sub-populations (also called demes or islands), which can become isolated if gene flow between them is limited. As biodiversity declines, it is crucial to understand the impact of fragmentation and gene flow on the evolutionary dynamics of bottlenecked populations and, in particular, their adaptive potential. In this review, we have shown that there are many theoretical studies investigating this impact. However, theoretical predictions are rarely directly tested or mostly with microbiology experiments whose design does not allow comparison with mathematical models. A stronger collaboration between theory and experiment (e.g., Marrec, Lamberti, and Bitbol 2021 combined with P. P. Chakraborty, Nemzer, and Kassen 2023) would lead to a better understanding of fragmentation and gene flow on the evolutionary dynamics of meta-populations. Also, more experiments with diploid organisms would enable better comparison with wild populations (e.g., Bakker et al. 2010).

3.3 Relative importance of these processes

3.3.1 Summary of the previous parts

We have reviewed in the previous parts how the different evolutionary processes can be disrupted during a bottleneck and how these processes shape the adaptive potential of bottlenecked populations (see Figure 3.1).

As expected, most of these mechanisms are predicted to decrease the adaptive potential following a bottleneck. Bottlenecks limit adaptation from *de novo* mutations by reducing the mutation supply and the fixation probability of beneficial mutations. More importantly for sexual populations, they also limit adaptation from standing genetic variation. Indeed, molecular variation is decreased due to

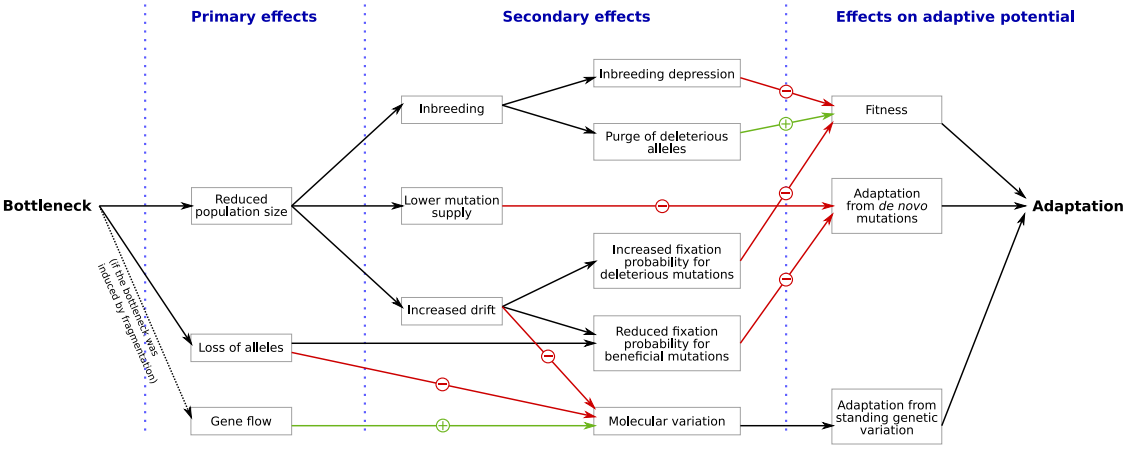


Figure 3.1 – Potential effects of a bottleneck on adaptive potential. Summary of the main effects of bottlenecks on the adaptive potential of a population, as described in this review. The existence and relative importance of these different effects vary across populations, bottleneck characteristics, and environments. Black arrows represent a causal effect, green arrows represent a positive effect, and red arrows represent a negative effect.

smaller population sizes, increased drift, and the loss of rare alleles during bottlenecks. In addition, the fitness of a bottlenecked population may decrease due to the accumulation of deleterious mutations and genetic load in general, which for sexual diploids includes inbreeding depression. A population with lower fitness will struggle to survive future environmental changes that may increase its probability of extinction.

Conversely, only two mechanisms can mitigate the negative impacts of bottlenecks. When the population is part of a meta-population, gene flow can restore some of the lost genetic variation by introducing new variation. In addition, in the case of sexual diploid populations, inbreeding could, under some conditions, facilitate the purge of deleterious alleles and, thus, reduce the masked genetic load.

These findings show that knowledge transfer from microbial to endangered wild populations is possible. A collaboration between microbiology and conservation biology would be fruitful if microbial experiments were adapted to include more characteristics of these endangered populations. For example, one could use (facultative) sexual micro-organisms, such as yeast, or include standing genetic variation in evolution experiments.

3.3.2 Relative importance of each of these processes

Several evolutionary processes come into play when trying to predict the impact of a bottleneck. As some of these processes have opposite effects, a major concern is to estimate the relative importance of these processes to predict the fate of a population. Even when two processes negatively affect the adaptive potential of populations, it may be useful to know which one is predominant to determine the conditions threatening the persistence of wild populations precisely. Determining

these conditions would help identify the key priorities in population management. In the following of this review, we discuss the relative contribution of the evolutionary processes seen above.

Selection vs. genetic drift

One of the major concerns during a bottleneck is the increase of genetic drift. The positive or negative aspect of genetic drift depends on whether the alleles are deleterious or beneficial.

In Section 3.2.2 about genetic drift, we have seen that a bottleneck reduces the fixation probability of beneficial alleles and increases the chance that they are lost compared to a fixed-size population.

The impact of a bottleneck on deleterious alleles is more complex, as inbreeding can facilitate their purge by natural selection. As previously said, little is known about the conditions required for selection to overcome drift. Even if these conditions were known, we would still have to choose if the management priority is to purge deleterious alleles, which requires inbreeding, or to restore genetic diversity, which requires outcrossing. Conservation biology often deals with small endangered wild populations that have already experienced severe bottlenecks. For such populations, the loss of genetic diversity may be a major concern, and the impact of the purge is minor, which has been confirmed in wild populations (Bouzat 2010) (but see Van Heerwaarden et al. 2008). For example, wild populations of elephants in South Africa underwent a severe bottleneck due to widespread hunting. Microsatellite comparisons of current wild populations with museum specimens of this elephant population before the bottleneck confirmed the loss of genetic diversity (Whitehouse and Harley 2001). In such wild populations, the response to selection can be expected to be less effective, as the probability of having beneficial mutations is low (Frankham 2009). A reduced response to selection after a bottleneck has already been highlighted in experimental fish populations. For example, populations of *Heterandria formosa* having undergone a bottleneck during their demographic history showed a 50% slower response to selection for heat tolerance than populations having not undergone a bottleneck (Klerks et al. 2019). An experiment showed that housefly populations that faced a short bottleneck followed by a period of expansion had better fitness and lower genetic load than populations kept at a constant size with a similar expected inbreeding score (Reed and Bryant 2001). Reed and Bryant (2001) concluded that, when managing endangered wild populations, the priority is to act on the cause of decline to promote rapid expansion and avoid inbreeding.

For adaptive polymorphisms such as the Major Histocompatibility Complex (MHC), the predominance of selection over drift probably depends on the duration of the bottleneck. The MHC is a set of polymorphic genes essential to the adaptive immune system of vertebrates and can be particularly affected by bottlenecks. The potential loss of diversity at this locus is of great concern as it is associated with increased disease susceptibility (Sommer 2005). In the meta-analysis (Sutton et al. 2011), the adaptive polymorphism of the MHC was shown to be significantly

reduced after a bottleneck, and even more so than neutral polymorphisms (by 15%). One possible explanation found by Sutton et al. (2011) is that negative frequency-dependent selection is an important force shaping pre-bottleneck Major Histocompatibility Complex diversity, resulting in a high frequency of very rare alleles. As these rare alleles are more at risk of being lost during a bottleneck, this would explain the greater reduction in Major Histocompatibility Complex diversity. The authors concluded that diversifying selection cannot counter genetic drift in recently bottlenecked populations. However, this conclusion can be mitigated as the authors did not find a significant effect on Major Histocompatibility Complex diversity for short-scale bottlenecks. For instance, in a water vole population undergoing a 4-month bottleneck, the Major Histocompatibility Complex diversity was greatly reduced during this period but recovered in a few generations to reach the pre-bottleneck level (Oliver and Piertney 2012).

Loss of heterozygosity vs. loss of genetic variation

Another open question is to identify the mechanism causing the more significant reduction in the adaptive potential of bottlenecked populations between the loss of heterozygosity and the loss of genetic variation. In population genetics theory, heterozygosity determines the evolutionary potential, particularly the short-term response to selection (Falconer 1960). Accordingly, *Drosophila* populations that faced intense or diffuse bottlenecks leading to the same level of heterozygosity showed no difference in their response to selection (England et al. 2003). Whereas both bottleneck regimes were expected to yield different allelic diversities, the measured allelic diversities were not significantly different (England et al. 2003).

On the other hand, Ørsted, Hoffmann, et al. (2019), focusing on *Drosophila* populations having experienced different levels of inbreeding, showed that molecular diversity was more strongly correlated to adaptation than was the expected inbreeding coefficient. This result highlights the importance of molecular diversity for adaptation. It provides a way to summarize the history of a population, which seems more relevant than keeping track of population sizes.

However, to our knowledge, the methods used to restore heterozygosity are the same as those used to restore diversity and consist of outcrossing (i.e., crossing the population with individuals from other populations and/or expanding the population size).

3.4 Thoughts for future research directions

In the future, conservation biology could benefit even more from microbiology by maintaining a close link between the two fields. With a reverse approach, evolution experiments using microorganisms could directly address the conservation biology needs. We make the following suggestions:

1. Include ecological factors within experimental evolution studies.

2. Include selective history when considering the demographic history of populations.
3. Include bottleneck characteristics and demographic history.

3.4.1 Testing the influence of ecological factors on population response

As discussed in this review, most microbiology studies investigating the adaptive potential of bottlenecked populations have taken an evolutionary perspective without considering ecological factors. However, natural populations evolve in interaction with their biotic and abiotic environment. Theoretical predictions could be biased without considering these ecological factors. For instance, Nordstrom et al. (2023) showed through stochastic individual-based simulations that considering population growth with negative density dependence (i.e., intraspecific competition) or density independence leads to different outcomes of evolutionary rescue. This prediction regarding the impact of density dependence vs. independence was empirically tested and confirmed with flour beetles (Olazcuaga et al. in prep.). More precisely, Olazcuaga et al. showed that the effect of negative density dependence varies depending on whether the populations have experienced a bottleneck in their demographic history. Osmond and de Mazancourt (2013) proved with an adaptive dynamic model that interspecific competition can favor evolutionary rescue by increasing the strength of selection and speeding up adaptation. Following Olazcuaga et al.'s example, examining how interspecific competition affects the probability of rescue in bottlenecked populations would be valuable.

3.4.2 Testing the influence of selective and demographic history on population response

In this review, we focused on the effects of bottlenecks, which entail random reductions in population size ("random bottleneck" hereafter), rather than selective bottlenecks, which involve non-random decreases in population size. Wild populations can experience both random and selective bottlenecks. Random bottlenecks can occur due to fragmentation, whereas selective bottlenecks are more likely to occur when adapting to a drastic environmental change, such as pollutants or antibiotic resistance. These selective bottlenecks can result in U-shaped population size curves during evolutionary rescue processes (Gomulkiewicz and Holt 1995). Selective bottlenecks, as random bottlenecks, can impact how populations respond to future stress. A few studies tested how adaptation to a first environmental change, which was associated with a decrease in population size, impacted the response to future adaptation to a new environment (Lachapelle, Colegrave, and Bell 2017; O'Connor, Fugère, and Gonzalez 2020; Samani and Bell 2016; Gonzalez and Bell 2013) using microorganisms: *Chlamydomonas reinhardtii*, *Pseudomonas fluorescens*, *Saccharomyces paradoxus*, and *Saccharomyces spp*, respectively). Adaptation to a new environmental change would be favored for populations that have

already undergone similar stress in their demographic history (Lachapelle, Colegrave, and Bell 2017) (see O'Connor, Fugère, and Gonzalez 2020 for a change in the speed of future adaptation). Conversely, if the stress is different, adaptation would be less likely (Lachapelle, Colegrave, and Bell 2017). This impact of the first dissimilar stress makes sense since the response to the selection of the first stress would reduce genetic variability (Carlson, Cunningham, and Westley 2014). Additionally, populations that have evolved under first stress during their demographic history seem to have a higher probability of extinction when they experience new and different stress (Lachapelle, Colegrave, and Bell 2017; Samani and Bell 2016; Gonzalez and Bell 2013). An increase in genetic load is expected during selective bottleneck (Stewart et al. 2017), which could explain this result. However, whether these deleterious mutations can be purged as efficiently as in a random bottleneck is unclear. Furthermore, the mean frequency of mutations and the genetic load can change in a complex way during a selective bottleneck, in contrast to a random bottleneck (Dussex et al. 2023). Overall, the evidence that adaptive bottlenecks increase the probability of extinction and impact the probability of adaptation suggests that the processes involved differ from those occurring in a random bottleneck or are not as straightforward as expected. The impacts of random bottlenecks *versus* selective bottlenecks have been studied theoretically in infection models and host-pathogen infection processes (as reviewed in Abel et al. 2015, e.g., Moxon and Kussell 2017 and De Ste Croix et al. 2020). However, how a selective bottleneck, compared to a random bottleneck, impacts the response to future stress has never been empirically studied. Integrating demographic and selective history can improve predictions of population response to different stresses.

Finally, it is essential to note that bottlenecks exist on a gradient and cannot be categorized into two binary categories. Selective and random bottlenecks represent the extreme points of this gradient, where the relative contribution of drift and selection varies. Moreover, we have discussed that natural selection can play a role in a random bottleneck process, challenging the assumption that a random bottleneck is entirely random. As a first step, it would be useful to compare the effects of selective and random bottlenecks on the probability of adaptation to future environmental conditions. Then, it would be important to study how the contribution of genetic drift and selection during demographic history influences the response of bottlenecked populations.

3.4.3 Testing the influence of bottleneck characteristics on future population response

In this review, we have focused on the impact of a single bottleneck on population response, except when considering microbial experiments that usually involve multiple bottlenecks. However, the demographic history of natural populations is never restricted to a single bottleneck (Hohenlohe, Funk, and Rajora 2021; Gladstone et al. 2022). Therefore, it is essential to consider the entire demographic history of natural populations, including the intensity and frequency of these bot-

tlenecks.

Microbiology informs us about the impact of the intensity and frequency of bottlenecks, which can be useful to conservation biology. Microbiologists can control the characteristics of the bottleneck, such as its frequency, intensity, and duration (LeClair and Wahl 2018). The impact of bottleneck intensity has been tested in microbial experimental evolution. The adaptive pathways differ depending on whether the bottleneck is weak or intense (Garoff et al. 2020; Vogwill et al. 2016; Mahrt et al. 2021). Overall, empirical studies on the evolution of antibiotic resistance have observed a negative correlation between bottleneck severity and adaptive response (Garoff et al. 2020; Huseby et al. 2017; Mahrt et al. 2021). In addition, Mahrt et al. (2021) showed a decrease in parallel evolution with increasing bottleneck intensity. This result suggests that after experiencing strong bottlenecks, resistance evolves through a wider range of genetic mechanisms, likely due to increased genetic drift. Theoretical studies of bottleneck characteristics suggest that smaller population sizes before or after the bottleneck constrain evolutionary paths, thus limiting the supply of beneficial mutations and adaptation (Gamblin, Gandon, et al. 2023). In addition, Wein and Dagan (2019) pointed out that while bottleneck intensity is a factor in determining population evolvability, selective conditions during evolution can play a more significant role. Mahrt et al. (2021) notably examined the interaction between bottleneck intensity and intensity of selective pressure. The application of the effect of bottleneck intensity to natural populations remains unclear. Olazcuaga et al. (2023) conducted an experiment demonstrating that *Tribolium castaneum* populations responds similarly to environmental change, regardless of the intensity of the bottleneck they experienced in their demographic history. England et al. (2003) also found no difference in adaptive potential between *Drosophila melanogaster* populations that underwent an intense or diffuse bottleneck designed to produce similar inbreeding levels.

The duration for which populations can recover, which is the time between two bottlenecks, also affects the adaptive potential of populations. For instance, Moxon and Kussell (2017) showed that increasing the severity of bottlenecks or reducing the growth period leads to faster adaptation during pathogen microbial infection. Natural microbial populations that experience frequent bottlenecks, such as pathogens, can adapt to changing environmental conditions. A theoretical study predicted a high probability that some mutations acquired during growth in a given host will be passed to the next one in viruses (Sigal et al. 2018). These results could apply to conservation biology since frequent bottlenecks are commonly observed in wild populations (Hohenlohe, Funk, and Rajora 2021). Recent genomic approaches have been used to determine the timing and nature of past population bottlenecks by detecting changes in the shape of the deleterious variation landscape (see Bortoluzzi et al. 2019 for an application with chicken populations as well as Cornuet and Luikart 1996 and Peery et al. 2012 for classical approaches).

An important area for future research is to investigate whether the cumulative effects of multiple bottlenecks are additive or synergistic. This effect could be studied experimentally in microbiology and then applied to natural populations. In theory, multiple bottlenecks will not have the same impact on the population's

ability to adapt from *de novo* mutations and from standing genetic variation. What matters for adaptation from *de novo* mutation is the current population census size, which is the result of the last bottleneck only. What matters for adaptation from standing genetic variation is the genetic diversity of the population, which results from past variations in population size (i.e., from the last common ancestor of the population to the present). Genetic diversity is proportional to the effective population size, which is usually computed as the harmonic mean of past population sizes for populations of varying sizes (Crow and Kimura 2009; Otto and Whitlock 1997; B. Charlesworth 2009). However, other parameters, such as population structure and selection, can also impact the effective population size. As these parameters can vary between different environments, comparisons between experimental and wild populations must be made with caution. Advances in population genomics applied to conservation biology are very useful in this case and are a fruitful avenue of research (Hohenlohe, Funk, and Rajora 2021). Moreover, even if genetic diversity should theoretically correlate with the response to selection, this effect is rarely observed in the wild due to interference from other biological mechanisms (e.g., plasticity or coevolution, see Pujol et al. 2018). We would need experiments with populations undergoing bottlenecks of different severity and duration to assess the differential impacts of such bottlenecks on genetic variation. Indeed, some studies have observed that genetic variation quickly recovers after a short bottleneck, with examples of both short-generation (water vole) and long-generation (white-tailed eagle) species (Oliver and Piertney 2012; L. F. Keller, Jeffery, et al. 2001; Hailer et al. 2006). Moreover, low initial genetic variation does not seem to be limiting for the adaptation of invaders (see the review Bock et al. 2015). Thus, there is likely to be a threshold of severity and duration above which a population struggles to recover.

To conclude, in this section and throughout the review, we have proposed several research directions and suggested new experiments that could help to understand the adaptation of bottlenecked populations. The new knowledge gained from these experiments could ultimately be integrated into existing methods for detecting species most at risk of extinction due to climate change (reviewed in Hoffmann and Sgrò 2011).

3.5 Perspectives

Our review has shown that the impact of bottlenecks on the evolutionary dynamics of populations is a topic that spans several fields, such as microbiology and conservation biology, and has inspired theoretical, and empirical works. Our review emphasizes that these fields share a common goal and are not as distinct as previously thought. We believe that an improved collaboration between these fields will lead to a better understanding of how bottlenecks affect the evolutionary dynamics of populations.

Similar to our review, Alexander et al. (2014) showed how seemingly unrelated fields address the evolution of declining populations. Specifically, Alexander et al.

(2014) emphasized that evolutionary rescue is a research topic in medicine (e.g., drug resistance evolution in patients undergoing chemotherapy) and conservation biology (e.g., survival of species undergoing habitat deterioration). Similar to our review, Alexander et al. (2014) pointed out that integrating different fields could accelerate our scientific knowledge.

We hope that these synthesis reviews will pave the way for empirical studies that combine different fields. Given the challenges of the 21st century, such as the loss of biodiversity, it would be highly valuable to employ approaches that enhance our comprehension of biological processes and our ability to forecast the reaction of natural populations to environmental change.

Authors' contributions

— Jasmine Gamblin: Conceptualization (Equal), Visualization (Lead), Writing - original draft (35% - Lead), Writing - review & editing (Equal).

— Loïc Marrec: Conceptualization (Equal), Writing - original draft (20% - Supporting), Writing - review & editing (Equal).

— Laure Olazcuaga: Project administration (Lead), Supervision (Lead), Conceptualization (Lead), Writing – original draft (45% - Lead), Writing - review & editing (Equal).

Contexte du projet

Ce projet de revue est né après ma rencontre avec Laure Olazcuaga et Loïc Marrec lors du symposium « Limits to adaptation » de l'édition 2022 du congrès de l'European Society for Evolutionary Biology. Ce symposium était co-organisé par Laure, et Loïc et moi y avons présenté nos travaux. Le *Journal of Evolutionary Biology (JEB)* proposant à chaque symposium d'écrire une revue ou une issue spéciale en rapport avec son thème, nous avons eu l'idée de mettre nos différentes connaissances en commun pour écrire cette revue. En effet, Laure a étudié l'adaptation de populations animales suite à un goulot d'étranglement (Olazcuaga et al. 2023), tandis que Loïc et moi avons travaillé sur la modélisation d'expériences d'évolution microbienne incluant des goulots d'étranglement (Gamblin, Gandon et al. 2023 ; Marrec et Bank 2023).

Malheureusement, le manuscrit que nous avons écrit n'était pas encore assez abouti pour être accepté dans *JEB*. Nous allons donc travailler à son amélioration afin de pouvoir le soumettre à un autre journal.

Références du Chapitre 3

Abel, S., P. S. Z. Wiesch, B. M. Davis et M. K. Waldor (2015). Analysis of Bottlenecks in Experimental Models of Infection. *PLOS Pathogens* **11**.

Agrawal, A. F. et M. C. Whitlock (2011). Inferences About the Distribution of Dominance Drawn From Yeast Gene Knockout Data. *Genetics* **187**, 553-566.

Alexander, H. K., G. Martin, O. Y. Martin, O. Martin et S. Bonhoeffer (2014). Evolutionary Rescue: Linking Theory for Conservation and Medicine. *Evolutionary Applications* **7**, 1161-1179.

Allendorf, F. W. (1986). Genetic Drift and the Loss of Alleles versus Heterozygosity. *Zoo Biology* **5**, 181-190.

Ament-Velásquez, S. L., C. Gilchrist, A. Rêgo, D. P. Bendixsen, C. Brice, J. M. Grosse-Sommer, N. Rafati et R. Stelkens (2022). The Dynamics of Adaptation to Stress from Standing Genetic Variation and de Novo Mutations. *Molecular Biology and Evolution* **39**, msac242.

Bakker, J., M. van Rijswijk, F. J. Weissing et R. Bijlsma (2010). Consequences of Fragmentation for the Ability to Adapt to Novel Environments in Experimental Drosophila Metapopulations: ESF-ConGen Meeting on Integrating Population Genetics and Conservation Biology. *Conservation Genetics* **11**, 435-448.

Barnosky, A. D., N. Matzke, S. Tomiya, G. O. U. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, B. Mersey et E. A. Ferrer (2011). Has the Earth's Sixth Mass Extinction Already Arrived? *Nature* **471** (7336), 51-57.

Barrett, R. D. H. et D. Schluter (2008). Adaptation from Standing Genetic Variation. *Trends in Ecology & Evolution* **23**, 38-44.

Barrett, S. C. H. et J. R. Kohn (1991). Genetic and Evolutionary Consequences of Small Population Size in Plants: Implications for Conservation. *Genetics and Conservation of Rare Plants.* Sous la dir. de D. A. Falk et K. E. Holsinger. New York : Oxford University Press, 3-30.

Bell, G., V. Fugère, R. Barrett, B. Beisner, M. Cristescu, G. Fussmann, J. Shapiro et A. Gonzalez (2019). Trophic Structure Modulates Community Rescue Following Acidification. *Proceedings of the Royal Society B: Biological Sciences* **286**, 20190856.

Bell, G. et A. Gonzalez (2011). Adaptation and Evolutionary Rescue in Metapopulations Experiencing Environmental Deterioration. *Science (New York, N.Y.)* **332**, 1327-1330.

Bertorelle, G., F. Raffini, M. Bosse, C. Bortoluzzi, A. Iannucci, E. Trucchi, H. E. Morales et C. van Oosterhout (2022). Genetic Load: Genomic Estimates and Applications in Non-Model Animals. *Nature Reviews Genetics* **23** (8), 492-503.

Boakes, E. H., J. Wang et W. Amos (2006). An Investigation of Inbreeding Depression and Purging in Captive Pedigreed Populations. *Heredity* **98**, 172-182.

Bock, D. G., C. Caseys, R. D. Cousens, M. A. Hahn, S. M. Heredia, S. Hübner, K. G. Turner, K. D. Whitney et L. H. Rieseberg (2015). What We Still Don't Know about Invasion Genetics. *Molecular Ecology* **24**, 2277-2297.

Bortoluzzi, C., M. Bosse, M. F. L. Derks, R. P. M. A. Crooijmans, M. A. M. Groenen et H.-J. Megens (2019). The Type of Bottleneck Matters: Insights into the Deleterious Variation Landscape of Small Managed Populations. *Evolutionary Applications* **13**, 330-341.

Bouzat, J. L. (2010). Conservation Genetics of Population Bottlenecks: The Role of Chance, Selection, and History. *Conservation Genetics* **11**, 463-478.

Burke, M. K. (2023). Embracing Complexity: Yeast Evolution Experiments Featuring Standing Genetic Variation. *Journal of Molecular Evolution*.

Burke, M. K., G. Liti et A. D. Long (2014). Standing Genetic Variation Drives Repeatable Experimental Evolution in Outcrossing Populations of Saccharomyces Cerevisiae. *Molecular Biology and Evolution* **31**, 3228-3239.

Campos, P. R. A. et L. M. Wahl (2010). The Adaptation Rate of Asexuals: Deleterious Mutations, Clonal Interference and Population Bottlenecks. *Evolution* **64**, 1973-1983.

Carlson, S. M., C. J. Cunningham et P. A. Westley (2014). Evolutionary Rescue in a Changing World. *Trends in Ecology & Evolution* **29**, 521-530.

Ceballos, G., P. R. Ehrlich, A. D. Barnosky, A. García, R. M. Pringle et T. M. Palmer (2015). Accelerated Modern Human–Induced Species Losses: Entering the Sixth Mass Extinction. *Science Advances* **1**, e1400253.

Chakraborty, P. P., L. R. Nemzer et R. Kassen (2023). Experimental Evidence That Network Topology Can Accelerate the Spread of Beneficial Mutations. *Evolution Letters* **7**, 447-456.

Chakraborty, R. et M. Nei (1982). Genetic Differentiation of Quantitative Characters between Populations or Species: I. Mutation and Random Genetic Drift. *Genetical Research* **39**, 303-314.

Charlesworth, B. (2009). Effective Population Size and Patterns of Molecular Evolution and Variation. *Nature Reviews Genetics* **10**, 195-205.

Charlesworth, D. et B. Charlesworth (1987). Inbreeding Depression and Its Evolutionary Consequences. *Annual Review of Ecology and Systematics* **18**, 237-268.

Chiucchi, J. E. et H. L. Gibbs (2010). Similarity of Contemporary and Historical Gene Flow among Highly Fragmented Populations of an Endangered Rattlesnake. *Molecular Ecology* **19**, 5345-5358.

Clarke, D. K., E. A. Duarte, A. Moya, S. F. Elena, E. Domingo, E Domingo, J J Holland, J. J. Holland et J. H. Holland (1993). Genetic Bottlenecks and Population Passages Cause Profound Fitness Differences in RNA Viruses. *Journal of Virology* **67**, 222-228.

Clayton, G. et A. Robertson (1955). Mutation and Quantitative Variation. *The American Naturalist*.

Cornuet, J. M. et G. Luikart (1996). Description and Power Analysis of Two Tests for Detecting Recent Population Bottlenecks from Allele Frequency Data. *Genetics* **144**, 2001-2014.

Crnokrak, P. et S. C. H. Barrett (2002). Perspective: Purging the Genetic Load: A Review of the Experimental Evidence. *Evolution* **56**, 2347-2358.

Crnokrak, P. et D. A. Roff (1995). Dominance Variance: Associations with Selection and Fitness. *Heredity* **75** (5), 530-540.

Crow, J. et M. Kimura (2009). *An Introduction to Population Genetics Theory*. Blackburn Press.

Czuppon, P., F. Blanquart, H. Uecker et F. Débarre (2021). The Effect of Habitat Choice on Evolutionary Rescue in Subdivided Populations. *The American Naturalist* **197**, 625-643.

Dayton, J. et P. Szczys (2021). Metapopulation Connectivity Retains Genetic Diversity Following a Historical Bottleneck in a Federally Endangered Seabird. *Ornithological Applications* **123**.

De Ste Croix, M., J. Holmes, J. J. Wanford, E. R. Moxon, M. R. Oggioni et C. D. Bayliss (2020). Selective and Non-Selective Bottlenecks as Drivers of the Evolution of Hypermutable Bacterial Loci. *Molecular Microbiology* **113**, 672-681.

DeRose, M. A. et D. A. Roff (1999). A Comparison of Inbreeding Depression in Life-history and Morphological Traits in Animals. *Evolution* **53**, 1288-1292.

De Visser, J. A. G. M. et D. E. Rozen (2005). Limits to Adaptation in Asexual Populations. *Journal of Evolutionary Biology* **18**, 779-788.

Dussex, N., H. E. Morales, C. Grossen, L. Dalén et C. van Oosterhout (2023). Purging and Accumulation of Genetic Load in Conservation. *Trends in Ecology & Evolution* **0**.

Elena, S. F. et R. E. Lenski (2003). Evolution Experiments with Microorganisms: The Dynamics and Genetic Bases of Adaptation. *Nature Reviews Genetics* **4**, 457-469.

Ellstrand, N. et D. Elam (2003). Population Genetic Consequences of Small Population Size: Implications for Plant Conservation. *Annual Review of Ecology and Systematics* **24**, 217-242.

England, P. R., G. H. R. Osler, L. M. Woodworth, M. E. Montgomery, D. A. Briscoe et R. Frankham (2003). Effects of Intense versus Diffuse Population Bottlenecks on Microsatellite Genetic Diversity and Evolutionary Potential. *Conservation Genetics* **4**, 595-604.

Falconer, D. S. (1960). Introduction to Quantitative Genetics. *Introduction to quantitative genetics.*

Fitzpatrick, S. W. et B. N. Reid (2019). Does Gene Flow Aggravate or Alleviate Maladaptation to Environmental Stress in Small Populations? *Evolutionary Applications* **12**, 1402-1416.

Frankham, R. (2009). Inbreeding in the Wild Really Does Matter. *Heredity* **104**, 124-124.

Frankham, R., J. Ballou, K. Ralls, M. Eldridge, M. Dudash, C. Fenster, R. Lacy et P. Sunnucks (2017). *Genetic Management of Fragmented Animal and Plant Populations*. Oxford University Press.

Frankham, R., D. Briscoe et J. Ballou (2002). *Introduction to Conservation Genetics*. Cambridge University Press.

Frankham, R. (2015). Genetic Rescue of Small Inbred Populations: Meta-Analysis Reveals Large and Consistent Benefits of Gene Flow. *Molecular Ecology* **24**, 2610-2618.

Frankham, R., K. Lees, M. E. Montgomery, P. R. England, E. H. Lowe et D. A. Briscoe (1999). Do Population Size Bottlenecks Reduce Evolutionary Potential. *Animal Conservation* **2**, 255-260.

Franklin, I. R. et R. Frankham (1998). How Large Must Populations Be to Retain Evolutionary Potential? *Animal Conservation* **1**, 69-70.

Fuerst, P. A. et T. Maruyama (1986). Considerations on the Conservation of Alleles and of Genic Heterozygosity in Small Managed Populations. *Zoo Biology* **5**, 171-179.

Gamblin, J., S. Gandon, F. Blanquart et A. Lambert (2023). Bottlenecks Can Constrain and Channel Evolutionary Paths. *Genetics* **224**. Sous la dir. de K. Jain, iyad001.

Garoff, L., F. Pietsch, D. L. Huseby, T. Lilja, G. Brandis et D. Hughes (2020). Population Bottlenecks Strongly Influence the Evolutionary Trajectory to Fluoroquinolone Resistance in Escherichia Coli. *Molecular Biology and Evolution* **37**, 1637-1646.

Gladstone, N. S., N. L. Garrison, T. Lane, P. D. Johnson, J. Garner et N. V. Whelan (2022). Population Genomics Reveal Low Differentiation and Complex Demographic Histories in a Highly Fragmented and Endangered Freshwater Mussel. *Aquatic Conservation: Marine and Freshwater Ecosystems* **32**, 1235-1248.

Gomulkiewicz, R. et R. D. Holt (1995). When Does Evolution by Natural Selection Prevent Extinction? *Evolution; international journal of organic evolution* **49**, 201.

Gonzalez, A. et G. Bell (2013). Evolutionary Rescue and Adaptation to Abrupt Environmental Change Depends upon the History of Stress. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120079.

Goodman, S. J., H. B. Tamate, Rebecca Wilson, R. L. Wilson, J. Nagata, S. Tatsuzawa, G. M. Swanson, J. M. Pemberton et D. R. McCullough (2001). Bottlenecks, Drift and Differentiation: The Population Structure and Demographic History of Sika Deer (Cervus Nippon) in the Japanese Archipelago. *Molecular Ecology* **10**, 1357-1370.

Grossen, C., F. Guillaume, L. F. Keller et D. Croll (2020). Purging of Highly Deleterious Mutations through Severe Bottlenecks in Alpine Ibex. *Nature Communications* **11**, 1001-1001.

Gyllenberg, M. et I. Hanski (1992). Single-Species Metapopulation Dynamics: A Structured Model. *Theoretical Population Biology* **42**, 35-61.

Hailer, F., B. Helander, A. O. Folkestad, S. A. Ganusevich, S. Garstad, P. Hauff, C. Koren, T. Nygård, V. Volke, C. Vilà et H. Ellegren (2006). Bottlenecked but Long-Lived: High Genetic Diversity Retained in White-Tailed Eagles upon Recovery from Population Decline. *Biology Letters* **2**, 316-319.

Haldane, J. B. S. (1927). A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society* **23**, 838-844.

Handel, A. et D. E. Rozen (2009). The Impact of Population Size on the Evolution of Asexual Microbes on Smooth versus Rugged Fitness Landscapes. *BMC Evolutionary Biology* **9**, 236-236.

Hanski, I. et O. Ovaskainen (2003). Metapopulation Theory for Fragmented Landscapes. *Theoretical Population Biology* **64**, 119-127.

Hedrick, P. W. et A. Garcia-Dorado (2016). Understanding Inbreeding Depression, Purging, and Genetic Rescue. *Trends in Ecology & Evolution* **31**, 940-952.

Heffernan, J. M. et L. M. Wahl (2002). The Effects of Genetic Drift in Experimental Evolution. *Theoretical Population Biology* **62**, 349-356.

Heilbron, K., M. Toll-Riera, M. Kojadinovic et R. C. MacLean (2014). Fitness Is Strongly Influenced by Rare Mutations of Large Effect in a Microbial Mutation Accumulation Experiment. *Genetics* **197**, 981-990.

Hindersin, L. et A. Traulsen (2014). Counterintuitive Properties of the Fixation Time in Network-Structured Populations. *Journal of The Royal Society Interface* **11**, 20140606.

Hoffmann, A. A. et C. M. Sgrò (2011). Climate Change and Evolutionary Adaptation. *Nature* **470** (7335), 479-485.

Hoffmann, A. A., C. M. Sgrò et T. N. Kristensen (2017). Revisiting Adaptive Potential, Population Size, and Conservation. *Trends in Ecology and Evolution* **32**, 506-517.

Hohenlohe, P. A., W. C. Funk et O. P. Rajora (2021). Population Genomics for Wildlife Conservation and Management. *Molecular Ecology* **30**, 62-82.

Houchmandzadeh, B. et M. Vallade (2011). The Fixation Probability of a Beneficial Mutation in a Geographically Structured Population. *New Journal of Physics* **13**, 073020.

Hufbauer, R. A., M. Szűcs, E. Kasyon, C. Youngberg, M. J. Koontz, C. Richards, T. Tuff et B. A. Melbourne (2015). Three Types of Rescue Can Avert Extinction in a Changing Environment. *Proceedings of the National Academy of Sciences* **112**, 10557-10562.

Huseby, D. L., F. Pietsch, G. Brandis, L. Garoff, A. Tegehall, A. Tegehall et D. Hughes (2017). Mutation Supply and Relative Fitness Shape the Genotypes of Ciprofloxacin-Resistant Escherichia Coli. *Molecular Biology and Evolution* **34**, 1029-1039.

Ingvarsson, P. K. (2001). Restoration of Genetic Variation Lost – the Genetic Rescue Hypothesis. *Trends in Ecology & Evolution* **16**, 62-63.

Jangjoo, M., S. F. Matter, J. Roland et N. Keyghobadi (2016). Connectivity Rescues Genetic Diversity after a Demographic Bottleneck in a Butterfly Population Network. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 10914-10919.

Kawecki, T. J., R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri et M. C. Whitlock (2012). Experimental Evolution. *Trends in ecology & evolution*.

Keller, L. F., K. J. Jeffery, P. Arcese, M. A. Beaumont, W. M. Hochachka, J. N. M. Smith et M. W. Bruford (2001). Immigration and the Ephemerality of a Natural Population Bottleneck: Evidence from Molecular Markers. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **268**, 1387-1394.

Keller, L. F., Lukas F. Keller et D. M. Waller (2002). Inbreeding Effects in Wild Populations. *Trends in Ecology and Evolution* **17**, 230-241.

Keyghobadi, N. (2007). The Genetic Implications of Habitat Fragmentation for Animals. *Canadian Journal of Zoology* **85**, 1049-1064.

Kimura, M. (1955). Random Genetic Drift in Multi-Allelic Locus. *Evolution* **9**, 419.

Kirkpatrick, M. et P. Jarne (2000). The Effects of a Bottleneck on Inbreeding Depression and the Genetic Load. *The American Naturalist* **155**, 154-167.

Kleinman-Ruiz, D., M. Lucena-Perez, B. Villanueva, J. Fernández, A. P. Saveljev, M. Ratkiewicz, K. Schmidt, N. Galtier, A. García-Dorado et J. A. Godoy (2022). Purging of Deleterious Burden in the Endangered Iberian Lynx. *Proceedings of the National Academy of Sciences* **119**.

Klerks, P. L., G. Athrey, Giridhar N. R. Athrey, G. N. R. Athrey et P. L. Leberg (2019). Response to Selection for Increased Heat Tolerance in a Small Fish Species, With the Response Decreased by a Population Bottleneck. *Frontiers in Ecology and Evolution* **7**.

Lachapelle, J., N. Colegrave et G. Bell (2017). The Effect of Selection History on Extinction Risk during Severe Environmental Change. *Journal of Evolutionary Biology* **30**, 1872-1883.

Lande, R. (1988). Genetics and Demography in Biological Conservation. *Science* **241**, 1455-1460.

Lande, R., S. Engen, B.-E. Sæther et B.-E. Saether (1998). Extinction Times in Finite Metapopulation Models with Stochastic Local Dynamics. *Oikos* **83**, 383.

LeClair, J. S. et L. M. Wahl (2018). The Impact of Population Bottlenecks on Microbial Adaptation. *Journal of Statistical Physics* **172**, 114-125.

Leigh, D. M., H. E. L. Lischer, F. Guillaume, C. Grossen et T. Günther (2021). Disentangling Adaptation from Drift in Bottlenecked and Reintroduced Populations of Alpine Ibex. *Molecular Ecology Resources* **21**, 2350-2363.

Lieberman, E., C. Hauert et M. A. Nowak (2005). Evolutionary Dynamics on Graphs. *Nature* **433**, 312-316.

López-Cortegano, E., E. Moreno et A. García-Dorado (2021). Genetic Purging in Captive Endangered Ungulates with Extremely Low Effective Population Sizes. *Heredity* **127**, 433-442.

Lopez-Fanjul, C. et A. Villaverde (1989). Inbreeding Increases Genetic Variance for Viability in Drosophila Melanogaster. *Evolution* **43**, 1800-1804.

Lynch, M., J. Conery et R. Burger (1995). Mutation Accumulation and the Extinction of Small Populations. *The American Naturalist* **146**, 489-518.

Lynch, M. et W. G. Hill (1986). Phenotypic Evolution by Neutral Mutation. *Evolution* **40**, 915-935.

Mahrt, N., A. Tietze, S. Künzel, S. Franzenburg, C. Barbosa, G. Jansen et H. Schulenburg (2021). Bottleneck Size and Selection Level Reproducibly Impact Evolution of Antibiotic Resistance. *Nature Ecology and Evolution* **5**, 1233-1242.

Marad, D. A., S. W. Buskirk et G. I. Lang (2018). Altered Access to Beneficial Mutations Slows Adaptation and Biases Fixed Mutations in Diploids. *Nature Ecology and Evolution* **2**, 882-889.

Marrec, L. et C. Bank (2023). Evolutionary Rescue in a Fluctuating Environment: Periodic versus Quasi-Periodic Environmental Changes. *Proceedings of the Royal Society B: Biological Sciences* **290**, 20230770.

Marrec, L., I. Lamberti et A.-F. Bitbol (2021). Toward a Universal Model for Spatially Structured Populations. *Physical Review Letters* **127**, 218102.

Marsden, C. D., D. Ortega-Del Vecchyo, D. P. O'Brien, J. F. Taylor, O. Ramirez, C. Vilà, T. Marques-Bonet, R. D. Schnabel, R. K. Wayne et K. E. Lohmueller (2016). Bottlenecks and Selective Sweeps during Domestication Have Increased Deleterious Genetic Variation in Dogs. *Proceedings of the National Academy of Sciences* **113**, 152-157.

McDonald, M. J., D. P. Rice et M. M. Desai (2016). Sex Speeds Adaptation by Altering the Dynamics of Molecular Evolution. *Nature* **531** (7593), 233-236.

Melbourne, B. A. et A. Hastings (2008). Extinction Risk Depends Strongly on Factors Contributing to Stochasticity. *Nature* **454**, 100-103.

Milot, E., H. Weimerskirch, P. Duchesne et L. Bernatchez (2007). Surviving with Low Genetic Diversity: The Case of Albatrosses. *Proceedings of the Royal Society B: Biological Sciences* **274**, 779-787.

Moxon, R. et E. Kussell (2017). The Impact of Bottlenecks on Microbial Survival, Adaptation, and Phenotypic Switching in Host-Pathogen Interactions. *Evolution; international journal of organic evolution* **71**, 2803-2816.

Muller, H. J. (1932). Some Genetic Aspects of Sex. *The American Naturalist* **66**, 118-138.

— (1964). The Relation of Recombination to Mutational Advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1**, 2-9.

Nei, M., T. Maruyama et R. Chakraborty (1975). The Bottleneck Effect and Genetic Variability in Populations. *Evolution; international journal of organic evolution* **29**, 1.

Nordstrom, S. W., R. A. Hufbauer, L. Olazcuaga, L. F. Durkee et B. A. Melbourne (2023). How Density Dependence, Genetic Erosion and the Extinction Vortex Impact Evolutionary Rescue. *Proceedings of the Royal Society B: Biological Sciences* **290**, 20231228.

O'Connor, L. M. J., V. Fugère et A. Gonzalez (2020). Evolutionary Rescue Is Mediated by the History of Selection and Dispersal in Diversifying Metacommunities. *Frontiers in Ecology and Evolution* **8**.

Olazcuaga, L., B. Lincke, S. DeLacey, L. F. Durkee, B. A. Melbourne et R. A. Hufbauer (2023). Population Demographic History and Evolutionary Rescue: Influence of a Bottleneck Event. *Evolutionary Applications* **16**, 1483-1495.

Oliver, M. et S. B. Piertney (2012). Selection Maintains MHC Diversity through a Natural Population Bottleneck. *Molecular Biology and Evolution* **29**, 1713-1720.

Ørsted, M., A. A. Hoffmann, E. Sverrisdóttir, K. L. Nielsen et T. N. Kristensen (2019). Genomic Variation Predicts Adaptive Evolutionary Responses Better than Population Bottleneck History. *PLOS Genetics* **15**.

Ørsted, M., E. Yashiro, A. A. Hoffmann et T. N. Kristensen (2022). Population Bottlenecks Constrain Host Microbiome Diversity and Genetic Variation Impeding Fitness. *PLOS Genetics* **18**, e1010206.

Osmond, M. M. et C. de Mazancourt (2013). How Competition Affects Evolutionary Rescue. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **368**, 20120085.

Otto, S. P. et M. C. Whitlock (1997). The Probability of Fixation in Populations of Changing Size. *Genetics* **146**, 723-733.

Peery, M. Z., R. Kirby, B. N. Reid, R. Stoelting, E. Doucet-Bëer, S. Robinson, C. Vasquez-Carrillo, J. N. Pauli et P. J. Palsboll (2012). Reliability of Genetic Bottleneck Tests for Detecting Recent Population Declines. *Molecular Ecology* **21**, 3403-3418.

Pollak, E. (1966). On the Survival of a Gene in a Subdivided Population. *Journal of Applied Probability* **3**, 142-155.

Przeworski, M., G. Coop et J. D. Wall (2005). The Signature of Positive Selection on Standing Genetic Variation. *Evolution* **59**, 2312-2323.

Pujol, B., S. Blanchet, A. Charmantier, E. Danchin, B. Facon, P. Marrot, F. Roux, I. Scotti, C. Teplitsky, C. E. Thomson et I. Winney (2018). The Missing Response to Selection in the Wild. *Trends in Ecology & Evolution* **33**, 337-346.

Reed, D. H. et E. H. Bryant (2001). Fitness, Genetic Load and Purging in Experimental Populations of the Housefly. *Conservation Genetics* **2**, 57-61.

Roelke, M. E., J. S. Martenson et S. J. O'Brien (1993). The Consequences of Demographic Reduction and Genetic Depletion in the Endangered Florida Panther. *Current Biology* **3**, 340-350.

Roff, D. A. et K. Emerson (2006). Epistasis and Dominance: Evidence for Differential Effects in Life-History Versus Morphological Traits. *Evolution* **60**, 1981-1990.

Rousselle, M., P. Simion, M.-K. Tilak, E. Figuet, B. Nabholz et N. Galtier (2020). Is Adaptation Limited by Mutation? A Timescale-Dependent Effect of Genetic Diversity on the Adaptive Substitution Rate in Animals. *PLOS Genetics* **16**, e1008668.

Sachdeva, H., O. Olusanya et N. Barton (2022). Genetic Load and Extinction in Peripheral Populations: The Roles of Migration, Drift and Demographic Sto-

chasticity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **377**, 20210010.

Samani, P. et G. Bell (2016). The Ghosts of Selection Past Reduces the Probability of Plastic Rescue but Increases the Likelihood of Evolutionary Rescue to Novel Stressors in Experimental Populations of Wild Yeast. *Ecology Letters* **19**. Sous la dir. de T. Coulson, 289-298.

Schenk, M. F., M. P. Zwart, S. Hwang, P. Ruelens, E. Severing, J. Krug et J. A. G. M. de Visser (2022). Population Size Mediates the Contribution of High-Rate and Large-Benefit Mutations to Parallel Evolution. *Nature Ecology and Evolution*.

Sigal, D., Jennifer N.S. Reid, J. N. Reid et L. M. Wahl (2018). Effects of Transmission Bottlenecks on the Diversity of Influenza A Virus. *Genetics* **210**, 1075-1088.

Sommer, S. (2005). The Importance of Immune Gene Variability (MHC) in Evolutionary Ecology and Conservation. *Frontiers in Zoology* **2**, 16.

Soulé, M. (1986). *Conservation Biology: The Science of Scarcity and Diversity*. Oxford University Press, Incorporated.

— (1987). *Viable Populations for Conservation*. Cambridge University Press.

Stewart, G. S., M. R. Morris, A. B. Genis, M. Szűcs, B. A. Melbourne, S. J. Tavener et R. A. Hufbauer (2017). The Power of Evolutionary Rescue Is Constrained by Genetic Load. *Evolutionary Applications* **10**, 731-741.

Sutton, J. T., S. Nakagawa, B. C. Robertson et I. G. Jamieson (2011). Disentangling the Roles of Natural Selection and Genetic Drift in Shaping Variation at MHC Immunity Genes. *Molecular Ecology* **20**, 4408-4420.

Swindell, W. R. et J. L. Bouzat (2005). Modeling the Adaptive Potential of Isolated Populations: Experimental Simulations Using Drosophila. *Evolution* **59**, 2159-2169.

— (2006). Inbreeding Depression and Male Survivorship in Drosophila: Implications for Senescence Theory. *Genetics* **172**, 317-327.

Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123**, 585-595.

Tajima, F. (1996). The Amount of DNA Polymorphism Maintained in a Finite Population When the Neutral Mutation Rate Varies among Sites. *Genetics* **143**, 1457-1465.

Tallmon, D., G. Luikart et R. Waples (2004). The Alluring Simplicity and Complex Reality of Genetic Rescue. *Trends in Ecology & Evolution* **19**, 489-496.

Tomasini, M. et S. Peischl (2020). When Does Gene Flow Facilitate Evolutionary Rescue? *Evolution* **74**, 1640-1653.

Tomasini, M. et S. Peischl (2022). The Role of Spatial Structure in Multi-deme Models of Evolutionary Rescue. *Journal of Evolutionary Biology* **35**, 986-1001.

Uecker, H., S. P. Otto et J. Hermisson (2014). Evolutionary Rescue in Structured Populations. *The American Naturalist* **183**, E17-35.

Van Heerwaarden, B., Y. Willi, T. N. Kristensen et A. A. Hoffmann (2008). Population Bottlenecks Increase Additive Genetic Variance but Do Not Break a Selection Limit in Rain Forest Drosophila. *Genetics* **179**, 2135-2146.

Vázquez-García, I., F. Salinas, J. Li, A. Fischer, B. Barré, J. Hallin, A. Bergström, E. Alonso-Perez, J. Warringer, V. Mustonen et G. Liti (2017). Clonal Heterogeneity Influences the Fate of New Adaptive Mutations. *Cell Reports* **21**, 732-744.

Vogwill, T., R. L. Phillips, D. R. Gifford et R. C. MacLean (2016). Divergent Evolution Peaks under Intermediate Population Bottlenecks during Bacterial Experimental Evolution. *Proceedings of The Royal Society B: Biological Sciences* **283**, 20160749.

Wahl, L. M., P. J. Gerrish et I. Saika-Voivod (2002). Evaluating the Impact of Population Bottlenecks in Experimental Evolution. *Genetics* **162**, 961-971.

Wein, T. et T. Dagan (2019). The Effect of Population Bottleneck Size and Selective Regime on Genetic Diversity and Evolvability in Bacteria. *bioRxiv*, 726158.

Whigham, P. A., G. C. Dick et H. G. Spencer (2008). Genetic Drift on Networks: Ploidy and the Time to Fixation. *Theoretical Population Biology* **74**, 283-290.

Whitehouse, A. M. et E. H. Harley (2001). Post-bottleneck Genetic Diversity of Elephant Populations in South Africa, Revealed Using Microsatellite Analysis. *Molecular Ecology* **10**, 2139-2149.

Whiteley, A. R., S. W. Fitzpatrick, W. C. Funk et D. A. Tallmon (2015). Genetic Rescue to the Rescue. *Trends in Ecology & Evolution* **30**, 42-49.

Whitlock, M. C. (2003). Fixation Probability and Time in Subdivided Populations. *Genetics* **164**, 767-779.

Willi, Y., J. Van Buskirk, B. Schmid et M. Fischer (2007). Genetic Isolation of Fragmented Populations Is Exacerbated by Drift and Selection. *Journal of Evolutionary Biology* **20**, 534-542.

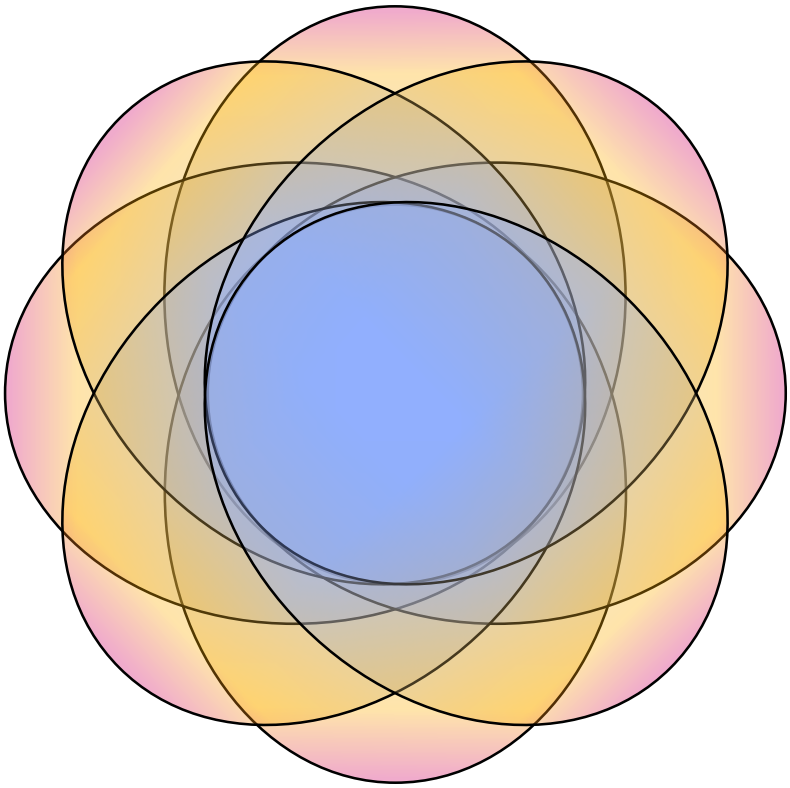
Willi, Y., T. Tregenza, J. Van Buskirk et A. A. Hoffmann (2006). Limits to the Adaptive Potential of Small Populations. *Annual Review of Ecology, Evolution, and Systematics* **37**, 433-458.

Xie, X., Y. Yang, Q. Ren, X. Ding, P. Bao, B. Yan, X. Yan, J. Han, P. Yan et Q. Qiu (2018). Accumulation of Deleterious Mutations in the Domestic Yak Genome. *Animal Genetics* **49**, 384-392.

Yu, W., B. Wu, X. Wang, Z. Yao, Y. Li et Y. Liu (2020). Scale-Dependent Effects of Habitat Fragmentation on the Genetic Diversity of Actinidia Chinensis Populations in China. *Horticulture Research* **7**.

Évolution des pangénomes bactériens

Chapitre 4



4

Persistent, Private and Mobile genes: a model for gene dynamics in evolving bacterial pangenomes

Sommaire

4.1 Introduction

Many bacterial species present an impressive diversity in terms of gene content. Due to pervasive intra- and inter-specific horizontal gene transfer (HGT), the number of genes present in a given species is often much higher than the number of genes contained in a typical genome of this species, and can potentially be very large. As multiple genomes from the same species began to be available around the year 2005, the observation that each new sequenced genome contained new genes lead Tettelin, Masignani, et al. (2005) to coin the term 'pangenome' for the set of all genes present in a species (Tettelin and Medini 2020). Genes from a pangenome are usually classified as 'core' (if they are present in every genome) or 'accessory' (if they are not) (Page et al. 2015). The diversity and organization of bacterial genes could be used to understand the mechanisms governing their evolutionary dynamics. Understanding the dynamics of accessory genes is also of interest for public health as most genes conferring higher virulence or antimicrobial resistance belong to this category.

4.1.1 Patterns and processes of pangenome evolution

In the genomic era, very rich information is available to study the patterns created by pangenome evolution: tens of thousands of sequenced genomes in certain bacterial species, including hundreds of completely assembled genomes. These patterns can be found in various summaries of the data. To name a few, the mechanisms of pangenome evolution probably influences the distribution of gene frequencies, the presence/absence patterns of genes, the diversity carried by gene families and the order of genes along a chromosome. For example, the gene frequency spectrum (GFS) - which is the histogram of gene frequencies in a set of genomes - often presents a 'U' shape that seems ubiquitous in bacterial pangenomes, whereby more genes are rare and very frequent than present at intermediate frequencies (Figure 4.5a). While these patterns can be precisely described and compared between different populations (Cummins et al. 2022; Botelho et al. 2023), an open problem is to identify the main processes generating them. To this end, we seek to develop a minimal model describing gene dynamics in a pangenome, that is able to reproduce the observed patterns. We start with three biologically relevant types of evolutionary gene dynamics, partially based on previous models of pangenome evolution, and tested several versions of our model to retain the one that was best able to reproduce these patterns.

4.1.2 Existing models of bacterial pangenome evolution

Most models of bacterial gene evolution study gene gains and losses along a reference species tree, usually obtained using the alignment of core genome sequences. A simple model is the classical Markovian model of binary character evolution along a phylogenetic tree, in which a gene can be gained ($0 \to 1$) and lost ($1 \to 0$) at constant rates along the tree lineages (Pagel 1997). Cohen and Pupko (2010) used

this model to detect horizontally transferred gene families, by allowing the rates to vary across gene families and using stochastic mapping to map gain and loss events on branches. Using a similar model that allowed rate variation across time and lineages, and considering small genome chunks rather than genes as the evolutionary units, Didelot, Darling, and Falush (2008) inferred variations of genomic flux across lineages in several bacterial species and correlated it to their lifestyle. This model class was later called the Finitely Many Genes (FMG) model by Zamani-Dahaj et al. (2016). Indeed, this model assumes that a finite pool of genes is available to the population, which is fixed from the beginning of the species' history, and remains constant throughout evolution.

Around 2012, alternative models emerged to account for gene immigration from other species. The Infinitely Many Genes (IMG) model (Baumdicker, Hess, and Pfaffelhuber 2012) assumes that new genes are imported at a constant rate along the phylogeny, then are lost at a constant rate and cannot be regained once lost. In this model, imported genes are picked from an infinite pool. Haegeman and Weitz (2012) presented a similar model with a fixed genome size, thus imposing that a gene importation must coincide with a gene loss. These models were originally designed to explain the U-shaped gene frequency spectrum (GFS) of pangenomes, and were fitted using exclusively the GFS.

Zamani-Dahaj et al. (2016) compared these two approaches and estimated the proportion of genes that were better explained by one gain (IMG model) versus multiple gains (FMG model), finding that both types of gene dynamics are present in 40 genomes of cyanobacteria. The two models are now implemented in the software Panaroo to allow for gain and loss rates estimation (Tonkin-Hill et al. 2020).

There are several limitations to the aforementioned approaches. First, some assumptions of existing models are limiting: the IMG model does not account for intra-species gene transfers, while the FMG model does not account for inter-species transfers. When comparing both types of gene dynamics, Zamani-Dahaj et al. (2016) find that under the FMG model, 15% of genes have patterns that are better explained by multiple gains, suggesting that they have been transferred. These genes are typically not well fitted by the IMG model. Meanwhile, the FMG model assumes that all genes from the pangenome are already present in the gene pool of the species from the root of the phylogeny. In reality, many genes are imported into the focal species over its evolutionary history, sometimes very recently compared to the whole evolutionary history.

We also identify methodological limitations. On the one hand, studies aiming at fitting the GFS use a very restricted fraction of the available data (the gene frequencies). As a result, they are not able to reproduce correctly the level of parsimony observed in gene presence/absence patterns (see Supplementary Figures 4.15b and 4.15c). On the other hand, while Zamani-Dahaj et al. (2016) reproduce the GFS of the 40 cyanobacteria species thanks to several types of gene dynamics and allowing variability in parameters, the best model has five categories of genes. With much larger genomic datasets available, it is necessary to define parsimonious models that lend themselves more easily to interpretation. In this direction, we

suggest a middle ground approach. We fix a priori the number of categories and parameters based on biological observations of gene dynamics, and check how well we are able to reproduce some of the global patterns observed in the data to decide whether more complex models are needed.

4.1.3 Persistent, Private and Mobile gene dynamics

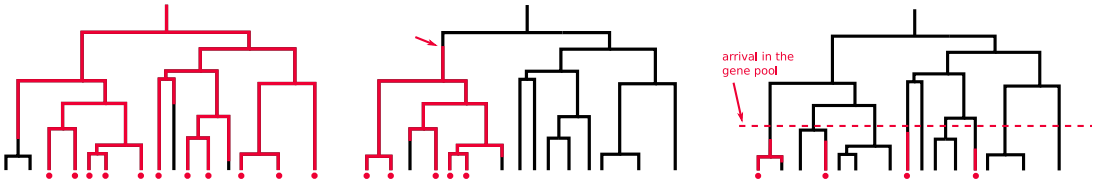
In the following, we list three qualitatively distinct gene behaviors that are relevant for bacterial populations. Our goal is to design a model of pangenome evolution accounting for these three types of gene dynamics.

First, we model the dynamics of genes that are in most circumstances essential to the bacteria, and thus are present in the ancestor and in the majority of sequenced genomes (Figure 4.1a). We use the adjective 'persistent' for these genes, as we expect them to largely overlap the set of genes usually designated as persistent (i.e., genes being present in a certain percentage of genomes, typically between 95 and 100%: Perrin and Rocha 2021). We hypothesize that the inferred loss rates of these genes will be very small. A loss can happen for example when a taxon colonizes a new niche where this gene is no longer essential, or by gene redundancy following the acquisition of a new gene performing the same function.

Second, we model genes that are specific to a given clade (Figure 4.1b). This specificity can have several explanations: a single event of transfer into the focal species, adaptation to a particular environment (e.g. bovine-adapted *Escherichia coli* lineages: Arimizu et al. 2019), defense against a phage specific to this clade (e.g. in *Vibrio crassostreae*: Piel et al. 2021), or epistatic interactions with genes already present in this clade (Whelan, R. J. Hall, and McInerney 2021; McInerney 2022; Beavan, Domingo-Sananes, and McInerney 2024). We call such genes 'private'. Some of these genes are already present in the ancestor of the population, while others appear along the species tree either by *de novo* gene birth or by inter-species transfer.

Finally, we model the dynamics of 'mobile' genes, i.e. genes that undergo frequent intra-species HGT such as transposons, prophages or genes located on plasmids. These genes have highly non-monophyletic presence/absence patterns at the leaves of the species tree (Figure 4.1c). A mobile gene is first introduced into the gene pool of the population by inter-species transfer, which is a rare event, and then is able to spread in the population through frequent intra-species HGT (Tenaillon et al. 2010). For example, the $bla_{\text{CTX-M}}$ gene family encodes CTX-M β-lactamases allowing their host to tolerate β-lactam antibiotics. Originally found as a chromosomal gene in genus *Kluyvera*, it has introgressed several times in other *Enterobacteriaceae* species and has rapidly spread within these species on various mobile genetic elements (D'Andrea et al. 2013).

In this paper we introduce the PPM model for pangenome evolution. It includes three generic types of evolutionary gene dynamics along a species phylogeny, designed to model the aforementioned behaviors: (i) Persistent genes are present in the ancestral genome and can only be lost, (ii) Private genes are gained once by some ancestral lineage, are lost sometimes and undergo no further transfers, and



(a) A Persistent gene **(b)** A Private gene **(c)** A Mobile gene

Figure 4.1 – The three behaviors captured by our model: (a) Persistent genes are present at the root of the phylogeny and rarely lost (b) Private genes arrive once in the phylogeny, transmit vertically and are occasionally lost, while (c) Mobile genes arrive once in the gene pool of the population, then undergo intra-species HGT (thus multiple gains and losses along the phylogeny).

(iii) Mobile genes undergo intra-species transfers following their arrival in the gene pool. We inferred the parameters of our model on a dataset of 902 *Salmonella enterica* genomes by fitting the presence/absence patterns of genes. Using these parameter values, we simulated genomes evolving according to our model and compared the obtained patterns to the ones from observed data. By assigning each gene to its most likely category, we were able to study the distribution of the three categories along the chromosome and on plasmids, as well as the gene functions present in each category.

4.2 The model

In the Persistent-Private-Mobile (PPM) model, genes evolve on the species phylogeny along which they can be gained and lost. The following events are accounted for: gene gain and loss, inter- and intra-specific horizontal gene transfer, sequencing and bioinformatics errors. For the sake of concision, we index the three gene categories of the model as follows: 0 for the Persistent category, 1 for the Private category and 2 for the Mobile category. Parameters referring to each of these categories are indexed accordingly.

— N_0 Persistent genes are present in the ancestral genome. They are lost at constant rate l_0 along the phylogeny.

— N_1 Private genes are present in the ancestral genome. At constant rate i_1, new Private genes arrive on branches of the phylogeny. Once they are present, they are lost at constant rate l_1 along the phylogeny. In the following, we fixed N_1 to be equal to the expected number of Private genes at stationarity, which is i_1/l_1 (Huson and Steel 2004). This category follows exactly the IMG model (Baumdicker, Hess, and Pfaffelhuber 2012).

— Mobile genes enter the gene pool of the population at constant rate i_2 across time/along the tree height. Once they are in the gene pool, Mobile genes can be gained and lost at constant rates g_2 and l_2 by all lineages of the population phylogeny below their arrival date. Gains are interpreted as transfers from the pool to a tree lineage. The fact that we only model this type of transfers

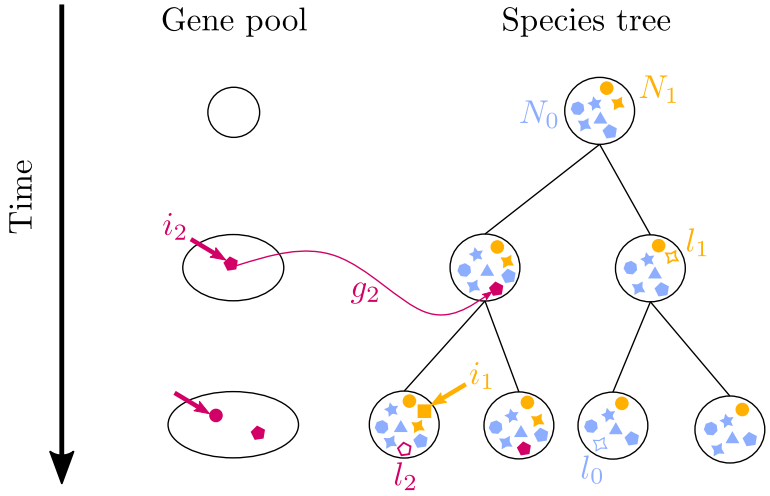


Figure 4.2 – Schematic description of genomes evolving according to the PPM model for pangenome evolution. Persistent genes are represented in blue, Private genes in orange and Mobile genes in Burgundy. Different shapes represent different gene functions. The following events are represented, for Persistent genes: loss (rate l_0); for Private genes: immigration on the tree (rate i_1), loss (rate l_1); and for Mobile genes: immigration into the gene pool (rate i_2), gain (rate g_2) and loss (rate l_2).

> (rather than transfers between lineages) is consistent with the predominance of so-called 'transfers from the dead': assuming that the number of sequenced genomes is small compared to the total number of individuals in the species (which is usually the case), a transferred gene has much more chance to come from a non-sampled ('dead') lineage than from another lineage of the reconstructed tree (Szöllosi, Tannier, Lartillot, et al. 2013).

These events are represented in Figure 4.2.

Lastly, our model accounts for errors in the presence/absence patterns of genes at the leaves of the phylogeny. Pangenome datasets can be subject to several sources of errors (Tonkin-Hill et al. 2020). These sources include sequencing errors (Salzberg 2019) and bioinformatics errors (Denton et al. 2014). Moreover, the genuine biological process of rapid gain and loss events of genes on the leaves of the phylogeny could be accounted for by our 'error' term. For example if a bacterium having very recently lost an essential gene is sequenced, it would probably not have left descent over several generations of replication *in vivo* (which is the case of bacteria at internal nodes of the phylogeny). This illustrates that gain and loss events at the leaves are happening at an evolutionary scale which is different from that of the rest of the phylogeny. Our model contains three error parameters ϵ_0, ϵ_1 and ϵ_2. ϵ_0 and ϵ_1 are the false negative rates for Persistent and Private genes, that is the probability that a gene was lost just before sampling or was not detected because of incorrect sequencing or annotation. ϵ_2 is both the false positive and false negative rate for Mobile genes. A false positive would be a gene gained just before sampling or spuriously detected. We assume that the false positive rate is zero for Persistent and Private genes, as these genes can be gained only once.

4.3 Results

4.3.1 Simulation study

In order to assess the accuracy of our inference procedure (described in the Methods Section 4.5), we simulated sets of presence/absence patterns and checked if we were able to infer the parameters that generated them. We randomly picked a 200-leaf tree, and 100 sets of values for parameters $N_0, l_0, i_1, l_1, i_2, g_2, l_2, \epsilon_0, \epsilon_1$ and ϵ_2. For each set of values we simulated genes evolving along the tree, and used the resulting presence/absence patterns to compute the maximum likelihood parameters. More details on this simulation study are given in the Supplementary Information, Section 4.6.2. Figure 4.3 show the relations between the true parameters (i.e. those used for simulations) and the inferred ones.

Most parameters show good correlations, with relative errors below 10% (or log-scale error below 0.1). In particular this is true for parameters governing the number of genes in each category: N_0, i_1 and i_2. The error rates of the Private and Mobile categories (ϵ_1 and ϵ_2) show higher relative errors (128% and 76%). The gain and loss rates for the Mobile category are not very well estimated with log-scale errors of 0.52 and 0.21, meaning they can be estimated up to a factor of respectively 3 and 1.5. This is due to very poor inference when l_2 is more than 100 times bigger than g_2 (Figure 4.3h). The average proportion of genes correctly assigned to their category is 97% across the 100 simulated sets (Figure 4.3l).

4.3.2 Maximum likelihood estimates of parameters

We inferred the Maximum Likelihood parameters of our model on a dataset of 902 *Salmonella enterica* genomes, and found the following estimates:

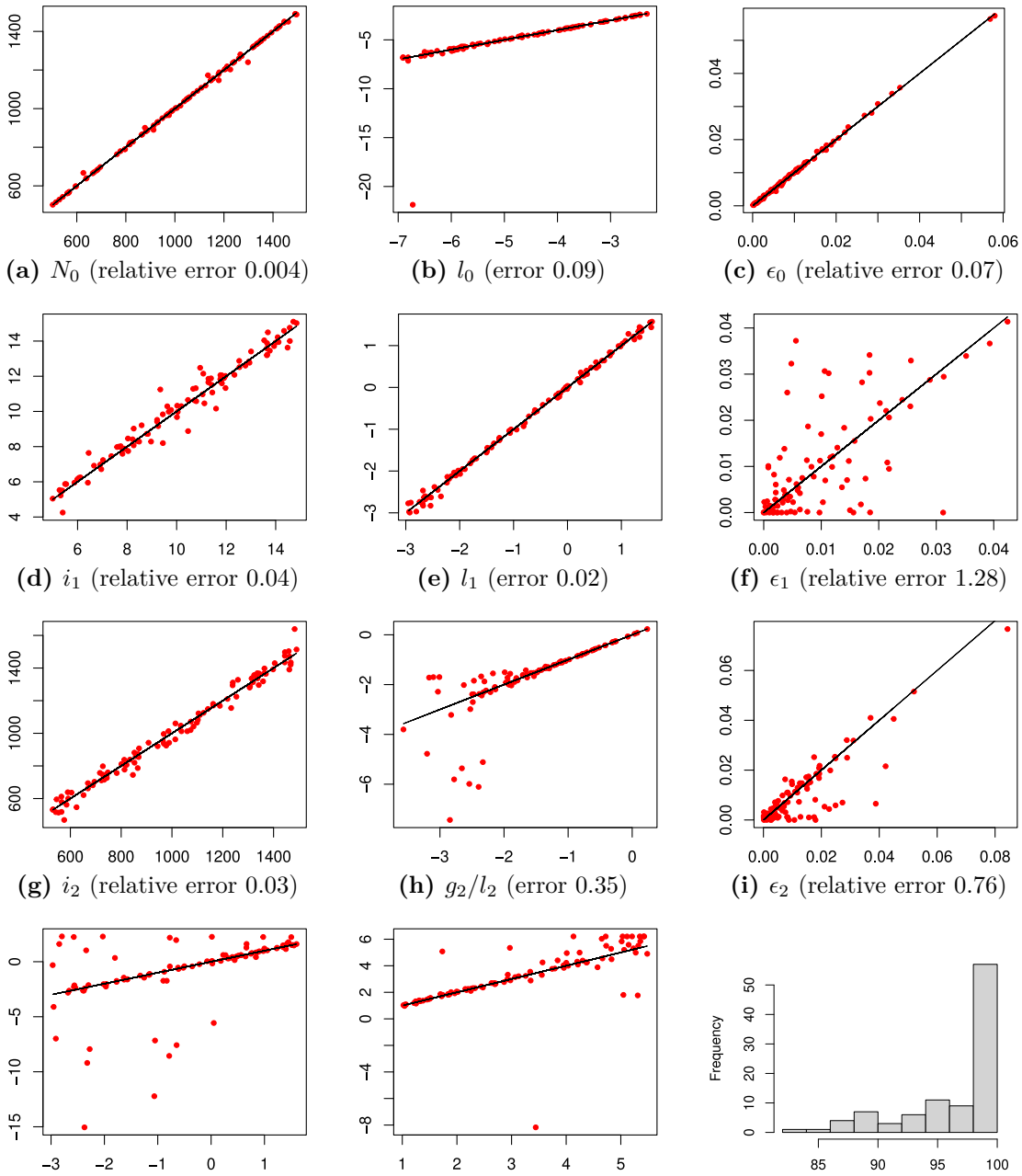
$$\left.\begin{array}{lll} \hat{N}_0 & = 4,014 & \text{CI: } [3894, 4134] \\ \hat{l}_0 & = 2.25 & \text{CI: } [2.18, 2.32] \\ \hat{\epsilon}_0 & = 0.00213 & \text{CI: } [0.00208, 0.00218] \end{array}\right\} \begin{array}{l} 4,007 \text{ observed} \\ \text{Persistent genes} \end{array} \tag{4.1}$$

$$\left.\begin{array}{lll} \hat{i}_1 & = 45,548 & \text{CI: } [45150, 45946] \\ \hat{l}_1 & = 152.4 & \text{CI: } [150.8, 154.0] \\ \hat{\epsilon}_1 & = 0.0188 & \text{CI: } [0.0182, 0.0194] \end{array}\right\} \begin{array}{l} 26,797 \text{ observed} \\ \text{Private genes} \end{array} \tag{4.2}$$

$$\left.\begin{array}{lll} \hat{i}_2 & = 462,793 & \text{CI: } [448832, 475463] \\ \hat{g}_2 & = 222.8 & \text{CI: } [220.0, 225.6] \\ \hat{l}_2 & = 12,858 & \text{CI: } [12717, 12999] \\ \hat{\epsilon}_2 & = 0.00225 & \text{CI: } [0.00220, 0.00230] \end{array}\right\} \begin{array}{l} 14,951 \text{ observed} \\ \text{Mobile genes} \end{array} \tag{4.3}$$

From these estimates, we computed several quantities of interest. The predicted number of genes in the ancestral genome is $\hat{N}_0 + \hat{i}_1/\hat{l}_1 = 4,313$ (CI: $[4187, 4439]$), close to the mean number of genes per genome in the dataset (4,556).

The expected number of genes of each category in the observed pangenome is computed using the forthcoming set of equations (4.6), and we find an expected



(a) N_0 (relative error 0.004) **(b)** l_0 (error 0.09) **(c)** ϵ_0 (relative error 0.07)

(d) i_1 (relative error 0.04) **(e)** l_1 (error 0.02) **(f)** ϵ_1 (relative error 1.28)

(g) i_2 (relative error 0.03) **(h)** g_2/l_2 (error 0.35) **(i)** ϵ_2 (relative error 0.76)

(j) g_2 (error 0.52) **(k)** l_2 (error 0.21) **(l)** Proportion of genes assigned to the correct category

Figure 4.3 – Results of the cross-validation study on a 200-leaf tree. (a-k) show correlations between true values (x-axis) and point estimates (y-axis) of parameters used to simulate sets of presence/absence patterns. Correlations for gain and loss parameters ($l_0, l_1, g_2/l_2, g_2$ and l_2) are plotted on a log scale. The legends of subplots display the mean relative error for each parameter (or mean error for log-scale plots). (l) is a histogram showing the proportion of patterns that were assigned to the correct category by our method across the 100 simulated sets.

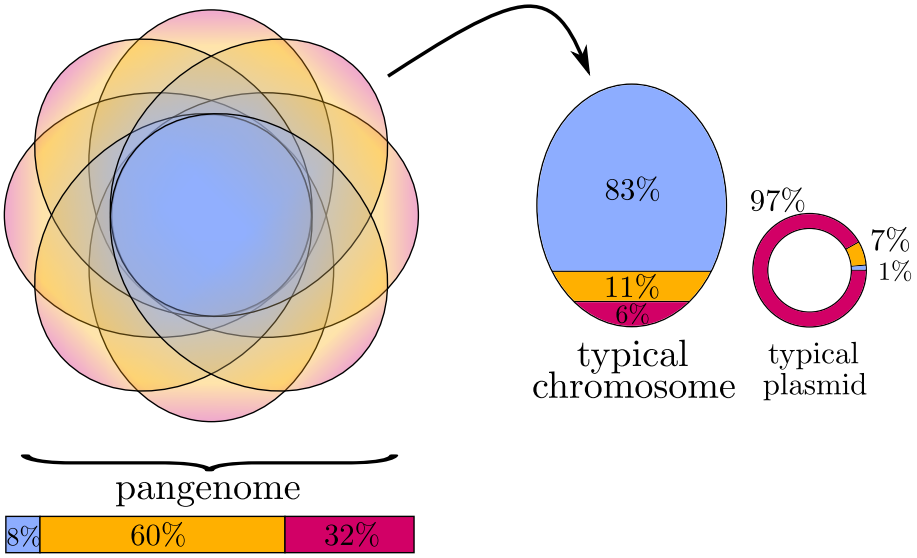


Figure 4.4 – Proportion of the Persistent, Private and Mobile categories in our *Salmonella enterica* dataset as inferred by our model. Colors indicate gene categories: blue for Persistent, orange for Private and Burgundy for Mobile. The observed pangenome is composed of 8% of Persistent genes, 60% of Private genes and 32% of Mobile genes. A typical chromosome carries 83% of Persistent genes, 11% of Private genes and 6% of Mobile genes, while a typical plasmid carries 1% of Persistent genes, 7% of Private genes and 97% of Mobile genes.

number of 4,007 Persistent genes, 26,797 Private genes and 14,951 Mobile genes. This means that the observed pangenome is composed of 8% of Persistent genes, 60% of Private genes and 32% of Mobile genes, as shown on Figure 4.4. We also assigned each gene to its most likely category, and computed the proportion of each category present in an average chromosome and in an average plasmid (Figure 4.4). As Persistent genes are typically present in many genomes, they represent a small proportion of the observed pangenome (8%) but a high proportion of the genes carried by a chromosome (83%). The opposite holds for Private and Mobile genes: they compose a small proportion of the chromosome (11% and 6%, respectively) but the majority of the pangenome. A typical genome from our dataset carries 4,556 genes, of which 3701 are Persistent, 504 are Private and 352 are Mobile.

The loss rates are ordered as we expected, increasing from Persistent to Private to Mobile gene categories (although from the simulation study we know that the absolute value of l_2 may not be very well inferred). The values can be compared with the average branch length of the tree, which is 5×10^{-4} substitution per site. This length corresponds to of the order of a few thousand years, if we divide by the estimated mutation rate of 10^{-7} substitution per site per year for *Salmonella enterica* (Duchêne et al. 2016). During this time interval, Persistent, Private and Mobile genes have a probability 0.001, 0.07 and 0.98 to be lost respectively. Additionally, a Mobile gene has a probability 0.02 to be gained if it was absent. All these probabilities were computed using the forthcoming transition matrix (4.15).

During these thousand years, a given genome is expected to gain 21 Private

genes that may result from *de novo* gene birth and direct transfers from other species, and 722 genes from the pool (if we assume that we are in the middle of the tree and the pool has around $i_2 \times H/2$ genes, where H is the tree height). The large probability of loss and large number of new gains of Mobile genes over thousand years implies that Mobile genes completely turnover over that period of time. By construction, a Private gene is introduced only once in the phylogeny. On the other hand, Mobile genes can be introduced several times from the gene pool to lineages of the phylogeny. We computed the expected number of distinct introductions for a Mobile gene using forthcoming equation (4.23), and found that they are introduced 25 times on average.

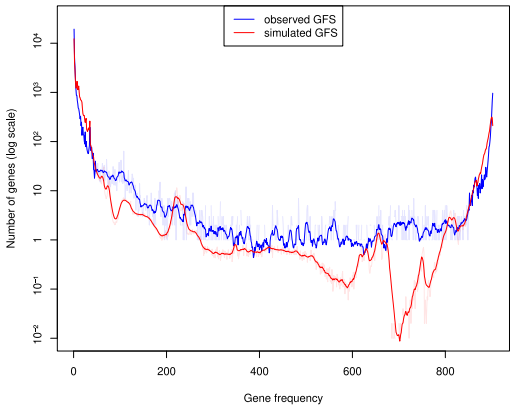
The error rate is quite small for the Persistent and Mobile categories (around 0.2%), and more important for the Private category (around 2%). This may indicate that the presence/absence patterns at the leaves are primarily driven by the underlying processes and not by random error.

4.3.3 Goodness of fit on two summary statistics: the GFS and parsimony plot

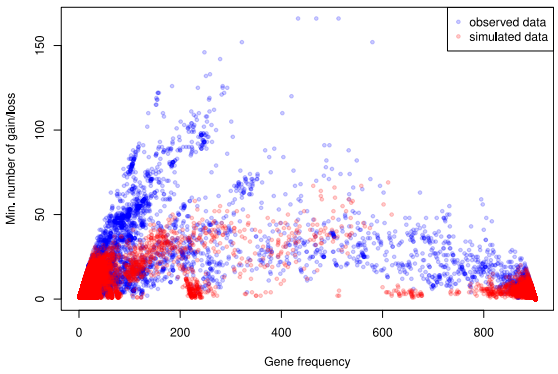
While we fitted our model using the presence/absence matrix, we verified that it is also able to reproduce two important patterns of summary statistics characterizing the observed data. To that end, we simulated a pangenome using our model with the maximum likelihood parameter values inferred on the *Salmonella enterica* dataset, and compared patterns from observed data versus simulated data.

The first pattern is the gene frequency spectrum (GFS), i.e. the distribution of gene frequencies in the observed pangenome. We chose the GFS as a summary statistic because it was used to fit pangenome evolution models in several previous studies. We plotted on Figure 4.5a the observed GFS (in blue) as well as the simulated GFS (in red). While the two curves do not perfectly correspond, in particular with a deficit of medium-high frequency genes in our model, the model qualitatively reproduces the U-shape found in the data.

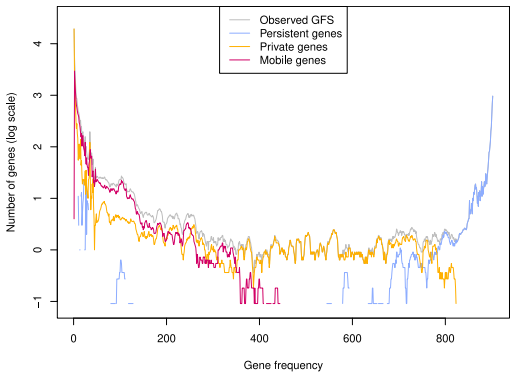
The second pattern is the parsimony plot, which shows the parsimony of gene patterns as a function of their frequency. The y-axis is the minimal number of gains and losses needed along the species tree to explain a gene's presence/absence pattern, thus small values correspond to very parsimonious patterns. We represented on Figure 4.5b the observed parsimony plot in blue and the simulated one in red. Our model reproduces the qualitative patterns of this plot, but the presence/absence patterns it generates are more parsimonious than in the observed data. In particular, the maximum score reached by simulated genes is around 50 while some observed genes go as high as 150. For a more detailed inspection of this parsimony vs. frequency plot, we plotted separate plots for each category on Supplementary Figure 4.13.



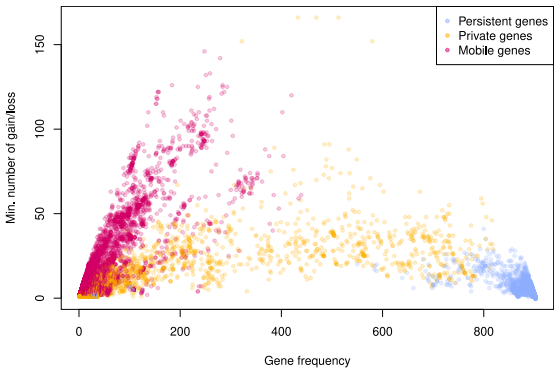
(a) Fit of the gene frequency spectrum



(b) Fit of the parsimony vs. frequency plot



(c) Gene frequency spectra of the three categories



(d) Parsimony vs. frequency plot, colored according to assigned categories

Figure 4.5 – Two multivariate summary statistics of pangenomic data that we use in order to assess the goodness-of-fit of our model, here plotted for a *Salmonella enterica* dataset of 902 genomes. (a) and (c) show the histogram of gene frequencies, called the Gene Frequency Spectrum (GFS). Curves are smoothed for better visualization; raw curves are shown in transparency on (a). (b) and (d) show the parsimony vs. frequency plot: the y-axis is the minimum number of gain and loss events needed along the species tree to explain the presence/absence pattern of a gene with frequency given on x-axis. (a) and (b) show the fit of these two statistics with our model. Observed data are plotted in blue, while simulated data are in red. Simulated data were obtained by simulating a set of genes evolving according to our model with Maximum Likelihood parameter estimates. The simulated GFS is an average over 100 simulations. (c) and (d) show the gene frequency spectrum and parsimony vs. frequency plot colored according to assigned categories.

4.3.4 Inferred categories

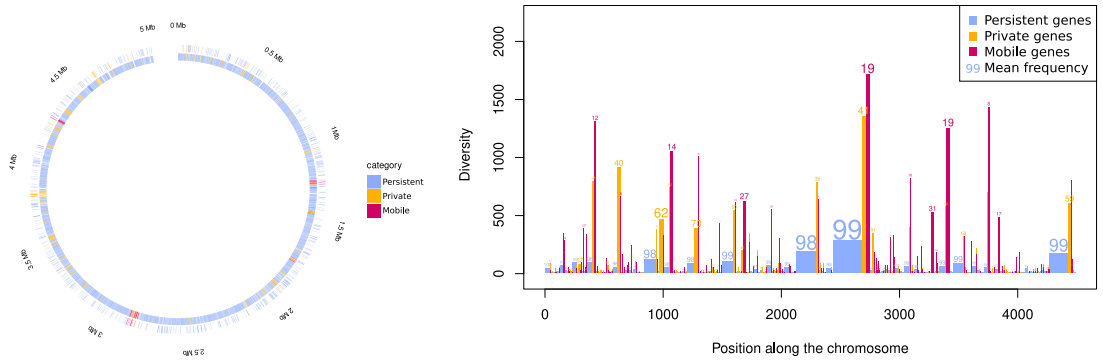
Using the ML parameter estimates, we computed for each gene the category with highest posterior probability. We assigned each gene to a category–Persistent, Private or Mobile. We then compared different characteristics of the three categories, such as frequency, parsimony, enrichment in certain gene functions and position on the genome. The ternary plot on Supplementary Figure 4.11 show that the assigned categories are clear-cut, with genes clustered around the corners of the triangle.
Figures 4.5c and 4.5d show the GFS and parsimony plot colored according to assigned categories. Genes assigned to the Persistent category have typically very high frequency, i.e. they are positioned on the right of these plots. Surprisingly, we also spot a few genes at low frequency that are assigned to the Persistent category. Taking a closer look at their presence/absence patterns revealed that these genes are present at high frequency in two or more clades. As a consequence, they are not well-fitted by the Private or the Mobile category. Genes assigned to the Private category present a wide range of frequencies, from very low (the majority of singletons is 'Private') to intermediate and high frequencies. While we could have expected singletons to be assigned preferentially to the Mobile category or to be encompassed by the error term, it is not the case as Mobile genes have a tendency to be introduced several times (and thus, to be present in several genomes) and the error term is constrained by other patterns. Genes assigned to the Mobile category have low to intermediate frequencies and represent the majority of genes having low parsimony (i.e., in the upper part of the parsimony plot).

Gene position

We studied the spatial distribution of the three gene categories on the bacterial chromosome and on plasmids. On Figure 4.4 is represented the relative proportion of the three categories on a typical chromosome versus on a typical plasmid from the *Salmonella enterica* dataset. As expected, we observe that chromosomes carry a majority of Persistent genes (83%), with some Private and Mobile genes (11% and 6%, respectively). On the other hand, plasmids carry almost exclusively Mobile genes (97%), which is consistent with their easy transfer.

We looked more closely as the spatial distribution of genes along a chromosome, starting arbitrarily with the first genome of the dataset, represented on Figure 4.6a. Gene are colored according to their assigned category (Persistent in blue, Private in orange, Mobile in Burgundy). In accordance with the relative category proportions described above, we observe a majority of Persistent genes on the chromosome. Moreover, Private and Mobile genes are not uniformly distributed but appear clustered.

To investigate whether the structure in gene category along the chromosome and the hotspots of private and mobile genes are conserved across strains, we used the 770 core genes as reference positions. We identified 641 genomes from the *Salmonella enterica* dataset that had the exact same ordering of core genes. For



(a) Chromosome of genome n°1 **(b)** Diversity of gene categories along the chromosome

Figure 4.6 – Distribution of gene categories along the chromosome. (a) shows gene categories along the chromosome of the first genome (chosen arbitrarily) of the *Salmonella enterica* dataset. (b) shows the diversity of gene categories along the *Salmonella enterica* chromosome, computed over 641 genomes sharing the same ordering of core genes. The position along the chromosome refers to the order of genes. Each bar represents genes from a given category present in a given interval between core genes. The width of the bar is the mean number of genes of this category present in this interval, and the height is the number of unique genes of this category present in this interval. The number above the bar is the mean frequency of these genes.

each genome and each interval between two core genes, we computed the number of genes of each category that were present in this interval. We illustrate the diversity and distribution of gene categories along the chromosome of these 641 genomes with conserved synteny (Figure 4.6b). Hotspots of private and mobile genes that are conserved across the genomes appear all along the chromosome. These hotspots include from one gene to a few dozen genes with low frequency, but across all genomes carry a substantial gene diversity. Notably, the six most diverse hotspots of mobile genes each carry more than 1000 unique mobile genes across the 641 genomes. Mean diversity of Private genes in an interval is strongly correlated with mean diversity of Mobile genes in the same interval (Pearson correlation coefficient 0.91). The correlation is weaker for the mean number of genes of each of these categories in a given interval (0.58). The concentration of accessory genes in a few hotspots is in agreement with a study from Oliveira, Touchon, Cury, et al. (2017), where the authors find that genes imported by HGT are concentrated in hotspots representing around 1% of the genome in a dataset comprising 80 bacterial species.

Gene functions

In order to study the gene functions present in each category, we selected genes that were annotated with a Cluster of Orthologous Genes (COG) database function (Galperin et al. 2021). A total of 2,743 Persistent genes, 5,613 Private genes and 2,638 Mobile genes were annotated with a function, representing respectively 70%, 21% and 18% of these categories. From this we built an association table between assigned gene category and COG function, illustrated on Figure 4.7. Overall,

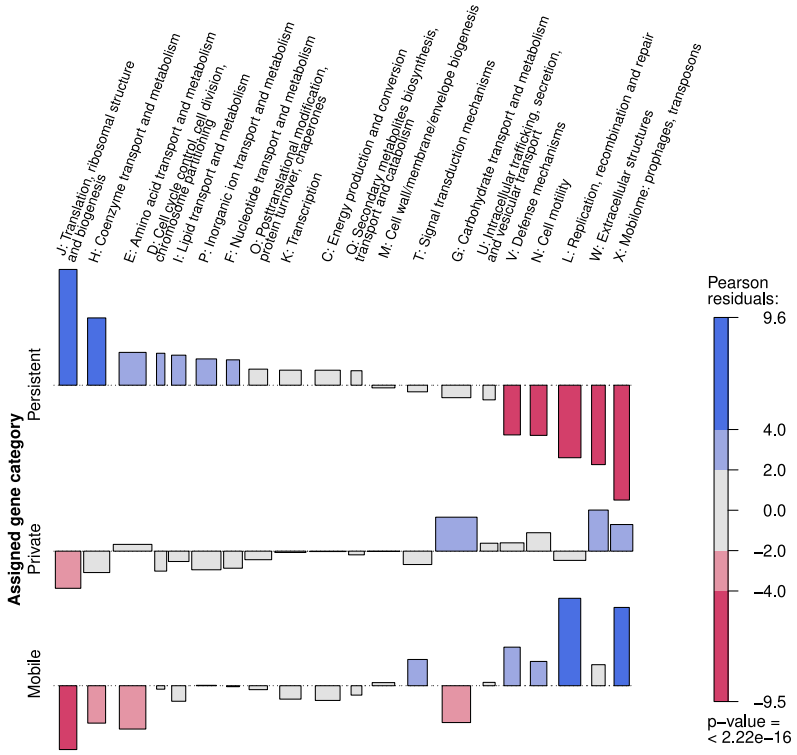


Figure 4.7 – Association between inferred gene categories (Persistent, Private, Mobile) and COG functions. Rectangle heights represent the Pearson residuals, i.e. the observed number of genes corresponding to a given pair of category and function, minus the expected number of such genes if category and function were independent, divided by the square root of the expected value. Rectangle widths are proportional to the square root of expected values, so that rectangles areas are proportional to the difference between observed and expected value. A positive residual indicates a positive association between a gene category and a gene function, while a negative one indicates the opposite.

there is a highly significant association between gene category and gene function ($p << 10^{-6}$).

In particular, there is a positive association between the Persistent category and key cellular functions such as translation (J), cell cycle control (D) and the transport and metabolism of coenzymes (H), amino acids (E), lipids (I), inorganic ions (P) and nucleotides (F). On the contrary, the Private and Mobile categories have nearly opposite associations to the Persistent genes. They are both positively associated with the mobilome (X). The Private category is also associated with carbohydrate transport and metabolism (G) and extracellular structures (W). In addition, the Mobile category is positively associated with replication, recombination and repair (L), defense mechanisms (V), signal transduction mechanisms (T) and cell motility (N). The fact that many genes involved in 'replication, recombination and repair' are inferred to be Mobile genes is probably due to transposases (enzymes helping moving transposons) annotated as recombinases.

4.4 Discussion

In this paper, we introduce the "Persistent, Private, Mobile" model for bacterial pangenome dynamics. This model proposes to classify bacterial genes into three qualitatively different dynamics on a phylogenetic tree, accounts for both inter- and intra- species gene transfer, and is scalable to large datasets. The model fitted to a dataset of 902 *Salmonella enterica* genomes reproduces the U-shaped gene frequency spectrum and the parsimony plot. Moreover, the category assignment is clear-cut and the behaviors that we intended to capture in the three gene classes are consistent with the inferred gain and loss parameters, as well as with the function and position of genes in the genome. This model could therefore be used for dynamics-aware gene classification.

4.4.1 A model for dynamics-aware category assignment

The PPM model was intended to model three gene categories, having different behavior and characteristics. The Persistent category was designed for essential/housekeeping genes present at high frequency. Indeed, we find that genes assigned to this category have a low loss rate and a high mean frequency (854). They constitute the majority of genes present on the chromosome. Moreover, this category is enriched in essential functions such as translation, cell cycle control and different metabolisms. The Private category was designed for accessory genes that are specific to a given clade. Genes assigned to this category have an intermediate loss rate and a small mean frequency (17). Interestingly, they constitute the majority of the observed pangenome, mainly because singletons are classified as 'Private' and therefore represent 70% of this category. The 'Private' category is enriched in genes linked to carbohydrate transport and metabolism, and to extracellular structures. These two associations could be explained by a lasting clade-specific adaptation to different nutrients and environments. Studies performed on *Salmonella enterica* have shown that different serovars display specific carbohydrate metabolism pathways (Seif et al. 2018) and adhesive mechanisms (e.g. pili and adhesins, see Wagner and Hensel 2011). This category is also enriched in genes from the mobilome (transposons, prophages), as might be expected for accessory genes. The Mobile category was designed for accessory genes undergoing frequent intra-species HGT. Genes assigned to this category have a high loss rate and a small mean frequency (22). They represent the overwhelming majority of genes found on plasmids. This category is positively associated with gene functions that are known to be highly transferred: prophages, transposons, transposases, defense mechanisms. It is also associated with signal transduction mechanisms and cell motility.

Overall, our classification identifies persistent genes, genes adapted to certain clades and highly mobile genes. This classification cannot be obtained solely based on the gene frequency, even when additionally considering parsimony (Figures 4.5c and 4.5d). For example, Private and Mobile genes can have similar frequency and parsimony (Figure 4.5d), but are actually organized differently on the species phy-

logeny. The strength of our classification of bacterial genes is that it is informed by patterns and relies on processes. By looking at the gene presence/absence patterns, it uses more information than classifications based on frequency. In addition, the fact that it uses the gene's inferred dynamics to classify it allows for a straightforward interpretation. Comparatively, the classification done by the tool PPanGGOLiN also uses the presence/absence matrix, but relies only implicitly on processes by assuming that neighboring genes have more chance to share the same category (Gautreau et al. 2020).

We envision several other applications of the PPM model. Once the classification is done, it can be used to detect insertion hotspots for accessory genes along the chromosome, as shown on Figures 4.6a and 4.6b. A preliminary study shows that the model would be able to infer a date of introduction for each Mobile gene, if applied to set of genomes from a recently emerged clade (see Supplementary Information Section 4.6.3 for details on arrival time inference). This would allow us to follow the temporal construction of the pangenome of the clade by progressive addition of genes.

4.4.2 Limitations

Our model has several limitations. First, it does not account for duplications. Although the method we use to generate gene families allows them to have multiple members per genome, we ignore duplicated genes by taking a binarized version of the presence/absence matrix Thus, duplicated genes are treated as one gene even if they are sometimes present in multiple copies. That being said, families containing at least one duplicate represent 8% of all families, and those containing at least two duplicates are only 3% of all families.

Second, we do not model the simultaneous transfer of multiples genes. Kloub et al. (2021) designed a method called HoMer to detect potential horizontal multi-gene transfers (HMGTs). In a dataset of 103 *Aeromonas* genomes, they detected that 8.5% of potential intra-species gene transfers were contained in an HMGT event (this percentage was much higher for inter-species transfers, of which 20% were estimated to be part of a HMGT). We expect this phenomenon to be particular relevant for Mobile genes, which are the majority of genes carried by plasmids. HMGT would be evident in our data as sets of mobile genes that are in strong linkage disequilibrium and therefore share similar frequency and similar parsimony.

Third, we do not explicitly include selection acting on gene presence/absence, although the interpretation of the dynamics of genes categories can involve selection: Persistent genes are under purifying selection, Private genes can interact epistatically with the genetic background of the clade, Mobile genes can be favorable in some environmental conditions but not in others

Fourth, we chose to formulate a parsimonious model with a small number of gene categories to interpret the classification resulting from the model in the light of genome organization, and gene function. Some developments of the PPM model could help increase the goodness-of-fit. In particular, in our data many genes have low to moderate frequency and non-parsimonious patterns (with more than 50

gain/loss events) and many genes have medium-large frequency (carried by around 600-800 genomes), features that our model can hardly reproduce (Figures 4.5a and 4.5b). One solution would be to split the Private and Mobile categories into subcategories with different rates. But if we look more closely at patterns that are not well fitted, we find that some of them are genes that are at high frequency in several clades, or at high frequency in one clade and low frequency in the rest of the tree. These could be better fitted if Private genes were allowed to be gained several times, or if genes were allowed to change categories over time: Private genes could become Persistent at some point (i.e., lower their loss rate) or Mobile could become Private in certain clades (i.e., be lost forever when they are lost, but at a lower rate). Allowing genes to change category would also be a way to ensure stationarity in the gene number of each category; in the current version of the model Persistent genes can only decrease in number through time and Mobile genes can only increase in number.

Last, the importance of gene transfers in bacteria render two questions inevitable when working with bacterial phylogenies: how reliable are phylogenetic reconstructions, and what do they represent? Studies have shown that a bacterial species tree can be reliably inferred even in the presence of gene transfers and conversions (in *Escherichia coli*: Touchon, Hoede, et al. 2009), and that phylogenetic distance is well correlated with gene content (in *Acinetobacter*: Touchon, Cury, et al. 2014, in Archea: Wolf et al. 2016). In our *Salmonella enterica* dataset, we evidence a clear phylogenetic signal in the gene patterns by comparing the parsimony of observed patterns to that of patterns simulated without tree (Supplementary Figure 4.10). Phylogenetic trees are usually supposed to represent parentage relationships between individuals, or clonal relationship in the case of bacteria. A different view has recently emerged, suggesting that reconstructed bacterial trees reflect more the structure of recombination inside the population than actual clonal relationships (Sakoparnig, Field, and van Nimwegen 2021). While the present study does not bring new arguments supporting either of these views, we believe that the PPM model remains relevant even in the latter one. In this view, the fact that Private genes are restricted to a given clade would not be a consequence of vertical inheritance (i.e., no transfers) but of transfers remaining inside a certain subpopulation, represented by a clade in the core tree. However, the way to properly address the question of incongruent gene histories lies in the field of gene tree/species tree reconciliation methods. These methods rely on models of genome evolution that usually allow for gene duplication, transfer and loss and are either likelihood-based (Szöllosi, Boussau, et al. 2012; Szöllosi, Tannier, Daubin, et al. 2015) or parsimony-based (Jacox et al. 2016; Bansal et al. 2018). They are able to exploit the signal contained in the genes' molecular sequences and to reconstruct their whole history, but are very computationally intensive and thus not adapted for comparative pangenomics or gene classification on big datasets.

In conclusion, the PPM model classifies gene dynamics into three qualitatively different behaviors, infers relevant rates, and identifies hotspots of gene exchanges along the genome. Applied to the large number of bacterial whole-genome sequences, it could significantly improve our understanding of bacterial pangenome

dynamics and evolution.

4.5 Methods

4.5.1 *Salmonella enterica* dataset

We applied our model to a dataset of 902 *Salmonella enterica* genomes and one *Salmonella bongori* as outgroup, downloaded from the Refseq database. We chose to work with complete genome assemblies, as draft assemblies can cause an artificial increase of the gene number due to fragmented genes (Denton et al. 2014). We thus selected all complete *Salmonella enterica* genomes that were available in this database, which yielded 1374 genomes (genomes downloaded the 07/09/2023). We performed the pangenome analysis using the PanACoTA pipeline (Perrin and Rocha 2021). After quality filtering 903 genomes were kept (with a maximal distance of 0.1 between 2 genomes). 46,146 gene families were detected. The persistent genome, that we defined as gene families having exactly one member in at least 99% of genomes, is composed of 2,887 families. The species tree was reconstructed using sequences from the persistent genome with IQTree. We used an ultrametric version of this tree obtained with LSD2 (To et al. 2016), which is shown on Supplementary Figure 4.9. This tree has a height of $H = 0.03$ and a total branch length of $L = 0.84$ substitution per site.

The data that we use in order to fit our model is the reconstructed species tree, as well as the binary presence/absence matrix describing which gene is present in which genome. As the gene clustering step of the pipeline allows gene families to have more than one member per genome, the presence/absence matrix is in fact non-binary. However, our model does not account for gene duplication, thus we use the binarized version of this matrix.

4.5.2 Likelihood computation

In the following, we describe how to compute the likelihood of the data according to our model. Recall that the Persistent, Private and Mobile categories are indexed by 0, 1 and 2, respectively. Let us first introduce some variables. The number of genes in the pangenome of the bacterial population described by the model is

$$N = N_0 + N_1 + I_1 + I_2 \tag{4.4}$$

where $I_1 \sim Pois(i_1 L)$ and $I_2 \sim Pois(i_2 H)$. L is the total length of the species tree $\mathcal{T}$, H is its height. Remember that we chose to fix $N_1 = i_1/l_1$, which is the mean number of Private genes in a genome at equilibrium (Huson and Steel 2004). The number of genes observed at the leaves is:

$$N^{obs} = N_0^{obs} + N_1^{obs} + I_1^{obs} + I_2^{obs} \tag{4.5}$$

where

$$\begin{aligned}
\mathbb{E}[N_0^{obs}] &= p_0^{obs}(root)N_0 \\
\mathbb{E}[N_1^{obs}] &= p_1^{obs}(root)\frac{i_1}{l_1} \\
\mathbb{E}[I_1^{obs}] &= \int_{\mathcal{T}} i_1 p_1^{obs}(x)\,\mathrm{d}x \\
\mathbb{E}[I_2^{obs}] &= \int_{t=0}^{H} i_2 p_2^{obs}(t)\,\mathrm{d}t
\end{aligned} \tag{4.6}$$

with $p_0^{obs}(root)$ the probability for a Persistent gene to be observed a the leaves, $p_1^{obs}(x)$ the probability for a Private gene that arrived at position x on the tree to be observed at the leaves, and $p_2^{obs}(t)$ the probability for a Mobile gene that arrived in the population at time t to be observed at the leaves. It can be noted that the integrals used to compute $\mathbb{E}[I_1^{obs}]$ and $\mathbb{E}[I_2^{obs}]$ are not of the same nature. In the first one, we integrate over every possible position in the tree, while in the second one we integrate over time. In the following, we will approximate $N_0^{obs}, N_1^{obs}, I_1^{obs}$ and I_2^{obs} by their expectation.

Although immigrant genes from the Private category (resp. from the Mobile category) share the same realization of the random variable I_1 (resp. I_2), we make the simplifying assumption that all genes are independent and compute a pseudo-likelihood of observed data:

$$\mathcal{L}(\text{data};\Theta) = \prod_{r_g \text{ observed}} \mathbb{P}_\Theta(r_g|g \text{ observed})^{\#r_g \text{ observations}} \tag{4.7}$$

where $\Theta = (N_0, i_1, l_1, i_2, l_2, g_2, \epsilon_0, \epsilon_1, \epsilon_2)$ and r_g is the pattern of presence/absence of gene g. Then for each observed pattern we have 4 components to compute. The four components correspond to the three categories, with the Private category having two components (for genes that were present in the ancestral genome and those that were not).

$$\begin{aligned}
&\mathbb{P}_\Theta(r_g|g \text{ observed}) = \\
&\quad \mathbb{P}_\Theta(\text{cat}_g = 0|g \text{ observed})\mathbb{P}_\Theta(r_g|\text{cat}_g = 0, g \text{ observed}) \\
&\qquad + \mathbb{P}_\Theta(\text{cat}_g = 1, g \in root|g \text{ observed})\mathbb{P}_\Theta(r_g|\text{cat}_g = 1, g \in root, g \text{ observed}) \\
&\qquad + \mathbb{P}_\Theta(\text{cat}_g = 1, g \notin root|g \text{ observed})\mathbb{P}_\Theta(r_g|\text{cat}_g = 1, g \notin root, g \text{ observed}) \\
&\qquad + \mathbb{P}_\Theta(\text{cat}_g = 2|g \text{ observed})\mathbb{P}_\Theta(r_g|\text{cat}_g = 2, g \text{ observed})
\end{aligned} \tag{4.8}$$

Likelihood for Persistent genes

For a Persistent gene g, we denote by $\mathbb{P}^{(0)}_{\text{root}}(r_g)$ the probability to observe presence/absence pattern r_g given that gene g was present at the root (which is always the case for Persistent genes). The (0) exponent is used to indicate that this probability is computed for a Persistent gene, i.e. a gene from category 0. This quantity only depends on the loss rate l_0 and the false negative rate ϵ_0, and can be computed using Felsenstein's pruning algorithm (Felsenstein 1981). Formally, we have:

$$\mathbb{P}^{(0)}_{\text{root}}(r_g) := \mathbb{P}_\Theta(r_g|\text{cat}_g = 0) \tag{4.9}$$

Thus:

$$\begin{aligned}
&\mathbb{P}_\Theta(\text{cat}_g = 0 | g \text{ observed}) \mathbb{P}_\Theta(r_g | \text{cat}_g = 0, g \text{ observed}) \\
&= \frac{N_0^{obs}}{N^{obs}} \times \mathbb{P}^{(0)}_{root}(r_g | \text{ obs}) \\
&= \frac{N_0 p_0^{obs}(root)}{N^{obs}} \times \frac{\mathbb{P}^{(0)}_{root}(r_g)}{p_0^{obs}(root)} \\
&= \frac{N_0}{N^{obs}} \times \mathbb{P}^{(0)}_{\text{root}}(r_g)
\end{aligned} \tag{4.10}$$

The first equality is obtained by writing the definition of each of the two factors. The second equality is obtained using Equation (4.6) and the fact that we approximate N_0^{obs} by its expectation.

Likelihood for Private genes

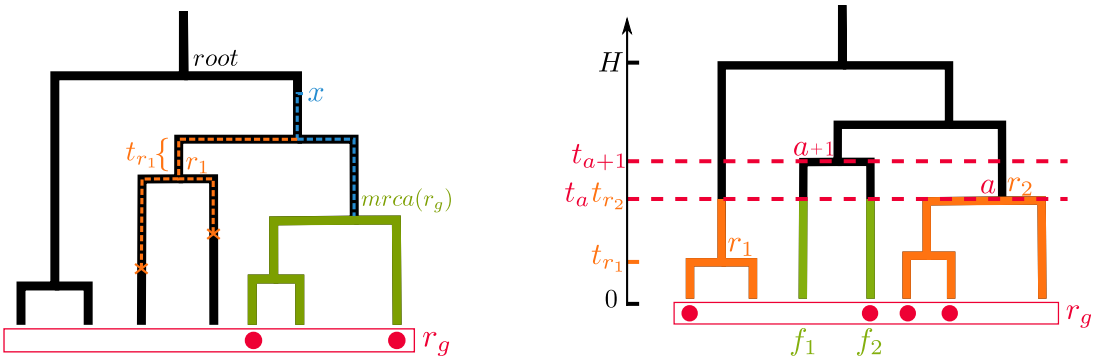
For Private genes, the term corresponding to ancestral genes is similar to the one for Persistent genes:

$$\begin{aligned}
&\mathbb{P}_\Theta(\text{cat}_g = 1, g \in root | g \text{ observed}) \mathbb{P}_\Theta(r_g | \text{cat}_g = 1, g \in root, g \text{ observed}) \\
&\qquad = \frac{i_1}{l_1 N^{obs}} \mathbb{P}^{(1)}_{\text{root}}(r_g)
\end{aligned} \tag{4.11}$$

For an imported Private gene g, we have to integrate over all possible tree positions where this gene could have appeared. These possible positions are located on the path between the most recent common ancestor of the leaves carrying this gene (denoted by $mrca(r_g)$) and the root of the tree. This is due to the fact that Private genes can be gained only once.

$$\begin{aligned}
&\mathbb{P}_\Theta(r_g | \text{cat}_g = 1, g \notin root) \\
&= \frac{1}{i_1 L} \int_{mrca(r_g)}^{root} \mathbb{P}^{(1)}_x(r_g)\, i_1 \mathrm{d}x \\
&= \frac{1}{L} \int_{mrca(r_g)}^{root} p^{(1)}_{1\to 1}(x, mrca(r_g)) q_1(x, mrca(r_g)) \mathbb{P}^{(1)}_{mrca(r_g)}(r_g)\, \mathrm{d}x \\
&= \frac{1}{L} \mathbb{P}^{(1)}_{mrca(r_g)}(r_g) \int_{mrca(r_g)}^{root} e^{-l_1(x - mrca(r_g))} \prod_{\substack{T_i \text{ subtree between} \\ x \text{ and } mrca(r_g) \\ \text{with root } r_i}} \left((1 - e^{-l_1 t_{r_i}}) + e^{-l_1 t_{r_i}} \mathbb{P}^{(1)}_{r_i}\left(0|_{T_i}\right) \right) \mathrm{d}x
\end{aligned} \tag{4.12}$$

Here, the first equality integrates over all possible positions for the appearance of the focal gene on the tree, i.e. on the path between $mrca(r_g)$ and $root$. L is the total tree length, and $\mathbb{P}^{(1)}_x(r_g)$ is the probability to observe pattern r_g knowing that gene g was present at position x in the tree.
The second equality decomposes $\mathbb{P}^{(1)}_x(r_g)$ into the probability that g survive between its arrival at position x and $mrca(r_g)$ (denoted by $p^{(1)}_{1\to 1}(x, mrca(r_g))$), times the probability that g was lost in all subtrees branching on the path between x and $mrca(r_g)$ (denoted by $q_1(x, mrca(r_g))$), times the probability to observe pattern r_g



(a) Likelihood for an imported Private gene

(b) Likelihood of a Mobile gene

Figure 4.8 – Schematic explanation of how to compute the likelihood of imported genes. (a) For imported Private genes we have to compute the probability to observe pattern r_g given that g was present at $mrca(r_g)$ (in green), times the integral over x of the probability to survive between x and $mrca(r_g)$ (in blue) times the probability to be lost in all trees branching between x and $mrca(r_g)$ (in orange). See Equation (4.12). (b) For Mobile genes we have to sum over all time intervals of the tree (i.e., intervals between 2 nodes) the product of the probabilities to observe pattern r_g restricted to subtrees created by cutting the tree at this interval (in orange), times the product of probabilities to observe pattern r_g restricted to single leaves (in green). See Equation (4.16).

given that g was present at $mrca(r_g)$. See Figure 4.8a for a graphical explanation. The third equality is obtained by factoring out $\mathbb{P}^{(1)}_{mrca(r_g)}(r_g)$ and writing the expressions for $p_{1\to 1}$ and q_1. $q_1(x, mrca(r_g))$ is the product for each subtree T_i branching between x and $mrca(r_g)$ of the probability that g is lost before the root r_i of T_i, plus the probability that g survive until r_i and that the null pattern (denoted by 0) is observed at the leaves of T_i.

It follows that:

$$
\begin{aligned}
&\mathbb{P}_\Theta(\text{cat}_g = 1, g \notin root | g \text{ observed}) \mathbb{P}_\Theta(r_g | \text{cat}_g = 1, g \notin root, g \text{ observed}) \\
&= \frac{I_1^{obs}}{N^{obs}} \times \mathbb{P}_\Theta(r_g | \text{cat}_g = 1, g \notin root, g \text{ observed}) \\
&= \frac{\int_{\mathcal{T}} i_1 p_1^{obs}(x)\, \mathrm{d}x}{N^{obs}} \times \frac{\mathbb{P}_\Theta(r_g | \text{cat}_g = 1, g \notin root)}{\frac{1}{L}\int_{\mathcal{T}} p_1^{obs}(x)\, \mathrm{d}x} \\
&= \frac{i_1}{N^{obs}} \mathbb{P}^{(1)}_{mrca(r_g)}(r_g) \int_{mrca(r_g)}^{root} e^{-l_1(x - mrca(r_g))} \prod_{\substack{T_i \text{ subtree between} \\ x \text{ and } mrca(r_g) \\ \text{with root } r_i}} \left((1 - e^{-l_1 t_{r_i}}) + e^{-l_1 t_{r_i}} \mathbb{P}^{(1)}_{r_i}\left(0|_{T_i}\right) \right) \mathrm{d}x
\end{aligned}
\tag{4.13}
$$

To obtain the second equality, we used the approximation of I_1^{obs} by its expectation as computed in Equation (4.6), and the fact that $\mathbb{P}_\Theta(g \text{ observed} | \text{cat}_g = 1, g \notin root) = \frac{1}{L}\int_{\mathcal{T}} p_1^{obs}(x)\, \mathrm{d}x$.

Likelihood for Mobile genes

For a Mobile gene, we have to integrate over the tree height for possible appearance times in the gene pool of the population. Once the gene is present in the gene pool, it evolves following a 2-state Markov model (0: absent and 1: present) with transition rates g_2 and l_2. The rate matrix is:

$$Q = \begin{pmatrix} -g_2 & g_2 \\ l_2 & -l_2 \end{pmatrix} \tag{4.14}$$

Thus the transition matrix for a time duration of t is:

$$P(t) = \begin{pmatrix} p^{(2)}_{0\to0}(t) & p^{(2)}_{0\to1}(t) \\ p^{(2)}_{1\to0}(t) & p^{(2)}_{1\to1}(t) \end{pmatrix} = e^{Qt} = \begin{pmatrix} \frac{l_2}{g_2+l_2} + \frac{g_2}{g_2+l_2}e^{-(g_2+l_2)t} & \frac{g_2}{g_2+l_2} - \frac{g_2}{g_2+l_2}e^{-(g_2+l_2)t} \\ \frac{l_2}{g_2+l_2} - \frac{l_2}{g_2+l_2}e^{-(g_2+l_2)t} & \frac{g_2}{g_2+l_2} + \frac{l_2}{g_2+l_2}e^{-(g_2+l_2)t} \end{pmatrix} \tag{4.15}$$

In the following time flows backward, i.e. greater time is further away in the past.

$$\begin{aligned}
&\mathbb{P}_\Theta(r_g|\text{cat}_g = 2) \\
&= \int_{t=0}^{H} \frac{1}{H}\mathbb{P}^{(2)}_t(r_g)\,\mathrm{d}t \\
&= \frac{1}{H}\sum_{a=1}^{2G-2}\int_{t=t_a}^{t_{a+1}} \mathbb{P}^{(2)}_t(r_g)\,\mathrm{d}t \\
&= \frac{1}{H}\sum_{a=G}^{2G-2}\int_{t=t_a}^{t_{a+1}} \prod_{T_i}\mathbb{P}^{(2)}_t\left(r_g|_{T_i}\right)\mathrm{d}t \\
&= \frac{1}{H}\sum_{a=G}^{2G-2}\int_{t=t_a}^{t_{a+1}} \prod_{\substack{T_i \text{ subtree} \\ \text{with root } r_i}} \left(p^{(2)}_{0\to1}(t-t_{r_i})\mathbb{P}^{(2)}_{r_i}\left(r_g|_{T_i}\right) + p^{(2)}_{0\to0}(t-t_{r_i})\mathbb{P}^{(2)}_{\bar{r_i}}\left(r_g|_{T_i}\right)\right) \\
&\qquad\qquad \times \prod_{f_i \text{ leaf}} p^{(2)}_{0\to r_g|_{f_i}}(t-t_{f_i})\,\mathrm{d}t
\end{aligned} \tag{4.16}$$

The first equality integrates over all possible heights where the gene was imported in the gene pool of the population. $\mathbb{P}^{(2)}_t(r_g)$ is the probability to obtain pattern r_g given that g arrived at height t of the tree.
The second equality decomposes the integral, summing the probability that gene g arrived between node a and node $a+1$ over all nodes of the tree. We assume here that nodes are ordered by increasing height, i.e. $t_1 = 0$ and $t_{2G-1} = H$. As $\mathcal{T}$ is ultrametric, we have in fact $t_1 = t_2 = ... = t_G = 0$.
In the third equality, $\mathbb{P}^{(2)}_t(r_g)$ is decomposed into a product over all subtrees obtained by cutting the phylogenetic tree at time t (the time of appearance of the gene).
In the fourth equality, we distinguish between subtrees that are a single leaf (second product) and other subtrees (first product), see Figure 4.8b. In the first product, the sum represents the fact that the gene can either be gained by the focal lineage from time t to t_{r_i} (with probability $p^{(2)}_{0\to1}(t)$), or not be gained (with probability

$p^{(2)}_{0\to 0}(t)$). $\mathbb{P}^{(2)}_{r_i}(r_g|_{T_i})$ is the probability to obtain pattern r_g restricted to the leaves of T_i given that g was present at root r_i of subtree T_i, while $\mathbb{P}^{(2)}_{\bar{r}_i}(r_g|_{T_i})$ is the probability to obtain this pattern given that is was absent at root r_i. These two quantities can be computed using Felsenstein's pruning algorithm (Felsenstein 1981). $p^{(2)}_{i\to j}(t)$ is the probability for a Mobile gene to switch from state i to state j during a time interval of length t, with $i, j \in \{0, 1\}$. We have:

$$\begin{aligned} p^{(2)}_{0\to 1}(t) &= \frac{g_2}{g_2+l_2} - \frac{g_2}{g_2+l_2}e^{-(g_2+l_2)t} \\ p^{(2)}_{0\to 0}(t) &= \frac{l_2}{g_2+l_2} + \frac{g_2}{g_2+l_2}e^{-(g_2+l_2)t} \\ p^{(2)}_{0\to r_g|_{f_i}}(t) &= \begin{cases} p^{(2)}_{0\to 1}(t)(1-\epsilon_2) + p^{(2)}_{0\to 0}(t)\epsilon_2, & \text{if } r_g|_{f_i} = 1 \\ p^{(2)}_{0\to 1}(t)\epsilon_2 + p^{(2)}_{0\to 0}(t)(1-\epsilon_2), & \text{if } r_g|_{f_i} = 0 \end{cases} \end{aligned} \tag{4.17}$$

with ϵ_2 the rate of false positives and false negatives for Mobile genes. It follows that:

$$\begin{aligned} &\mathbb{P}_\Theta(\text{cat}_g = 2|g \text{ observed})\mathbb{P}_\Theta(r_g|\text{cat}_g = 2, g \text{ observed}) \\ &= \frac{I_2^{obs}}{N^{obs}} \times \mathbb{P}_\Theta(r_g|\text{cat}_g = 2, g \text{ observed}) \\ &= \frac{\int_0^t i_2 p_1^{obs}(t)\,\mathrm{d}t}{N^{obs}} \times \frac{\mathbb{P}_\Theta(r_g|\text{cat}_g = 2)}{\frac{1}{H}\int_0^t p_2^{obs}(t)\,\mathrm{d}t} \\ &= \frac{i_2}{N^{obs}} \sum_{a=G}^{2G-2} \int_{t=t_a}^{t_{a+1}} \prod_{T_i \text{ subtree}} \left(p^{(2)}_{0\to 1}(t-t_{r_i})\mathbb{P}^{(2)}_{r_i}\left(r_g|_{T_i}\right) + p^{(2)}_{0\to 0}(t-t_{r_i})\mathbb{P}^{(2)}_{\bar{r}_i}\left(r_g|_{T_i}\right)\right) \\ &\qquad \times \prod_{f_i \text{ leaf}} p^{(2)}_{0\to r_g|_{f_i}}(t-t_{f_i})\,\mathrm{d}t \end{aligned} \tag{4.18}$$

To obtain the second equality, we used the approximation of I_2^{obs} by its expectation as computed in Equation (4.6), and the fact that $\mathbb{P}_\Theta(g \text{ observed}|\text{cat}_g = 2) = \frac{1}{H}\int_0^t p_2^{obs}(t)\,\mathrm{d}t$.

4.5.3 Inference method

Optimization method

In order to fit our model on the *Salmonella enterica* dataset, we computed the Maximum Likelihood parameters of the model:

$$\hat{\Theta} = \underset{\Theta}{\text{argmax}}\, \mathcal{L}(\text{data}; \Theta) \tag{4.19}$$

Substituting the expectations of Equation (4.6) in Equation (4.5), we observe that the 10 parameters ($N_0, l_0, i_1, l_1, i_2, g_2, l_2, \epsilon_0, \epsilon_1$ and ϵ_2) are linked by the following relation:

$$N^{obs} = p_0^{obs}(root)N_0 + \frac{i_1}{l_1}p_1^{obs}(root) + \int_{\mathcal{T}} i_1 p_1^{obs}(x)\,\mathrm{d}x + \int_{t=0}^{H} i_2 p_2^{obs}(t)\,\mathrm{d}t \qquad (4.20)$$

(given that $p_0^{obs}(root)$ depends on l_0 and ϵ_0, $p_1^{obs}(root)$ and $p_1^{obs}(x)$ depend on l_1 and ϵ_1, and $p_2^{obs}(t)$ depends on g_2, l_2 and ϵ_2). N^{obs} being the total number of observed genes in the dataset, we know its value. This relation is a constraint that we have to put on parameters to ensure that the expected number of observed genes predicted by the model coincides with the number of actually observed genes. Hence, we have only 9 degrees of freedom in the model. For example, we can express i_2 as a function of other parameters:

$$i_2 = \frac{N^{obs} - p_0^{obs}(root)N_0 - \frac{i_1}{l_1}p_1^{obs}(root) - \int_{\mathcal{T}} i_1 p_1^{obs}(x)\,\mathrm{d}x}{\int_{t=0}^{H} p_2^{obs}(t)\,\mathrm{d}t} \qquad (4.21)$$

We performed likelihood maximization over the 9 free parameters using the Nelder-Mead algorithm, which is a derivative-free optimization method. As this optimization procedure offers no guarantee to find the global maximum, we tried three different approaches to have some confidence in the result. We performed 5 optimizations with different, randomly chosen initial conditions and took the maximum over these 5 replicates. Then, we designed two methods to find a good starting point for the optimization. The first one consists in making 10 optimizations on subsamples consisting in 1/10 of the gene families, and then use the median of inferred parameters as starting point for the whole dataset. The second one consists in assigning genes a priori to a gene category (based only on the frequency and parsimony of the presence/absence pattern), performing the optimization separately for each category (thus, with fewer parameters) and using the separately inferred parameters as starting point for the complete optimization. Of course, this last method was only possible because we had gained some knowledge on the model and the data while studying previous versions of the model. All the above approaches yielded the same result, which make us confident to have found the global maximum. Optimizations that converged to the global maximum took from 6 to 15 hours to run on 50 cores, depending on the starting point. The time complexity is linear in the number of gene families and quadratic in the number of genomes. The inference tool will be available at `https://github.com/JasmineGamblin/PPMmodelPangenome`. More details on the optimization method (including the boundaries imposed on parameters and the way to draw random starting points) are given in the Supplementary Information Section 4.6.1.

Category assignment

Once the MLE of parameters was computed, we assigned each gene to the category with highest posterior probability, i.e. gene g with pattern r_g was assigned to category c such that:

$$c = \underset{c=0,1,2}{\operatorname{argmax}}\, \mathbb{P}_{\hat{\Theta}}(\mathrm{cat}_g = c | r_g) \tag{4.22}$$

Expected number of gains for a Mobile gene

Let G_g be the number of introductions for a Mobile gene g. By 'introduction', we mean a transfer between the gene pool and a lineage of the tree where this gene has never been present. We can compute G_g as follows, integrating over all possible arrival times in the gene pool and summing over all possible branches where an introduction could happen:

$$\begin{aligned}
&\mathbb{E}(G_g | \mathrm{cat}_g = 2) \\
&\qquad = \frac{1}{H} \int_0^H \mathbb{E}(G_g | T_g = t)\, \mathrm{d}t \\
&\qquad = \frac{1}{H} \int_{t=0}^H \sum_{\substack{T_i \text{ subtree} \\ \text{below } t}} \sum_{\substack{(b,c) \\ \text{branch of } T_i}} \mathbb{E}(\mathbb{1}_{g \text{ introduced in } (b,c)})\, \mathrm{d}t \\
&\qquad = \frac{1}{H} \int_0^H \sum_{\substack{T_i \text{ subtree} \\ \text{below } t}} \sum_{\substack{(b,c) \\ \text{branch of } T_i}} \int_{t_c}^{t_b} g_2 e^{-g_2(t-u)}\, \mathrm{d}u\, \mathrm{d}t
\end{aligned} \tag{4.23}$$

Conditionally on g having arrived at time t in the pool, the probability that it is introduced on a branch (b, c) below this time t is $\int_{t_c}^{t_b} g_2 e^{-g_2(t-u)}\, \mathrm{d}u$. Indeed, $g_2\, \mathrm{d}u$ is the probability that g is gained during time interval $\mathrm{d}u$ and $e^{-g_2(t-u)}$ is the probability that no gain happened before time u.

4.6 Supplementary Information

4.6.1 Detailed inference procedure

The likelihood of the data is computed as explained in the Main Text Section 4.5.2. The optimization of the log-likelihood is done by the Nelder-Mead algorithm. Details of this optimization are described below.

Starting point of the optimization algorithm

As explained in the Main Text Section 4.5.3, the starting point of the optimization is either randomly selected, or chosen via a pre-processing step. Random starting points were drawn from the following distributions:

— $N_0 \sim \mathcal{U}[M/3, M]$
— $l_0 \sim \mathcal{E}(L/5)$
— $i_1 \sim \mathcal{U}[\frac{N^{obs}}{5L}, \frac{N^{obs}}{L}]$
— $l_1 \sim \mathcal{E}(L/50)$
— $g_2 \sim \mathcal{E}(L/50)$

— $l_2 \sim \mathcal{E}(L/500)$
— $\epsilon_0, \epsilon_1, \epsilon_2 \sim \mathcal{E}(100)$

The symbol $\mathcal{U}$ denotes the uniform distribution and the symbol $\mathcal{E}$ denotes the exponential distribution. The parameters of these distributions depend on dataset characteristics, in order to automatically adapt to new datasets. More precisely, they depend on the mean number of genes per genome (M), the number of observed genes (N^{obs}) and the total branch length of the tree (L).

Exploration of parameter space

To prevent the algorithm from exploring implausible regions of the parameter space, we placed the following constraints on the ranges of parameters, listed below:

— $N_0 \in [0, 2M]$
— $l_0 \in [0, 100/L]$
— $i_1 \in [0, 3N^{obs}/L]$
— $l_1 \in [0, 1000/L]$
— $g_2 \in [0, 1000/L]$
— $l_2 \in [0, 50000/L]$
— $\epsilon_0, \epsilon_1, \epsilon_2 \in [0, 0.2]$

As explained in the Methods Section 4.5.3, the estimate of parameter i_2 is deduced from the estimates of other parameters and N^{obs}, and so has to be computed at each evaluation of the log-likelihood. If the returned value for i_2 is negative, the function evaluation returns a very low value ($-K_2 + i_2$ where K_2 is taken equal to 10^9) to drive the algorithm away from this parameter configuration. We use a threshold value of $k_1 = 10^{-4}$ on variations of the log-likelihood to terminate the optimization procedure. The values of k_1 and K_2 may have to be adapted to the size of the dataset, as bigger datasets have lower likelihood. We verified that all of the parameters inferred for the *Salmonella enterica* data lied away from these boundaries.

Accelerating the likelihood computation

Computing the likelihood of our dataset given the PPM model takes about 2 minutes on 50 cores. This is due to the large number of genes in our dataset (46,146 genes, compressed into 15,192 unique presence/absence patterns), the size of our tree (902 leaves) and the fact that we have to integrate over all possible arrival times and all subtrees when computing the likelihood of Mobile genes. In order to reduce this time as much as possible, we implemented a tree data structure to store intermediate results of pruning algorithms. With this implementation, the multiple calls to the pruning algorithm requested to compute the likelihoods of Private and Mobile genes (see Main Text Sections 4.5.2 and 4.5.2) never do the same computation twice during one evaluation of the dataset likelihood. Moreover, the dataset likelihood computation is easy to parallelize as the genes likelihoods are independent from each other.

Number of free parameters to infer

There are 10 parameters in the PPM model as presented in the Main Text. However, as explained in Section 4.5.3, i_2 can be expressed as a function of N^{obs} and the other parameters, leaving 9 free parameters to infer. To speed up the optimization algorithm, we start by maximizing the likelihood assuming that all error rates are equal ($\epsilon_0 = \epsilon_1 = \epsilon_2$), i.e. with 7 free parameters. Then, we maximize the likelihood of the complete model using the previous estimate as starting point.

Avoiding underflow

During this procedure, we always compute the log-likelihood instead of the likelihood to avoid underflow (i.e., numbers that are smaller than the available precision and thus treated as zero), which are common when working with large phylogenies. Products featured in the likelihood computation are easily replaced by sums of log-likelihoods. However, sums in the likelihood require use of the so-called 'log-sum-exp trick'. Let $a = \log(x)$ and $b = \log(y)$. If we try to compute $\log(x + y) = \log(e^a + e^b)$, there is a risk of underflow if x and y are very small numbers. Instead, we compute $\log(x + y) = a + \log(1 + e^{b-a})$. Here, we do not have to compute x or y directly but only their ratio $\frac{y}{x} = e^{b-a}$, which is less likely to underflow.

4.6.2 Detailed simulation study

We conducted a simulation study to assess the accuracy of the inference procedure. To that end, we drew a random tree and simulated genes evolving along that tree for various sets of parameter values. Then we compared the inferred parameters to their true values on these simulated datasets (correlations between true and inferred parameters are shown in the Main Text Figure 4.3).

Drawing a random phylogenetic tree

We generated a 200-leaf tree by drawing its topology from the beta-splitting trees distribution with $\beta = -1/2$ (D. Aldous 1996; D. J. Aldous 2001). The branching times were drawn uniformly in $[0, 1]$ and we imposed $H = 1$. The code used to generate the tree will be available at `https://github.com/JasmineGamblin/PPMmodelPangenome`.

Drawing the parameter values

We drew 100 sets of parameter values using the following distributions ($\log\mathcal{U}$ denotes the log-uniform distribution):

— $N_0 \sim \mathcal{U}[500, 1500]$
— $l_0 \sim \log\mathcal{U}[0.1/L, 10/L]$
— $i_1 \sim \mathcal{U}[500/L, 1500/L]$

— $l_1 \sim \log\mathcal{U}[5/L, 500/L]$
— $i_2 \sim \mathcal{U}[500/H, 1500/H]$
— $g_2 \sim \log\mathcal{U}[5/L, 500/L]$
— $l_2 \sim \log\mathcal{U}[250/L, 25000/L]$
— $\epsilon_0, \epsilon_1, \epsilon_2 \sim \mathcal{E}(100)$

These distributions are designed to yield an average of a thousand genes from each category, and to span a wide range of gain and loss rates.

Inference

For each of these 100 datasets, we inferred the parameter values using the following procedure (which is also the procedure that we used on the *Salmonella enterica* dataset):

— Run five optimizations with different random starting points, using the model version where all error rates are equal (7 free parameters).
— Take the best parameter estimates from the previous step as starting point of the optimization on 9 free parameters.

4.6.3 Arrival time of Mobile genes

As we model explicitly the importation of Mobile genes into the gene pool through time, we can in theory infer the arrival dates of Mobile genes from their presence/absence pattern. Conditional on the presence/absence pattern r_g of Mobile gene g, the distribution of its arrival time T_g is computed as follows:

$$\begin{aligned}
p(T_g = t \,|\, r_g, \mathrm{cat}_g = 2) & \\
&= \frac{1}{H}\frac{\mathbb{P}(r_g \,|\, T_g = t, \mathrm{cat}_g = 2)}{\mathbb{P}(r_g \,|\, \mathrm{cat}_g = 2)} \\
&= \frac{1}{H\mathbb{P}(r_g \,|\, \mathrm{cat}_g = 2)}\mathbb{P}^{(2)}_t(r_g) \\
&= \frac{1}{H\mathbb{P}(r_g \,|\, \mathrm{cat}_g = 2)} \prod_{\substack{T_i \text{ subtree} \\ \text{with root } r_i \\ \text{below time } t}} \left(p^{(2)}_{0\to 1}(t - t_{r_i})\mathbb{P}^{(2)}_{r_i}\left(r_g|_{T_i}\right) + p^{(2)}_{0\to 0}(t - t_{r_i})\mathbb{P}^{(2)}_{\bar{r_i}}\left(r_g|_{T_i}\right)\right) \\
&\qquad \times \prod_{f_i \text{ leaf}} p^{(2)}_{0\to r_g|_{f_i}}(t - t_{f_i})
\end{aligned} \tag{4.24}$$

$r_g|_{T_i}$ denotes the restriction of pattern r_g to the leaves of subtree T_i, and $\mathbb{P}^{(2)}_{\bar{r_i}}(r_g|_{T_i})$ is the probability to obtain this pattern given that gene g is was absent from node r_i. The products in the last equality run over all subtrees (including leaves) formed by cutting the tree at height t.

Using maximum likelihood estimates of the parameters inferred on the *Salmonella enterica* dataset, we simulated 14,827 Mobile genes and computed the distribution

of their arrival time conditional on their pattern. An arbitrary subset of the resulting distributions are shown in Figure 4.12a. The distributions are completely flat between height 0.0015 and 0.03, i.e. between $H/20$ and H. However, there are differences between genes in the time period close to the leaves, between height 0 and height $1.5e^{-4} = H/200$. In Figure 4.12b is represented the correlation between true arrival time (which was recorded during simulations) and inferred arrival time (computed as the mode of the conditional distribution from Equation (4.24)). We see a somewhat satisfying correlation for genes imported very recently (between 0 and $1e^{-4}$), but earlier arrival times seem completely uncorrelated with the inferred arrival times. This is likely due to the high values of gain and loss rates for the Mobile category: the high turnover of these genes erases the signal of the introduction date very quickly, and thus the model is able to infer it approximately only for very recently acquired genes. In the light of this, the PPM model should be able to infer arrival dates of Mobile genes on datasets for which the gain and loss rates are not too high compared to tree height. In fact, we can even compute the timescale at which the gene patterns reach stationary state: according to the transition matrix (4.15) shown in Main text, the characteristic time is $1/(g_2 + l_2)$, i.e. roughly $0.8e^{-4}$. This value is coherent with the above observations on simulated data.

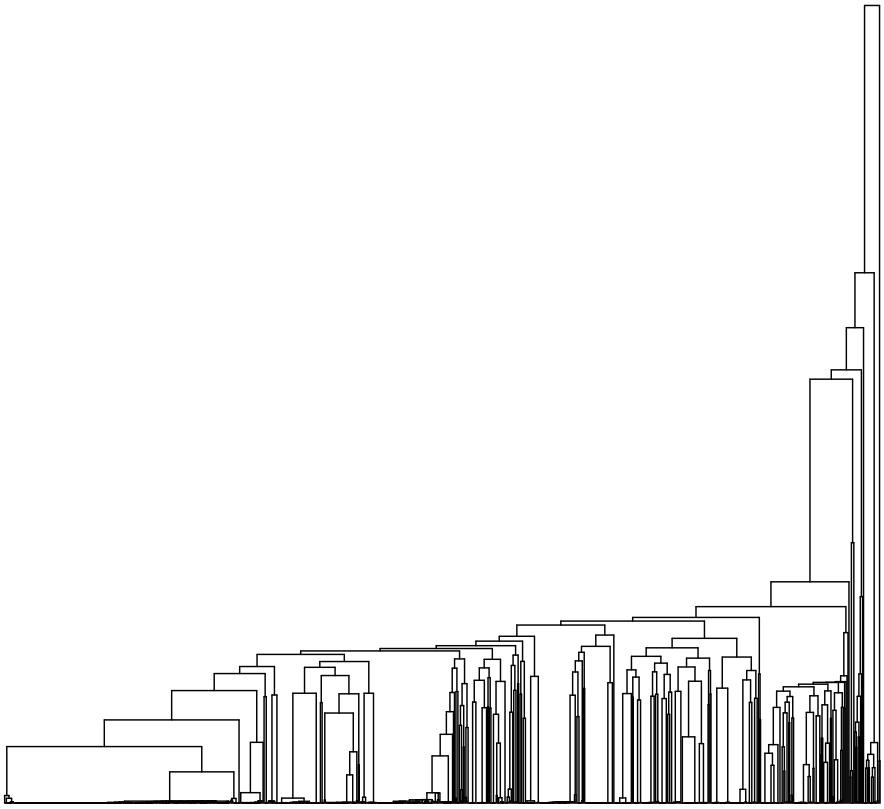


Figure 4.9 – Phylogenetic tree of our 902 *Salmonella enterica* genomes. It was inferred using sequences of genes present on at least 99% of genomes, with the software IQ-TREE. An outgroup (*Salmonella bongorii*) was used to root the tree and is not shown on the picture.

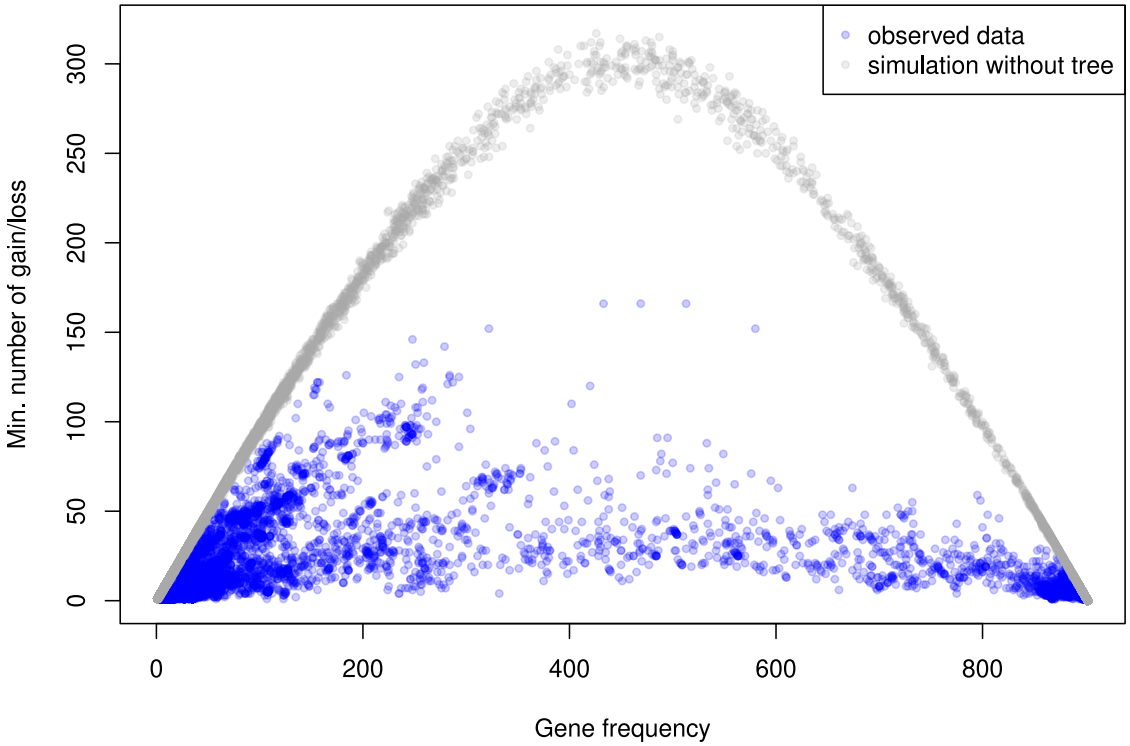


Figure 4.10 – Blue points correspond to the parsimony vs. frequency plot of the *Salmonella enterica* dataset. The y-axis is the minimum number of gain and loss events needed along the species tree to explain the presence/absence pattern of a gene with frequency given on x-axis. Grey points correspond to gene patterns that were randomly simulated without using the species tree (but using the same frequency distribution as in the data).

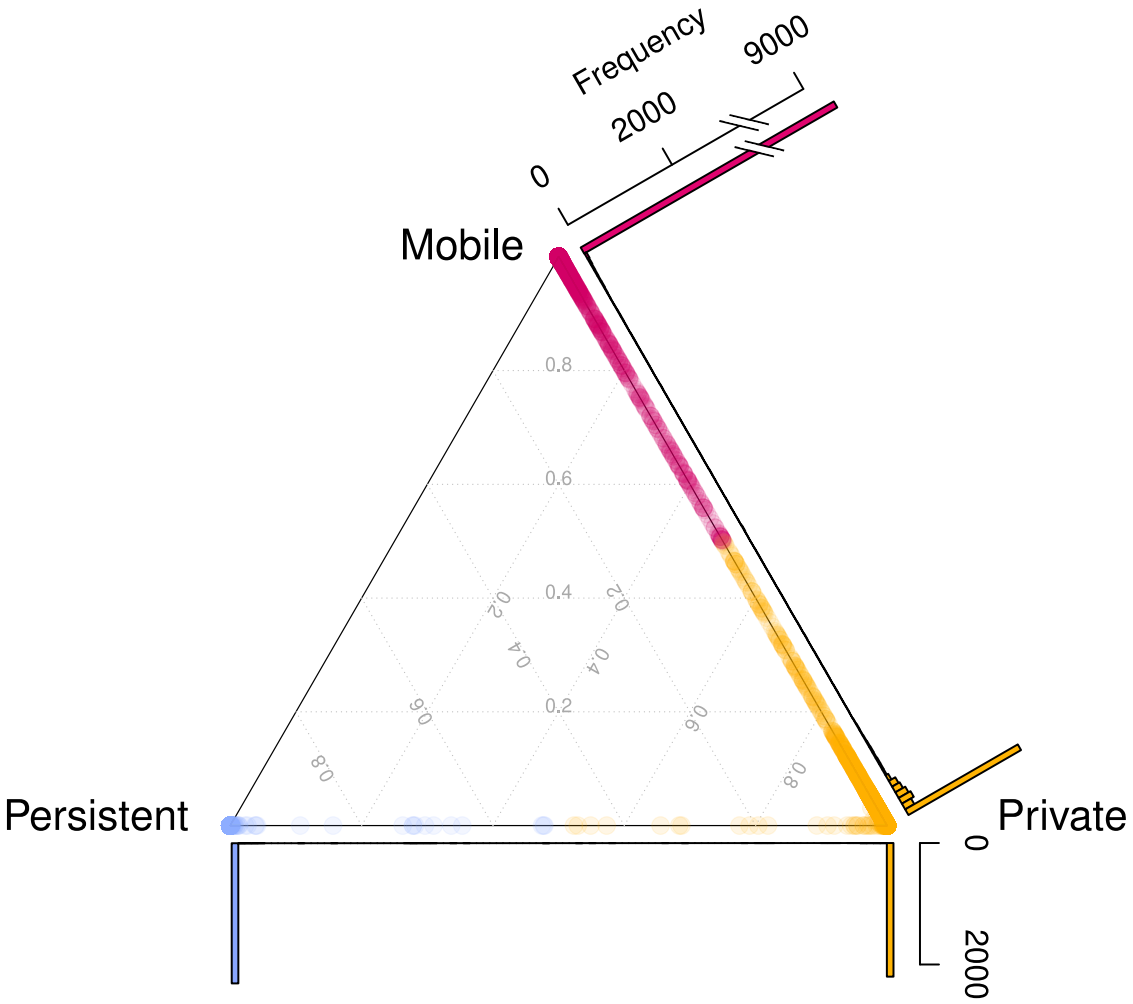
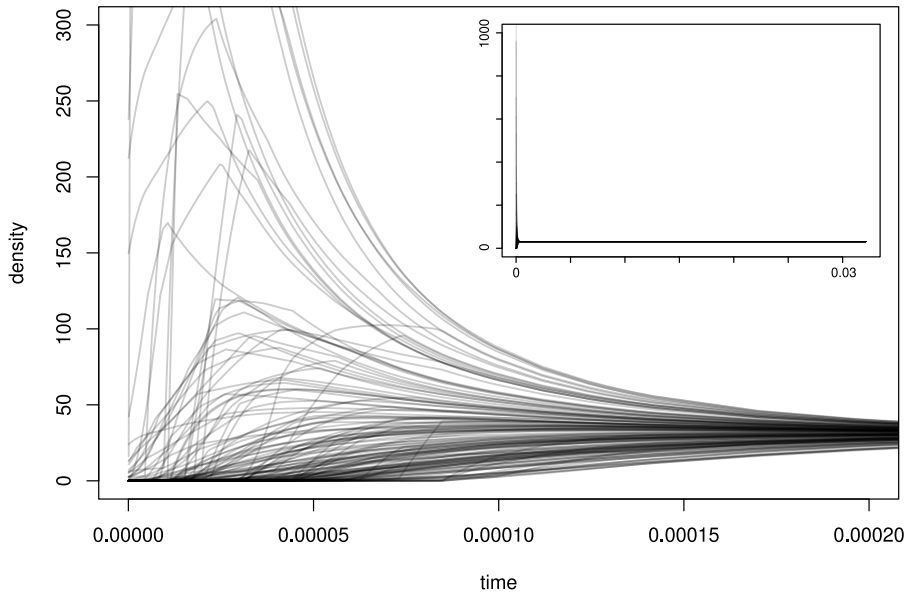
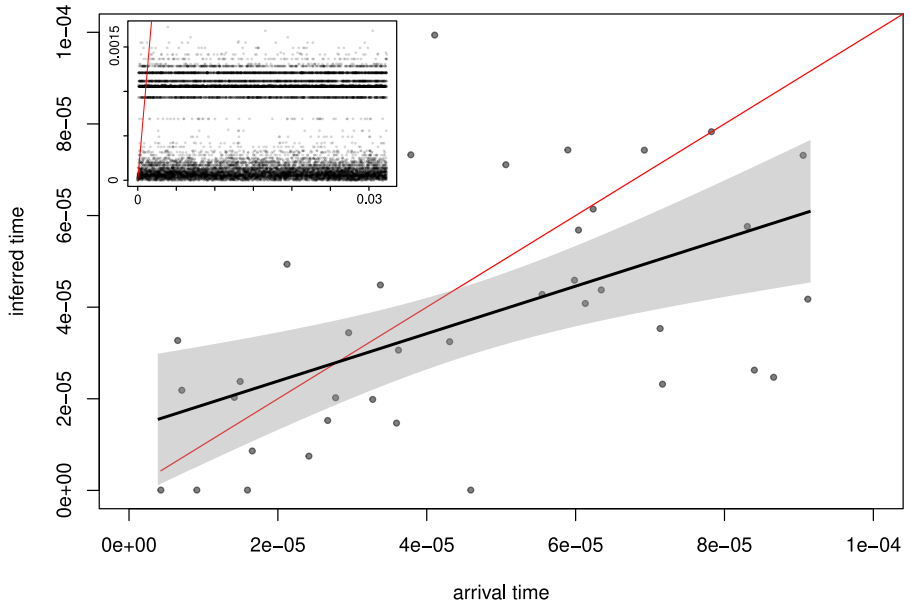


Figure 4.11 – Ternary plot showing category assignment for genes from the *Salmonella enterica* dataset. Each unique pattern is represented by a point, whose coordinates are the respective probabilities for this pattern to belong to each of the three categories (Persistent, Private and Mobile). Histograms show the distributions of points along two of the edges.

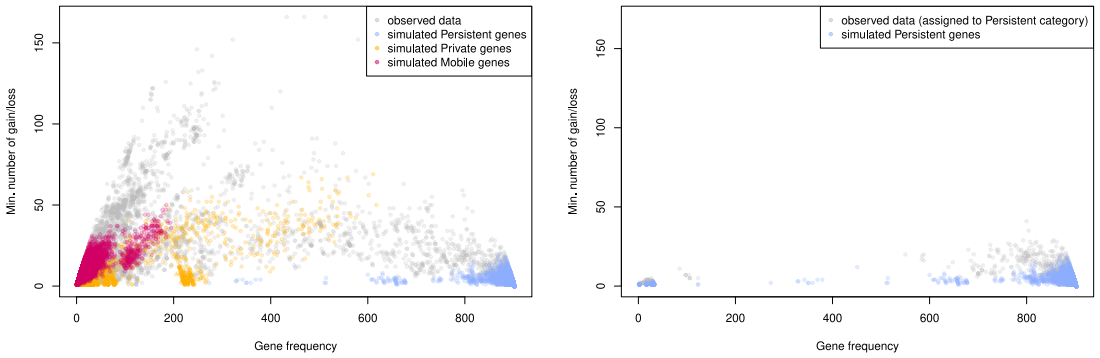


(a) Inferred distribution of arrival time



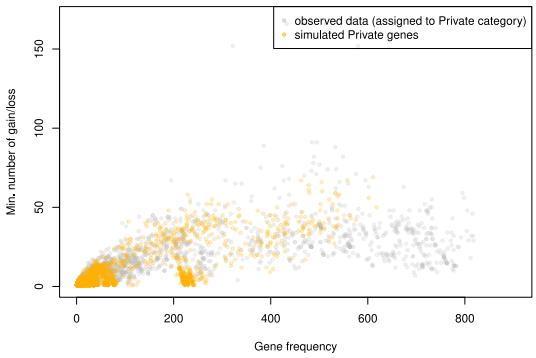
(b) Inference of arrival time

Figure 4.12 – (a) shows the distributions of arrival time conditional on gene patterns for every 100^{th} gene of the 14,827 simulated Mobile genes. The main plot is an enlargement of the full distributions shown in the upper right box. (b) shows the correlation between true arrival time (recorded during simulations) and inferred arrival time (computed as the mode of the conditional distribution) for 14,827 simulated Mobile genes. The line $y = x$ is plotted in red for comparison. The main plot is an enlargement of the full picture shown in the upper left box.

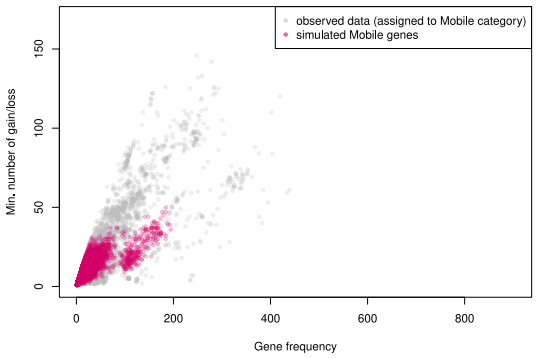


(a) Fit of the parsimony plot

(b) Persistent genes



(c) Private genes



(d) Mobile genes

Figure 4.13 – (a) Fit of the parsimony plot with observed data in gray and simulated data colored according to gene category. (b) Showing only Persistent genes (for observed data: genes assigned to this category; for simulated data: category used for simulation), (c) Showing only Private genes. (d) Showing only Mobile genes.

Historique du projet

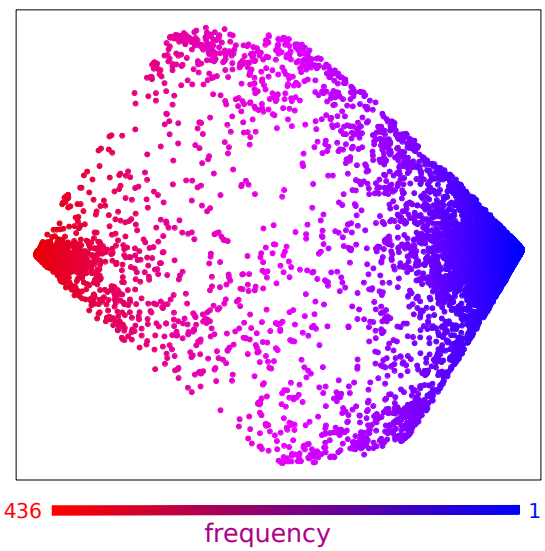
Étude préalable de données pangénomiques

Ce projet de modélisation de l'évolution des pangénomes bactériens a débuté par l'étude d'un ensemble de 436 génomes d'*Escherichia coli* commensales récoltés par l'unité de recherche IAME (notamment utilisé par Marin et al. 2022). Après avoir tracé la distribution des fréquences de gènes (ou GFS) et vérifié qu'elle présentait bien une forme en U, nous avons cherché d'autres moyen d'étudier les répartitions des gènes dans ce jeu de données. Nous avons eu l'idée de calculer pour chaque gène le nombre minimum de gains et pertes le long de la phylogénie nécessaires pour expliquer sa répartition aux feuilles, puis de tracer ce score en fonction de la fréquence. Ceci a donné la figure de parcimonie en fonction de la fréquence.

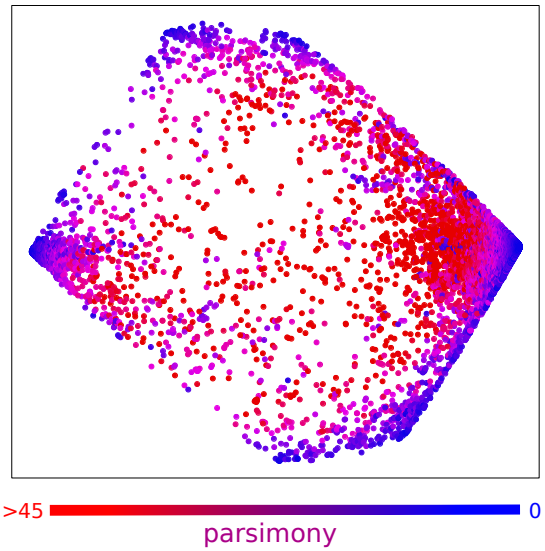
Nous avons ensuite voulu représenter les différentes répartitions comme un nuage de points en deux dimensions, en utilisant la méthode du positionnement multidimensionnel (en anglais *multidimensional scaling*, T. F. Cox et M. A. A. Cox 2000). Étant donné des points dans un espace à n dimensions, cette méthode permet de représenter ces points dans un espace à $d < n$ dimensions (en général $d = 2$ ou 3) tout en conservant au mieux les distances entre points. Dans notre cas, les points sont les répartitions de gènes, qui sont des vecteurs binaires en dimension $n = 436$ (le nombre de génomes). Utiliser une distance euclidienne entre ces vecteurs ne semble pas approprié, car il semble naturel de vouloir corriger la distance par les relations phylogénétiques : on voudrait considérer que les répartitions de deux gènes présents dans deux clades proches sont plus similaires que celles de deux gènes présents dans deux clades éloignés. Pour chaque gène, nous avons reconstruit les états ancestraux à tous les nœuds de l'arbre d'espèce par maximum de parcimonie. Puis la distance entre deux répartitions a été calculée comme la distance euclidienne entre les vecteurs de présence/absence non pas aux feuilles, mais à tous les nœuds de l'arbre. Le nuage de points obtenu est représenté en Figure 4.14. Nous avons ensuite cherché quelles caractéristiques des répartitions permettaient le mieux de prédire leur position dans le nuage de points. Après différents essais, nous avons trouvé que la fréquence, la parcimonie et la spécificité au clade gauche ou droit de l'arbre étaient fortement corrélés à la position. Cela a confirmé que la fréquence et la parcimonie des répartitions étaient des propriétés importantes à étudier pour les pangénomes.

Test des modèles existants

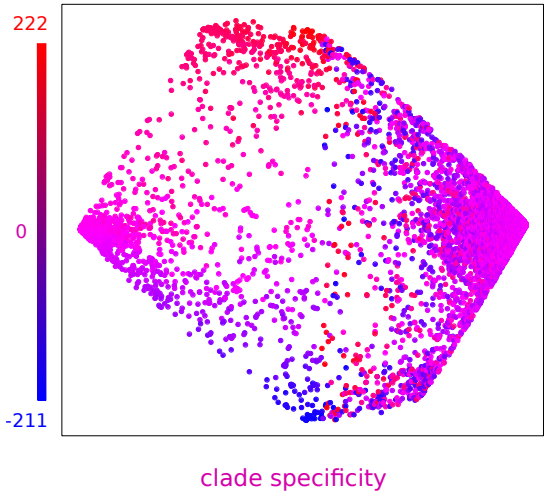
Reproduction de statistiques multivariées. Nous avons utilisé ces données pour tester des modèles existants d'évolution de pangénome. Nous nous somme surtout concentrés sur le modèle développé par Haegeman et Weitz (2012) pour expliquer le spectre de fréquence génique (GFS) en forme de U. Ce modèle suppose un nombre fixe de gènes par génome, qui sont remplacés par un nouveau gène (tiré d'un pool infini) à un taux θ. Nous avons quelque peu adapté ce modèle, notamment pour prendre en compte l'évolution le long d'un arbre fixé. En effet, le



(a) Fréquence



(b) Parcimonie (nombre minimum d'événements de gains et pertes)



(c) Spécificité au clade gauche (différence de fréquence entre les clades gauche et droit de l'arbre d'espèce)

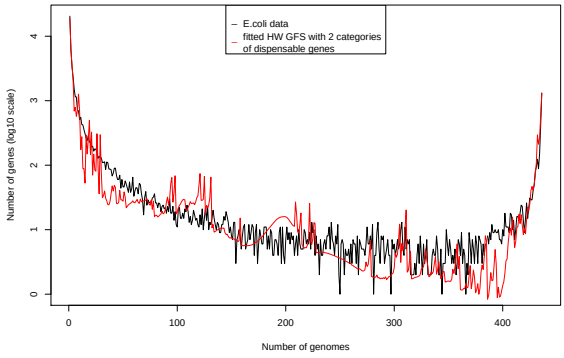
Figure 4.14 – Représentation en deux dimensions par positionnement multidimensionnel des répartitions de gènes d'un ensemble de 436 génomes d'*E. coli* commensales. Le code couleur représente la valeur de chacune des trois caractéristiques corrélant avec le positionnement : (a) la fréquence du gène, (b) la parcimonie de la répartition, et (c) la prédominance dans le clade gauche par rapport au clade droit de l'arbre d'espèce.

modèle original suppose que la population évolue selon un processus de Moran et la généalogie correspond donc à un coalescent de Kingman. D'autre part, nous avons utilisé une version du modèle comportant deux classes de gènes accessoires (renouvelés aux taux θ_1 et θ_2) et une classe de gènes *core* immuables. L'ajustement des paramètres directement sur le GFS donne un résultat plutôt convaincant (Figure 4.15a). Cependant, nous avons aussi mis au point une méthode alternative pour estimer les paramètres à partir des répartitions de gènes, qui portent une information plus riche que le GFS. En simulant des données avec les paramètres estimés par cette méthode, on s'aperçoit que le modèle n'est pas capable de reproduire le GFS (Figure 4.15b) ni le graphe de parcimonie en fonction de la fréquence (Figure 4.15c).

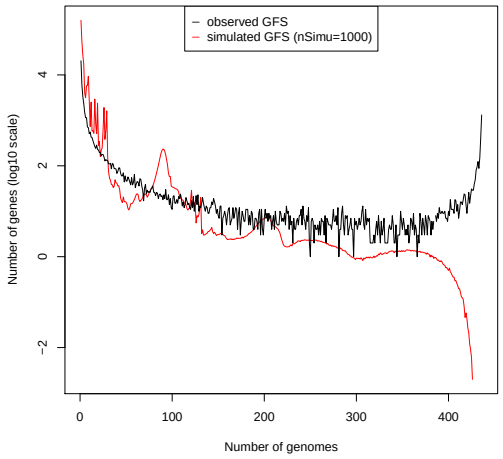
Nombre d'introductions par gène. La limite de ce modèle (ainsi que de celui développé par Baumdicker, Hess et Pfaffelhuber 2012) nous semblait être le fait de n'autoriser qu'une seule introduction par gène dans la phylogénie. C'est une hypothèse assez forte étant donné l'importance des transferts horizontaux chez les bactéries. Pour étudier cette question, nous avons calculé pour chaque famille de gènes le nombre d'introductions inféré par maximum de parcimonie. La distribution du nombre d'introductions par famille montre qu'en dehors des gènes singletons – au nombre de 20 500 – une grande partie des familles ont une répartition plutôt cohérente avec plusieurs introductions (Figure 4.16a). C'est ce qui nous a poussé à développer un modèle comportant une catégorie de gènes pouvant être gagnés plusieurs fois le long de la phylogénie.

Nombre de classes de gènes. Une autre question importante était pour nous le nombre de classes de gènes distinctes nécessaires pour décrire les données, que ces classes correspondent à des dynamiques différentes ou à des valeurs de paramètres différentes. Pour étudier cela, nous avons premièrement inféré les taux de renouvellement individuels des gènes, en utilisant le modèle de Haegeman et Weitz fitté sur les répartitions. La distribution obtenue est très large et donc difficile à représenter, avec des valeurs de θ allant de 0,2 à 25 000. La majorité des valeurs sont tout de même comprises entre 0 et 150, et en se limitant à cet intervalle on observe une distribution bimodale, avec un mode proche de 0 et un autre situé autour de 70 (Figure 4.16b). Ce résultat rejoint les conclusions de l'étude menée par Wolf et al. (2016) : une classe de gènes semble avoir un taux de renouvellement extrêmement élevé, une autre des taux plus faibles et distribués autour d'une valeur (ici 70). Nous ajoutons à cela une classe de gènes au taux de renouvellement très faible, correspondant probablement à des gènes aux fonctions essentielles. Cela trace donc les contours de trois classes de gènes avec des dynamiques potentiellement assez différentes.

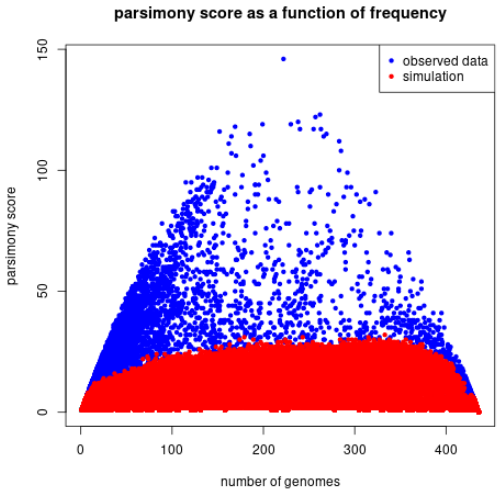
Deuxièmement, nous avons utilisé le logiciel de phylogénie IQ-TREE pour faire de la sélection de modèle sur le modèle FMG et déterminer le nombre de classes de gènes optimal (chaque classe ayant des taux de gain et de perte différents). Ce modèle suppose un processus de gains et pertes de gènes le long d'une phylogénie,



(a) Fit direct du GFS

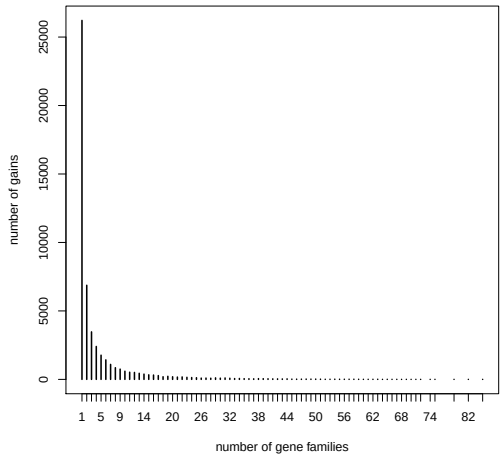


(b) GFS simulé

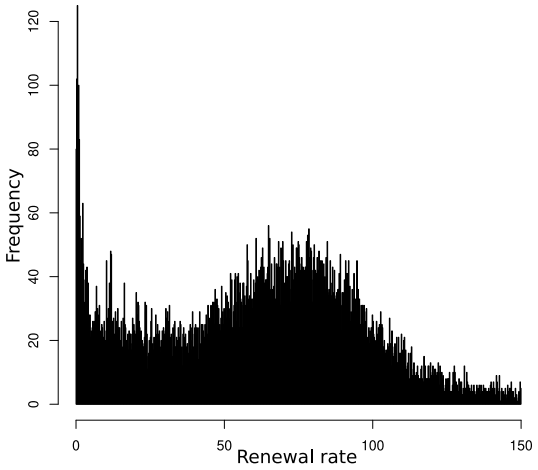


(c) Parcimonie simulée

Figure 4.15 – Tests du modèle de Haegeman et Weitz. (a) comparaison entre le GFS observé de 436 génome d'*E. coli* (en noir) et le GFS ajusté prédit par le modèle (en rouge). (b) comparaison entre le GFS observé et le GFS simulé selon le modèle avec les paramètres estimés à partir des répartitions de gènes observées. (c) comparaison de la figure de parcimonie en fonction de la fréquence pour les données observées (en bleu) et les données simulées selon le modèle avec les paramètres estimés à partir des répartitions de gènes observées (en rouge).



(a) Distribution du nombre d'introductions



(b) Distribution des taux de renouvellement

Figure 4.16 – Résultats de deux études préliminaires sur 436 génomes d'*E. coli*. (a) distribution du nombre d'introductions par famille de gènes, inféré par maximum de parcimonie. (b) distribution du taux de renouvellement par famille de gènes, inféré individuellement pour chaque famille en utilisant le modèle de Haegeman et Weitz. Distribution coupée à 150.

dans lequel tous les gènes sont contenus dans un pool fixé au départ. Nous avions d'emblée disqualifié ce modèle car l'hypothèse que le pangénome conserve la même composition durant toute l'histoire évolutive d'une espèce nous paraissait peu vraisemblable (c'est cette réflexion qui nous a menés à imaginer la catégorie Mobile, avec une introduction des gènes au fur et à mesure dans le pool génétique). Cependant, des outils d'inférence pour ce modèle étant déjà disponibles, nous avons voulu savoir si il permettait de distinguer qualitativement différentes classes de gènes. Le résultat retourné a été d'utiliser 10 classes, ce qui était aussi le nombre maximum de classes testé. Comme nous ne souhaitions pas multiplier le nombre de classes pour garder un modèle facilement interprétable, nous avons décidé de conserver une seule classe de gènes par type de dynamique, quitte à les diviser par la suite en fonction de l'application.

Construction du modèle PPM

Après ces premiers résultats, nous avons mis en place le modèle PPM comportant les 3 types de dynamiques Persistante, Privée et Mobile. Plusieurs ajustements ont eu lieu pour chacune des catégories avant d'aboutir à la version présentée dans ce chapitre.

La catégorie Persistante était à l'origine une catégorie *core*, c'est-à-dire composée de gènes présents dans exactement tous les génomes. La Figure 4.17 montre les différentes options envisagées ensuite, et leur effet sur la figure de parcimonie : (a) des gènes présents partout mais pouvant par erreur n'être pas détectés, (b) des

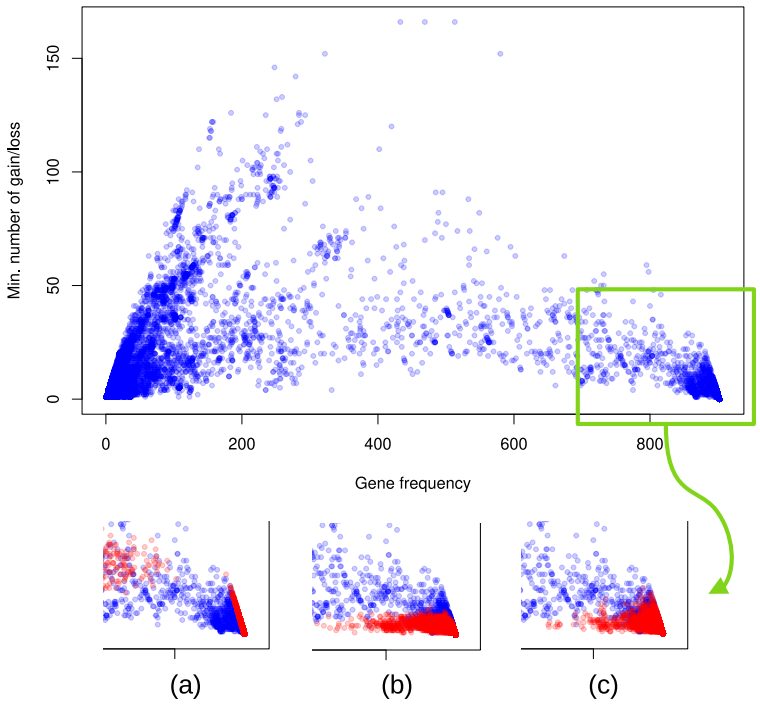
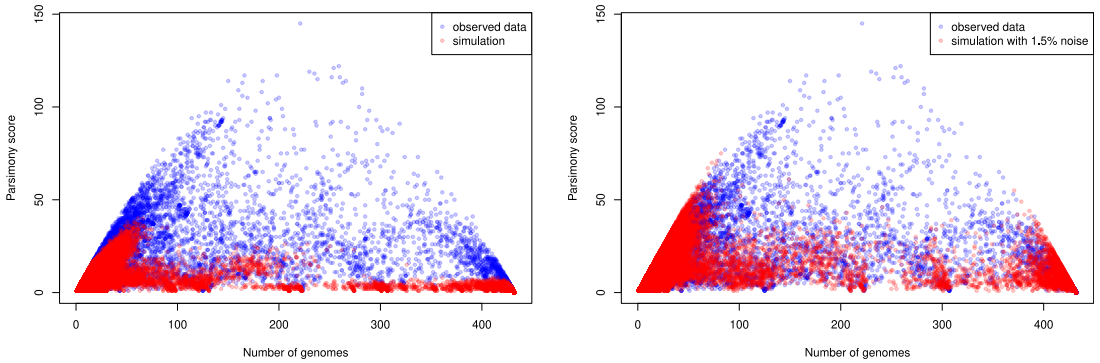


Figure 4.17 – Trois tentatives de reproduire la figure de parcimonie pour la catégorie Persistante. (a) modèle avec taux d'erreur non nul, (b) modèle avec taux de perte non nul, (c) modèle avec taux de perte et taux d'erreur non nuls. Données : 902 génomes de *Salmonella enterica*.

gènes avec un taux de perte constant, et (c) une combinaison des deux options précédentes (taux de perte + taux d'erreur). On constate que seule l'option (c) permet de reproduire la forme triangulaire de la partie droite de la figure, qui correspond aux gènes à haute fréquence et donc à la majorité des gènes de la catégorie Persistante. C'est donc cette option que nous avons retenue.

La catégorie Privée correspond au modèle IMG de Baumdicker, Hess et Pfaffelhuber (2012). À l'origine, le nombre N_1 de gènes privés présents à la racine était un paramètre libre. Dans un soucis de réduction du nombre de paramètres et après avoir constaté que cela n'affectait pas les performances du modèle de manière significative, nous avons fixé ce paramètre à sa valeur attendue à l'équilibre, i_1/l_1 (Huson et Steel 2004). C'est d'ailleurs aussi ce que font les auteurs du modèle IMG.

La catégorie Mobile avait également au début un paramètre N_2 réglant le nombre de gènes mobiles dans le génome ancestral. En faisant une première étude simulation, nous avons constaté qu'il y avait un soucis d'identifiabilité entre N_2 et i_2 (réglant le nombre de gènes importés après l'ancêtre commun), et avons donc choisi de supprimer ce paramètre. Avec le recul apporté par l'étude que nous avons faite par la suite sur l'inférence du temps d'introduction des gènes mobiles (Section 4.6.3), nous comprenons que ce problème est dû aux taux de gain et perte très élevés pour cette catégorie. Nous avons aussi réfléchi à des solutions pour que cette catégorie possède une distribution stationnaire comme la catégorie Privée, et



(a) Modèle PPM sans erreur

(b) Simulation du modèle PPM avec 1,5% d'erreur

Figure 4.18 – Simulation du modèle PPM avec et sans erreur. (a) figure de parcimonie comparant les données observées et les données simulées par le modèle PPM sans erreur. (b) figure de parcimonie comparant les données observées et les données simulées avec les même paramètres mais en ajoutant une erreur exponentiellement distribuée. Données : 436 génomes d'*E. coli*.

donc à des moyen de modéliser la disparition de gènes du pool. Finalement nous avons décidé de ne pas implémenter ces changements afin de ne pas trop complexifier le modèle, mais la question demeure de savoir qui d'un pool de gène de taille constante (en espérance) et d'un pool augmentant linéairement avec le temps correspond le mieux à la réalité biologique. Cela dépend probablement de l'échelle de temps considérée.

Enfin, nous avions initialement ajoutée la possibilité d'une erreur aléatoire aux feuilles dans l'espoir que le modèle génère des répartitions moins parcimonieuses. La Figure 4.18a montre la figure de parcimonie pour le jeu de données d'*E. coli* et les données simulées par le modèle PPM sans erreur avec les paramètres inférés sur les données. La Figure 4.18b montre en rouge des données simulées avec les mêmes paramètres, mais en ajoutant une erreur aléatoire aux feuilles. Dans cette simulation, le taux d'erreur pour chaque gène est tiré dans une distribution exponentielle dons la moyenne est 1,5%. On constate que ces données simulées correspondent beaucoup mieux aux données observées. Par simplicité, nous avons ajouté dans le modèle un taux d'erreur différencié selon les catégories mais qui et fixe au lieu d'être distribué. Bien que l'ajout de l'erreur améliore considérablement la vraisemblance du modèle, le taux d'erreur inféré est très faible pour les catégories Persistante et Mobile (environ 0,2%) et ne produit donc pas l'effet escompté sur la parcimonie des répartitions simulées avec le modèle.

La plupart des analyses présentées ici ont été réalisées sur un jeu de données composé de 436 génomes d'*E. coli*. Pour la version finale du modèle, nous avons choisi d'utiliser un jeu de données composé de 902 génomes complets de *Salmonella enterica* issus de la base de données Refseq. Cela est dû au fait que nous souhaitions travailler avec des génomes complètement assemblés, et avec une espèce moins fortement structurée qu'*E. coli*. Ainsi la dynamique inférée a plus chance d'être

représentative de l'espèce au lieu de refléter une moyenne entre des dynamiques propres aux différents phylogroupes.

Références du Chapitre 4

Aldous, D. (1996). Probability Distributions on Cladograms. *Random Discrete Structures*. Sous la dir. de D. Aldous et R. Pemantle. T. 76. The IMA Volumes in Mathematics and Its Applications. New York, NY : Springer New York.

Aldous, D. J. (2001). Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science* **16**, 23-34.

Arimizu, Y., Y. Kirino, M. P. Sato, K. Uno, T. Sato, Y. Gotoh, F. Auvray, H. Brugere, E. Oswald, J. G. Mainil, K. S. Anklam, D. Döpfer, S. Yoshino, T. Ooka, Y. Tanizawa, Y. Nakamura, A. Iguchi, T. Morita-Ishihara, M. Ohnishi, K. Akashi, T. Hayashi et Y. Ogura (2019). Large-Scale Genome Analysis of Bovine Commensal Escherichia Coli Reveals That Bovine-Adapted E. Coli Lineages Are Serving as Evolutionary Sources of the Emergence of Human Intestinal Pathogenic Strains. *Genome Research* **29**, 1495-1505.

Bansal, M. S., M. Kellis, M. Kordi et S. Kundu (2018). RANGER-DTL 2.0: Rigorous Reconstruction of Gene-Family Evolution by Duplication, Transfer and Loss. *Bioinformatics* **34**, 3214-3216.

Baumdicker, F., W. R. Hess et P. Pfaffelhuber (2012). The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biology and Evolution* **4**, 443-456.

Beavan, A., M. R. Domingo-Sananes et J. O. McInerney (2024). Contingency, Repeatability, and Predictability in the Evolution of a Prokaryotic Pangenome. *Proceedings of the National Academy of Sciences* **121**, e2304934120.

Botelho, J., L. Tüffers, J. Fuss, F. Buchholz, C. Utpatel, J. Klockgether, S. Niemann, B. Tümmler et H. Schulenburg (2023). Phylogroup-Specific Variation Shapes the Clustering of Antimicrobial Resistance Genes and Defence Systems across Regions of Genome Plasticity in *Pseudomonas Aeruginosa*. *eBioMedicine* **90**, 104532.

Cohen, O. et T. Pupko (2010). Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution* **27**, 703-713.

Cox, T. F. et M. A. A. Cox (2000). *Multidimensional Scaling*. CRC Press. 332 p.

Cummins, E. A., R. A. Hall, C. Chris, J. O. McInerney et A. McNally (2022). Distinct Evolutionary Trajectories in the Escherichia Coli Pangenome Occur within Sequence Types. *Microbial genomics* **8**.

D'Andrea, M. M., F. Arena, L. Pallecchi et G. M. Rossolini (2013). CTX-M-type β-Lactamases: A Successful Story of Antibiotic Resistance. *International Journal of Medical Microbiology*. Special Issue Antibiotic Resistance **303**, 305-317.

Denton, J. F., J. Lugo-Martinez, A. E. Tucker, D. R. Schrider, W. C. Warren et M. W. Hahn (2014). Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLOS Computational Biology* **10**, e1003998.

Didelot, X., A. E. Darling et D. Falush (2008). Inferring Genomic Flux in Bacteria. *Genome Research* **19**, 306-317.

Duchêne, S., K. E. Holt, F.-X. Weill, S. Le Hello, J. Hawkey, D. J. Edwards, M. Fourment et E. C. Holmes (2016). Genome-Scale Rates of Evolutionary Change in Bacteria. *Microbial Genomics* **2**, e000094.

Felsenstein, J. (1981). Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution* **17**, 368-376.

Galperin, M. Y., Y. I. Wolf, K. S. Makarova, R. Vera Alvarez, D. Landsman et E. V. Koonin (2021). COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens. *Nucleic Acids Research* **49**, D274-D281.

Gautreau, G., A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, A. Perrin, C. Médigue, A. Calteau, S. Cruveiller, C. Matias, C. Ambroise, E. P. C. Rocha et D. Vallenet (2020). PPanGGOLiN: Depicting Microbial Diversity via a Partitioned Pangenome Graph. *PLOS Computational Biology* **16**, e1007732.

Haegeman, B. et J. S. Weitz (2012). A Neutral Theory of Genome Evolution and the Frequency Distribution of Genes. *BMC Genomics* **13**, 196-196.

Huson, D. H. et M. Steel (2004). Phylogenetic Trees Based on Gene Content. *Bioinformatics* **20**, 2044-2049.

Jacox, E., C. Chauve, G. J. Szöllosi, Y. Ponty et C. Scornavacca (2016). ecceTERA: Comprehensive Gene Tree-Species Tree Reconciliation Using Parsimony. *Bioinformatics* **32**, 2056-2058.

Kloub, L., S. Gosselin, M. S. Fullmer, J. Graf, J. P. Gogarten, J. P. Gogarten et M. S. Bansal (2021). Systematic Detection of Large Scale Multi Gene Horizontal Transfer in Prokaryotes. *Molecular Biology and Evolution* **38**, 2639-2659.

Marin, J., Olivier Clermont, G. Royer, M. Mercier-Darty, J.-W. Decousser, O. Tenaillon, E. Denamur et F. Blanquart (2022). The Population Genomics of Increased Virulence and Antibiotic Resistance in Human Commensal Escherichia Coli over 30 Years in France. *Applied and Environmental Microbiology*.

McInerney, J. O. (2022). Prokaryotic Pangenomes Act as Evolving Ecosystems. *Molecular Biology and Evolution*.

Oliveira, P. H., M. Touchon, J. Cury et E. P. C. Rocha (2017). The Chromosomal Organization of Horizontal Gene Transfer in Bacteria. *Nature Communications* **8** (1), 841.

Page, A. J., C. A. Cummins, M. Hunt, V. K. Wong, S. Reuter, M. T. Holden, M. Fookes, D. Falush, J. A. Keane et J. Parkhill (2015). Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* **31**, 3691-3693.

Pagel, M. (1997). Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **255**, 37-45.

Perrin, A. et E. P. C. Rocha (2021). PanACoTA: A Modular Tool for Massive Microbial Comparative Genomics. *NAR Genomics and Bioinformatics* **3**, lqaa106.

Piel, D., M. Bruto, Y. Labreuche, F. Blanquart, S. Chenivesse, S. Lepanse, A. James, R. Barcia-Cruz, J. Dubert, B. Petton, E. Lieberman, K. M. Wegner, F. A. Hussain, K. M. Kauffman, M. F. Polz, D. Bikard, S. Gandon et F. Le Roux (2021). Genetic Determinism of Phage-Bacteria Coevolution in Natural Populations. *bioRxiv*.

Sakoparnig, T., C. Field et E. van Nimwegen (2021). Whole Genome Phylogenies Reflect the Distributions of Recombination Rates for Many Bacterial Species. *eLife* **10**. Sous la dir. d'A. Nourmohammad et A. M. Walczak, e65366.

Salzberg, S. L. (2019). Next-Generation Genome Annotation: We Still Struggle to Get It Right. *Genome Biology* **20**, 92.

Seif, Y., E. Kavvas, J.-C. Lachance, J. T. Yurkovich, S.-P. Nuccio, X. Fang, E. Catoiu, M. Raffatellu, B. O. Palsson et J. M. Monk (2018). Genome-Scale Metabolic Reconstructions of Multiple Salmonella Strains Reveal Serovar-Specific Metabolic Traits. *Nature Communications* **9**, 3771.

Szöllosi, G. J., B. Boussau, S. S. Abby, E. Tannier et V. Daubin (2012). Phylogenetic Modeling of Lateral Gene Transfer Reconstructs the Pattern and Relative Timing of Speciations. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17513-17518.

Szöllosi, G. J., E. Tannier, V. Daubin et B. Boussau (2015). The Inference of Gene Trees with Species Trees. *Systematic Biology* **64**.

Szöllosi, G. J., E. Tannier, N. Lartillot et V. Daubin (2013). Lateral Gene Transfer from the Dead. *Systematic Biology* **62**, 386-397.

Tenaillon, O., D. Skurnik, B. Picard et E. Denamur (2010). The Population Genetics of Commensal Escherichia Coli. *Nature Reviews Microbiology* **8**, 207-217.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C.

Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli et C. M. Fraser (2005). Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome". *Proceedings of the National Academy of Sciences* **102**, 13950-13955.

Tettelin, H. et D. Medini, éd. (2020). *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Cham : Springer International Publishing.

To, T.-H., M. Jung, S. Lycett et O. Gascuel (2016). Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology* **65**, 82-97.

Tonkin-Hill, G., N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J. A. Lees, R. A. Gladstone, S. Lo, C. Beaudoin, R. A. Floto, S. D. Frost, J. Corander, S. D. Bentley et J. Parkhill (2020). Producing Polished Prokaryotic Pangenomes with the Panaroo Pipeline. *Genome Biology* **21**, 180.

Touchon, M., J. Cury, E.-J. Yoon, L. Krizova, G. C. Cerqueira, C. Murphy, M. Feldgarden, J. Wortman, D. Clermont, T. Lambert, C. Grillot-Courvalin, A. Nemec, P. Courvalin et E. P. Rocha (2014). The Genomic Diversification of the Whole Acinetobacter Genus: Origins, Mechanisms, and Consequences. *Genome Biology and Evolution* **6**, 2866-2882.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. L. Bouguénec, M. Lescat, S. Mangenot, V. Martinez-Jéhanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Médigue, E. P. C. Rocha et E. Denamur (2009). Organised Genome Dynamics in the Escherichia Coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics* **5**, e1000344.

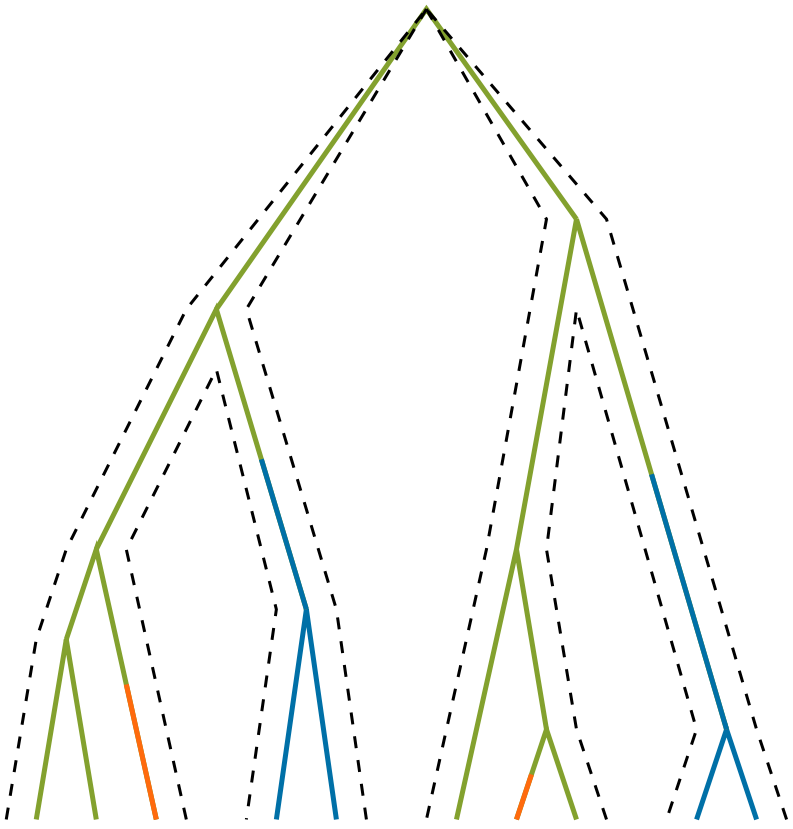
Wagner, C. et M. Hensel (2011). Adhesive Mechanisms of Salmonella Enterica. *Bacterial Adhesion: Chemistry, Biology and Physics*. Sous la dir. de D. Linke et A. Goldman. Dordrecht : Springer Netherlands, 17-34.

Whelan, F. J., R. J. Hall et J. O. McInerney (2021). Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. *Molecular Biology and Evolution* **38**, 3697-3708.

Wolf, Y. I., K. S. Makarova, A. E. Lobkovsky et E. V. Koonin (2016). Two Fundamentally Different Classes of Microbial Genes. *Nature Microbiology* **2**, 1-6.

Zamani-Dahaj, S. A., M. Okasha, J. Kosakowski et P. Higgs (2016). Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Molecular Biology and Evolution* **33**, 1843-1857.

Chapitre 5



5

Discussion

Sommaire

Cette discussion me permet dans un premier temps de tirer les conclusions des deux parties distinctes de ma thèse. Dans un second temps, je présente les développements prévus à court terme pour la deuxième partie de mon travail de thèse, ainsi que les perspectives inspirées par ce travail concernant les dynamiques d'échange de gènes entre bactéries.

5.1 Conclusion de la Partie I : Impacts évolutifs des goulots d'étranglement de population

5.1.1 Chemins évolutifs contraints

Grâce à notre modèle pour les goulots d'étranglement périodiques présenté dans le Chapitre 2, nous avons montré que la démographie de la population contraint les chemins évolutifs empruntés : une faible dilution et un cycle bref favorisent l'adaptation par la mutation bénéfique la plus fréquente, tandis qu'une forte dilution et un cycle long favorisent l'adaptation par la mutation la plus avantageuse. Nous avons aussi prédit le nombre de dilutions nécessaires pour voir apparaître des doubles mutants, en fonction de paramètres démographiques (abondance minimale et maximale au cours d'un cycle). Tandis qu'augmenter l'abondance maximale permet toujours d'obtenir des doubles mutants plus tôt, augmenter l'abondance minimale a un effet non-monotone sur leur temps d'apparition. Il existe un ratio de dilution optimal permettant de maximiser la vitesse d'adaptation, que nous avons calculé. Nos prédictions éclairent la conception d'expériences d'évolution, précisent dans quelles conditions différents types de mutations conférant l'antibiorésistance évoluent et plus généralement démontrent le rôle clé de la démographie sur l'adaptation des populations microbiennes.

5.1.2 Lien avec la distribution de Luria-Delbrück

Les calculs que nous avons effectués pour calculer le nombre de mutants à la fin d'un cycle de croissance nous placent dans la longue lignée des travaux portant sur la distribution de Luria-Delbrück (Zheng 1999). Ce modèle extrêmement simple – une population en croissance exponentielle produisant un seul type de mutants – induit une distribution du nombre de mutants à un temps t fixé assez compliquée à exprimer analytiquement (Angerer 2001). Son utilité pour comprendre l'évolution microbienne et la croissance des tumeurs a mené à de nombreux développements. Plusieurs formulations du modèle ont été étudiées, différant par leur degré de stochasticité : la première formulation de Luria et Delbrück (1943) prend en compte uniquement la stochasticité des mutations (la souche initiale ainsi que les clones mutants ont une croissance déterministe), celle de Lea et Coulson (1949) introduit une dynamique stochastique pour les mutants, et enfin celle de Bartlett (1978) est entièrement stochastique (dynamique de la souche initiale et des mutants). La manière la plus pratique de calculer la distribution de Luria-Delbrück est d'utiliser un

algorithme qui permet de calculer les termes successifs par récurrence (Ma, Sandri et Sarkar 1992).

Plusieurs travaux ont étudié différents comportements limites de cette distribution, notamment dans le cas où la taille de population est grande et le taux de mutation petit. P. Keller et Antal (2014) notent B le nombre de mutants au temps auquel la souche initiale atteint une taille N, μ le taux de mutation et $\theta = N\mu$. Ils étudient les limites successives

$$\lim_{\substack{N\to\infty,\mu\to 0\\ \theta \text{ const.}}} B =: V \qquad \text{puis} \qquad \lim_{\theta\to\infty} \frac{V}{a} =: Z, \tag{5.1}$$

où a est un facteur d'échelle dépendant de θ assurant la convergence en loi, que les auteurs trouvent égal à $\theta^{\frac{1}{\gamma}}$ (γ étant le ratio du taux de croissance de la souche initiale sur celui du mutant). Cela signifie que dans la limite où le taux de mutation est faible et l'influx mutationnel est grand, le nombre de mutants est équivalent au facteur d'échelle multiplié par une quantité distribuée d'ordre 1.

Gerrish (2008) a étudié la distribution du nombre de mutants après une dilution (cela permettant d'améliorer la précision d'une méthode de calcul du taux de mutation, Crane, S. M. Thomas et Jones 1996). Dans le Chapitre 2, nous étudions l'évolution d'une population sur plusieurs cycles de croissance-dilution dans le cas où θ est grand. Nous avions donc uniquement besoin de calculer la valeur du facteur d'échelle pour connaître l'ordre de grandeur du nombre de mutants après la première dilution ainsi qu'aux cycles suivants. Dans notre cas $N = n^\alpha$ et $\mu = 1/n$, nous retrouvons donc comme P. Keller et Antal (2014) – mais par une méthode plus directe, voir Appendice 2.5.10 – un facteur d'échelle égal à $n^{(\alpha-1)\frac{r_{10}}{r_{00}}}$ (avec r_{10} le taux de croissance de notre mutant d'intérêt et r_{00} celui de la souche initiale). Les prédictions faites en utilisant uniquement le facteur d'échelle devant la distribution Z – qui est à queue épaisse avec tous ses moments infinis – sont probablement un peu grossières, mais les simulations que nous avons réalisées montrent tout de même une bonne adéquation avec nos prédictions pour $n \geq 10^9$.

5.1.3 Recommandations expérimentales

Sur la base du modèle que nous avons développé, nous recommandons un taux de dilution optimal permettant de maximiser la vitesse d'adaptation (Sections 2.4.3 et 2.5.10). Cependant, ce raisonnement a été fait en supposant que le volume de culture et la durée d'un cycle étaient fixés. Cela est généralement le cas, pour des raisons d'ordre pratique : on dispose d'un certain espace et type de matériel pour faire l'expérience, et il est pratique de devoir faire les dilutions chaque jour à heure fixe (donc, un cycle de 24h).

Si au contraire on laisse tous les paramètres expérimentaux varier (durée du cycle, taux de dilution, abondance maximale de la souche initiale), et que l'on cherche à minimiser le temps d'apparition des doubles mutants et non plus leur cycle d'apparition, les conclusions sont assez différentes. Faisons un calcul grossier. D'après la Section 2.5.8, les doubles mutants apparaissent à partir des mutants '10'

au cycle k,

$$k \approx 1 + \frac{\delta - \frac{r_{10}}{r_{00}}(\alpha - 1)}{(\frac{r_{10}}{r_{00}} - 1)(\alpha - \beta)} \tag{5.2}$$

Ils apparaissent donc à un temps t_{11} approximativement égal à

$$t_{11} \approx kt_n \approx \frac{\ln n}{r_{00}} \left(\alpha - \beta + \frac{\delta - \frac{r_{10}}{r_{00}}(\alpha - 1)}{\frac{r_{10}}{r_{00}} - 1} \right) \tag{5.3}$$

où t_n est la durée d'un cycle, n est l'inverse du taux des mutations faiblement avantageuses, et $-\delta$ est le logarithme en base n du taux des mutations fortement avantageuses. Ces paramètres sont des paramètres démographiques qu'on suppose fixés. D'autre part les paramètres expérimentaux sur lesquels on peut jouer sont α, le logarithme de l'abondance maximale de la souche initiale, et $\beta - \alpha$, le logarithme du taux de dilution. Pour minimiser le temps d'apparition t_{11} et donc la vitesse d'adaptation par unité de temps, il faudrait alors maximiser l'abondance maximale ainsi que le taux de dilution. Le fait de considérer également la production de doubles mutants à partir des simples mutants '01' produit le même résultat.

Ces recommandations correspondraient à cultiver la population microbienne dans la cuve la plus grande possible, avec des dilutions les plus faibles et les plus rapprochées possibles. Autrement dit, il faudrait utiliser un très grand chémostat pour maximiser la vitesse d'adaptation ! Cette conclusion semble intuitive : une grande population permet un afflux de mutations plus important, et se passer des dilutions permet d'éviter de perdre des mutations bénéfiques. Malgré le calcul rapide présenté ci-dessus, affirmer que notre modèle prédit une évolution plus rapide en chémostat serait une extrapolation un peu trop grande. En effet, notre modèle ne prend pas en compte explicitement la quantité de ressource disponible et suppose qu'elle est toujours en abondance grâce au renouvellement du milieu lors des dilutions. Lorsque le taux de dilution tend vers 1, la quantité de milieu nutritif ajoutée au moment du transfert tend vers 0 et ces conditions ne sont plus vérifiées.

Cependant, une étude ultérieure a démontré que le taux de dilution optimal était effectivement de 1 en utilisant des simulations avec modélisation explicite de la ressource (Delaney, Letten et Engelstädter 2023). Les auteurs de cette étude ont également vérifié que le taux d'évolution (calculé comme le taux de mutations bénéfiques qui vont atteindre la fixation) simulé dans le cadre d'un chémostat prolonge continûment celui observé dans les simulations de passages en série lorsque le taux de dilution tend vers 1.

Bien sûr, ces recommandations se heurtent à des considérations pratiques. Premièrement l'espace disponible, d'autant plus qu'il faut en général réaliser en parallèle plusieurs réplicats de l'expérience. Deuxièmement l'utilisation d'un chémostat, qui est apparemment plus délicat à utiliser que la méthode des passages en série (Gresham et Dunham 2014). À l'équilibre, le taux de réplication doit être égal au flux sortant (lui-même égal au flux entrant), une question essentielle est donc de savoir si il est expérimentalement possible d'avoir un flux assez élevé pour que la population se réplique à la même vitesse qu'en phase exponentielle. Dans tous les

cas, l'accélération par un facteur 10 à 100 prédite par Delaney, Letten et Engelstädter (2023) en passant d'un protocole standard de passages en série à un chémostat pourrait augmenter l'intérêt pour cette technique de culture.

5.1.4 Intérêt pour la génomique de la conservation

Dans le Chapitre 3, nous avons passé en revue les différents mécanismes par lesquels un goulot d'étranglement affecte le potentiel adaptatif d'une population. Ces différents mécanismes sont résumés Figure 3.1. Tout au long de cette revue nous comparons ce qui est prédit et observé dans les expériences microbiennes, à ce qui est prédit et observé chez les populations animales sauvages. L'optique était d'évaluer à quel point les expériences microbiennes peuvent informer la génétique de la conservation, et de formuler des recommandations pour les rendre plus pertinentes.

À l'issue de cette étude, nos recommandations principales sont d'effectuer des expériences avec des eucaryotes unicellulaires tels que la levure *Saccharomyces cerevisiae*, et d'inclure dans les expériences de la diversité génétique initiale au lieu de débuter avec une souche unique, comme préconisé par Burke (2023). Cela permettrait de réaliser des expériences à plus grande échelle (en termes de nombre d'individus et de rapidité d'évolution) qu'avec des animaux tels que *Drosophila melanogaster*, tout en conservant des caractéristiques fondamentales des animaux que sont la reproduction sexuée et la possibilité de s'adapter à partir de la diversité génétique existante.

Nous mettons ainsi en lumière l'importance du dialogue entre ces deux champs d'étude, dans la continuité de Alexander et al. (2014).

5.2 Conclusion de la Partie II : Évolution des pangénomes bactériens

5.2.1 Le modèle PPM pour la dynamique des gènes bactériens

Le Chapitre 4 présente le modèle Persistant-Privé-Mobile (PPM) que nous avons développé pour étudier les différents modes de propagation des gènes dans une espèce bactérienne. Nous avons fait le choix d'un modèle parcimonieux et explicable, qui décrit trois types de dynamiques de gènes le long d'une phylogénie : les gènes persistants sont présents dans le génome ancestral et rarement perdus, les gènes privés sont importés une seule fois et ne subissent aucun transfert dans l'espèce focale (ils sont spécifiques à un clade), et les gènes mobiles subissent de nombreux transferts suite à leur arrivée dans le pool génétique de l'espèce focale. J'ai implémenté une méthode d'inférence afin de pouvoir calculer les paramètres du modèle maximisant la probabilité de données génomiques observées, et appliqué cette méthode à un premier jeu de données composé de 902 génomes de *Salmonella enterica*, une espèce pathogène majeure de l'homme. Le modèle PPM est capable de reproduire les tendances générales de statistiques multivariées importantes des pangénomes,

et permet une classification des gènes selon leur dynamique la plus probable. Ces deux points sont discutés dans les deux sections suivantes.

5.2.2 Représentation des données pangénomiques

Un enjeu important dans l'étude des pangénomes est la représentation des données. En particulier, concevoir des statistiques multivariées décrivant les données permet d'évaluer la pertinence de différents modèles. C'est donc un besoin vital au stade actuel, où la question de savoir quelles sont les forces principales régissant la distribution des gènes dans une espèce bactérienne est loin d'être tranchée (voir Section 5.4.1). Ce besoin a notamment été évoqué par Baumdicker et Kupczok (2023).

Les premières statistiques utilisées pour décrire un pangénome bactérien sont apparues en même temps que le terme « pangénome » dans Tettelin, Masignani et al. (2005) : ce sont les courbes de raréfaction du génome *core* et d'enrichissement du pangénome, qui ont ensuite été couramment utilisées dans les analyses pangénomiques. Quelques années plus tard, Lapierre et Gogarten (2009) remarquent que l'information contenue dans ces courbes est exactement la même que dans la distribution des fréquences de gènes (appelée parfois GFS pour *gene frequency spectrum*) : on peut reconstruire l'une à partir de l'autre et inversement. Ils recommandent donc l'utilisation de cette dernière, car le tracé des courbes de raréfaction demande d'énumérer toutes les façons d'ordonner n génomes avec n la taille du jeu de données, ce qui représente $n!$ possibilités. Cependant seule la courbe moyenne est équivalente au GFS, il est donc tout de même possible de tracer les courbes pour un petit échantillon des $n!$ possibilités pour avoir un aperçu de la variabilité.

Par la suite, le GFS a été utilisé comme statistique permettant de calibrer les modèles d'évolution de pangénome (Baumdicker, Hess et Pfaffelhuber 2012 ; Haegeman et Weitz 2012). Lobkovsky, Wolf et Koonin (2013) s'en servent pour comparer différents modèles et rejeter le modèle supposant un même taux de renouvellement pour tous les gènes. Bolotin et Hershberg (2015) utilisent le GFS pour comparer les processus à l'œuvre dans des espèces bactériennes clonales (c'est-à-dire qui recombinent très peu, comme *Yersinia pestis* ou *Mycobacterium tubercolosis*) par rapport à des espèces recombinantes – représentant la majorité des espèces étudiées –, les premières n'ayant presque aucun gène à basse fréquence.

Graphe de parcimonie en fonction de la fréquence. Nous avons introduit une nouvelle statistique permettant de représenter la répartition des gènes dans un ensemble de génomes : le graphe de parcimonie en fonction de la fréquence. Chaque répartition est représentée par un point, dont l'abscisse correspond à la fréquence du gène et l'ordonnée au nombre minimum de gains et pertes nécessaires le long de l'arbre de parenté des génomes pour générer cette répartition. C'est une statistique en deux dimensions, qui contient donc plus d'information que le GFS (celui-ci peut être retrouvé en projetant le nuage de points sur l'axe des abscisses). Elle peut être utilisée pour comparer visuellement les pangénomes de différentes populations, ou estimer l'adéquation d'un modèle comme nous le faisons. Pour calculer un écart

entre des répartitions observées et simulées, on pourrait imaginer calculer à la fois l'écart entre les distributions de fréquences et celui entre les distributions de scores de parcimonie (obtenu en projetant le nuage de point sur l'axe des ordonnées). Elle a l'inconvénient de reposer sur la connaissance de la phylogénie de l'échantillon, qui peut être sujette à des erreurs de reconstruction et même à différentes interprétations (voir Section 5.4.2). Il faudrait tester la sensibilité de ce graphe aux variations de topologie de l'arbre. Même en dehors de ces considérations, on peut simplement voir le score de parcimonie d'un gène comme indiquant si le gène est préférentiellement présent dans des génomes proches génétiquement ou pas.

Lors des études préliminaires pour ce chapitre, nous avions aussi travaillé sur une autre représentation en deux dimensions de l'ensemble des répartitions de gènes. Celle-ci était basée sur une mesure de distance entre deux répartitions prenant en compte leur dissimilarité corrigée par les relations phylogénétiques entre génomes (Figure 4.14). Cette représentation était moins interprétable que le graphe de parcimonie en fonction de la fréquence, mais a confirmé l'importance de ces deux statistiques (parcimonie et fréquence) pour caractériser les données.

Au-delà des statistiques permettant de décrire les répartitions des gènes dans un pangénome, se pose aussi la question de représenter un pangénome cette fois du point de vue des génomes : comment représenter un ensemble de génomes en tenant compte de l'ordre des gènes sur un chromosome, de la présence de plasmides, de la grande variabilité en contenu génique ? Une méthode intuitive consiste à utiliser un graphe, dont les nœuds représentent les gènes. Deux gènes sont reliés par une arête s'ils se trouvent côte à côte dans au moins un génome du jeu de données. C'est par exemple la méthode proposée par l'outil PPanGGOLiN (Figure 5.1). Dans le but de représenter les points chauds d'insertion des gènes accessoires le long du chromosome, nous avons imaginé la Figure 4.6b. Elle résulte d'une méthode similaire à celle établie par Oliveira, Touchon, Cury et al. (2017), où les gènes *core* servent de repères pour définir des intervalles le long du chromosome qui sont communs à tous les génomes de l'échantillon. Ces deux types de représentation peuvent sûrement être encore améliorées en terme d'interprétabilité, en particulier pour de grands jeux de données. Par ailleurs, Wolf et al. (2016) ont utilisé des statistiques centrées sur les génomes et non sur les gènes pour ajuster et comparer plusieurs modèles : deux courbes montrant respectivement la similarité de contenu génique et la similarité de l'ordre des gènes en fonction de la distance phylogénétique.

5.2.3 Classification des gènes d'un pangénome

Comme expliqué dans la Section 4.4.1, la classification des gènes d'un pangénome a longtemps reposé sur leur fréquence. Un atout majeur du modèle PPM est de pouvoir réaliser une classification informée par la dynamique supposée des gènes. Une étude des fonctions des gènes assignés à chacune des trois catégories a apporté d'importants éléments de validation du modèle en montrant que ces catégories semblaient pertinentes du point de vue fonctionnel et mécanistique. Cette

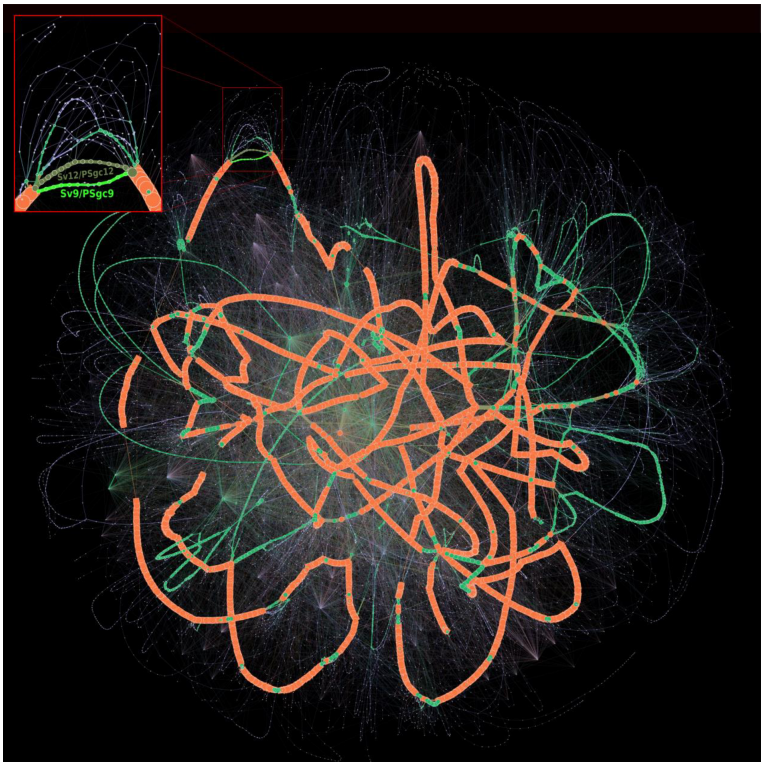


Figure 5.1 – Graphe du pangénome de 3117 génomes d'*Acinetobacter baumannii*, réalisé avec l'outil PPanGGOLiN. L'épaisseur d'une arête correspond au nombre de génomes où l'on retrouve ces deux gènes côte à côte, et l'épaisseur d'un nœud à sa fréquence. Nœuds et arêtes sont colorées en fonction de l'appartenance des gènes aux catégories *persistent* (orange), *shell* (vert) et *cloud* (bleu) définies par l'outil (figure tirée de Gautreau et al. 2020).

étude a notamment montré que la catégorie Persistante contient une surreprésentation de gènes portant des fonctions essentielles (traduction de l'ADN, métabolisme, contrôle du cycle cellulaire), tandis que les gènes Mobiles comportent une surreprésentation de transposons, prophages et mécanismes de défense, qui sont en effet bien connus pour se transmettre par transferts horizontaux. Les fonctions surreprésentées dans la catégorie Privée suggèrent quant à elles des mécanismes d'adaptation à différents types de nutriments et d'environnements (métabolisme des glucides, structures extracellulaires).

Réciproquement, une classification des gènes sur la base de leur dynamique peut apporter des indices sur le mode de diffusion de gènes peu connus. Cela nous a aussi permis d'étudier la répartition des différents types de gènes le long du chromosome et sur les plasmides. Grâce à cela, nous avons identifié les points chauds d'insertion des gènes accessoires (Privés et Mobiles) le long du chromosome de *Salmonella enterica*. De plus, la classification informée par la dynamique a potentiellement un intérêt important en pangénomique comparative, pour donner un aperçu facile à interpréter des types de dynamiques à l'œuvre dans différentes populations ou espèces.

5.3 Développements prévus

5.3.1 Finalisation de l'outil d'inférence pour le modèle PPM

Notre modèle d'évolution de pangénomes bactériens est accompagné d'un outil d'inférence permettant de calculer les paramètres maximisant la vraisemblance sur des données génomiques, puis de catégoriser les gènes selon leur dynamique la plus probable. Cet outil pourrait être très utile dans l'analyse comparative de pangénomes bactériens, la classification de gènes et de manière générale la compréhension des mécanismes d'évolution des pangénomes. Cependant, son utilisation par d'autres équipes de recherche ne sera possible que si nous mettons à disposition un outil facile à manier et efficace en terme de temps de calcul. Actuellement il faut une dizaine d'heure sur 50 cœurs pour estimer les paramètres sur un ensemble de 902 génomes dont le pangénome contient environ 46 000 familles, et il faut relancer plusieurs fois le processus pour être sûr d'atteindre l'optimum global. Je pense que ces performances peuvent être améliorées, et projette donc de travailler à une optimisation supplémentaire du code. La première mesure que je compte prendre est d'utiliser un profiler sur le code pour identifier les pistes d'amélioration.

Ensuite, il serait intéressant d'améliorer les méthodes que nous avons déjà commencé à mettre en place pour trouver un bon point de départ à l'optimisation (Section 4.5.3), car un bon point de départ permet de réduire le temps d'exploration de l'espace des paramètres et augmente les chances de trouver l'optimum global. La solution consistant à faire des optimisations préalables sur des sous-ensembles des données et à prendre la médiane des paramètres inférés (pour ne pas donner trop de poids aux valeurs aberrantes) semble prometteuse, mais d'autres tests sont nécessaires pour déterminer combien de sous-ensembles sont nécessaires et de quelle taille.

D'autre part, j'aimerais implémenter des méthodes existantes permettant d'accélérer le calcul de vraisemblance le long d'un arbre. Une méthode couramment utilisée consiste à pré-calculer la vraisemblance aux cerises de l'arbre (nœuds supportant deux feuilles) pour toutes les combinaisons possibles d'états aux feuilles. Nous utilisons déjà l'optimisation consistant à traiter une seule fois les gènes ayant une répartition identique à toutes les feuilles de l'arbre. Cependant, une méthode plus avancée introduite par Kobert, Stamatakis et Flouri (2017) permet de détecter également les gènes ayant une répartition identique aux feuilles d'un sous-arbre, afin de ne réaliser qu'une seule fois les calculs de vraisemblance pour ce sous-arbre.

Le but est ensuite de proposer notre méthode d'inférence sous la forme d'un logiciel en ligne de commande ainsi que d'un package R (interfacé avec le code C++) bien documentés, afin de la rendre accessible au plus grand nombre.

5.3.2 Étude de clades récents d'*Escherichia coli*

Nous avons l'intention d'appliquer notre méthode d'inférence à d'autres jeux de données, notamment des ensembles de génomes d'*Escherichia coli* d'un même

clone ou « sequence type » (ST). Cette étude se fera en collaboration avec l'équipe d'Olivier Tenaillon à l'hôpital Cochin, qui dispose d'une base de données nettoyée de 60 000 génomes d'*E. coli* (Vigué et al. 2022). Cela nous permettra de mieux comprendre les origines et l'évolution du pangénome d'un clone, et de comparer les taux de transferts horizontaux de différents STs. Une des questions que nous nous posons concerne les différences potentielles en termes de dynamique de pangénome entre des clones résistants et virulents comme *E. coli* ST131 – qui posent un problème de santé publique majeur – et les autres. L'étude Cummins et al. (2022) a ouvert la voie sur ce sujet en étudiant la composition du pangénome d'une vingtaine de STs, et a identifié des différences entre les fonctions présentes dans le génome *core* des ST131 et ST10 (un ST généraliste) par rapport aux autres.

D'autre part, le fait d'utiliser des séquences de clones ayant récemment émergé nous permettra de tester une autre fonctionnalité de notre modèle. Comme discuté Section 4.6.3, le fait que le modèle PPM inclue de manière explicite le temps d'introduction des gènes mobiles devrait permettre d'inférer ces dates. Cependant, notre étude préliminaire sur des données simulées avec les paramètres du jeu de données *Salmonella enterica* a montré qu'il n'y avait pas de signal dans ces données pour inférer cette variable. Cela est probablement dû au fait que la profondeur de l'arbre est trop importante par rapport au temps caractéristique de renouvellement des gènes mobiles. Ainsi, nous avons de bonnes raisons de penser que notre modèle serait capable d'inférer le temps d'arrivée des gènes mobiles le long de la phylogénie d'un clade récent. Cela permettrait donc de retracer le processus de constitution du pangénome d'un clade au cours du temps.

5.3.3 Pistes d'amélioration du modèle PPM

Après l'étude du modèle PPM présentée Chapitre 4, nous nous sommes demandé comment déterminer les pistes d'amélioration les plus prometteuses. Une idée est d'étudier les gènes dont les répartitions sont les plus mal prises en compte, c'est-à-dire celles ayant une faible probabilité d'avoir été générée par le modèle. Nous pourrions étudier deux aspects : est-ce qu'on retrouve parmi ces gènes des répartitions dont ont sait qu'elles sont mal prises en compte (par exemple un gène présent à haute fréquence dans plusieurs clades assez éloignés, ou un gène présent à haute fréquence dans un clade et sporadiquement en dehors) ? Et est-ce que ces gènes encodent des fonctions qui suggèrent un avantage sélectif particulièrement important, ou pointent vers un mécanisme de transfert particulier ?

5.4 Perspectives

5.4.1 Les gènes accessoires, sélectionnés ou non ?

Personne ne conteste a priori le fait que certain gènes accessoires bactériens confèrent un avantage à leur porteuse, au vu des fonctions encodées par une partie de ces gènes (par exemple des mécanismes de défense). Dans une revue sur

les liens entre transferts horizontaux et adaptation chez les bactéries, B. J. Arnold, Huang et Hanage (2022) distinguent plusieurs mécanismes de sélection, et expliquent notamment que dans certains cas le transfert est le seul moyen pour un gène avantageux de se fixer dans une population. En effet, en présence de sélection balancée très forte sur un locus, un gène avantageux ne pourra se fixer dans toute la population qu'en parvenant à s'insérer dans des génomes comportant des allèles différents à ce locus (Takeuchi et al. 2015). Du polymorphisme peut être protégé en cas de sélection fréquence-dépendante négative (Corander et al. 2017 ; Azarian et al. 2020) par exemple due à la prédation par les phages (Rodriguez-Valera et al. 2009), ou parce que la population étudiée recouvre plusieurs niche écologiques.

Le débat entre neutralistes et sélectionnistes consiste donc à discuter l'importance de la sélection dans la dynamique des gènes accessoires, de manière similaire à celui concernant les mutations ponctuelles. On peut cependant arguer qu'il existe une différence entre un gène acquis par transfert horizontal et une mutation ponctuelle : a priori le gène vient d'une bactérie donneuse présente dans le même environnement et chez qui ce gène n'avait pas encore été perdu, ce n'est donc pas une mutation aléatoire parmi toutes celles possibles. Ce débat a notamment vu l'utilisation d'arguments invoquant la taille efficace par chacun des partis. En effet, la taille efficace des populations bactériennes est positivement corrélée à la taille de leur pangénome. Andreani, Hesse et Vos (2017) interprètent cette corrélation comme un signe d'évolution neutre, faisant un parallèle entre diversité du pangénome et diversité nucléotidique. McInerney, McNally et O'Connell (2017) y voient plutôt un signe que la plupart des gènes accessoires sont avantageux, puisque les populations à grande taille efficace sont censées mieux répondre à la sélection naturelle. Shapiro (2017) met en garde contre l'utilisation du concept de taille efficace pour les pangénomes, étant donné qu'il n'est pas clairement défini pour un gène présent dans plusieurs espèces. Douglas et Shapiro (2024) proposent de prendre les pseudogènes comme référence de neutralité, et trouvent qu'ils présentent des annotations fonctionnelles significativement différentes des autres gènes accessoires, suggérant qu'une proportion importante des ces derniers est sous sélection positive.

Le GFS a également été utilisé pour montrer sa compatibilité (Haegeman et Weitz 2012) ou son incompatibilité (Lobkovsky, Wolf et Koonin 2013) avec un modèle neutre. B. J. Arnold, Huang et Hanage (2022) suggèrent que prendre en compte d'autres aspects des données pourrait permettre de mieux discriminer entre différents modèles, comme discuté en Section 5.2.2. Ils proposent d'autres options, par exemple des séries temporelles pour étudier les variations de fréquences géniques dans un environnement donné, de prendre en compte la liaison génétique, les mutations synonymes et non-synonymes pour détecter la sélection intra-gène, les données écologiques et enfin de prendre en compte le domaine bactérien dans son entièreté. Par exemple la base de données MicrobeAtlas rend exploitable les informations concernant le milieu de prélèvement des génomes procaryotes archivés par le NCBI, ce qui a permis de mettre en lumière l'importance de différent facteurs sur les taux d'échange de gènes entre espèces bactériennes : co-occurrence dans le même milieu, abondance et dispersion géographique des espèces (Dmitrijeva et al. 2024).

De notre côté, nous avons réfléchi à plusieurs options pour incorporer l'information contenue dans les séquences des gènes accessoires à notre modèle. Calculer une vraisemblance prenant en compte en plus de la répartition des gènes le contenu de leurs séquences nécessite de considérer le couplage entre leur phylogénie et celle de l'espèce. Pour chaque gène et chaque jeu de paramètres (taux d'immigration, gain et perte) il est possible de simuler des arbres de gènes correspondant à notre modèle, conditionnellement à la configuration de présence/absence du gène aux feuilles de l'arbre d'espèce. Nous pourrions ensuite confronter ces arbres simulés à l'information contenue dans les séquences, et ce de deux manières différentes : 1) en comparant la distribution d'arbres simulés à la distribution postérieure d'arbres du gène inférée de manière bayésienne à partir des séquences, ou 2) en calculant la vraisemblance des séquences conditionnellement aux arbres simulés. Dans les deux cas, il serait ensuite possible d'optimiser sur les paramètres afin de trouver le meilleur couplage pour chacun des gènes. Le défi serait alors de trouver des méthodes économes en temps de calcul pour réaliser ces optimisations.

5.4.2 Que représentent les phylogénies bactériennes ?

Comme mentionné précédemment (Sections 1.1.3 et 4.4.2), il existe aussi un débat concernant la signification des phylogénie bactériennes reconstruites à partir d'alignements de séquences. Tandis que ces phylogénies étaient traditionnellement interprétées comme représentant des relations clonales entre individus (à quelques erreurs de reconstruction près), l'étude de Sakoparnig, Field et van Nimwegen (2021) avance des arguments suggérant que ces phylogénies refléteraient en fait des relations de recombinaison plus ou moins préférentielle entre individus. Ils montrent notamment par des simulations que pour des ratio taux de recombinaison sur taux de mutation comparables à ceux inférés chez de nombreuses espèces bactériennes recombinantes, le signal porté par les mutations héritées de manière clonale est en grande partie effacé par les recombinaisons, en particulier pour les branches les plus profondes de l'arbre. D'autre part, ils étudient la distribution des fréquences de SNPs partagés par exactement deux souches parmi un ensemble de 92 génomes d'*E. coli*, comme un moyen d'étudier la fréquence de recombinaison entre les différentes lignées. Ils trouvent que cette distribution suivant approximativement une loi de puissance n'est compatible ni avec une hypothèse de structure clonale avec un faible « bruit » dû à la recombinaison, ni avec une hypothèse de recombinaison libre entre toutes les lignées, et serait insuffisamment expliquée par une hypothèse de recombinaison préférentielle entre lignées apparentées.

Quelle que soit l'interprétation mécanistique la plus juste (qui a de grande chance de varier selon l'espèce considérée), il n'en demeure pas moins que ces phylogénies forment un regroupement hiérarchique des individus en fonction de leur proximité génomique et ont à ce titre une utilité dans les analyses pangénomiques. Cela a été évoqué plus tôt en Section 5.2.2 : des statistiques comme la parcimonie des répartitions ou la forme de la relation entre similarité du contenu génique et distance phylogénétique capturent un signal intéressant présent dans les données

(Touchon, Cury et al. 2014 ; Wolf et al. 2016), même si l'on ne sait pas ce qui l'a produit .

Il serait cependant extrêmement intéressant de pouvoir déterminer les facteurs principaux qui forment ces groupes de similarité génomique. Est-ce que ce sont les relations clonales, la recombinaison préférentielle, ou est-ce que les deux se recoupent largement (recombinaison préférentielle entre génomes apparentés) ? D'autres facteurs pouvant expliquer une recombinaison préférentielle avec des lignées plus éloignées sont la coexistence dans un même environnement, ou des événements contingents tels que l'acquisition d'un système de restriction-modification (pouvant favoriser la recombinaison homologue avec les lignées portant un système apparenté, Oliveira, Touchon et Rocha 2016).

5.4.3 Modèles couplés

Les deux points discutés précédemment questionnent tous deux les facteurs influençant la dynamique d'échange de gènes entre bactéries – en ce qui concerne les transferts de gènes accessoires pour le premier, et la recombinaison homologue de gènes *core* pour le second. Dans cette section, je passe en revue certaines idées de modélisation qui pourraient permettre de répondre à ces questions. En effet, l'identification et la quantification de ces différents facteurs passent a priori par l'exploitation de nouvelles dimensions des données et aussi par des modèles capables de coupler différents mécanismes.

De tels modèles existent déjà. On peut notamment citer un modèle couplant évolution moléculaire du génome *core* et processus de gain et perte de gènes accessoires (Marttinen et al. 2015), un modèle d'échange de gènes à l'échelle de communautés bactériennes (Niehus et al. 2015), et des modèles phylodynamiques destinés aux bactéries pathogènes, couplant l'évolution moléculaire avec la démographie des infections et ayant le potentiel d'intégrer des métadonnées telles que la présence de résistance aux antibiotiques ou l'environnement du prélèvement (passés en revue par Ingle, Howden et Duchene 2021).

Un modèle phylo-écologique. Un modèle « phylo-écologique » couplant l'effet de la proximité génomique et de la niche écologique sur les échanges de gènes accessoires pourrait s'inspirer du modèle que nous avons développé pour les gènes mobiles. Pour rappel, ce modèle suppose que des gènes arrivent à taux constant dans le pool génétique de la population, et les gènes présents dans le pool peuvent être gagnés et perdus à taux constants le long des lignées de la population. On pourrait imaginer l'évolution de génomes (en terme de contenu génique) le long d'un arbre clonal divisé en clades ayant des pools génétiques distincts (pools de clade), avec un caractère discret de niche qui évolue aussi le long de la phylogénie et donne accès à des pools génétiques distincts (pools de niche, voir Figure 5.2). Les gènes sont gagnés le long des branches de la phylogénie. Une lignée qui gagne un gène l'ajoute au pool de son clade et de sa niche ; si elle change de niche elle ajoute tous ses gènes au pool de la nouvelle niche. Les lignées gagnent uniquement des gènes du pool de leur niche, et préférentiellement ceux qui sont aussi dans le pool

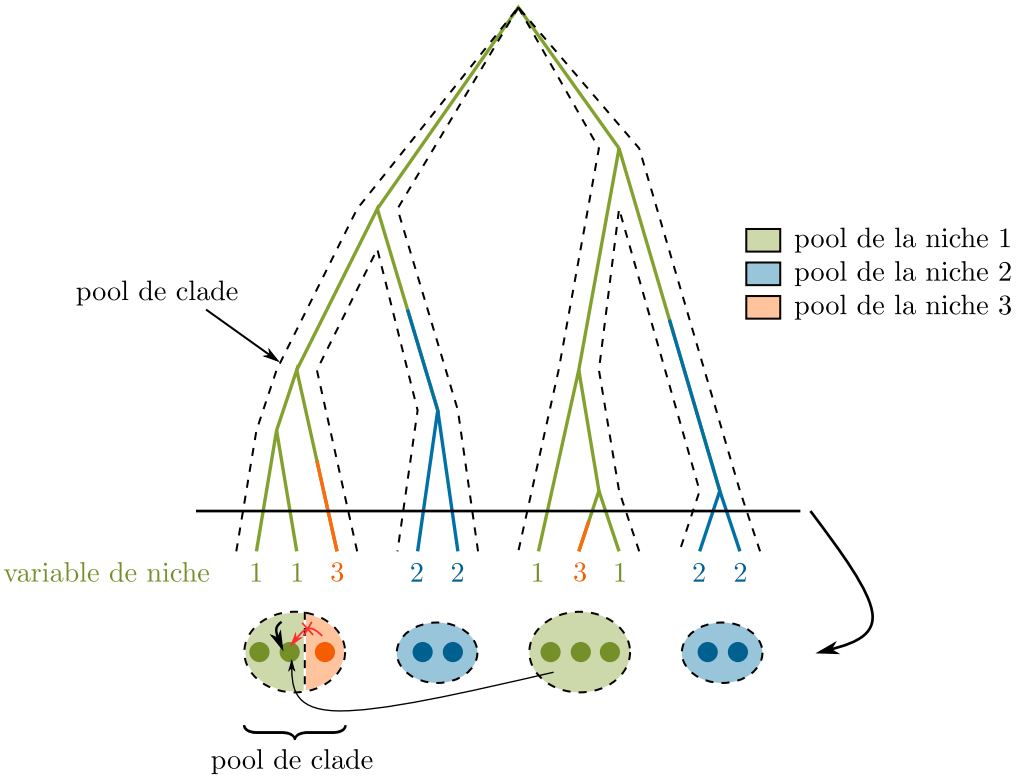


Figure 5.2 – Illustration d'une idée de modèle phylo-écologique. Les lignes colorées représentent des lignées bactériennes. Les couleurs (ainsi que les entiers de 1 à 3) représentent trois niches écologiques différentes entre lesquelles les lignées peuvent se déplacer. Les ovales colorés situés sous l'arbre illustrent les recoupements entre les pools génétiques propres aux clades (délimités par des lignes noires pointillées) et les pools génétiques propres aux niches (colorés en vert, bleu et orange selon la niche). Les possibilités de transferts de gènes sont illustrés pour la deuxième lignée en partant de la gauche : cette lignée peut recevoir uniquement des gènes issus du pool de la niche 1 (en vert). Elle reçoit préférentiellement les gènes se trouvant à la fois dans le pool de sa niche et le pool de son clade (phénomène illustré par l'épaisseur plus importante de cette flèche).

de leur clade. Les gènes pourraient être soit délétères, soit neutres, soit avantageux (dans une seule niche ou inconditionnellement), ce qui serait modélisé par des taux de perte différenciés. Ce modèle permettrait de simuler différents scénarios en faisant varier la subdivision en clades, la proportion de gènes de chaque catégories, et l'importance de l'échange préférentiel entre individus d'un même clade. Pour l'inférence sur des jeux de données il faudra disposer de métadonnées indiquant la niche écologique dans laquelle chaque génome a été prélevé, et déterminer une méthode d'inférence adaptée.

Ce genre de modèle (dont il reste à prouver qu'il soit facilement manipulable) pourrait montrer tout l'intérêt du modèle de pool génétique. Ce modèle est à la fois moins lourd et plus pertinent qu'un modèle considérant des échanges explicites entre lignées (car aussi grands que deviennent les jeux de données, les génomes séquencés ne représentent toujours qu'une infime partie des populations bacté-

riennes). Encore plus intéressant, il permet de modéliser explicitement la constitution d'un pool génétique de niche au fur et à mesure de l'arrivée de nouvelles lignées.

Un modèle permettant de trancher sur les facteurs influençant la formation de différents clusters génomiques au sein d'une espèce (les clades) paraît plus compliqué à mettre en place. Pour simuler des gènes *core* recombinant en fonction de la distance génomique entre individus, chaque échange nécessite de connaître l'identité des lignées donneuse et receveuse. On ne peut donc pas recourir à la méthode du pool génétique. Il faudrait alors maintenir en permanence une matrice de distance génomique entre toutes les lignées, en la mettant à jour à chaque événement de recombinaison et potentiellement à chaque mutation sur n'importe quelle lignée ! Ajouter des caractères discrets influant aussi sur les taux de recombinaison tels que la niche écologique ou la présence de système de restriction-modification ne semble pas trop compliqué. Cela paraît très lourd mais se simule peut-être à petite échelle. On pourrait alors simuler l'évolution du génome *core*, inférer un arbre à partir des données simulées et mesurer comment les différents paramètres remodèlent (ou pas) les clusters génomiques.

Références du Chapitre 5

Alexander, H. K., G. Martin, O. Y. Martin, O. Martin et S. Bonhoeffer (2014). Evolutionary Rescue: Linking Theory for Conservation and Medicine. *Evolutionary Applications* **7**, 1161-1179.

Andreani, N. A., E. Hesse et M. Vos (2017). Prokaryote Genome Fluidity Is Dependent on Effective Population Size. *The ISME Journal* **11**, 1719-1721.

Angerer, W. P. (2001). An Explicit Representation of the Luria–Delbrück Distribution. *Journal of Mathematical Biology* **42**, 145-174.

Arnold, B. J., I.-T. Huang et W. P. Hanage (2022). Horizontal Gene Transfer and Adaptive Evolution in Bacteria. *Nature Reviews Microbiology* **20**, 206-218.

Azarian, T., P. P. Martinez, B. J. Arnold, X. Qiu, L. R. Grant, J. Corander, C. Fraser, N. J. Croucher, L. L. Hammitt, R. Reid, M. Santosham, R. C. Weatherholtz, S. D. Bentley, K. L. O'Brien, M. Lipsitch et W. P. Hanage (2020). Frequency-Dependent Selection Can Forecast Evolution in Streptococcus Pneumoniae. *PLOS Biology* **18**, e3000878.

Bartlett, M. S. (1978). *An Introduction to Stochastic Processes: With Special Reference to Methods and Applications*. CUP Archive. 412 p.

Baumdicker, F., W. R. Hess et P. Pfaffelhuber (2012). The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biology and Evolution* **4**, 443-456.

Baumdicker, F. et A. Kupczok (2023). Tackling the Pangenome Dilemma Requires the Concerted Analysis of Multiple Population Genetic Processes. *Genome Biology and Evolution* **15**, evad067.

Bolotin, E. et R. Hershberg (2015). Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biology and Evolution* **7**, 2173-2187.

Burke, M. K. (2023). Embracing Complexity: Yeast Evolution Experiments Featuring Standing Genetic Variation. *Journal of Molecular Evolution*.

Corander, J., C. Fraser, M. U. Gutmann, B. Arnold, W. P. Hanage, S. D. Bentley, M. Lipsitch et N. J. Croucher (2017). Frequency-Dependent Selection in Vaccine-Associated Pneumococcal Population Dynamics. *Nature Ecology & Evolution* **1**, 1950-1960.

Crane, G. J., S. M. Thomas et M. E. Jones (1996). A Modified Luria-Delbrück Fluctuation Assay for Estimating and Comparing Mutation Rates. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **354**, 171-182.

Cummins, E. A., R. A. Hall, C. Chris, J. O. McInerney et A. McNally (2022). Distinct Evolutionary Trajectories in the Escherichia Coli Pangenome Occur within Sequence Types. *Microbial genomics* **8**.

Delaney, O., A. D. Letten et J. Engelstädter (2023). Frequent, Infinitesimal Bottlenecks Maximize the Rate of Microbial Adaptation. *Genetics* **225**, iyad185.

Dmitrijeva, M., J. Tackmann, J. F. Matias Rodrigues, J. Huerta-Cepas, L. P. Coelho et C. von Mering (2024). A Global Survey of Prokaryotic Genomes Reveals the Eco-Evolutionary Pressures Driving Horizontal Gene Transfer. *Nature Ecology & Evolution*, 1-13.

Douglas, G. M. et B. J. Shapiro (2024). Pseudogenes Act as a Neutral Reference for Detecting Selection in Prokaryotic Pangenomes. *Nature Ecology & Evolution* **8**, 304-314.

Gautreau, G., A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, A. Perrin, C. Médigue, A. Calteau, S. Cruveiller, C. Matias, C. Ambroise, E. P. C. Rocha et D. Vallenet (2020). PPanGGOLiN: Depicting Microbial Diversity via a Partitioned Pangenome Graph. *PLOS Computational Biology* **16**, e1007732.

Gerrish, P. J. (2008). A Simple Formula for Obtaining Markedly Improved Mutation Rate Estimates. *Genetics* **180**, 1773-1778.

Gresham, D. et M. J. Dunham (2014). The Enduring Utility of Continuous Culturing in Experimental Evolution. *Genomics*. Experimental Evolution and the Use of Genomics **104** (6, Part A), 399-405.

Haegeman, B. et J. S. Weitz (2012). A Neutral Theory of Genome Evolution and the Frequency Distribution of Genes. *BMC Genomics* **13**, 196-196.

Ingle, D. J., B. P. Howden et S. Duchene (2021). Development of Phylodynamic Methods for Bacterial Pathogens. *Trends in Microbiology* **29**, 788-797.

Keller, P. et T. Antal (2014). Mutant Number Distribution in an Exponentially Growing Population. *Journal of Statistical Mechanics: Theory and Experiment*.

Kobert, K., A. Stamatakis et T. Flouri (2017). Efficient Detection of Repeating Sites to Accelerate Phylogenetic Likelihood Calculations. *Systematic Biology* **66**, 205-217.

Lapierre, P. et J. P. Gogarten (2009). Estimating the Size of the Bacterial Pan-Genome. *Trends in Genetics* **25**, 107-110.

Lea, D. E. et C. A. Coulson (1949). The Distribution of the Numbers of Mutants in Bacterial Populations. *Journal of Genetics* **49**, 264-285.

Lobkovsky, A. E., Y. I. Wolf et E. V. Koonin (2013). Gene Frequency Distributions Reject a Neutral Model of Genome Evolution. *Genome Biology and Evolution* **5**, 233-242.

Luria, S. E. et M. Delbrück (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491-511.

Ma, W. T., G. vH Sandri et S. Sarkar (1992). Analysis of the Luria–Delbrück Distribution Using Discrete Convolution Powers. *Journal of Applied Probability* **29**, 255-267.

Marttinen, P., N. J. Croucher, M. U. Gutmann, J. Corander et W. P. Hanage (2015). Recombination Produces Coherent Bacterial Species Clusters in Both Core and Accessory Genomes. *Microbial Genomics* **1**, e000038.

McInerney, J. O., A. McNally et M. J. O'Connell (2017). Why Prokaryotes Have Pangenomes. *Nature microbiology* **2**, 17040.

Niehus, R., S. Mitri, A. G. Fletcher et K. R. Foster (2015). Migration and Horizontal Gene Transfer Divide Microbial Genomes into Multiple Niches. *Nature Communications* **6**, 8924.

Oliveira, P. H., M. Touchon, J. Cury et E. P. C. Rocha (2017). The Chromosomal Organization of Horizontal Gene Transfer in Bacteria. *Nature Communications* **8** (1), 841.

Oliveira, P. H., M. Touchon et E. P. C. Rocha (2016). Regulation of Genetic Flux between Bacteria by Restriction–Modification Systems. *Proceedings of the National Academy of Sciences* **113**, 5658-5663.

Rodriguez-Valera, F., A.-B. Martín-Cuadrado, B. Rodriguez-Brito, L. Pasic, T. F. Thingstad, F. Rohwer et A. Mira (2009). Explaining Microbial Population Genomics through Phage Predation. *Nature Precedings*, 1-1.

Sakoparnig, T., C. Field et E. van Nimwegen (2021). Whole Genome Phylogenies Reflect the Distributions of Recombination Rates for Many Bacterial Species. *eLife* **10**. Sous la dir. d'A. Nourmohammad et A. M. Walczak, e65366.

Shapiro, B. J. (2017). The Population Genetics of Pangenomes. *Nature microbiology* **2**, 1574-1574.

Takeuchi, N., O. X. Cordero, E. V. Koonin et K. Kaneko (2015). Gene-Specific Selective Sweeps in Bacteria and Archaea Caused by Negative Frequency-Dependent Selection. *BMC Biology* **13**, 20.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins, K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli et C. M. Fraser (2005). Genome Analysis of Multiple Pathogenic Isolates of Streptococcus

Agalactiae: Implications for the Microbial "Pan-Genome". *Proceedings of the National Academy of Sciences* **102**, 13950-13955.

Touchon, M., J. Cury, E.-J. Yoon, L. Krizova, G. C. Cerqueira, C. Murphy, M. Feldgarden, J. Wortman, D. Clermont, T. Lambert, C. Grillot-Courvalin, A. Nemec, P. Courvalin et E. P. Rocha (2014). The Genomic Diversification of the Whole Acinetobacter Genus: Origins, Mechanisms, and Consequences. *Genome Biology and Evolution* **6**, 2866-2882.

Vigué, L., G. Croce, M. Petitjean, E. Ruppé, O. Tenaillon et M. Weigt (2022). Deciphering Polymorphism in 61,157 Escherichia Coli Genomes via Epistatic Sequence Landscapes. *Nature Communications* **13**, 4030.

Wolf, Y. I., K. S. Makarova, A. E. Lobkovsky et E. V. Koonin (2016). Two Fundamentally Different Classes of Microbial Genes. *Nature Microbiology* **2**, 1-6.

Zheng, Q. (1999). Progress of a Half Century in the Study of the Luria–Delbrück Distribution. *Bellman Prize in Mathematical Biosciences* **162**, 1-32.