Escherichia coli diversity and evolution : perspectives from the study of 80,000 genomes

Lucile Vigué

Université Paris Cité

École Doctorale Bio Sorbonne Paris Cité (562)

Laboratoire "Infection, Anti-microbien, Modélisation, Évolution" (IAME), UMR_S 1137

Escherichia coli diversity and evolution: perspectives from the study of 80,000 genomes

Par Lucile VIGUÉ

Thèse de doctorat de Microbiologie

Dirigée par Olivier TENAILLON

Présentée et soutenue publiquement le 13 septembre 2023

Devant un jury composé de :

Laurent DURET Directeur de recherche, Université de Lyon 1, CNRS	*Rapporteur*
Isabel GORDO Principal Investigator, Institute Gulbenkian de Ciência	*Rapportrice*
Guillaume ACHAZ Professeur des universités, Université Paris Cité, Collège de France, CNRS	*Examinateur*
Nicolas BIERNE Directeur de recherche, Université de Montpellier, CNRS	*Examinateur*
Marie TOUCHON Chargée de recherche, Sorbonne Université, Institut Pasteur, CNRS	*Examinatrice*
Olivier TENAILLON Directeur de recherche, Université Paris Cité, INSERM	*Directeur de thèse*

Abstract / Résumé

Abstract

***Escherichia coli* diversity and evolution:**
perspectives from the study of 80,000 genomes

A commensal bacteria in the gut of humans and many vertebrates, *Escherichia coli* is also a deadly pathogen responsible for 950,000 deaths per year worldwide. As a generalist organism capable of adapting to different ecological niches, it is a species of choice for studying evolution on different time scales. Its status as a model organism in biology and its importance for human health have led to the sequencing of many strains worldwide. The aim of this thesis is to analyse the diversity present in 81,440 of these genomes and to understand how this diversity can inform us about the evolutionary processes at work in this species.

The 81,440 genomes collected cover the natural diversity of *Escherichia coli*. Strains isolated in humans and more specifically in a clinical context are largely represented. In particular, 11,000 of these genomes are *Shigella*, obligate pathogenic strains of primates that have adopted an intracellular lifestyle. To study these 81,440 genomes, I extracted the coding sequences and organised them in a database. A comparison of the core genomes of these strains allowed me to classify them into 240 clusters from which I was able to infer a global phylogeny of the species corrected for recombination.

In order to further analyse the mutational patterns, I used Direct-Coupling Analysis (DCA). This statistical physics approach allows to predict the effect of a mutation occurring in a gene and inducing an amino-acid change in the corresponding protein. By modelling the interactions between pairs of amino acids within the protein, DCA allows the genetic context in which the mutation occurs to be taken into account.

By applying DCA to thousands of *E. coli* core genes, I have shown that it can predict not only the native amino acids of this species but also the polymorphisms observed in it. DCA also predicts the probability of observing a mutation at a certain frequency. In doing so, it reveals differences in the efficiency of natural selection between different subpopulations of *E. coli*. In particular, natural selection appears to be much less effective in *Shigella* strains,

consistent with the reduced effective size of this population.

Genetic context was found to be key to the quality of the predictions made by DCA. This context is built up over long time scales by the addition of many weak interactions between amino acids. These do not affect all residues of a protein in the same way. DCA can predict the variability of these residues. In particular, between 30% and 50% of the sites in a protein are highly constrained by the genetic background of *E. coli*. A mutation at one of these sites will generally be deleterious if it occurs alone. These sites do not therefore tolerate polymorphisms. However, they can co-evolve over long time scales so that the amino acids observed there vary widely between species.

If individual residues of a protein can evolve at different rates, so can proteins. I have developed a selection test, based on the DCA, which allows genes to be compared with each other. In the short term, the essential genes are those under the strongest purifying selection pressure, while the level of expression determines the long-term rate of evolution. This test also detects inactivations of transcriptional factors, inactivations that appear to be selected in the short term but counter-selected in the longer term.

The present work demonstrates the interest of coupling the study of large genome databases with modelling approaches to understand the evolution of a species on different time scales.

Keywords *Escherichia coli, Population genetics, Recombination, Core genome, Polymorphism, Epistasis, Mutation effect prediction, Direct-Coupling Analysis, Inactivations.*

Résumé

Diversité et évolution d'*Escherichia coli* :
perspectives ouvertes par l'étude de 80 000 génomes

Bactérie commensale de l'intestin de l'homme et de nombreux vertébrés, *Escherichia coli* est aussi un pathogène mortel responsable de 950,000 morts par an dans le monde. Organisme généraliste capable de s'adapter à différentes niches écologiques, il s'agit d'une espèce de choix pour étudier l'évolution sur différentes échelles de temps. Son statut d'organisme modèle en biologie et son importance pour la santé humaine ont favorisé le séquençage de très nombreuses souches dans le monde entier. L'objectif de cette thèse est d'analyser la diversité présente dans 81 440 de ces génomes et de comprendre comment celle-ci peut nous informer sur les processus évolutifs à l'œuvre dans cette espèce.

Les 81 440 génomes rassemblés couvrent la diversité naturelle d'*Escherichia coli*. Les souches isolées chez l'humain et plus précisément dans un contexte clinique sont largement représentées. En particulier, 11 000 de ces génomes sont des *Shigella*, des souches pathogènes obligatoires des primates ayant adopté un mode de vie intra-cellulaire. Pour étudier ces 81 440 génomes, j'en ai extrait les séquences codantes que j'ai organisées dans une base de données. Une comparaison du core génome de ces souches m'a permis de les répartir en 240 clusters à partir desquels j'ai pu inférer une phylogénie globale de l'espèce corrigée pour la recombinaison.

Afin d'analyser plus en profondeur les profils mutationnels, j'ai employé le Direct-Coupling Analysis (DCA). Cette approche issue de la physique statistique permet de prédire l'effet d'une mutation survenant dans un gène et induisant un changement d'acide aminé dans la protéine correspondante. En modélisant les interactions entre paires d'acides aminés au sein de la protéine, le DCA permet de prendre en compte le contexte génétique dans lequel la mutation survient.

En appliquant le DCA à des milliers de core gènes d'*E. coli*, j'ai montré qu'il pouvait prédire les acides aminés natifs de cette espèce mais aussi les polymorphismes qui y sont observés. Le DCA prédit également la probabilité d'observer une mutation à une certaine fréquence. Ce faisant, il permet de mettre en évidence des différences d'efficacité de la sélection naturelle entre différentes sous-populations d'*E. coli*. En particulier, la sélection naturelle semble nettement moins efficace dans les souches de *Shigella*, en accord avec la taille efficace réduite de cette population.

Le contexte génétique s'est avéré clé dans la qualité des prédictions faites par le DCA. Ce contexte se construit sur des échelles de temps longues par l'addition de nombreuses interactions faibles entre acides aminés. Celles-ci n'affectent pas tous les résidus d'une protéine de la même manière. Le DCA permet de prédire la variabilité de ces résidus. En particulier, entre 30% et 50% des sites d'une protéine sont extrêmement contraints par le contexte génétique d'*E. coli*. Une mutation sur l'un de ces sites sera généralement délétère si elle survient seule. Ces sites ne tolèrent donc pratiquement pas de polymorphismes. Cependant ils peuvent coévoluer sur de longues échelles de temps de sorte que les acides aminés qui y sont observés varient largement d'une espèce à l'autre.

Si les différents résidus d'une protéine peuvent évoluer à différentes vitesses, il en est de même des protéines. J'ai développé un test de sélection, basé sur le DCA, permettant de comparer les gènes entre eux. À court terme les gènes essentiels sont ceux sous la plus forte pression de sélection purifiante tandis que le niveau d'expression détermine le taux d'évolution à long terme. Ce test détecte aussi des inactivations de régulateurs de la transcription, inactivations qui semblent sélectionnées à

court-terme mais contre-sélectionnées sur le plus long terme.

Le présent travail démontre l'intérêt de coupler l'étude de larges banques de génomes à des approches de modélisation pour comprendre l'évolution d'une espèce sur différentes échelles de temps.

Mots clés *Escherichia coli, Génétique des populations, Recombinaison, Core génome, Polymorphisme, Épistasie, Prédiction de l'effet des mutations, Direct-Coupling Analysis, Inactivations.*

Remerciements

En février 2020, j'arrivais à IAME pour un stage de master qui allait se poursuivre par un doctorat. Si le présent manuscrit retranscrit à peu près fidèlement le cheminement scientifique que j'y ai suivi, il ne rend nullement compte de toutes les personnes formidables que j'ai pu rencontrer chemin faisant.

En premier lieu, je souhaite très vivement remercier Olivier. Si je t'ai si naturellement demandé d'être mon directeur de thèse, c'est parce que je connaissais depuis longtemps tes excellentes qualités tant scientifiques qu'humaines. Tu m'as fait découvrir la génétique des populations lorsque, encore étudiante à l'X, je me disais confusément que je souhaiterais bien conjuguer informatique et biologie. Tu m'as accompagnée tout au long de mon projet de recherche de troisième année sans compter tes heures. Fidèle à ce que tu es : gentil, jovial, désintéressé et, bien évidemment, excellent scientifique. C'est comme cela que je t'ai retrouvé deux années plus tard comme superviseur de stage de master puis comme directeur de thèse. Nos nombreuses discussions scientifiques, ton enthousiasme communicatif, et nos déjeuners passés à refaire le monde vont me manquer.

Débuter dans la recherche académique en février 2020, n'est pas faire preuve d'un excellent sens du timing. Les confinements successifs, les différentes vagues de Covid-19 avec les nécessaires restrictions sanitaires associées, auraient facilement pu assombrir cette thèse. Mes remerciements vont donc tout naturellement vers Erick Denamur qui a largement fait en sorte que cela ne soit pas le cas. Merci de m'avoir accueillie si chaleureusement à IAME. Merci d'avoir veillé tout au long de ces années au bien-être de chacun d'entre nous. Ta connaissance encyclopédique d'*Escherichia coli*, et le plaisir évident que tu as à la partager, ont largement contribué à me faire m'intéresser à cette bactérie. Cela aura toujours été un plaisir d'échanger avec toi et j'espère que tu apprécieras ce manuscrit.

C'est à IAME que j'ai effectué la majeure partie de ma thèse. J'y ai trouvé un environnement de travail chaleureux et ouvert, et j'y ai rencontré nombre de personnes formidables. Je tiens en particulier à remercier tous les membres de l'équipe QEM et de l'équipe EVRest pour nos échanges scientifiques passionnants et nos discussions autour du déjeuner peut-être moins scientifiques mais tout aussi passionnantes. En particulier, Antoine pour ta bonne humeur et le plaisir que j'ai toujours eu à venir te parler de science ; Imane pour ta grande gentillesse et ton enthousiasme scientifique à toute épreuve ; Mélanie et Benoît pour m'avoir si bien accompagnée dans notre haine commune des tomates crues ; Mathilde pour ton énergie à soulever des montagnes dont BacteriaGame et Transmission sont de superbes exemples ; Hervé et Lionel pour avoir géré les serveurs avec brio mais surtout pour votre gentillesse, votre disponibilité et votre réactivité au moindre souci ; Claire H. et André pour m'avoir fait découvrir les joies du microbiote intestinal ; toutes celles et ceux qui ont rejoint ce bureau du fond du couloir et ont découvert, à leurs dépens, ma tendance au bavardage, en particulier Claire P. Un laboratoire n'irait nulle part sans celles et ceux qui en assurent la gestion administrative au quotidien. Et IAME est particulièrement chanceux en la matière, merci Myriam, Stefan et Houda

pour votre efficacité, votre gentillesse et votre disponibilité.

Je me dois une attention particulière pour deux personnes que j'ai rencontrées à IAME et qui ont joué un rôle majeur dans mon parcours. En premier lieu, Alaksh. Nous avons collaboré sur certains projets et j'ai eu largement l'occasion d'y apprécier tes compétences scientifiques et la simplicité avec laquelle tu rends service sans rien attendre en retour. Mais c'est bien au-delà de cela, la certitude que j'ai toujours eu de pouvoir échanger avec toi sur n'importe quel sujet, de trouver un interlocuteur ouvert, à l'écoute, empathique. En second lieu, mes remerciements vont tout naturellement vers Zoya. Nous sommes arrivées à IAME à quelques mois d'écart l'une de l'autre et notre expérience commune du doctorat nous a rapprochées. Avoir pu parler de nos vies avant, pendant et après le doctorat, de nos joies et de nos passages à vide, tout cela a eu beaucoup de valeur à mes yeux. Tu le sais déjà mais tu es bien plus qu'une collègue, une véritable amie.

Arriver dans un nouveau laboratoire en milieu de dernière année de doctorat n'est pas évident. Mais c'est sans compter l'équipe géniale que j'ai rejointe en venant à l'Institut Cochin : Magia, Sophiane, Maureen, Justine, Alan, Juliette, Marie-Florence, Chantal, Sébastien, Flavia, José, Erika, Célia, Caroline, Adèle, Maïra, Laura, Auguste, Hugo, Anne, Arnaud. Excusez-moi pour les coups de fil parfois incongrus auxquels vous avez dû répondre, je vous avais caché mon expertise avancée en médecine. Ces quelques pages sont un peu courtes pour vous remercier individuellement mais sachez que vous avez chacun contribué à embellir la fin de ma thèse. Je suis venue chaque jour à Cochin avec le sourire et c'est grâce à vous. Ivan, c'est à Bath que je t'ai rencontré et que j'ai tout de suite apprécié ta gentillesse (et ton goût pour l'histoire de l'Europe !). Une fois à Cochin, j'ai constaté l'étendue de tes connaissances sur les mécanismes de réparation de l'ADN mais j'ai surtout pu apprécier l'attention que tu portes à chacun. Merci pour l'accueil que tu m'as réservé !

Une heureuse surprise de ma thèse aura été d'intégrer le programme Jeunes Talents L'Oréal-UNESCO Pour les Femmes et la Science. Un grand merci à la Fondation L'Oréal pour son accompagnement tout au long des derniers mois. Je tiens aussi à remercier chacune des 34 autres lauréates. Nos échanges, souvent très personnels, ont été précieux dans mon parcours. Et pour celles qui vivent ou sont de passage à Paris, c'est toujours un plaisir de se retrouver.

La thèse n'est qu'une étape d'un parcours qui débute bien plus tôt. Je ne saurais dire en quelques mots la chance d'être née et d'avoir grandi dans une famille aimante, heureuse, joyeuse. Merci papa pour les cours de physique option (très) jeune public, merci Pierre pour le pendant en mathématiques. Merci maman, Hélène, Marie-Gabrielle, Isabelle et Ségolène pour avoir toujours été à mes côtés, m'avoir soutenue et encouragée dans mes choix. Vous avez chacune été des exemples qui m'ont aidé à grandir. Merci à mes neveux et nièces : Marius, William, Antoine, Juliette, Charlotte, Raphaël, Manon et Gaspard. Cela fait huit ans que je m'émerveille de chacun de vos progrès.

Il y a un monde en dehors de la thèse, et les amis qui le peuplent. Il y a ceux que l'on connaît depuis (pratiquement) toujours. Merci Pernelle, Camille, Marie, Pauline et Julie pour votre présence à mes côtés depuis si longtemps. Nos aventures aux quatre coins de la France puis au Rwanda ont forgé la femme que je suis aujourd'hui. Il y a aussi les amis que l'on rencontre plus tard mais qui ne comptent pas moins : Antoine, Morgane, Arthur, Lucie, Camille, David, Dimitri O., Dimitri K., Gladys, Gauthier, Jean-Côme, Mathieu, Matthieu, Philippe, Emeline, Amalia (bien évidemment), Romain et Teven. Toujours prêts à débattre de n'importe quel sujet et toujours partants pour se retrouver, votre présence a été un vrai bol d'air.

Dimitri, tu sais tout ce que tu représentes pour moi. Tu es à mes côtés depuis maintenant sept

ans. Tu m'as aidée et soutenue bien avant que je ne démarre une thèse et bien plus encore depuis. Merci pour tout cela et bien davantage.

Enfin, puisque la forme n'est autre que le fond qui refait surface, il me reste à remercier Philippe pour le généreux partage de son template LaTeX et Dimitri pour sa relecture typographique, si la typographie de cette thèse laisse encore à désirer ce n'est nullement une carence de ta part mais simplement que je n'ai pas su tirer tout le bénéfice de tes excellents conseils.

Preamble

Throughout my PhD research, I focused on understanding the diversity and evolution of *E. coli* species. To achieve this, I studied a vast collection of over 80,000 genomes of *E. coli* and *Shigella*. To analyze these genomes, I organized more than 400 million coding sequences into a database. For further investigation, I employed Direct-Coupling Analysis (DCA), an unsupervised machine learning method inspired by statistical physics. DCA allowed me to model amino-acid sequences and capture pairwise epistatic couplings, which helped infer the effects of the mutations observed in *E. coli*.

The first three chapters of this manuscript serve as the introduction. Chapter 1 provides an overview of *E. coli* as a species, discussing its ecology and importance in public health. Chapter 2 delves into the evolution of *E. coli*, highlighting how advances in the field of population genetics have enhanced our understanding of the species' diversity. Chapter 3 explores two practical tools used in evolutionary studies: experimental evolution and protein mutational landscapes. It specifically focuses on three important evolutionary topics that proved relevant to my research: the impact of metabolism on niche adaptation, the transition from commensalism to pathogenicity, and the development of antibioresistance.

The next four chapters form the core of my work. Chapter 4 describes the methodology I employed to analyse over 80,000 genomes and effectively organize 400 million coding sequences into a database. Chapters 5, 6, and 7 delve into different applications of Direct-Coupling Analysis for interpreting the diversity observed in these genomes. Chapter 5 demonstrates how individual mutations can be predicted and analyzed. Chapter 6 focuses on studying the variability of amino-acid sites both within *E. coli* and across distant species. Chapter 7 examines signatures of natural selection at the gene level. Finally, Chapter 8 serves as the conclusion, summarizing the key findings of this manuscript and suggesting potential avenues for future exploration.

The work presented in this manuscript has resulted in two publications:

- Vigué, L., Croce, G., Petitjean, M., Ruppé, E., Tenaillon, O., and Weigt, M. (2022). Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes. Nature Communications, 13(1), 4030.

- Vigué, L., and Tenaillon O. Predicting the effect of mutations to investigate recent events of selection across 60,472 *Escherichia coli* strains. Accepted for publication at Proceedings of the National Academy of Sciences.

The findings of the first publication are discussed in chapters 5 and 6. These are the result of a fruitful collaboration with Giancarlo Croce and Martin Weigt. The second publication is discussed in chapter 7. I would also like to acknowledge the contributions of Marie Petitjean and Étienne Ruppé for providing the 80,000 genomes that I have analyzed in this research.

In addition to the research presented in this manuscript, I have also contributed to the analysis of 16S data from microbiota experiments. These contributions have resulted in two other publications:

- Hobson, C. A., Vigué, L., Magnan, M., Chassaing, B., Naimi, S., Gachet, B., ... and Birgy, A. (2022). A Microbiota-Dependent Response to Anticancer Treatment in an *In Vitro* Human Microbiota Model: A Pilot Study With Hydroxycarbamide and Daunorubicin. Frontiers in Cellular and Infection Microbiology, 618.
- Hobson, C. A., Vigue, L., Naimi, S., Chassaing, B., Magnan, M., Bonacorsi, S., ... and Tenaillon, O. (2022). MiniBioReactor Array (MBRA) *in vitro* gut model: a reliable system to study microbiota-dependent response to antibiotic treatment. JAC-Antimicrobial Resistance, 4(4), dlac077.

You can find all four articles in the Appendix section. Furthermore, I have been involved in analyzing the dynamics of *de novo* mutations in experimental evolution. Although this work is not yet submitted for publication, it is expected to lead to a future publication.

Table of contents

List of Figures

List of Tables

Acronyms

AIEC	Adherent-Invasive *E. coli*
CDE	Context-Dependent Entropy
CIE	Context-Independent Entropy
DAEC	Diffusely Adherent *E. coli*
DCA	Direct-Coupling Analysis
DNA	Deoxyribonucleic acid
EAEC	Enteroaggregative *E. coli*
ECOR	*E. coli* Reference Collection
EHEC	Enterohemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
EPEC	Enteropathogenic *E. coli*
ETEC	Enterotoxigenic *E. coli*
ExPEC	Extraintestinal Pathogenic *E. coli*
GLASS	Gene-Level Amino-acid Score Shift
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
HPI	High Pathogenicity Island
HR	Homologous Recombination
IND	Independent sites method
InPEC	Intestinal Pathogenic *E. coli*
IPR	Inverse Participation Ratio
IS	Insertion Sequence
LTEE	Long-Term Evolution Experiment
MEPS	Minimum Effective Processing Segment
MGE	Mobile Genetic Element
MLEE	Multilocus Enzyme Electrophoresis
MLST	Multilocus sequence typing
MMR	DNA mismatch repair
MSA	Multiple Sequence Alignment
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
PMN	Polymorphonuclear leukocytes
sIgA	Secretory immunoglobulin A
SNP	Single Nucleotide Polymorphism
ST	Sequence Type
UTI	Urinary Tract Infection

Chapter 1

Ecology of *E. coli* and its role in health and disease

1.1 *E. coli* in public health and research

Escherichia coli is a Gram-negative facultative anaerobic bacteria (Tenaillon et al. 2010). The ease with which it is grown and manipulated in the laboratory has contributed to making it a model species in life science research. Although most humans carry *E. coli* asymptomatically in their gut, this bacteria can also act as a deadly pathogen. Approximately 950,000 people die each year from *E. coli* intra and extra-intestinal infections, making *E. coli* the second leading cause of death among bacterial pathogens (Ikuta et al. 2022).

In 2019, the World Health Organization declared antimicrobial resistance to be among the top ten global public health threats facing humanity (EClinicalMedicine 2021). As the bacteria with the highest number of resistance-related and attributable deaths, *E. coli* is closely monitored (Murray et al. 2022). The emergence and spread of antibiotic-resistant clones in this species involves complex dynamics that are not yet fully understood. One of these clones is ST131, which increased rapidly in frequency in the early 2000s before plateauing at intermediate levels (Kallonen et al. 2017). Meanwhile, other drug-sensitive clones have managed to remain stable. Therefore, *E. coli* provides an excellent model for studying the dynamics of antibiotic resistance in bacterial populations.

In nature, *E. coli* exhibits versatile behaviour, with a wide range of isolation sources, ecological interactions and lifestyles. As a generalist organism, it inhabits a variety of ecological niches, including the gut of various species, water, soil and sediments. It also covers the full range of possible interactions with its host: from commensalism and even mutualism to facultative and obligate pathogenicity in the case of *Shigella* and enteroinvasive *E. coli* (EIEC) (Khalil et al. 2018). *Shigella* and EIEC have also adopted an intracellular lifestyle and limited their host range to primates. For this reason, *E. coli* represents an excellent model to study adaptation to different ecological niches and transition from commensalism to pathogenicity.

The *E. coli* population is structured into phylogroups referred to as A, B1, B2, C, D, E, F and G. More distant strains, called *Escherichia* clades and numbered C-I to C-V, form a genetic continuum within the genus *Escherichia* (Walk et al. 2009). *Escherichia* clade I is generally thought to belong to the species *E. coli*, while the other clades are considered too distant, but the subject remains contro-

versial (Clermont et al. 2013; Cobo-Simón et al. 2023). The striking phenotypic and clinical differences between *Shigella* and *E. coli* led to their being studied as two distinct species (Touchon et al. 2009). They were even assigned to separate genera in the 1940s in order to establish a strict separation between an obligate pathogen and what was thought to be solely a commensal (Lan et al. 2002). The discovery of EIEC and advances in genomics in the following decades led to a reconsideration of this separation. A new consensus gradually emerged that *Shigella* forms a subgroup of *E. coli* strains with a specific set of virulence factors and an intracellular lifestyle that make them obligate pathogens of primates.

1.2 First descriptions of *E. coli* and *Shigella* species

The common coli bacterium was isolated and described in 1885 (Friedmann 2014). It was later named *Escherichia coli* after its discoverer, the Bavarian physician Theodor Escherich. Theodor Escherich became interested in faecal bacteria after studying a cholera epidemic in Naples in 1884. At that time, many direct causal links between specific pathogens and diseases were being discovered. Escherich therefore naturally looked for the causative agent of infantile diarrhoea, one of the main causes of death at the time. However, he soon concluded that the intensity of the changes observed in the microbial composition of the gut during the course of the disease made it impossible to identify a single causative agent. He therefore chose the opposite approach and focused his work on the stools of healthy individuals, mainly infants and new-borns. He believed that a better understanding of the bacteria found in a healthy gut would help to explain the physiology of digestion, the factors that determine the onset of an intestinal disease and might even give clues as to how to treat it. The novelty of his approach lies not only in his intuition that the 'normal' gut microbial flora must be studied to understand the disease. Indeed, he was also among the first to use new technologies, in particular microscopy and Robert Koch's techniques for culturing and characterising bacteria. His research led him to focus on *Bacterium coli commune*, a bacteria regularly found and often predominant among the aerobic faecal flora of healthy individuals. He demonstrated that this bacteria could grow in the absence of oxygen, a condition similar to that of the gut.

Shigella is the causative agent of bacillary dysentery. This disease, known since Antiquity, is characterised by diarrhoea often accompanied by blood and mucus originating from the disruption of the intestinal epithelium. In 1898, Kiyoshi Shiga described *Bacillus dysenteriae* after isolating it from the faeces of a patient during a dysentery outbreak in Japan (Bensted 1956; Lan et al. 2002). Ten years earlier, in 1888, Chantemesse and Widal had already isolated an organism that they believed caused epidemic dysentery. This organism was later found to be identical to *Bacillus dysenteriae*. However, Chantemesse and Widal had described it so poorly that the discovery of *Bacillus dysenteriae* is generally attributed to Kiyoshi Shiga, after whom it was renamed *Shigella*. In the following years, other strains of *Shigella* were recurrently isolated during dysentery epidemics. When cultured, some were shown to release a toxin that became known as Shiga toxin. Another name often associated with early studies of *Shigella* is that of the German bacteriologist Walther Kruse. His work, published in 1907, was based on serological tests to characterise different types of *Shigella*, one of which was a late lactose fermenter and was later referred to as *Shigella sonnei*, while the other types were incapable of fermenting lactose. Compared to Kruse's work, which was well known in Germany, the English-speaking countries remained behind in the study of *Shigella*. It was not until the outbreak of the First

World War, and the intense burden caused by dysentery among soldiers, that British scientists began to actively study this bacteria. In 1929, *Shigella boydii* was isolated in India by the British bacteriologist John Boyd and, in 1940, a *Shigella* genus consisting of four species—*S. dysenteriae*, *S. flexneri*, *S. boydii* and *S. sonnei*—was formally recognised.

1.3 *E. coli* in the wild

E. coli is a generalist organism that lives in the gut of vertebrates. It can also be sampled from environmental sources: soil, water and sediments. Its population size is estimated at 10^{20} (Tenaillon et al. 2010).

The prevalence of *E. coli* in the gut of vertebrates varies from complete absence in some species to over 90% in humans (Figure 1.1) (Gordon et al. 2003). It also reaches higher densities in human faeces—typically 10^7 to 10^9 colony-forming units per gram of faeces, three orders of magnitude higher than in domestic animals (Tenaillon et al. 2010). A wide range of factors were found to be significant predictors of *E. coli* in a host species: climate, body mass, diet and host phylogeny. However, these predictive factors tend to be correlated with one another, making it difficult to determine the exact role played by each. Animals frequently exposed to human activities are also more likely to carry *E. coli* in their gut. This highlights the importance of *E. coli* transmission between hosts and the role of humans as a reservoir of *E. coli* for other species. Human exposure also impacts on the intra-host diversity of *E. coli*, with domestic animals having lower intra-host diversity than their wild counterparts.

It was long thought that *E. coli* could not survive and multiply for an extended period of time in the environment. Therefore, the presence of *E. coli* in the environment was interpreted as a marker of faecal contamination. However, it is now known that some *E. coli* are able to persist and grow in the environment. It is estimated that half of the *E. coli* population lives there (Tenaillon et al. 2010).

The frequent retrieval of closely related strains from distant locations indicates a global spread and rapid circulation of *E. coli*. Phylogenetically distant strains are also commonly co-isolated from the same source. However, this does not obscure the preferential association of certain phylogroups with certain ecological niches. Phylogroups A and B1 adopt a generalist lifestyle as they can frequently be found in the environment as well as in any vertebrate group. In comparison, phylogroups B2 and D have a more restricted host range, often limited to endothermic vertebrates. Phylogroup B2 is even more specialised, as it prevails in mammals with a hindgut fermentation chamber (Gordon et al. 2003). While B2 strains are under-represented in the environment, *Escherichia* clade I seems to favour this niche (Walk et al. 2009; Touchon et al. 2020). Yet, the preferential association of certain phylogroups with certain niches should not mask the absence of host-specific strains. We observe a differential pattern of phylogroup distribution between species rather than a direct association between certain ecological niches and certain phylogroups. In this respect, *Shigella* and EIEC stand out as pathogens restricted to primates (Mattock et al. 2017). A human-specific commensal *E. coli* clone has also been reported (Clermont et al. 2008), but the study was performed before the advent of high-throughput sequencing and would require to be confirmed now that more data is available on non-human *E. coli* reservoirs.

Human populations show different levels of intra-host *E. coli* diversity, with the highest levels in tropical regions. Striking geographical and temporal differences also appear in the distribution of

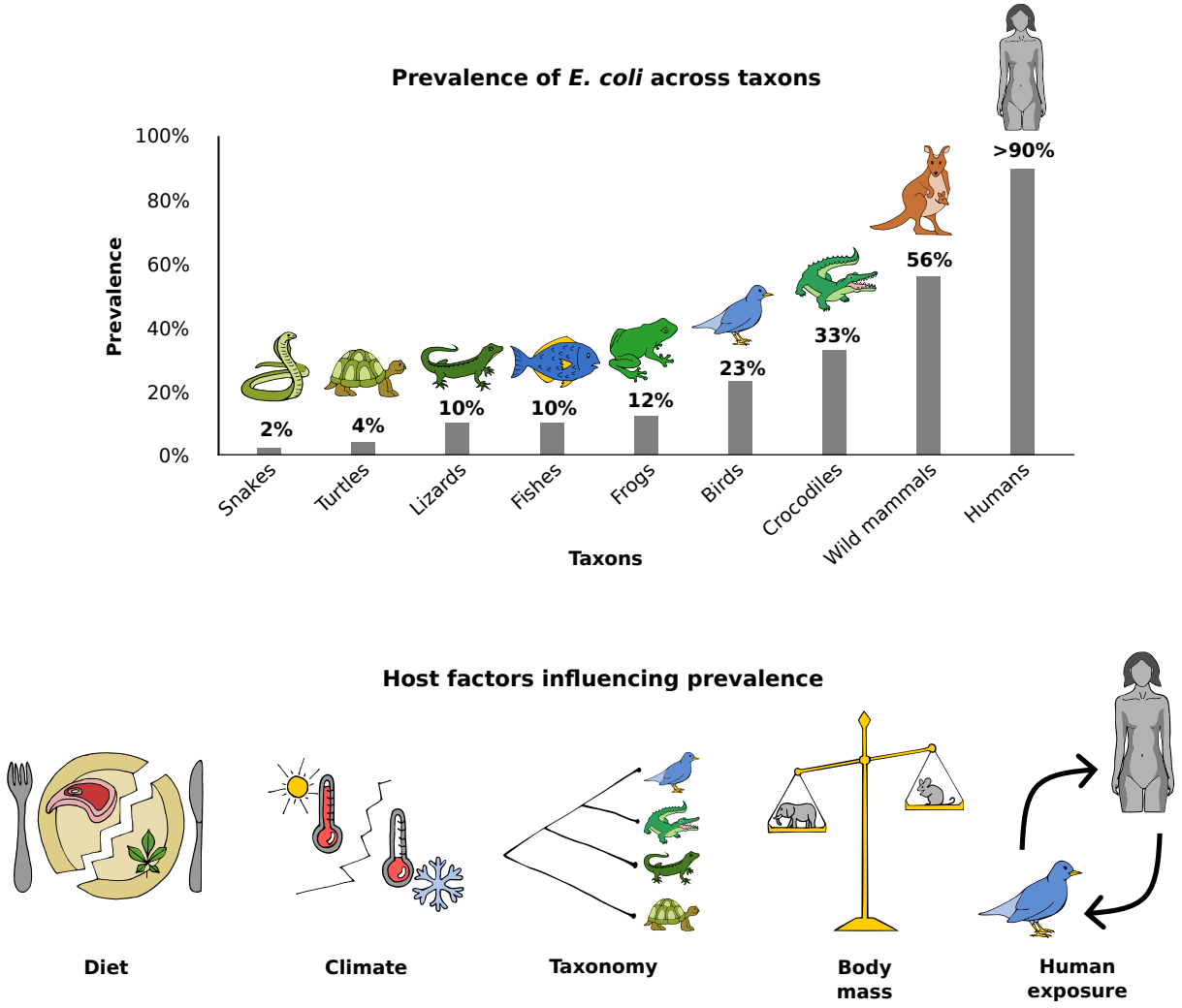

FIGURE 1.1: **Prevalence of *E. coli* across species and factors influencing this prevalence.**
All the data from non-human species are extracted from (Gordon et al. 2003). The estimation of the prevalence of *E. coli* in human comes from (Martinson et al. 2020).

phylogroups in humans. In the 1980s, A strains predominated in France, while B1 strains predominated in Mali. B2 strains were almost absent (around 2%) in Malian subjects, whereas they represented close to one fifth of Croatian commensal *E. coli* (Duriez et al. 2001). In France, the distribution of phylogroups has changed radically in just two decades: while B2 represented 10.5% and A 61% of the strains sampled in the 1980s, the former have increased almost threefold, reaching 29.5% in the 2000s, and the latter have more than halved, falling to 25.5% (Tenaillon et al. 2010). These temporal and geographical differences are generally attributed to lifestyle, including diet and hygiene, rather than host genetics. Indeed, comparison of *E. coli* isolated from French metropolitans, French metropolitans who had moved to French Guiana and local Wayampi Amerindians in French Guiana showed that French metropolitans who had moved to French Guiana showed intermediate characteristics between the other two populations (Skurnik et al. 2008). Climate or geographical location are unlikely to play an important role in explaining these differences, as they cannot account for the drastic temporal changes observed in metropolitan France between the 1980s and the 2000s.

1.4 *E. coli* in the human gut microbiota

The human body is home to approximately 10^{13} to 10^{14} bacteria, roughly the number of human cells (Sender et al. 2016). Of these 10^{13} to 10^{14} bacteria, 70% reside in the colon. Indeed, the gastrointestinal tract represents an excellent niche for micro-organisms as it provides them with a large surface area and an abundant source of nutrients. We estimate that 500 to 1000 different species can be found in an individual's gut at any given time. The bacterial group composition varies longitudinally along the gastrointestinal tract—biopsy samples from the small intestine and the colon show different bacterial composition—and latitudinally, with microbes present in the lumen differing from those found in the mucosal layer close to the intestinal epithelium (Sekirov et al. 2010).

Studies in germ-free animal models—animals without microbiota—have shown the crucial importance of the gut microbiota for host health. The presence of a gut microbiota benefits the host in a number of ways: directly, by protecting it from incoming pathogens, extracting calories from indigestible polysaccharides (Backhed et al. 2005) and sometimes playing a therapeutic role, and indirectly, by developing its immunity.

The very high density of microbes in the gut generates strong competition for space and food. Microorganisms also produce a range of antimicrobial compounds to prevent competitors from settling in the gut. This intense competition provides a strong barrier to colonisation for any newcomer, whether commensal or pathogenic. From the host's point of view, it is particularly beneficial in limiting the risk of infection by pathogenic species. The role of the gut microbiota extends beyond the prevention of pathogenic infections. Indeed, the success of gut microbiota transplantations in curing certain digestive tract infections has highlighted the therapeutic role of the gut microbiota. The success of certain drug-based therapies also seems to depend on the state of the microbiota (Sekirov et al. 2010; Moran et al. 2019).

The intestinal mucosa represents the largest surface area in contact with the external environment. The dense community of microbes close to the mucosa therefore represents the largest fraction of antigens that stimulate immune cells. Hence, the gut microbiota plays a key role in the education of the host immune system (Sekirov et al. 2010). It is thought that changes in diet and hygiene can reduce exposure to microorganisms, resulting in a disrupted and immature microbiota. This un-

derdeveloped microbiota would be unable to educate the host immune system, thus increasing the incidence of allergies and autoimmune disorders in the host (Moran et al. 2019).

In this landscape, *E. coli* represents the predominant aerobic organism of the gut microbial flora (Tenaillon et al. 2010; Martinson et al. 2020). However, it accounts for only a small fraction of this microbiota as anaerobes dominate aerobes by two or three orders of magnitude. *E. coli* is also one of the first bacteria to colonise newborns. It may contribute to depleting oxygen along the gastrointestinal mucosa, thus paving the way for colonisation of the gut by strict anaerobes. In the gut, it inhabits the large intestine, particularly the caecum and colon, where it lives in the mucosal layer. It regularly detaches from the mucosa with degraded mucus and falls into the lumen to be excreted in the faeces. As an early coloniser of the gut and inhabitant of the intestinal mucosa, *E. coli* is likely to play an important role in the education of the immune system. Its lipopolysaccharide-rich outer membrane, common to all gram-negative bacteria, stimulates the production of secretory immunoglobulin A (sIgA) by the host. These sIgA bind microbes, preventing uncontrolled microbial growth. This reduces the interaction between the gut microbiota and the mucosal immune system, thereby decreasing the host's response to its resident microbes and the risk of high inflammation (Sekirov et al. 2010).

Most individuals carry one or two resident strains of *E. coli* and a few transient strains at any given time (Martinson et al. 2020). Resident strains are those that remain in the gut for a long time, at least several weeks. By contrast, transient strains remain there for only a few days. In other words, resident strains are those that have successfully colonised the gut—they produce enough offspring to compensate for their evacuation through intestinal transit—whereas transient strains fail to colonise the gut, they are ingested and immediately lost. In a way, the latter provide a picture of an individual's daily exposure to *E. coli* present in their environment. The only context in which shared clones have been repeatedly reported is when people live together in the same household. These clones may be resident in one individual and transient in others, suggesting an important role for frequent exposure and close contact in the sharing of *E. coli* strains.

The presence of *E. coli* resident and transient strains in the human gut has been known since the mid-twentieth century (Sears et al. 1950; Sears et al. 1952). However, identifying the factors that determine whether a strain is resident or transient has proved difficult. Experimental attempts to colonise the gut by ingesting strains almost inevitably fail. Sometimes ingested strains are not even recovered as transient. Extreme procedures carried out on dogs, involving massive feeding, use of chemical means to clean the gut and rectal injections, have also failed to achieve successful colonisation (Sears et al. 1956). It was noted that mice treated with streptomicin as well as human neonates were easier to colonise, suggesting an important role for the host microbiota in preventing colonisation (Tenaillon et al. 2010; Lou et al. 2021). The advent of sequencing has given some insight into the factors determining residency by providing access to strain phylogroups and to their gene content. Phylogroups B2 and D tend to be over-represented among resident strains, although this relationship is far from systematic (Nowrouzian et al. 2005). Resident strains are also more likely to carry adhesion factors, iron acquisition systems and genes involved in sugar and amino-acid metabolism (Lou et al. 2021). It is important to note, however, that there is no single trait that determines residency: successful gut colonisation probably relies on a wide range of genes. The failure of most experimental attempts to colonise the gut, even with a previously resident strain, also suggests that even a very fit strain has little chance of successfully colonising the gut of an adult with a mature microbiota.

Resident strains do not stay in the gut forever and are eventually replaced. However, the factors that determine this change remain unclear. Changes in an individual's environment, such as travel, have sometimes been linked to changes in residency, but in other cases travel has not impacted the resident strains. Diarrhoea and artificial purging also failed to show an effect on residency (Sears et al. 1952; Sears et al. 1956). Some reports have linked antibiotic treatment to changes in resident strains, but it is also possible for a resident strain to persist despite antibiotic treatment, even when the strain is sensitive to the prescribed antibiotics (Martinson et al. 2020). Antibiotics are known to disrupt the gut microbiota, with effects that can still be observed long after treatment has ended. However, they have very different effects on different individuals (Andremont et al. 2019). For example, our lab conducted a study in an *in vitro* gut model that revealed a reduced impact of ceftriaxone on the composition of the microbiota when the microbiota exhibited high β-lactamase activity (Hobson et al. 2022).

1.5 *E. coli* in human disease

E. coli is generally described as a commensal, as it inhabits the human gut without harming its host. It benefits from the gut environment, including nutrients, protection from external stresses, transport and dissemination. In turn, it may also provide some benefits to the host, suggesting a degree of mutualism. Amongst these benefits, we have already mentioned its role in oxygen depletion of the neonatal gut, education of the immune system and protection against pathogenic newcomers. It is also capable of producing vitamin K, thus extending the metabolic capabilities of the host (Suvarna et al. 1998). However, the harmless commensal can also turn into a deadly pathogen.

Figure 1.2 summarises the range of diseases that *E. coli* can cause. They can be divided into two broad categories: intestinal and extra-intestinal diseases. The former are characterised by diarrhoea and can lead to fatal haemolytic uraemic syndrome (HUS) when the strain produces Shiga toxin. On the other hand, harmless commensals from the gut can translocate to another organ and cause an opportunistic infection there, resulting in extra-intestinal disease. In this case, the severity of the disease is mainly determined by the condition of the host: young, healthy people exhibit milder infections than older, immuno-compromised patients (Denamur et al. 2021; Katouli 2010).

Enteroinvasive *E. coli* (EIEC) and *Shigella* have an invasive pathotype. Infection with these strains occurs in several stages (Figure 1.3): the strain enters an M cell of the intestinal epithelium, is endocytosed by a resident macrophage, induces its death and proceeds to infect adjacent epithelial cells (Mattock et al. 2017; Pasqua et al. 2017). Different groups of *Shigella* have arisen independently at various times in history, as is the case with EIEC. However, they infect by very similar mechanisms and use the same set of genes, revealing a high degree of convergence (Wirth et al. 2006).

Strains involved in intestinal diseases harbour a restricted set of virulence genes with strong effects. In contrast to intestinal diseases, extra-intestinal virulence is multigenic: it relies on many genes with small effects, usually genes coding for adhesins, toxins, protectins and iron uptake systems. These genes have been identified experimentally through the use of animal models, such as the mouse model of sepsis (Denamur et al. 2021). In this model, *E. coli* strains exhibit two distinct behaviours: some kill most of the injected mice (killer phenotype) while others kill almost none (non-killer phenotype) (Picard et al. 1999). The B2 and D strains mainly belong to the first category and the A, B1 and E strains to the second. It should be noted that, like any model, the mouse model of sepsis

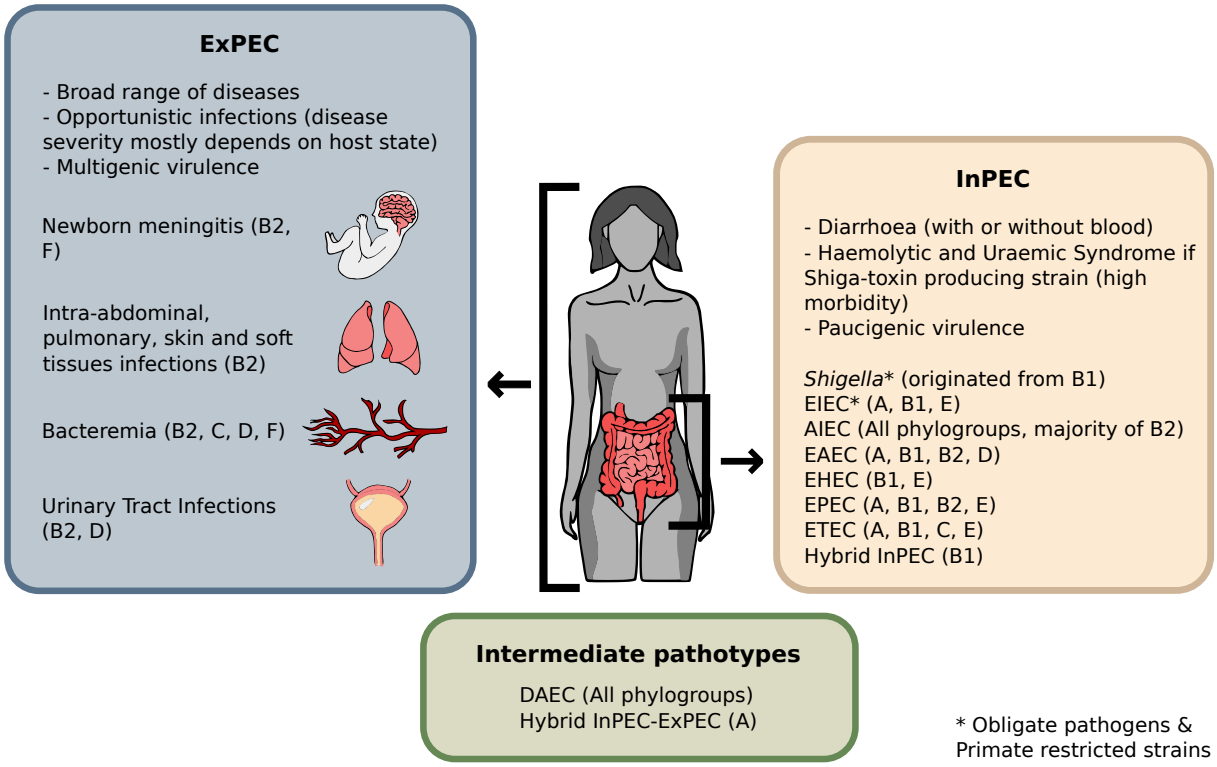

FIGURE 1.2: **Range of diseases caused by *E. coli*, main human pathotypes, and corresponding phylogroups.** ExPEC: Extraintestinal pathogenic *E. coli*; InPEC: Intestinal pathogenic *E. coli*; AIEC: adherent-invasive *E. coli*; DAEC: diffusely adherent *E. coli*; EAEC: enteroaggregative *E. coli*; EHEC: enterohemorrhagic *E. coli*; EIEC: enteroinvasive *E. coli*; EPEC: enteropathogenic *E. coli*; ETEC: enterotoxigenic *E. coli*. See (Denamur et al. 2021) for further details on these pathotypes.

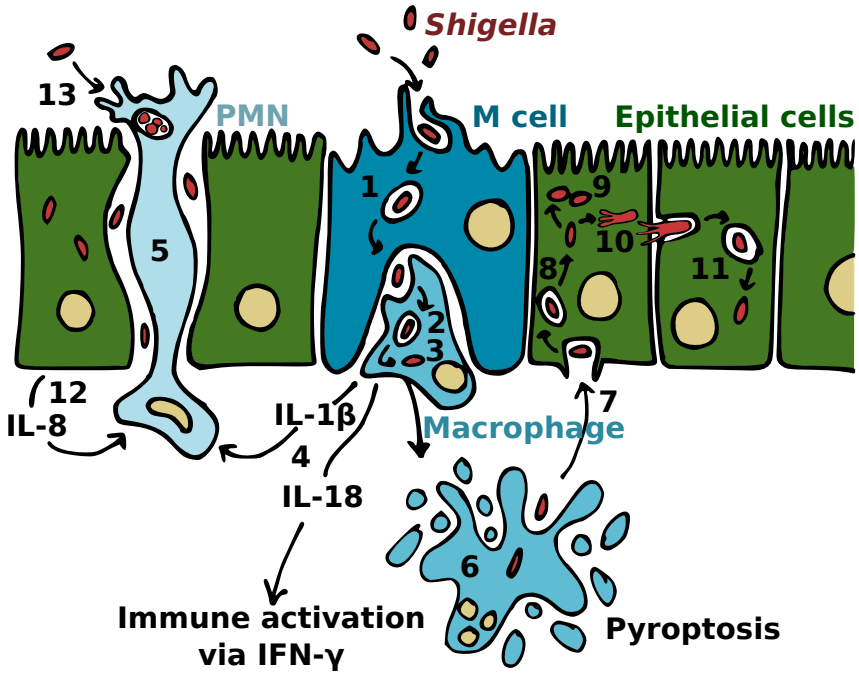

FIGURE 1.3: **Stages of *Shigella* infection (adapted from (Mattock et al. 2017)).**
Shigella colonizes the intestinal epithelium in two ways: entrance into M cells and destabilization of the epithelial barrier. *Shigella* enters into M cells by inducing membrane ruffling (1), before being endocytosed by a resident macrophage (2), evading the phagocytic vacuole (3) and inducing both the release of pro-inflammatroy signals (4) and the macrophage death (6). The release of pro-inflammatroy signals leads to the recruitment of polymorphonuclear leukocytes (PMN) that destabilize epithelial cell junctions thus allowing other strains to cross the epithelial barrier (5). Macrophage death and lysis (6) allows *Shigella* to invade epithelial cells from their basolateral membrane (7). The strain then escapes from its vacuole (8) and multiplies in the cell cytoplasm (9). Thanks to an actin tail, *Shigella* can move inside the cell and infect a neighbooring cell (10), there again it escapes its vacuole and enters the cytoplasm (11). Polymorphonuclear leukocytes clear the infection (13). NB: *EIEC* infection follows very similar steps to those described above.

cannot provide perfect information on the virulence of a strain. Indeed, in this model, the strains are directly injected into the blood at high concentration, so that we do not detect any virulence gene involved in the first stages of the infection.

E. coli, therefore, demonstrates a wide range of niches it inhabits and diseases it is capable of causing. Different strains exhibit distinct ecological traits and varying levels of virulence. This diversity, which has been widely observed by microbiologists, provides an excellent foundation for gaining insights into the evolution of this species.

Chapter 2

E. coli population genetics

2.1 *E. coli* population structure

The *E. coli* species inhabits a wide range of ecological niches and readily switches from commensalism to pathogenicity. This generalist behaviour requires genetic flexibility and frequent adaptations. As we have seen, different phylogroups tend to favour different niches and behave differently within a human host. To decipher the determinants of *E. coli* evolution, we therefore need to better characterise its population structure.

2.1.1 Pre-PCR era

The first attempts to elucidate the population structure of *E. coli* predate the discovery of the double helix model of DNA structure. They trace back to Kauffman who developed serotyping in 1940 (Kauffmann 1947). This method consists of characterising three surface antigens of the bacteria: O (somatic), K (capsular) and H (flagellar). Serotyping studies have led to two main discoveries. Firstly, the O, K and H antigens show non-random associations. These combinations of O, H and K alleles were called serotypes. Secondly, some serotypes can be found all over the world.

Another breakthrough came in the 1980s with the introduction of multilocus enzyme electrophoresis (MLEE) (Selander et al. 1986). MLEE characterises the relative electrophoretic mobility of several housekeeping enzymes. Each mobility variant corresponds to an allele of the enzyme. The haplotypes formed by the alleles observed at different loci can then be compared between strains. The closer two strains are, the more alleles they should share. Robert K. Selander and Howard Ochman, who pioneered the use of MLEE for bacterial species, also built the first *E. coli* reference collection (ECOR) (Ochman et al. 1984). ECOR was compiled by sampling strains from various continents and host species in an effort to capture the diversity of *E. coli*. MLEE studies performed on the ECOR collection revealed that the strains' haplotypes clustered into different groups. Strains from the same cluster can originate from distant locations or from different host species, which again confirms the existence of intense circulation of strains across the world and between species (Selander et al. 1980). The population structure was formalised by identifying four phylogroups: A, B1, B2 and D (Herzer et al. 1990; Johnson et al. 2001). Four remaining ECOR strains belong to what would later be designated as phylogroup E (Clermont et al. 2021).

During these years, work accumulated suggesting a close proximity between *E. coli* and *Shigella*.

In 1953, serotyping documented the sharing of identical O antigens between the two bacteria (Ewing 1953). Four years later, hybridisation experiments proved that they could recombine with each other (Luria et al. 1957). In the late 1960s, nucleic acid reassociation experiments confirmed the proximity of *E. coli* and *Shigella* by showing that the thermal stability of *E. coli-S. flexneri* DNA reassociation products was comparable to that of *E. coli-E. coli* (Brenner et al. 1969). MLEE performed on *Shigella* strains found that their electrophoretic types could fall into ECOR clusters (Ochman et al. 1983). More importantly, different *Shigella* strains could fall into different clusters, suggesting distinct emergences of different *Shigella* lineages from *E. coli* species.

2.1.2 PCR era

The advent of sequencing has opened a new era in population genetics. The 1990s saw the first complete sequence of a cellular genome when a team of researchers led by J. Craig Venter published the *Haemophilus influenzae* genome in 1995 (Fleischmann et al. 1995). In parallel with these efforts to sequence complete genomes, a new method has emerged for using smaller DNA sequences: multilocus sequence typing (MLST) (Enright et al. 1999). The spirit of MLST is very similar to that of MLEE, but instead of equating mobility variants with alleles, it directly uses the DNA sequences of genes to detect alleles. This improves resolution because it allows gene sequences that code for enzymes with similar electrophoretic mobility to be distinguished, typically when they differ only at synonymous sites. A haplotype based on alleles at different loci is called a sequence type (ST). There are two main MLST classifications in use today: the Warwick classification (Wirth et al. 2006) and the Institut Pasteur classification (Jaureguy et al. 2008).

Access to DNA sequences of housekeeping genes also made it possible to construct the first phylogenies (Lecointre et al. 1998). The phylogenies of the ECOR strains confirmed the first observations of the MLEE method: the phylogroups coincide with the phylogenetic clades (Wirth et al. 2006). Sampling of new strains allowed the phylogeny to be refined. Phylogroup D was found not to be monophyletic and was divided into phylogroups D and F (Jaureguy et al. 2008). In addition, a sister clade to phylogroups A, B1 and *Shigella* was named phylogroup C (Escobar-Páramo et al. 2004). The phylogenies were also anchored on an outgroup species, such as *Salmonella enterica*. This suggested that B2 was the most basal phylogroup, followed by D (Lecointre et al. 1998).

In parallel, Rolland *et al.* applied ribotyping to a dataset based on the ECOR collection, to which they had added strains of *Shigella* and enteroinvasive *E. coli* (EIEC) (Rolland et al. 1998). Their work showed that *Shigella* emerged several times from phylogroups B1 and D, while EIEC were even more widely distributed between phylogenetic groups A, B1 and B2.

Pre-PCR methods such as MLEE or ribotyping were already able to identify phylogroups. However, they were too complex and time-consuming to be used routinely in many laboratories. In this respect, the advent of triplex-PCR proved to be a game-changer, as it greatly democratised the identification of strain phylogroups (Clermont et al. 2000). Based on the presence-absence pattern of three DNA fragments (two genes and one anonymous DNA fragment), the phylogenetic group to which a strain belongs can be found with a very low misclassification rate (Figure 2.1.B). In 2013, triplex-PCR was transformed into quadruplex-PCR by adding a fourth DNA target (Clermont et al. 2013). This was done in order to follow the progress made in our understanding of the structure of *E. coli* phylogroups.

In 2009, a study of 20 complete genome sequences of *E. coli* and *Shigella* paved the way for

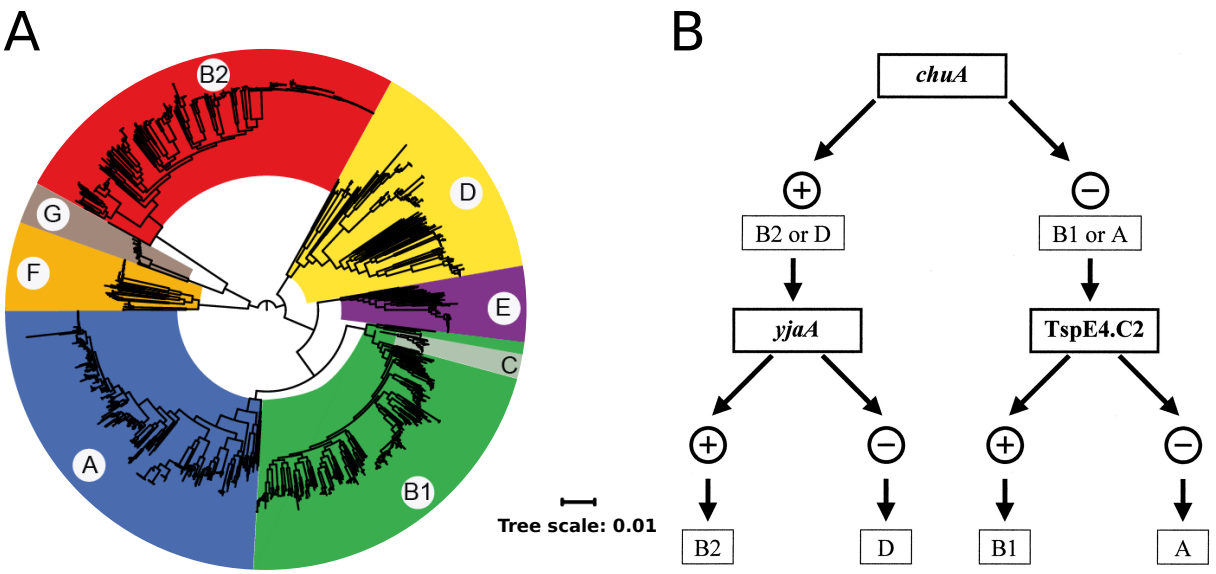

FIGURE 2.1: ***E. coli* phylogeny and phylogroups.**
A. Phylogeny of 1,294 *E. coli* strains isolated from diverse sources on the Australia continent, the different phylogroups are highlighted with specific colors (Figure from (Touchon et al. 2020)).
B. Decision tree used to assign a phylogroup to an *E. coli* strain in the original triplex-PCR method (Figure from (Clermont et al. 2000)).

genome-wide population genetics of *E. coli* (Touchon et al. 2009). Advances in sequencing over the past decade now make it possible to conduct similar studies with over a thousand strains (Touchon et al. 2020).

Currently, the phylogeny of *E. coli* is considered to consist of eight phylogroups (Figure 2.1.A) (Touchon et al. 2020). Phylogroups A, B1, B2 and D are the most common and oldest known. The other four (phylogroups C, E, F, G) are less common. Phylogroup G is the latest phylogroup to be discovered and includes strains intermediate between phylogroups F and B2 (Clermont et al. 2019). The notion of rarity or commonness of a phylogroup is of course very biased. The vast majority of *E. coli* strains sequenced so far are human commensals or pathogens, which leads to an obvious bias in the databases. The current picture of a species with eight phylogroups is far from fixed. Firstly, because sampling of new strains, especially environmental or non-human strains, will certainly lead to the identification of new phylogroups. One study has already suggested the existence of a phylogroup H (Lu et al. 2016), but so far only one strain of this phylogroup—found in the gut of *Marmota himalayana*—is known. Secondly, the criteria on the basis of which phylogroups are split remain controversial. A recent study advocates splitting some of the existing phylogroups into several, which would result in a total of 14 phylogroups (Abram et al. 2021). Whatever refinements in phylogeny may be made in the future, there is now a consensus that *E. coli* has a strong population structure with different subgroups that have emerged one after the other and coexist today. A quick look at the phylogeny may be enough to detect strong differences between these groups. Indeed, phylogroups D and F have very long terminal branches compared to the other phylogroups, suggesting that they follow distinct evolutionary dynamics.

As technologies have improved, more detailed knowledge of the population structure of *E. coli* has become available. The very first technology used, serotyping, is subject to significant bias because it is

based on surface antigens. These are under strong diversifying selection, hence their frequent evolution and exchange between strains. For example, strains of the same sequence type (ST) often exhibit different O:H combinations. The advent of MLST has clearly improved the classification of strains. However, the classification into STs is in itself somewhat arbitrary. Some old STs are very abundant and show great diversity, while others have emerged recently as a result of a single mutation and contain very few strains and almost no diversity. The decreasing cost of sequencing, year after year, has allowed better analysis of genomes and the study of larger data sets, resulting in a more accurate picture of the structure of the species. Another advantage of sequencing is that DNA sequences can be made available online. This means that a genome can be studied by anyone in the world, even if they do not have access to the strain. This allows large-scale computational studies, with publications based on the analysis of as many as 100,000 *E. coli* genomes (Abram et al. 2021). However, it is worth noting that although serotyping and MLEE were far from perfect, the advent of sequencing has mainly helped to refine the picture of *E. coli* population structure without really invalidating the observations made with these earlier methods.

2.2 *E. coli* genome

Thus, *E. coli* represents a highly stuctured species. In order to understand the determinants of this structure, we need to analyse its genome more in depth.

2.2.1 Genome size

A first and very basic statistic of genomes is their sizes. The *E. coli* genome is on average 5 Mb long, but the range of variation in genome size spans nearly 2 Mb between *E. coli* strains (Touchon et al. 2020)—with phylogroups A and B1 having smaller genomes than B2 and D (Bergthorsson et al. 1998).

The factors that determine the size of bacterial genomes and the role that selection plays in this process are not yet fully understood. The effective population size constitutes one of the main determinants of the intensity of natural selection. It is itself influenced by the lifestyle of the species. When examining prokaryotic species with different lifestyles—free-living bacteria, commensals, obligate animal pathogens and obligate intracellular organisms—there is a general trend towards decreasing effective population size and genome size. This suggests that when natural selection decreases, genomes shrink (Bobay et al. 2017b; Bobay et al. 2018).

Another interpretation may be that parasites need fewer functions to survive because they depend mainly on their host to provide them with the metabolic capabilities they lack. It has also been suggested that the reduction in genome size may increase the growth rate, so that smaller genomes are selected. However, studies using dN/dS—the ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitution per synonymous site—to estimate the strength of purifying selection have shown that genome size positively correlates with purifying selection (Bobay et al. 2018). Furthermore, among *E. coli* strains, differences in genome size do not translate into different growth rates (Bergthorsson et al. 1998). Overall, this suggests that although the loss of some unnecessary genes may be selected for in parasites, natural selection may not be strong enough in these species to maintain other useful functions (Mira et al. 2001).

These thoughts apply perfectly to the transition of *E. coli* to an intracellular lifestyle embodied

by *Shigella*. Indeed, *Shigella* loses genes more rapidly than other *E. coli* strains, acquires them more slowly and, once acquired, is also less likely to retain these new genes (Hershberg et al. 2007; Passel et al. 2008). At the same time, it also has more non-synonymous mutations, a pattern consistent with relaxed selection (Balbi et al. 2009). However, the multiple emergence of *Shigella* and EIEC throughout history allows the detection of some convergent gene losses that certainly reflect adaptation (Pupo et al. 2000).

In 2002, Dmitri Petrov suggested that a simple neutral model could explain the observed variations in genome size (Petrov 2002). In this framework, genome size reflects a balance between rare events of large DNA gains and many smaller losses. This model was developed on the basis of animal data. Bacterial genomes are of course very different from those of eukaryotic species, not least because they contain less non-coding DNA—these regions being more likely to be selectively neutral and therefore subject to frequent loss and acquisition. However, *E. coli* shows the same pattern of small and frequent gene losses balanced by larger and less frequent gains (Touchon et al. 2009). In this species, most differences in genome size tend to be inherited vertically. This is consistent with a neutral scenario. However, selection seems to be at play when focusing on the location of these gains and losses. Indeed, they tend to be correlated so that the origin of replication and the terminus of replication remain diametrically opposed (Bergthorsson et al. 1998).

2.2.2 Chromosomal organization

Gene acquisitions do not occur randomly along the genome. They tend to cluster at specific locations known as 'integration hotspots' (Touchon et al. 2009). These hotspots help to preserve the organisation of the rest of the genome. Indeed, conserved genes organize themselves in the same order in most *E. coli* strains, despite huge variations in the content of accessory genes.

Natural selection may act to preserve this organisation. Indeed, any perturbation can generate a fitness cost for multiple reasons. It could disrupt gene dosage by changing the distance between a gene and the origin of replication. It could also reverse some genes, triggering a conflict between DNA replication and transcription. If a rearrangement shifts the origin of replication, it will result in two chromosomal branches of unequal length between the origin and the terminus of replication, reducing the efficiency of DNA replication. In the long term, genome rearrangements can also be deleterious because they create a barrier to recombination: homologous recombination events will result in gene losses and duplications.

The number of genome rearrangements is reported to correlate with the number of insertion sequences (IS) (Touchon et al. 2009). This is expected because mobile elements trigger intra-genomic recombination. However, we have to consider these reports with caution as insertion sequences also make genome assemblies much more complex and we cannot exclude that some of these rearrangements may simply correspond to assembly artefacts. However, studies have independently reported an increase in IS accompanied by chromosomal rearrangements when species have evolved to an intracellular lifestyle (Moran et al. 2004). These include studies focusing on the transition from *E. coli* to *Shigella* (Jin et al. 2002; Touchon et al. 2009).

2.2.3 GC content

Another simple statistics used to describe genomes is their GC content, *i.e.* the proportion of guanines and cytosines in the DNA. The GC content of the genome varies considerably between bacterial species, ranging from only 13.5% in *Candidatus Zinderia insecticola* (McCutcheon et al. 2010) to 74% in *Micrococcus luteus* (Ohama et al. 1990). It correlates with effective population size and lifestyle, with obligate endosymbionts showing some of the lowest GC contents. *E. coli* stands in the middle with an average GC content of 50.6% (Touchon et al. 2020).

GC content shows little variation within *E. coli* strains, with a standard deviation of 0.14% (Touchon et al. 2020). However, it varies along the genome, particularly in integration hotspots and at the terminus of replication. The former is expected, as these hotspots contain DNA from distant species with different GC content. The observation of AT enrichment at the terminus of replication is not yet fully explained, but has been linked to a lower level of homologous recombination (Touchon et al. 2009).

It has long been thought that the variations in GC content observed between species reflected differences in mutational bias between organisms. This hypothesis was supported by the observation that 4-fold degenerate sites—assumed to be weakly selected for protein function—showed the greatest variation in GC content, while non-degenerate, highly constrained positions of the second codon showed the least variation (Rocha et al. 2010). This argued in favour of GC content variations being induced by neutral mechanisms.

However, several studies contradict this hypothesis. Indeed, most genomes—including GC-rich genomes—show a strong and consistent mutation bias in favour of AT for *de novo* mutations. This bias has been observed both experimentally in *E. coli* (Schaaper et al. 1991; J. Sargentini et al. 1994) and in nature by examining recently emerged mutations in different species (Hershberg et al. 2010; Hildebrand et al. 2010). It could be due to the spontaneous deamination of cytosines to uracils and 5-methylcytosine to thymine when the DNA is single-stranded, usually during transcription. Indeed, induction of transcription increases the frequency of these deaminations on the non-transcribed strand by a factor of four (Beletskii et al. 1996).

In this respect, *Shigella* has an even greater bias for AT mutations than *E. coli*. These are eliminated over time when they do not occur at a 4-fold degenerate site. Together with the observation of a higher dN/dS ratio in *Shigella*, these results support the hypothesis that enrichment in AT mutations reflects weaker purifying selection (Balbi et al. 2009).

If mutation bias were the only factor at play, GC content should reach a much lower equilibrium in many species, including *E. coli* and *Shigella*. Various hypotheses have been put forward to explain the GC content values observed in nature, they are reviewed in (Rocha et al. 2010). One of the most hotly debated issues concerns the role of recombination. Homologous recombination could counteract the AT bias of mutation in at least two ways. First, while most *de novo* GC to AT mutations are deleterious, selection may favour recombination to restore ancestral nucleotides. Second, the recombination process itself may create a GC bias. According to the biased gene conversion hypothesis, mismatches in recombination heteroduplexes are repaired in favour of G and C. This hypothesis, which does not require any action of selection, could explain the observed increase in GC content. In both cases, a GC-rich genome would be the signature of a highly recombinant species. In contradiction with this hypothesis, an increase in recombination rates does not necessarily correlate with an increase in GC

content. Furthermore, newly recombined alleles observed in nature seem to show an AT bias instead of a GC bias (Bobay et al. 2017a). Even if there were indeed a GC bias, it is not clear that recombination is sufficiently frequent to counteract the AT bias of *de novo* mutations. All attempts to study the role of biased gene conversion in nature are necessarily limited by the difficulty of correctly detecting recombination. Furthermore, even an association between the occurrence of recombination and an increase in GC content cannot be taken as evidence for the biased gene conversion hypothesis, as recombination is intrinsically linked to the efficiency of natural selection.

If no neutral process can explain the observed GC contents, an alternative would be that they are selected for. The observation that *E. coli* strains harbouring GC-rich versions of genes grow faster than their AT-rich counterparts supports an adaptive view of GC content (Raghavan et al. 2012). The hypothesis of a selective role for GC content challenges some of the foundations of modern population genetics. In particular, it suggests the absence of any neutral sites in the genome. Indeed, 4-fold degenerate sites that are weakly selected for protein function would then reflect selection for nucleotide composition. This also raises questions about the observed differences between species. Do differences in GC content reflect different local fitness optima or differences in the strength of natural selection to achieve the same global optimum? If high GC content is adaptive, it is also surprising that the mutational processes that take place in bacteria have not evolved to be more biased towards GCs (Rocha et al. 2010).

2.3 Genetic diversity in *E. coli* species

2.3.1 Mechanisms generating genetic diversity

Differences between strains can result from two main types of process: mutation and horizontal gene transfer.

Mutations occur naturally as a result of errors during DNA replication or when DNA is damaged and not properly repaired. In *E. coli*, estimates of the spontaneous mutation rate range from 10^{-4} to 3.7×10^{-3} mutations per genome per generation (Williams 2014). One important mechanism that keeps the mutation rate so low is DNA mismatch repair (MMR). MMR controls the fidelity of DNA replication by repairing errors in base incorporation, insertion or deletion. It also repairs some DNA damage. MMR involves several proteins, including MutS, which recognises mismatched base pairs and unpaired bases, and MutL, which recruits other proteins to repair the error once it has been detected. Therefore, loss of MMR by inactivation of *mutS* or *mutL* genes results in a 100-fold increase in transition rate and a 1000-fold increase in frameshift rate. The strain is then described as a mutator (Denamur et al. 2006).

In addition to mutation, horizontal gene transfer is an important source of genome diversification. To allow stable implantation of foreign DNA into a cell, the DNA must:

1. enter the cell,

2. establish itself as a self-replicating plasmid or by integrating the chromosome,

3. not be lost (through drift or counter-selection).

DNA uptake proceeds by transformation, conjugation, transduction or lysogenic conversion (Figure 2.2).

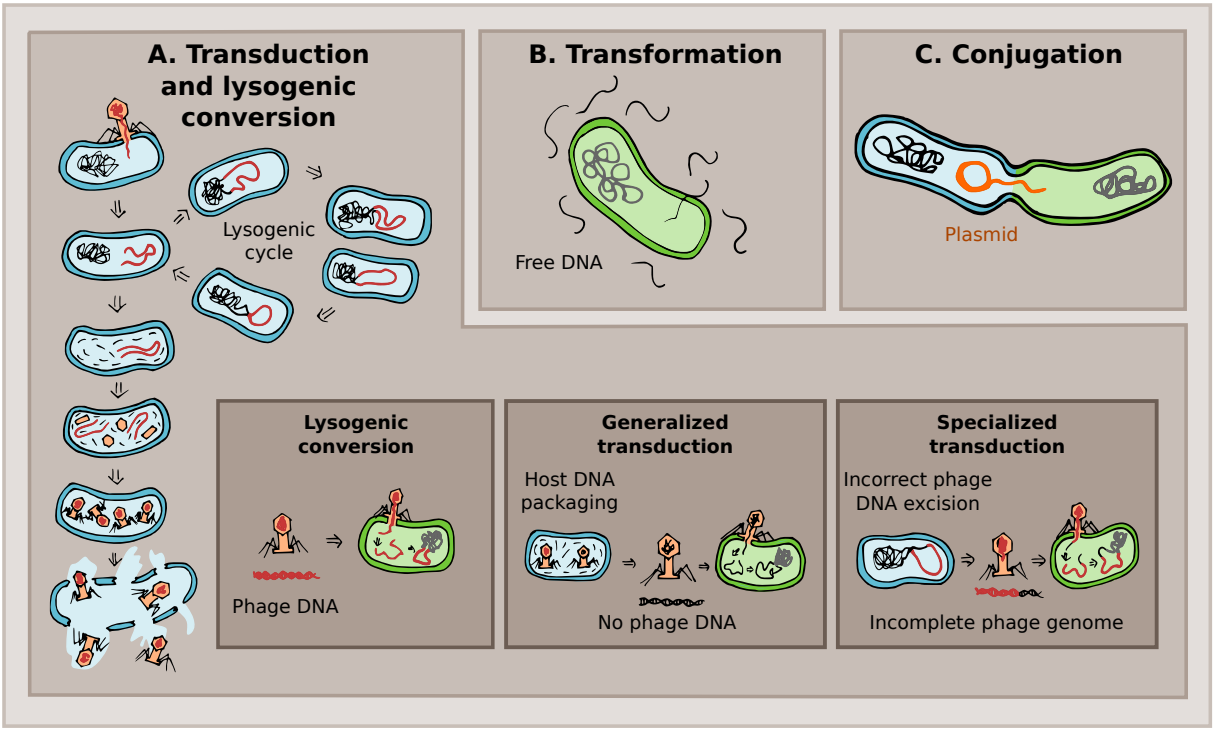

FIGURE 2.2: **The three main modes of foreign DNA acquisition by bacteria.**
Figure A is adapted from (Touchon et al. 2017).

Transformation is the direct uptake of extracellular DNA by the bacteria (Thomas et al. 2005). Transformation is the only one of the three mechanisms that allows the uptake of foreign DNA without the mediation of phages or mobile genetic elements, each having a restricted host range. Transformation requires the bacteria to be naturally competent, a condition that was long thought not to be met by *E. coli* (Mandel et al. 1970). However, some reports indicate a modest level of natural competence for *E. coli* when present in water or food (Baur et al. 1996; Bauer et al. 1999).

Conjugation requires physical contact between the donor and recipient cell and the formation of a pore through which DNA can pass (Thomas et al. 2005). The mobile genetic elements that code for these conjugative systems are carried on plasmids or on a DNA fragment that excises from the chromosome before transfer. Bacteria that code for closely related conjugative systems cannot initiate conjugation, a phenomenon called surface exclusion. The effect of the capsule has not yet been fully studied, but there is some evidence to suggest that conjugation between different serotypes is less efficient and that loss of the capsule increases the rate of conjugation (Haudiquet et al. 2021).

Transduction and lysogenic conversion are phage-mediated uptakes of DNA (Touchon et al. 2017). Lysogenic conversion occurs when the infectious phage integrates into the bacterial genome, resulting in the acquisition of the phage DNA. Transduction happens when the phage has mistakenly packaged host DNA from the bacteria where it was produced (see Figure 2.2.A), resulting in the acquisition of bacterial DNA by the recipient cell. Phages have a narrow host range. In particular, the presence or absence of a capsule and the type of capsule present will determine the types of phage that can infect the cell (Haudiquet et al. 2021).

Once it has entered the cell, the DNA must establish itself either as a plasmid or by integrating the chromosome. The first solution requires that it escapes restriction enzymes and can replicate independently (Thomas et al. 2005). The second requires an additional step of integration into the chromosome. Some mobile DNA fragments integrate into the chromosome by a series of mechanisms known as 'site-specific recombination' because they occur at specific locations in the DNA (Grindley et al. 2006). Homologous recombination (HR) is another mechanism of integration of foreign DNA into the chromosome. HR is a major mechanism of DNA repair. But because it uses a DNA template, it can also generate diversity if the template is not strictly identical to the damaged sequence. HR relies on the RecA protein (Bell et al. 2016). When DNA is damaged, RecA searches for a repair template by identifying a sequence that has perfect homology on a segment covering at least 23 to 27 base pairs: the minimum effective processing segment (MEPS) (Shen et al. 1986). The number of MEPS between two strains decreases exponentially with sequence divergence. This means that the probability of foreign DNA integrating into a chromosome by HR decreases exponentially with sequence divergence (Vulić et al. 1997). MEPS homology checking is under the control of MutS protein (Delmas et al. 2005). Therefore, inactivation of *mutS* gene will result in a recombinant phenotype (in addition to the mutator phenotype already described).

2.3.2 Observed diversity in *E. coli*

Two randomly sampled *E. coli* strains will show on average a 3% divergence in nucleotide composition (Touchon et al. 2009). This figure rises to 6.1% if the *E. coli* species boundaries are extended to include *E. coli* clade I (Cobo-Simón et al. 2023). Conversely, the diversity will be lower when comparing strains from the same phylogroup. Nucleotide diversity within a phylogroup is positively correlated with the distance from its most recent common ancestor: the more time that has passed, the more divergence has accumulated (Touchon et al. 2020).

When one compares the nucleotide sequences of two strains, most differences cluster in two regions (Milkman 1997; Tenaillon et al. 2010). These regions have become known as 'bastions of polymorphisms'. They correspond to the *rfb* operon which encodes the O surface antigen and to a region containing the *fim* and the *hsd* operons. The *fim* operon encodes Type I Pili involved in adhesion and is known to frequently change its transcriptional state. The *hsd* operon—host specificity genes—encodes type I restriction and modification systems that allow bacteria to detect and cut foreign DNA. Because of their function in self-recognition or in coding for proteins exposed at the cell surface, they are subject to strong diversifying selective pressures, which explains their high level of divergence.

In contrast to a relatively low level of overall nucleotide divergence between shared genes, the gene content between two strains diverges by more than 30% (Touchon et al. 2009). While an average *E. coli* strain harbours 4,700 genes, less than 2,500 of these were found in 99% of 1,294 isolates (Touchon et al. 2020). The remaining genes vary from strain to strain.

It has long been known that bacteria can acquire new genes that allow them to express new characteristics (Ochman et al. 2000). Antibiotic resistance genes are often carried by plasmids that strains exchange. Some other traits—such as the ability to ferment a sugar—are linked to the presence of specific genetic operons. A virulence plasmid is also involved in the invasive phenotype of *Shigella* (Sansonetti et al. 1982). However, the extent of variability in the gene repertoire was clearly overlooked before the advent of whole genome sequencing.

In 2005, Tettelin and colleagues asked the question: 'How many genomes are needed to fully describe a bacterial species?' (Tettelin et al. 2005). At the time, they were looking for a universal vaccine that would effectively target all strains of *Streptococcus agalactiae*. To do this, they needed to sequence enough strains to cover the full diversity of the species. Using a quantitative approach, they showed that while the number of genes shared between strains tended to stabilise at around 1,806 genes, the addition of any new strain led to the identification of new genes. They coined the term pan-genome to describe the genome of a species and came to the surprising conclusion that the pan-genome of *Streptococcus agalactiae* was open. The observations of Tettelin and colleagues on *Streptococcus agalactiae* were found to be true for most bacterial species, including *E. coli* (Touchon et al. 2009) (Figure 2.3).

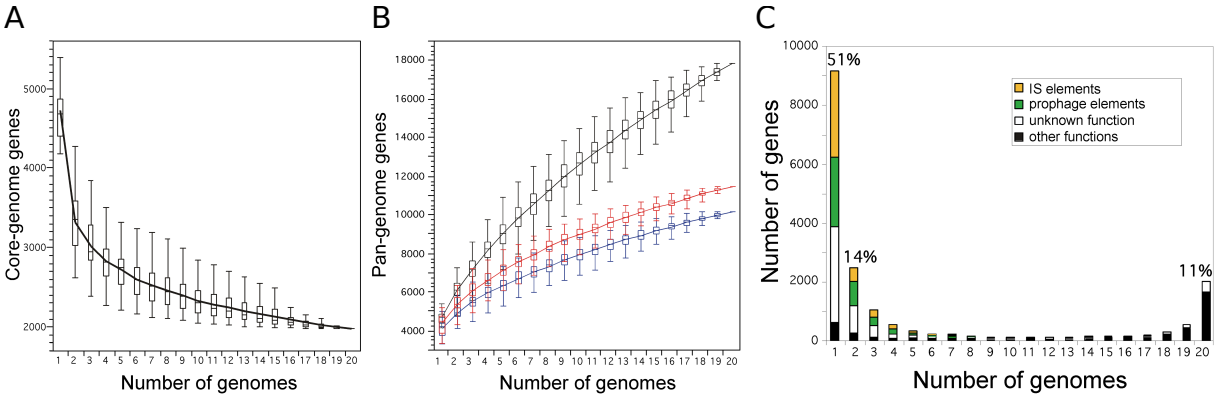

FIGURE 2.3: ***E. coli* core and pan-genome (Figures from (Touchon et al. 2009)).**
A. *E. coli* core genome size according to the number of genomes considered.
B. *E. coli* pan-genome size according to the number of genomes considered.
C. Frequencies of genes across all 20 genomes analysed.

We generally refer to genes shared by all members of a species as core genes and to others as accessory genes. In practice, when studying large datasets, the more flexible definition of a persistent genome that includes genes present in at least 95% or 99% of strains is preferred to that of a core genome that requires genes to be present in absolutely all strains (Touchon et al. 2020). Indeed, as the number of strains increases, the number of core genes should decrease to very low levels due to sequencing artefacts or very rare gene losses.

The pan-genome follows a U-shaped distribution (Figure 2.3.C), with a high peak of very low frequency genes and a lower peak of high frequency genes. The former consists mainly of singletons—genes present in a single genome—and the latter corresponds to the persistent and core genomes (Touchon et al. 2009).

The persistent genome contains many housekeeping genes. It has been suggested that the ancestral genome may be a better representation of the housekeeping functions of the species than a core or persistent genome (Touchon et al. 2009). But as it is less practical to infer, it is rarely used in practice. The size of the *E. coli* core genome decreases when *Shigella* strains are added, consistent with the multiple gene losses that have accompanied the emergence of *Shigella* subgroups.

Although they represent a large part of the species pan-genome, singletons constitute on average less than 1% of the genes in a single strain (Touchon et al. 2020). They include selfish DNA, such as transposable elements and prophage elements (Touchon et al. 2009), as well as some defence sys-

tems and cell envelope genes (Touchon et al. 2020). However, most of them have no clear functional assignment, making it difficult to examine their exact role in *E. coli* evolution. Singletons also tend to be smaller than other genes and are also more likely to be located near a contig edge. This suggests that some of them could correspond to pseudogenes and others to sequencing and assembly artefacts. However, the singletons do not explain the openness of the pan-genome. Even after removing them, the pan-genome remains open. The rest of the accessory genome—composed of genes at intermediate frequencies—shows an overrepresentation of genes involved in cell motility, intracellular trafficking and secretion, carbohydrate transport and metabolism, and secondary metabolism. But again, most genes have no clear functional assignment (Touchon et al. 2020).

Among the genes contributing to the accessory genome are the mobile genetic elements (MGEs). These elements are selfish fragments of DNA that can move within a genome or from one genome to another. They include prophages, plasmids and transposons. Genomes isolated from the same environment tend to have a more similar number of MGEs, although these MGEs are not necessarily the same. This could reflect a higher rate of MGE infection in certain niches or a higher probability of adapting to certain niches through MGE acquisition. The first hypothesis might be more likely as MGEs seem to be mainly a burden to the cell. Indeed, they are always present at low frequency and no MGE belongs to the persistent genome of *E. coli*. Furthermore, most MGE acquisitions are on terminal branches, suggesting that they are rapidly lost (Touchon et al. 2020).

The observation of rapid loss of MGEs raises questions about gene turnover rates. Phylogroups A and B1 have the smallest genomes and persistent genomes and carry fewer MGEs than other phylogroups. Surprisingly, they have the most diverse pan-genome and also carry more diverse MGEs. The smaller genomes could therefore reflect a higher turnover of genes rather than a lower rate of gene acquisition (Touchon et al. 2020).

One might expect the accessory genome to reflect adaptation to specific niches. However, the very low number of clade-specific genes somehow contradicts this view (Touchon et al. 2009). On the other hand, phylogroups still clearly emerge from a principal component analysis conducted on a gene family presence/absence matrix (Figure 2.4.A) (Touchon et al. 2020). In particular, phylogroups A and B1, which are closely related in the phylogeny of the species, are also grouped on the basis of the presence and absence of accessory genes: for these two phylogroups, relatedness in terms of the persistent genome translates into relatedness in gene content. In contrast, phylogroups D and F are quite divergent in species phylogeny but cluster in terms of gene repertoire. For these phylogroups, a common environment could determine the gene repertoire more than vertical inheritance. Gene repertoire relatedness (GRR) can help decipher the link between gene content and population structure. GRR measures the similarity between the gene repertoire of two genomes by dividing the number of common genes by the number of genes in the smaller genome. By comparing GRR to patristic distance—*i.e.* the sum of the branch lengths between two genomes over the phylogeny inferred from the persistent genome—we observe two main regimes (Figure 2.4.B (Touchon et al. 2020)):

- Strains that are very close in terms of persistent genome (low patristic distance) also share a large number of common genes (high GRR). This is normal, as they may have diverged recently and have not had time to fully change their genetic repertoire. However, the GRR varies very strongly with patristic distance, suggesting significant gene turnover.

- For the less related strains, there is not a very strong relationship between GRR and their re-

latedness in terms of persistent genome, as suggested by the very large variance around the regression line.

This paints a dynamic picture of the propensity of *E. coli* to acquire and lose genes. However, the precise function of this accessory genome remains enigmatic, particularly regarding the proportion of genes that genuinely contribute to strain adaptation compared to those that represent a more neutral diversity.

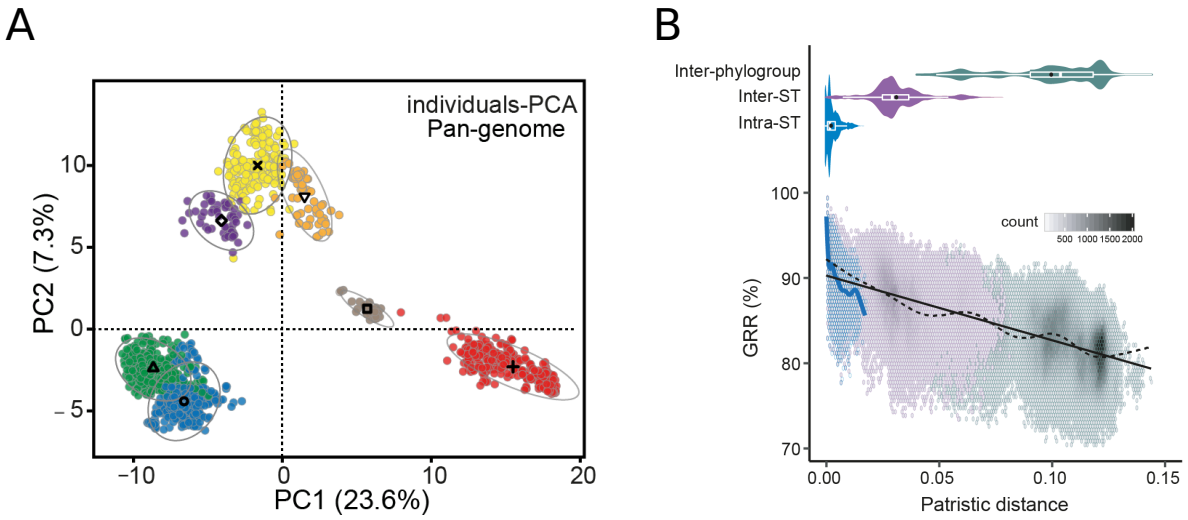

FIGURE 2.4: **Gene repertoire and strains relatedness (Figures from (Touchon et al. 2020)).**
A. Principal component analysis of the pan-genome, phylogroups are shown with the same color code as in Figure 2.1.A.
B. Top: Violin plots of the patristic distance computed between pairs of strains from the same ST (intra-ST, blue), from different ST (inter-ST, purple) and from different phylogroups (inter-phylogroup, green). Bottom: bivariate histogram of the association between Gene Repertoire Relatedness (GRR) and the patristic distance for pairs of strains. The linear fit is shown with a black solid line. The spline fit (generalized additive model) for the whole comparison is shown with a black dashed line and the one for the intra-ST comparison is shown with a blue solid line.

2.3.3 Roles of mutation and homologous recombination in *E. coli* evolution

As observed, *E. coli* exhibits a highly dynamic accessory genome. Nonetheless, despite being less diverse, its core genome is far from being static. Two primary processes—namely mutation and homologous recombination—contribute significantly to its evolution.

Role of mutation

Mutations can be broadly classified into three categories: neutral, deleterious and beneficial. Population genetics predicts that deleterious mutations will be purged by natural selection while beneficial mutations will get selected. In contrast to these, neutral mutations will be able to segregate thanks to random drift and contribute to the species diversification.

In the early days of genetics, most *de novo* mutations were assumed to be deleterious. Therefore, the spontaneous mutation rate had to be as low as possible to prevent them from occurring. Thus, a non-zero spontaneous mutation rate was thought to reflect a balance between the cost of creating deleterious mutations and the metabolic cost of reducing the mutation rate (Drake et al. 1998). The study of mutators has challenged this view by suggesting that beneficial mutations were not so rare and implying that we could witness adaptation occuring on far smaller time scales than previously thought.

Mutators are expected to arise naturally as a result of spontaneous mutations in MMR genes. However, it was soon observed that the rate of appearance of mutators in nature far exceeded the level expected from the mutation/selection equilibrium (Denamur et al. 2006). This suggested an adaptive role for high mutation rates under specific conditions. Indeed, mutation is not always detrimental, as it also feeds evolution with new variants on which natural selection can act (Williams 2014). Therefore, the cost of a high mutation rate will depend on the proportion of deleterious *de novo* mutations relative to those that are neutral or beneficial. These proportions change according to the stage of adaptation (Couce et al. 2015). For a maladapted strain, *e.g.* one that invades a new niche, a large proportion of mutations will be beneficial. However, once it has reached a fitness peak, most of the new mutations will be deleterious. Mutators often emerge in the early stages of adaptation to a new environment by hitchhiking with the beneficial mutations they produce. To do this, mutator cells must first reach an appreciable frequency in the population, a necessary condition for high mutation rates to produce beneficial mutations.

The phylogenetic study of the *mutS* gene in *E. coli* indicates that this species has been able to lose and reacquire *mutS* several times during its history (Denamur et al. 2000). As the loss of *mutS* gene generates a mutator phenotype, this suggests that *E. coli* has evolved through alternating phases of low and high mutation rates. Interestingly, the loss of *mutS* also triggers a highly recombinant phenotype that facilitates the reacquisition of a functional version of *mutS* through homologous recombination. Another interesting feature is the proximity of the *rpoS* and *mutS* genes on the *E. coli* chromosome. The loss of the former is beneficial to *E. coli* when the strain is faced with external stresses such as nutrient deprivation. Large deletion events involving both *rpoS* and *mutS* could be selected for due to the loss of *rpoS* (Bridier-Nahmias et al. 2021). This would allow the corresponding strains to reach sufficiently high frequencies for the mutator phenotype induced by the loss of *mutS* to start being beneficial. The highly recombinant phenotype of the strain would help reacquire a functional *mutS* allele once a high mutation rate is no longer beneficial.

In addition to the emergence of mutators, mutation rates can also increase in response to environmental signals. Indeed, the SOS response—activated in response to high stress—down-regulates the mismatch repair, leading to an increase in mutation rates (Saint-Ruf et al. 2006). As we can see, the field of population genetics has radically shifted from the view that most mutations are deleterious and counter-selected to an adaptive role for high mutation rates.

Role of recombination

We often classify recombination into two distinct categories: homologous recombination (HR) and horizontal gene transfer (HGT). The former replaces an existing DNA fragment with a homologous fragment. The latter introduces new genetic material into the genome. Although these two phenom-

ena can operate by the same mechanisms, they are usually studied separately, because of their very different impact on the evolution of species. They can also complement each other: an initially rare HGT event can spread in a species through homologous recombination at the flanking parts of the foreign DNA region. There is some evidence to suggest that this process is at work around recombination hotspots in bacterial species (Schubert et al. 2009; Oliveira et al. 2017). Indeed, the phylogenies of the core genes flanking these hotspots are often incongruent with the phylogeny of the species—a potential signature of frequent HR. We have already discussed how HGT can influence pan-genome dynamics. Hereafter, we will focus on the effect of homologous recombination on *E. coli* evolution.

Early studies of *E. coli* population genetics based on serotyping and MLEE all detected few haplotypes and strong linkage between alleles (Tenaillon et al. 2010). They concluded that *E. coli* was a clonal species. However, the bacterial mode of reproduction based on binary fission necessarily creates linkage between alleles (Feil et al. 2001). Even frequent recombination may not be sufficient to completely break this linkage. Thus, recombination studies based solely on linkage may have underestimated the evolutionary role of homologous recombination in *E. coli*.

Access to the first DNA sequences allowed the construction of the first phylogenetic trees, opening up new avenues to study the role of homologous recombination in *E. coli*. Indeed, if a gene has undergone recombination, its phylogeny should be incongruent with that of other genes. That is, strains that are clustered on the phylogeny of this gene could be distributed in different parts of the trees of other genes. For example, it was a study of phylogenetic incongruence that suggested the frequent loss and re-acquisition of *mutS* throughout the history of *E. coli* (Denamur et al. 2000).

While access to DNA sequences has undoubtedly provided a more detailed picture of human recombination in *E. coli*, we must bear in mind that DNA sequence analysis does not allow us to access the intrinsic rate of recombination that occurs in the wild. Firstly, because we only detect HR when it brings mutations: we cannot detect recombination events between too closely related strains. Secondly, because HR events pass through the filter of natural selection that eliminates deleterious changes and favours beneficial ones (Touchon et al. 2009). This is evident from the observation that bastions of polymorphisms in the *E. coli* genome—which are known to be subject to diversifying selection—show the highest signals of recombination. However, the more strains we sample, the more we have access to very recent recombination events that have not had time to pass through the filter of natural selection. This highlights one of the potential benefits of analysing large strain datasets.

The literature gives very different estimates of the size of recombination fragments, ranging from 50 bp (Touchon et al. 2009) to 1 kb (Milkman et al. 1993). As these sizes are inferred from direct analysis of DNA sequences, they tend to underestimate the size of the incoming DNA fragment. Firstly, because one side of the recombinant fragment may be perfectly identical to that of the recipient strain it replaces. Secondly, because frequent recombination events may overlap and create a mosaic of short recombinant fragments. For these reasons, the estimation of the relative weight of HR versus mutation in the contribution to new SNPs could also be questionable. However, it is estimated that both generate on average the same level of nucleotide diversity in *E. coli* (Didelot et al. 2012).

Role of mutation and recombination in sexual isolation: is *E. coli* clonal or panmictic?

The relative significance of mutation and recombination plays a crucial role in determining the population structure. When recombination is absent, new variants solely arise through mutation and are vertically inherited, resulting in a clonal population characterized by a perfectly tree-like structure (depicted on the left side of Figure 2.5). Conversely, when recombination greatly surpasses mutation in generating diversity, the species is said panmictic or sexual. In this scenario, each gene possesses its own history, and the species structure is better represented as a network rather than a tree (illustrated on the right side of Figure 2.5). Inferring the phylogeny of a panmictic species leads to a star-like tree with long terminal branches, as strains are, on average, equidistant from each other. Such a phylogeny provides very limited information about the species' history. Between these two extreme regimes, a phylogeny still emerges, albeit potentially disrupted by local recombination events.

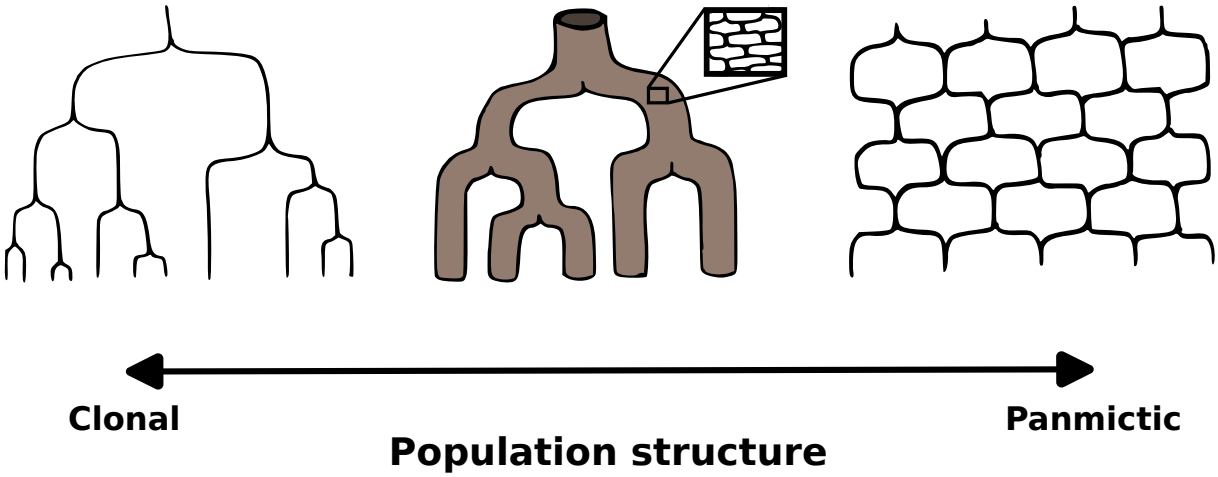

FIGURE 2.5: **Representation of possible population structures.**
These structures range from perfectly clonal—when no recombination is at play—to perfectly panmictic—when recombination strongly dominates. Between both regimes, a somehow structured population can emerge, even if recombination remains frequent within each branch. Figure adapted from (Smith et al. 1993).

Speciation begins when a barrier to recombination arises between strains that previously recombined freely. Sequence divergence forms an important barrier to successful homologous recombination. This is because the frequency of MEPS decreases exponentially with sequence divergence and, with it, the probability of successful homologous recombination. As mutation rate determines the speed at which two strains diverge, it therefore plays a crucial role in the establishment of sexual isolation (Fraser et al. 2007). It is counteracted by recombination: each time a HR event succeeds, the level of divergence decreases sharply (Figure 2.6.A). The balance between mutation and recombination thus defines two main regimes:

- The divergent regime where mutation increases the divergence between each pair of strains and recombination is too weak to counteract it. Sexually isolated groups regularly emerge within the population. The species is clonal.

- The metastable regime where recombination periodically mixes the strains, thus decreasing

the divergence created by mutation. The population retains its cohesion. The species is panmictic.

It should be noted that even when a population is in the metastable regime, stochastic events may prevent recombination for a sufficiently long period of time for some strains to escape and become sexually isolated.

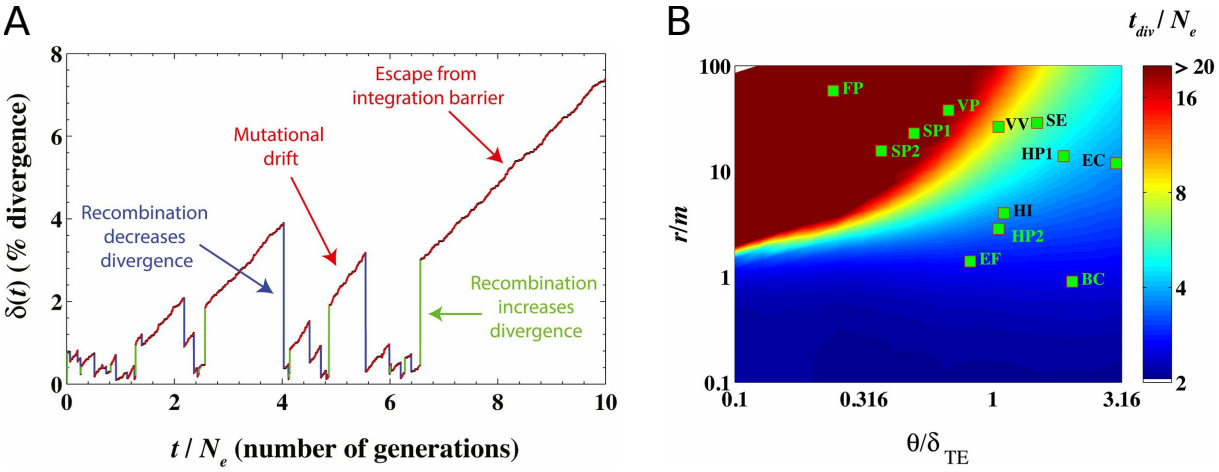

FIGURE 2.6: **Roles of recombination and mutation in the speciation process.**
A. Stochastic evolution of local divergence $\delta(t)$ between pair of strains. While divergence increases linearly due to mutation, it decreases sharply due to random events of recombination between the two strains up to the moment when divergence reaches a threshold above which recombination is no longer possible.
B. The landscape of divergent and metastable regimes according to r/m, the relative strength of recombination over mutation, and θ/δ_{TE}, with $\theta = 2\mu N_e$, the population diversity, and δ_{TE} the transfer efficiency. Bacterial species are mapped onto this landscape, *E. coli* is EC. Figures from (Dixit et al. 2017).

Using estimates of mutation and recombination rates in different bacterial species, Dixit and colleagues have shown that some belong to the former regime and others to the latter (Dixit et al. 2017). As for *E. coli*, it remains on the border between the two regimes but could be classified as divergent: even if recombination is high, it can still be considered a clonal species. The mismatch repair system could also play a role, as it controls both mutation and recombination rates. We know that *E. coli* has experienced episodes of MMR inactivation that have made the bacteria mutator and highly recombinant. This may have altered the balance between mutation and recombination and moved *E. coli* onto the landscape shown on Figure 2.6.B.

With *E. coli* so close to the boundary between clonal and panmictic species, it is not surprising that characterising its population structure remains controversial. Indeed, it all boils down to accurately estimating the mutation and recombination rates in the wild. And estimating these rates, especially the latter, is far from easy.

The observation that many phylogenies of individual genes are consistent with the species phylogeny obtained from the concatenation of all core genes supports the clonality of *E. coli*. Furthermore, coalescent simulations also supported the idea that the observed recombination rate did not obscure the phylogeny (Touchon et al. 2009). The global aspect of *E. coli* phylogeny also pleads in favour of clonality. Indeed, a freely recombining species should present us with a star-like phylogeny, whereas *E. coli*'s phylogeny is much more structured with clearly visible clades.

In 2021, Sakoparnig *et al.* challenged this widely accepted view (Sakoparnig et al. 2021). Using a dataset of 91 strains, they compared strains two by two, counting the occurrence of single nucleotide polymorphisms (SNPs) in a sliding window of 3 kb. Peaks of high SNP density were interpreted as recombination events. This allowed them to provide new estimates of the occurrence of recombination in *E. coli*. From these estimates, they concluded that recombination was too frequent to maintain a clonal structure. They justified the structured aspect of *E. coli* phylogeny by the existence of unequal recombination rates between different strains: strains belonging to the same clade would frequently recombine with each other but rarely with strains from other clades. However, these explanations suffer from some weaknesses. There is no clear mechanism that could underlie these supposedly unequal recombination rates. Indeed, different phylogroups may be found in the same environment, so that they are not geographically isolated and they are not sufficiently divergent for sequence divergence to prevent recombination. Furthermore, if phylogeny reflects recombination rates, it is not clear why regions known to recombine frequently (*e.g.* core genes flanking recombination hotspots) are the least consistent with the species phylogeny. The methodology followed in the paper may also present some limitations. Indeed, the recombinant regions are identified as those with a higher-than-average SNP density. This seems adequate for closely related strains where recombination is more likely to provide novelty. However, this may be misleading for more distantly related strains where the recombinant regions are more likely to be those with lower-than-average SNP density. We can also see some contradiction between estimating recombination rates based on the idea that recombination increases strain divergence and explaining the emergence of the structured phylogeny of *E. coli* based on the idea that recombination prevails mainly between very closely related strains.

In summary, we are still far from fully understanding the population structure of *E. coli*. But there is little doubt that any study of *E. coli* phylogeny must try to account for recombination.

2.3.4 Is this diversity neutral or selected?

Not all mutations that emerge in a population meet the same fate. Most disappear, some reach fixation and others persist for a long time at intermediate frequencies. Population genetics aims to identify the forces that decide these different fates.

The advent of the neutral theory in the late 1960s marked a major advance in this field. This theory—formulated by Motoo Kimura (Kimura 1968) and independently by Jack L. King and Thomas H. Jukes (King et al. 1969)—proposes that random drift, rather than natural selection, is responsible for most of the observed genetic diversity: this diversity would reflect random occurrences of equally adapted variants. In contrast to neutralists, adaptationists (Mayr 1983) consider that most differences between organisms result from adaptation to different environments or from balancing selection.

A comparison of the occurrence of non-synonymous mutations and synonymous mutations has enabled the predictions of these theories to be tested. Synonymous mutations are changes in the DNA that do not alter the protein sequence. They are therefore expected to have a much smaller effect than non-synonymous mutations, which induce an amino-acid change. According to neutralists, natural selection works mainly by eliminating harmful mutations rather than by promoting genetic diversity. They therefore predict that natural selection will eliminate more non-synonymous mutations than synonymous mutations. With the exception of some very specific loci that have been shown to be subject to diversifying selection, the predictions of the neutral theory are broadly consis-

tent with observed reality: in most of the genome, non-synonymous diversity segregates much less than synonymous diversity (Ohta et al. 1996).

In the 1970s, Tomoko Ohta refined the neutral theory by focusing on the role of slightly deleterious mutations (Ohta 1973). She suggested the existence of a class of near-neutral mutations whose effect on fitness is sufficiently small that natural selection cannot act effectively on them. These mutations—mainly slightly deleterious variants—follow a dynamic driven by genetic drift, as do neutral variants. A mutation is nearly neutral if its fitness s verifies: $|s| \sim 1/N_e$, with N_e the effective population size. It follows that natural selection will be more effective in large populations, while random drift will dominate the fate of mutations in small populations.

In recent years, the growing interest in the role played by historical contingency has revived the old 'neutral'-versus-'selective' debate (Starr et al. 2016). Evolutionary contingency arises when mutations that have reached fixation depend on permissive mutations that have occurred previously. Once fixed, they influence the fate of future mutations and become increasingly deleterious to eliminate—a phenomenon called entrenchment (Shah et al. 2015). The concept of contingency places epistasis at the forefront of molecular evolution. An amino acid that is neutral or beneficial in one genetic context may be deleterious in another due to epistatic interactions between residues (Breen et al. 2012). Similarly, a gene may be beneficial in one strain but deleterious in another due to interactions with other loci in the genome—a concept known as genome-wide epistasis. Characterising these epistatic interactions is therefore essential for analysing the diversity we observe within and between species and for understanding the extent to which contingency shapes molecular evolution.

Chapter 3

Studying *E. coli* evolution in practice

3.1 Experimental evolution

So far, we have seen how isolating *E. coli* strains from different ecological niches and studying their genomes could help decipher the evolution of this species. In doing so, we collect current isolates and aim to infer past events. A complementary approach to understanding the evolution of *E. coli* consists in following its adaptation to a new environment. This is the purpose of experimental evolution.

In experimental evolution, scientists introduce a strain into a controlled environment and track its evolution in real time. This type of setting makes it possible to test theoretical predictions of population genetics and to directly measure evolutionary parameters such as mutation rate or distribution of fitness effects. It is important to bear in mind that the environment in which we let the strain evolve is generally far less complex than a natural one. Thus, we should always take the findings of experimental evolution with care and not generalize them too quickly. That being said, two types of experimental settings have played a crucial role in improving our understanding of *E. coli* evolution: the long-term evolution experiment (LTEE) (Lenski et al. 1991) and mouse gut colonisation (Gordo et al. 2014).

The LTEE is an ongoing experiment launched by Richard Lenski in 1988. Twelve populations of *E. coli* REL606 grow at 37°C on DM25 medium. Each day, 1% of each population is transferred to a new flask. Every 500 generations, a sample of each population is frozen to serve as a 'fossil record'. Of the twelve populations, six can grow on arabinose (Ara+ variants) and the other six cannot (Ara- variants). This otherwise neutral marker allows for competition experiments between replicates to assess changes in fitness. Overall, the LTEE can be described as a very basic and stable environment with no ecological interactions between *E. coli* and other species.

The mouse gut represents a more realistic environment. In this context, a strain of *E. coli* will have to invade and stably colonise the gut of a mammal. In doing so, it will interact with the host immune system, evolve in a structured environment and sometimes compete with the resident microbiota. Different mouse models have been developed to address different biological questions. The germ-free mouse—a mouse without microbiota—mimics the colonisation of the gut of newborns during the early stages of life. As colonisation of a gut with a mature microbiota rarely succeeds, the streptomycin-treated mouse offers an alternative where *E. coli* will still have to compete with a resident microbiota, but with better chances of stably colonising the gut. This second model mimics

a more mature gut that has lost its colonisation resistance—a situation that occurs after antibiotic consumption or in the case of severe inflammation. Other models involve the choice of different host states (young or old (Barreto et al. 2020a; Barreto et al. 2020b), immunocompetent or immunocompromised (Barroso-Batista et al. 2015)) and help to decipher the determinants of *E. coli*'s interaction with its host.

Although the two types of environments offered by the LTEE and the mouse gut differ radically, a series of very similar observations have been made in both contexts. These common observations cover a range of theoretical questions from the predictability of evolution to the evolution of mutation rate, as well as the existence of clonal interference or the role of epistasis in evolution.

The early stages of evolution appear to be highly reproducible, with the same set of mutations occurring repeatedly from one replicate to another. Later stages of evolution can depend on the mutations that previously reached fixation in each sample. As these mutations may differ between replicates, the later stages of evolution will be more contingent and less reproducible (Tenaillon et al. 2016; Gordo et al. 2014).

Mutation is the currency of evolution. Without new mutations, a strain cannot adapt. The mutation rate is therefore a crucial factor in adaptation. Interestingly, the mouse gut and the LTEE have seen the emergence of mutators (Tenaillon et al. 2016; Ramiro et al. 2020). However, increasing the mutation rate is not the only way to adapt to a new niche. Indeed, in both contexts, mutators have emerged in some replicates, but not in all. In cases where strains remained non-mutators, they accumulated mutations at the same rate in the mouse gut and in the LTEE (Frazão et al. 2022).

When beneficial mutations emerge at low rates, they reach fixation one after the other. This evolutionary dynamic follows successive selective sweeps. But if many beneficial mutations emerge at the same time, they compete with each other: in the absence of recombination, we observe the Hill-Robertson effect (Hill et al. 1966). In bacteria, this phenomenon is known as clonal interference. In addition to preventing most of the beneficial mutations that arise in the population from reaching fixation, clonal interference also limits their rate of fixation. In the LTEE and in the mouse gut, pervasive clonal interference has been observed (Maddamsetti et al. 2015; Barroso-Batista et al. 2014b). Sometimes different mutations leading to the same phenotype compete with each other. This leads to fixation of the new phenotype without loss of genetic diversity.

Epistasis occurs in evolution when the effect of one mutation depends on the presence or absence of other mutations. In other words, epistasis reflects the interaction between different genetic loci. In the LTEE, a pattern of diminishing return epistasis emerges: mutations that fix early tend to have a greater beneficial impact on fitness than if they fix later (Maddamsetti et al. 2015). An even more radical case of negative epistasis occurs when mutations are mutually exclusive. In the mouse gut, *dcuB* and *focA* are two frequent targets of adaptation. However, they are never mutated together (Barroso-Batista et al. 2014b). This is probably due to their redundancy, as they both modulate anaerobic respiration: it is sufficient to mutate one of the two genes to obtain the desired phenotype, there is no additional benefit to mutate the second.

The processes of adaptation occurring in genomes during experimental evolution also leave a signature on sequences. Classical population genetics approaches such as those based on the comparison of synonymous and non-synonymous mutations have proven effective in detecting those signatures (Tenaillon et al. 2012; Tenaillon et al. 2016).

An important difference between the LTEE and the mouse gut is the role of ecological interactions.

By design, the LTEE aims to limit ecological interactions: *E. coli* is the only cultured species, a poor medium and strong bottlenecks also limit the emergence of sub-lineages that would adapt to specific niches. In contrast, an *E. coli* colonising a mouse gut will need to interact with the host and with other species in the resident microbiota, if any. When mice harbour a resident microbiota, evolution becomes less predictable (Barroso-Batista et al. 2020). Horizontal gene transfer events also occur, typically between the resident *E. coli* and the new strain (Frazão et al. 2019; Barreto et al. 2020a; Frazão et al. 2022). These can be triggered by phages or correspond to plasmids transferred by conjugation.

Although the design of the LTEE limits the emergence of ecological interactions, two clones (S and L) with different metabolic capabilities eventually emerged at generation 6,000 (Bull 2000; Maddamsetti et al. 2015). The L type was better at growing on fresh media and during this phase secreted a metabolite that was consumed by S, giving the latter an advantage during later stages of growth. Other coexisting phenotypes based on differences in sugar consumption were also observed in the mouse gut (Sousa et al. 2017).

Strikingly, many of the processes involved in the LTEE have also been observed during the evolution in the mouse gut. This degree of convergence between very different experimental settings suggests that experimental evolution can inform the evolutionary dynamics that occur in nature. Furthermore, the longest evolutionary experiment to date—the LTEE—has now reached a point where its time scale matches events that have occurred in nature, such as the expansion of the ST131 clone. Bridging the gap between evolution in the laboratory and evolution in the wild is likely to yield valuable insights into the dynamics of evolution.

3.2 Protein mutational landscapes

During the course of an evolution experiment, we often detect mutations that increase fitness. If we except trivial situations such as a premature stop codon that inactivate a gene, it is most of the time very difficult to guess what these mutations do. In other terms, it is challenging to relate a genotype to a phenotype.

Data-driven approaches may help to characterise the relationship between the genotype of an organism and its phenotype. Applied at the scale of a single protein rather than that of the whole organism, they generate 'protein mutational landscapes' (Figliuzzi et al. 2016). A protein mutational landscape is a function that takes an amino-acid sequence as input and outputs the level of functionality of the corresponding protein. It allows to investigate how mutations can affect this level of functionality. It thus informs about the paths that evolution may follow.

Protein mutational landscapes have great similarities with fitness landscapes (Figure 3.1). However, the latter are inferred at the organism level. They provide information on the effect of a given mutation on the survival and reproductive potential of the organism. The mutational landscape of a protein cannot provide such information, as the relationship between protein functionality and an organism's fitness depends on many factors. For example, an enzyme that degrades antibiotics may greatly improve the fitness of a bacteria exposed to antibiotics, but will be completely useless in an environment without antibiotics. A mutation that decreases the activity of this enzyme will then be highly deleterious in the former environment but neutral in the latter. However, if we focus on the core genome of *E. coli*, we can make the plausible assumption that most of the proteins it encodes are used at some point, so that the mutational landscapes of these proteins will be closely related to

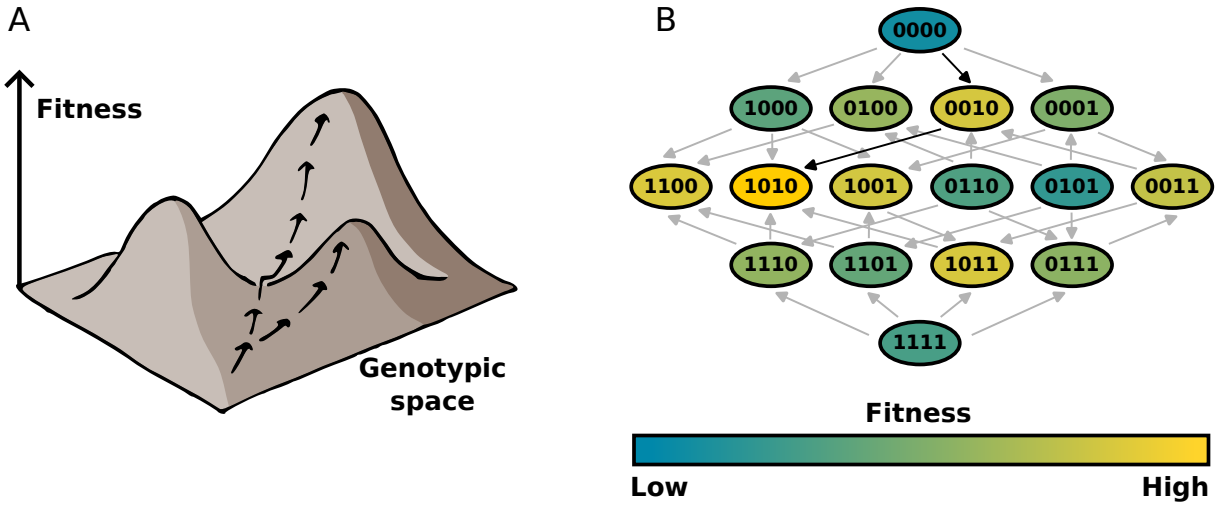

FIGURE 3.1: **The concept of fitness landscape (adapted from (Visser et al. 2014)).**
A. Illustration of a fitness landscape with three fitness peaks and two evolutionary paths.
B. Empirical fitness landscape of all possible combinations of four diallelic loci. Arrows link genotypes differing by a single mutation and point toward the most fit genotype of the pair. Black arrows show the shortest path between the wild type (0000) and global maximum.

the fitness landscape of the organism.

The pervasiveness of epistasis makes the construction of these landscapes a complex task. Indeed, the effect of a mutation often depends on the rest of the genetic background (Starr et al. 2016). This means that characterising the effect of three individual mutations may not be sufficient to guess the effect of the combination of these three mutations if they are introduced together in the same sequence. Epistasis may result from a non-linear relationship between a biological trait, such as the stability of a protein, and the observed phenotype. For example, slightly deleterious mutations may additively decrease the stability of a protein to a point where the protein can no longer fold. The first mutations will then have very little impact on the functionality of the protein compared to the last mutation that crosses the threshold between the folded and unfolded states of the protein. This type of epistasis is known as global or non-specific epistasis. It is opposed to specific epistasis which involves direct or indirect interactions between mutations, usually amino-acid sites that are in contact in the three-dimensional fold of the protein.

Protein mutational landscapes can be derived from experimental data or inferred by *in silico* methods. Deep mutational scans allow to characterise experimentally the effect of individual mutations (Jacquier et al. 2013). However, reconstruction of the adaptive landscape also requires characterising the effect of combinations of mutations. This quickly becomes experimentally unfeasible, as the number of mutation combinations grows exponentially with the number of sites studied. Therefore, only adaptive landscapes limited to a very small set of sites have been derived from experimental data (Schenk et al. 2013).

In silico approaches allow the modelling of much more complex mutational landscapes. Some of the most promising *in silico* approaches rely on evolutionary information to estimate the effect of mutations. Some proteins or protein domains have evolved independently in different species over millions of years. In the course of evolution, they have accumulated many mutations but have managed to remain functional: while beneficial and near-neutral mutations have been able to reach

fixation, most deleterious mutations have been eliminated by natural selection. In other words, when we sample homologous protein sequences from distant species, we have access to some of the local maxima of our landscape. Evolutionary methods use these amino-acid sequences observed in nature to try to infer the overall shape of the mutational landscape (Figliuzzi et al. 2016) (Figure 3.2). More concretely, they try to fit a function that takes an amino-acid sequence as input and outputs the probability of observing that sequence in nature. This function should give a very good estimate of the level of functionality of the protein, as protein sequences frequently observed in nature are likely to be functional, whereas non-functional proteins will be counter-selected and therefore associated with a very low probability of being observed in nature.

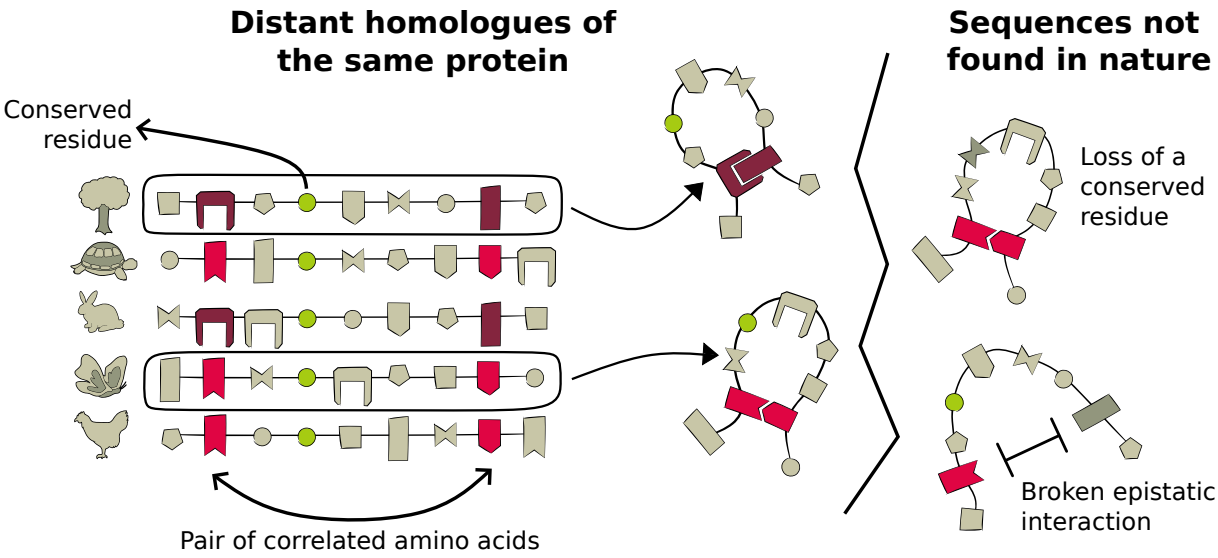

FIGURE 3.2: **Evolutionary methods to study protein sequences.**
Homologues of the same protein found in distant species can be aligned in a multiple sequence alignment (MSA). The study of this MSA allows to detect patterns of conservation (green) and of correlation between sites (red). These reflect functional constraints that limit the range of protein sequences that can be found in nature.

Most evolutionary approaches can be classified as independent sites (or non-epistatic) methods (IND). They focus on conservation patterns across homologous protein sequences (green residue on Figure 3.2). For example, if in all species a tryptophan is observed at locus 123 of the protein of interest, it is very likely that any mutation occurring at this locus will profoundly affect the functionality of the protein. On the other hand, if all twenty amino acids are represented at approximately equal frequencies at locus 321, we can assume that this locus is selectively neutral. In other words, any mutation occurring at locus 321 will be close to neutral. IND methods include SIFT (Sorting Intolerant From Tolerant) (Ng et al. 2003) and PolyPhen (Polymorphism Phenotyping) (Adzhubei et al. 2013). In addition to evolutionary information, PolyPhen also takes into account structural information. However, no IND method can take into account epistatic interactions between sites (red residues on Figure 3.2). In contrast to IND, Direct-Coupling Analysis (DCA) (Morcos et al. 2011) takes into account the above mentioned conservation patterns and also captures coevolutionary patterns between pairs of sites. Sites that interact epistatically will tend to coevolve: when a site mutates, its partner must mutate accordingly to maintain the interaction. This leaves a detectable signature of correlation between amino-acid sites when distant amino-acid sequences are compared with each

other. It is this signature that DCA uses to retrieve epistatic couplings between amino-acid sites. This epistatic approach makes DCA context-aware, as opposed to IND which is context-agnostic. DCA detects residues in contact in the three-dimensional fold of the protein, as these residue pairs correlate strongly (Morcos et al. 2011). DCA has also been used in a variety of other contexts: protein design (Russ et al. 2020), prediction of deep mutational scanning outcomes (Figliuzzi et al. 2016) and the study of amino-acid changes between two closely related genomes (Couce et al. 2017). In all these applications, it has consistently outperformed IND. There are epistatic approaches other than DCA that give very similar results (Riesselman et al. 2018; Laine et al. 2019). However, DCA is explicitly parameterised in terms of epistatic couplings and conservation, making it interpretable.

3.3 Bringing different approaches together to answer biological questions

Sequence analysis with the classical tools developed in population genetics, evolution experiments, and protein mutational landscapes, offer three complementary approaches to studying species' evolution. Each of these approaches sheds light on important evolutionary topics. Within this section, I will concentrate on three specific subjects that have proven relevant to my research: the role of metabolism in niche adaptation, the transition from commensalism to pathogenicity and the acquisition of antibioresistance.

3.3.1 The role of metabolism in niche adaptation

Evolutionary experiments provide valuable insight into the early targets of adaptation to new ecological niches. The regulation of metabolic pathways repeatedly emerges as one of these targets. In the LTEE and in the mouse gut model, *E. coli* benefits from the loss of specific metabolic operons or from their constitutive activation due to the loss of a repressor. For example, all twelve LTEE populations lost D-ribose catabolism during the first 2000 generations of the experiment. This loss improved the fitness of *E. coli* by 1% to 2% (Cooper et al. 2001). The *gat* operon is also lost very quickly during the colonisation of the mouse gut (Barroso-Batista et al. 2014b). In this case, a mutational hotspot could contribute to the emergence of this phenotype. However, other genes involved in metabolism are also frequent targets of adaptation to the mouse gut. For example, in the gut of a streptomicin-treated mouse, *E. coli* constitutively activates sorbitol metabolism (Barroso-Batista et al. 2014b). In another experiment, led with a natural isolate of *E. coli*, it is the *dgo* operon, involved in the galactonate pathway, that was constitutively activated due to the loss of its repressor *dgoR* (Lescat et al. 2017). After two weeks of colonising the gut of a germ-free mouse, *E. coli* loses *lrp*, a gene regulating amino-acid catabolism (Barroso-Batista et al. 2020). This loss enhances *E. coli*'s ability to compete for amino acids, particularly serine and threonine.

Evolutionary experiments have highlighted the key role of metabolism in niche adaptation. Interestingly, wild-type isolates of *E. coli* show differences in metabolic capacity that may correspond to adaptation to different niches. For example, strains from phylogroup B1—a phylogroup often found in the environment—carry genes involved in the degradation of rhamnose, sucrose, xylose, glycerate and tartrate (Touchon et al. 2020). These pathways are likely to play a role in colonising plants. In

contrast, phylogroup B2 is negatively associated with these traits. Metabolic genes present in all *E. coli* also display different transcriptional profiles between environmental and enteric strains (Walk et al. 2009). In the gut, *E. coli* feeds on mucus. Mucus is a complex growth medium composed of at least 12 sugars. Different strains of *E. coli* have different abilities to grow on each of these sugars (Foster-Nyarko et al. 2022). When *E. coli* invades the mouse gut, it also consumes these sugars in a hierarchical order (Chang et al. 2004).

Altogether, these elements give a dynamic picture of the metabolic capabilities of *E. coli*. These vary from strain to strain and change readily when a strain invades a new niche.

3.3.2 Transitioning from commensalism to pathogenicity

Mapping *Shigella* and enteroinvasive *E. coli* (EIEC) strains onto the species phylogeny reveals a striking pattern: multiple occurrences in unrelated parts of the tree (Wirth et al. 2006; Pasqua et al. 2017). This observation supports the hypothesis of multiple emergence of these pathotypes, suggesting a high level of convergence in the series of evolutionary events leading to the emergence of an invasive phenotype.

Again, experimental evolution can help to uncover the processes involved. When cultured in the presence of macrophages, commensal *E. coli* strains repeatedly increase their ability to survive intracellularly and to escape from macrophages (Proença et al. 2017). Both of these characteristics are essential to the invasive process that characterises bacillary dysentery. Experimental evolution thus shows that *E. coli* can acquire pathogenic traits as a direct result of its evolution under the pressure of the host's innate immune system. Evolved populations show a reduced ability to grow on single carbon sources, suggesting that adaptation to an intracellular lifestyle comes at the cost of more specialised behaviour (Azevedo et al. 2016). This latter finding is consistent with the restricted host range of EIEC and *Shigella* compared to the generalist behaviour of the rest of *E. coli*.

When analysing the patterns of gene loss and gain along the *E. coli* phylogeny, the main evolutionary event leading to the EIEC and *Shigella* pathotype is the horizontal acquisition of a pINV plasmid carrying many virulence determinants (Pasqua et al. 2017). Other pathogenicity islands may be acquired subsequently, but in a less systematic way. In addition to the acquisition of virulence determinants, the strains also lost other genes such as *ompT* or *cad*. *ompT* encodes a protease that reduces the invasive potential of the bacteria. Thus, pathoadaptation is not only achieved through the acquisition of virulence genes but also through targeted inactivations.

A greater challenge is to study the transition from commensalism to pathogenicity among ExPEC strains. Indeed, as they are not obligate pathogens and mainly cause opportunistic infections, the determinants of their virulence are less easy to identify (Denamur et al. 2021). Strains causing urinary tract infections (UTIs) are often also present in the patient's gut as the dominant *E. coli* in the faecal flora. Similarly, phylogroups B2 and D tend to be over-represented among gut-resident strains and among strains responsible for extra-intestinal infections. These observations support the prevalence theory which states that strains that successfully colonise the gut are also more likely to cause extra-intestinal infections, because when they reside in the gut they have ample opportunity to move to other organs. This view opposes the special pathogenicity theory, which supports the existence of specific genes that enhance the virulence potential of some *E. coli* strains over others (Katouli 2010).

The existence of such virulence factors can be studied in animal models. As mentioned in section

1.5, in a mouse model of sepsis, some *E. coli* strains exhibit a killer phenotype and others a non-killer phenotype (Picard et al. 1999). Interestingly, the genetic determinants associated with the killer phenotype may also promote colonisation of the gut. They include adhesion factors, siderophores, polysaccharide capsules and toxins. Adhesion factors may play a key role in promoting colonisation of the epithelial mucus layer by *E. coli*. It is known that iron scavenging is absolutely crucial for the survival of bacteria in the host. Capsules and toxins are likely to help bacteria resist predation by other species or infection by phages within the microbiota. It should be noted that extra-intestinal diseases most likely represent an evolutionary dead end for *E. coli*, as the strain will either rapidly kill its host or be eliminated by antibiotics and the host immune system. In addition, the strain will have very little opportunity to spread to another host during the course of the infection. This last element is still debated in the particular case of urinary tract infections, where the infectious strain may have some opportunities to spread through urine. In most cases, it is likely that the selection for virulence determinants comes from their association with improved survival in the gut. These observations contributed to the hypothesis that extra-intestinal virulence emerged as a by-product of adaptation to the gut (Le Gall et al. 2007). In other words, pathogenicity could derive from commensalism. This hypothesis links the theories of prevalence and special pathogenicity. Indeed, 'virulence determinants' would be selected to favour colonisation of the gut. Strains carrying these virulence determinants will therefore be more likely to stably inhabit the gut, giving them the opportunity to infect other organs—as predicted by the prevalence theory. Once infection has occurred, the same virulence determinants could in turn increase the severity of the disease by increasing survival in extra-intestinal compartments—as predicted by the special pathogenicity theory. The line between mutualism and pathogenicity can be very difficult to draw: the factors that make *E. coli* Nissle 1917—one of the most studied probiotics— so successful in preventing other *E. coli* strains from causing intestinal disease are very similar to the determinants of virulence mentioned above. Indeed, Nissle uses adhesion factors and siderophores to stably colonise the gut and out-compete these pathogenic strains (Foster-Nyarko et al. 2022).

During extra-intestinal infection, *E. coli* can acquire patho-adaptive traits. One of the most studied is the appearance of point mutations in the *fimH* gene coding for type I fimbriae adhesion (Denamur et al. 2021). These point mutations modulate the binding capacity of *E. coli* and have been shown to be crucial in urinary tract invasion. Some gene inactivations were also identified in diseases, such as that of *lrhA*, a type 1 fimbriae and flagellum repressor (Kisiela et al. 2017), or *rbsR*, the ribose operon repressor (Bridier-Nahmias et al. 2021). These results suggest that motility and sugar metabolism play an important role in the ability of a strain to invade an extra-intestinal compartment.

In summary, *E. coli* could increase its virulence through three main mechanisms (Denamur et al. 2021). Firstly, by horizontally acquiring pathogenicity islands: the pINV plasmid for *Shigella* and EIEC, the high pathogenicity island (HPI) encoding iron acquisition systems for ExPEC strains. Secondly, by inactivating certain genes: *ompT* or *cad* for *Shigella* and EIEC, *lrhA* or *rbsR* for ExPEC. Thirdly, by mutating genes, as was observed with *fimH*.

3.3.3 The acquisition of antibioresistance

As a commensal of the human gut, *E. coli* frequently faces antibiotic treatment, even when not the direct target. This context favours the acquisition of antibiotic resistance, which in turn could worsen

the prognosis of an infection when *E. coli* translocates to an extra-intestinal body site. Although mostly detected when *E. coli* causes disease, the acquisition of antibiotic resistance in this species could result from an adaptation to the human gut (Tedijanto et al. 2018).

The study of antibiotic resistance in *E. coli* strains isolated from wild animals highlights the importance of human exposure in the spread of resistance (Skurnik et al. 2006). Strains isolated from animals in an area devoid of humans show no resistance, and levels of resistance increase with human density. For *E. coli* species, humans and human activities are most likely the main reservoir of antibiotic resistance.

Antibiotic resistance involves a wide range of mechanisms as summarised in Figure 3.3. These include preventing access to the target of the antibiotic, altering the target or directly inactivating the antibiotics. At the molecular level, they can result from point mutations, acquisition of new genes by horizontal gene transfer or changes in the regulation of existing genes—typically by the inactivation of a repressor.

When grown in the presence of ciprofloxacin, *E. coli* readily acquires deleterious mutations in the *marR* and *acrR* genes (Praski Alzrigat et al. 2021). These two genes repress the AcrAB-TolC efflux pump. Their inactivation therefore leads to overexpression of these pumps, which expel the antibiotic from the cell. Inactivation of *marR* comes at a cost because this gene also regulates other pathways. In contrast, *acrR* only regulates the pump and its inactivation has no detectable growth cost. Evolutionary experiments to identify mutations that would compensate for the loss of *marR* found only mutations that restored growth by reducing resistance. This illustrates a trade-off between resistance and growth for mutations targeting *marR*. We observe this trade-off in the clinic. Indeed, clinical strains isolated from patients treated with ciprofloxacin show only mildly deleterious mutations in *marR*, but no complete inactivation. This contrasts with *in vitro* experiments that can easily select for complete inactivation of *marR* when *E. coli* is cultured in the presence of ciprofloxacin. The human body certainly represents a more complex environment where the acquisition of resistance occurs at a higher growth cost than that observed in *in vitro* assays.

TEM-β-lactamases are enzymes that hydrolyse penicillins and cephalosporins (Salverda et al. 2010). In 1963, researchers identified the first TEM variant—TEM-1—in a patient carrying a penicillin-resistant bacteria. Since the 1980s, we have witnessed a 'β-lactamase cycle' in which each new antibiotic brought to market led to the emergence of new β-lactamases that cause resistance to that antibiotic. For this reason, β-lactamases have been extensively studied and used as a model for evolution. Figliuzzi and colleagues showed that DCA-based mutational landscapes can accurately capture the genetic context of TEM-1 and predict the extent of resistance that arises from variants obtained through large-scale mutagenesis (Figliuzzi et al. 2016).

Since the first β-lactamase variants emerged in the second half of the twentieth century, we now have to cope with even more worrying variants: the extended-spectrum β-lactamases (ESBLs) (Bezabih et al. 2021). ESBLs are a class of β-lactamases that hydrolyse a wide range of antibiotics. They are carried on plasmids that can spread horizontally among bacteria. Moreover, they tend to be acquired from strains that already have resistance to other classes of antibiotics, leading to the emergence of multidrug-resistant clones that pose a significant therapeutic threat. An example of such a clone is *E. coli* ST131 (Nicolas-Chanoine et al. 2014). This B2 clone is of great concern because of its tendency to infect extra-intestinal body sites. Most ST131s are resistant to fluoroquinolones due to point mutations in the chromosomal genes *gyrA* and *parC*. In addition, some lineages also produce ESBL—

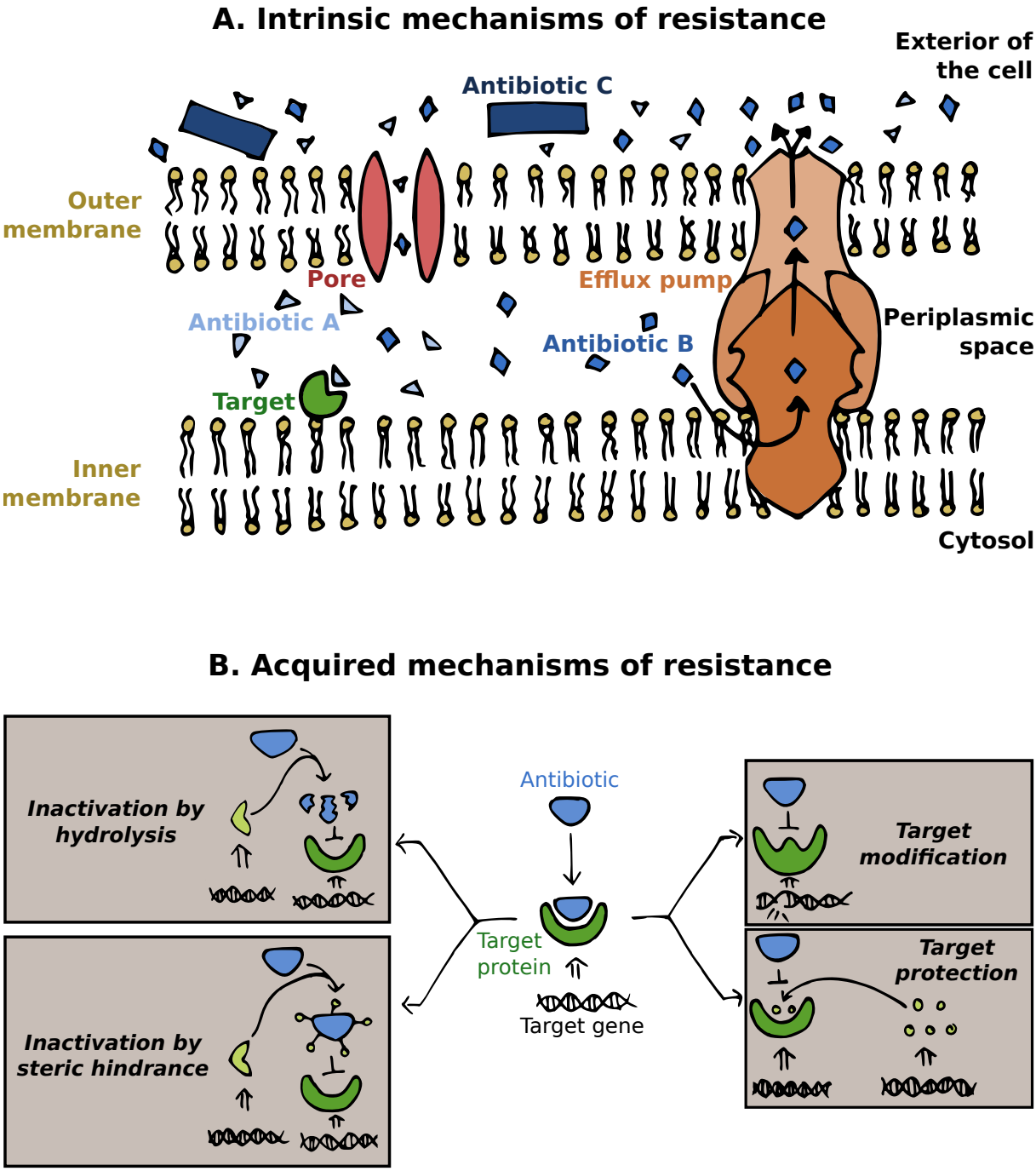

FIGURE 3.3: **Mechanisms of antibiotic resistance (adapted from (Blair et al. 2015)).**
A: Of the three antibiotics A, B and C, only A is able to reach its target. Antibiotic C cannot enter the cell while B reaches the periplasmic space but is pumped out before binding to its target.
B: Bacteria can acquire resistance to a specific antibiotic through two main types of mechanisms: changes to the target and interaction with the antibiotic. The former can proceeds by a mutation in the target gene that prevents the binding of the antibiotic to the target protein or by the action of an auxiliary gene that allows to chemically modify the target protein without changing its amino-acid sequence. The latter involves the production of an enzyme that hydrolyses the antibiotic or modifies its structure to prevent it from binding to its target.

typically CTX-M-15 which confers resistance to cefotaxime—while non-ESBL lineages have often acquired other plasmids that confer resistance to ampicillin and amoxicillin. The evolutionary success of ST131—which went almost undetected in the early 2000s before rapidly increasing in frequency and stabilising at an intermediate level—suggests that this clone has found a way to compensate for the fitness costs associated with multi-resistance.

If multidrug resistant *E. coli* clones with no associated costs were to emerge, one would expect them to invade the population and replace the drug-sensitive clones. However, a longitudinal study of strains isolated from bacteremia over a period of 10 years showed that ST131 stabilised, while the drug-sensitive ST73 clone managed to remain stable and dominant (Kallonen et al. 2017). To explain their results, the authors invoke the existence of frequency-dependent negative selection, but they have no other elements to support their hypothesis. What is certain is that the dynamics of resistance spread are complex and not yet well understood. The maintenance of drug-susceptible clones in an era of intensive antibiotic use is reminiscent of the conflicting results regarding the role of antibiotics on the residence of strains in the gut. While some studies have reported cases where antibiotic use coincided with a change in residency, others have not observed such a change. In particular, one study reported the case of a resident strain that remained for a year despite being exposed to and sensitive to tetracycline (Martinson et al. 2020). This shows that *in vitro* resistance tests cannot encapsulate the complexity of mechanisms that allow a strain to survive antibiotic treatment. These mechanisms include resistance, tolerance, protection by biofilm or cooperation with other members of the microbiota that can confer collective antibiotic resistance by degrading antibiotics.

In the three topics we have explored so far, it is evident that both in natural and experimental evolution, a diverse range of mutations can contribute to *E. coli* adaptation. These mutations encompass not only the acquisition of new traits but also numerous instances of gene losses, particularly involving transcriptional regulators.

3.4 Objectives of this thesis

While the 1990s saw the first complete genome sequences, the 2000s witnessed the rise of the first population genetic studies based on the comparison of a few genomes. At that time, acquiring new genetic data required a lot of work, time and money, thereby limiting the size of the datasets used for the studies. In the 2010s, the advent of high-throughput sequencing dramatically changed this. With over 250,000 *E. coli* genomes already available on Enterobase (Zhou et al. 2019), collecting large genetic datasets is no longer a challenge. The question is: is it worth it? What can be discovered with so much data that could not be studied with a handful of carefully selected high-quality genomes? Although this question may seem a bit provocative, we should keep in mind that the notion of a pan-genome was discovered with only 8 genomes of *S. agalactiae* (Tettelin et al. 2005) and that one of the reference studies on the genetic diversity of *E. coli* and *Shigella* is based on the analysis of 20 genomes (Touchon et al. 2009).

Three main motivations underlie the use of such large data sets:

1. In large samples, rare events become frequent. An event that has a 0.1% chance of occurring will rarely be observed in a dataset of 20 genomes, but on average should be found 80 times in a dataset of 80,000 genomes. This gives us statistical power to detect and analyse these events.

It also means that a large dataset gives a complete picture of the true genetic diversity of *E. coli* in nature: an event never observed in 80,000 genomes should be extremely rare in nature.

2. A large number of genomes allows the construction of much more detailed phylogenies. A phylogeny of 20 genomes has a few branches with many events occurring on each branch. In contrast, a phylogeny of 80 000 genomes has many more branches, giving us a much more detailed view of the history of species and a clearer idea of the order in which events took place.

3. In line with 2., a larger dataset also allows access to much more recent events. Indeed, once you have sampled the genomes of most of the major clades, any new genome you add to your dataset will be closely related to one of those already included and will therefore only differ in recent mutation, recombination, deletion or duplication events. These recent events are interesting because they give a picture of what happens in the wild before selection has time to act, *e.g.* events of DNA acquisition by HGT before they have been filtered out by selection. Paradoxically, they can also give access to some short-term dynamics of selection, typically the first stages of adaptation to a new ecological niche.

In this manuscript, we aim to explore the natural diversity among more than 80,000 *E. coli* strains. To do so, we will take advantage of modeling approaches—particularly DCA—to interpret and make sense of this observed diversity. Our main focus is to gain insights into how this diversity builds up over various time scales. Specifically, we seek to answer questions such as: Are the short-term dynamics of drift and selection comparable to the long-term ones? Can evolutionary approaches based on the study of distant species be applied to predict diversity within a species? How does the genetic context evolve over time? To address these questions, we will investigate different genetic scales, ranging from individual mutations to the variability of amino-acid sites, the intensity of natural selection on various genes, and the diversity of gene repertoires among closely related strain clusters.

Chapter 4 of this manuscript outlines the process of analyzing and organizing 80,000 *E. coli* strains into a database. In Chapter 5, we delve into the prediction and analysis of individual mutations. Chapter 6 is dedicated to studying the variability of amino-acid sites, both within the *E. coli* species and across distant species. Moving on to Chapter 7, we explore the signatures of natural selection at the gene level. Finally, Chapter 8 serves as the conclusion, discussing the main results presented in this manuscript and providing suggestions for future research directions.

Chapter 4

Building a database of 81,440 *E. coli* and *Shigella* genomes

4.1 Motivation

The goal of this thesis is to explore the natural diversity of a single species, *E. coli*, through the analysis of more than 80,000 genomes. However, working at such a scale also presents some difficulties. Firstly, large databases are inherently biased. Unlike some *E. coli* collections that have been carefully constructed to reflect the true diversity of the species in nature, public databases tend to be crowded with specific clones of clinical interest and also lack isolates from non-human sources. Secondly, the quality of the data available in public databases is variable, with high-quality complete genomes alongside poorly sequenced and assembled contig files. These data therefore need to be carefully cleaned before being analysed. Last but not least, most current bioinformatics softwares do not scale well with the size of the datasets, typically if their complexity is quadratic.

In the present work, we decided to build a database of genes present in 80,000 *E. coli* and *Shigella* strains. Working at the gene level allows us to circumvent most of the difficulties related to algorithmic complexity. Indeed, due to the low level of nucleotide diversity in *E. coli*, we rarely observe more than 1000 distinct sequences of a single gene in our 80,000 strains. This means that we can perform most of our analyses on these fewer distinct sequences and weight them according to their representation in the 80,000 genomes. Gene sequences are also valuable for population genetics studies because they offer us a range of tools for analysing evolutionary dynamics, *e.g.* comparison of synonymous and non-synonymous mutations, protein mutational landscapes, etc. We can also annotate them and try to infer their functions, which allows us to interpret the results in a more biologically relevant way.

The upcoming chapter will include very technical sections as it outlines the procedures undertaken to construct a database comprising 81,440 strains of *E. coli* and *Shigella*.

4.2 Identifying homologous genes

81,440 genomes were downloaded from Enterobase (Zhou et al. 2019). These genomes were initially provided as assembled contig files. The number of contigs within each file varied from 1 to 6,624 with a median value of 226 (Figure 4.1.B).

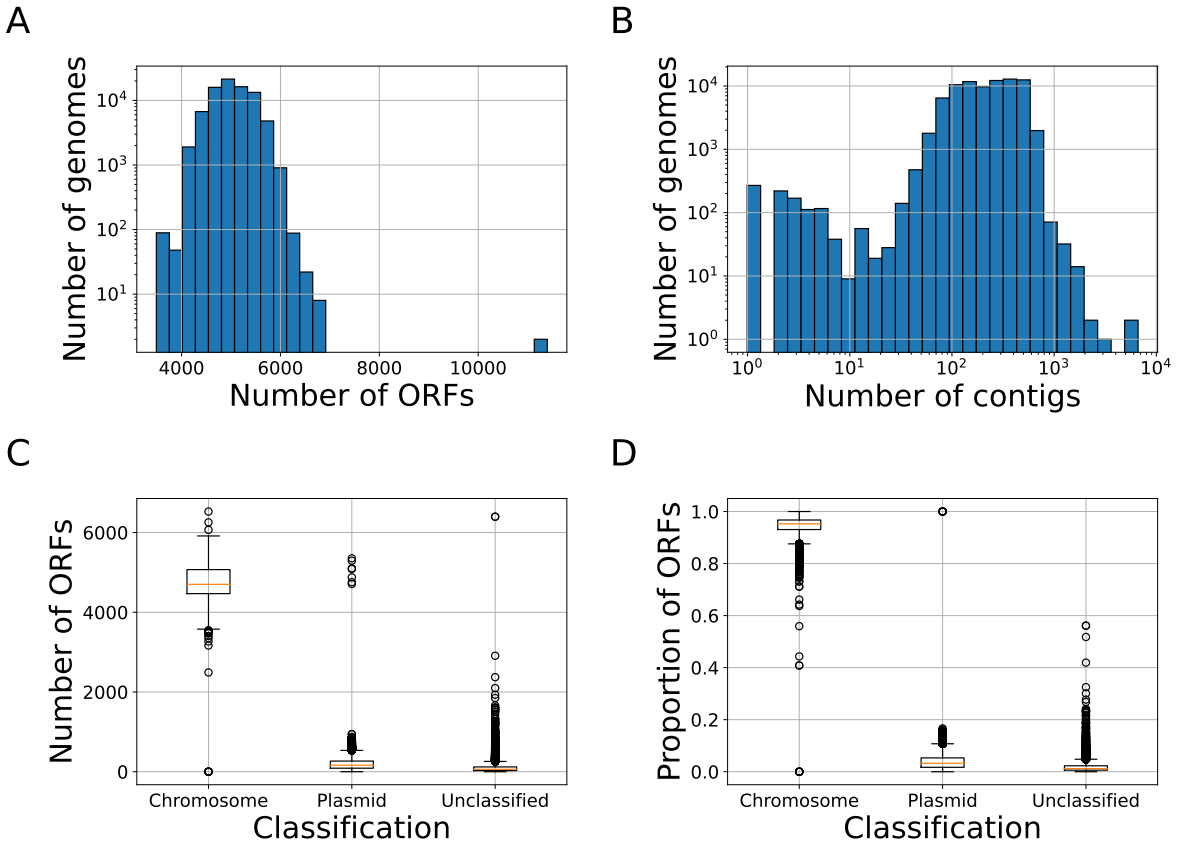

FIGURE 4.1: **Open reading frames (ORFs) and contigs in each of the 81,440 genomes.**

A. Distribution of the number of ORFs—as identified by Prodigal (Hyatt et al. 2010)—per genome.

B. Distribution of the number of contigs per genome.

C. Number of ORFs per genome located on contigs classified by PlaScope (Royer et al. 2018) as chromosomal, plasmidic or not classified.

D. Proportion of ORFs per genome located on contigs classified by PlaScope (Royer et al. 2018) as chromosomal, plasmidic or not classified.

To determine the phylogroup of each genome, we first processed these contig files using both the *in silico* ClermonTyping and the Mash genome-clustering methods (Ondov et al. 2016; Beghain et al. 2018). According to the metadata, 70,301 of these genomes are *E. coli* and the remaining 11,139 are *Shigella*. However, since the metadata could be inaccurate or incomplete at times, we employed ShigEiFinder (Zhang et al. 2021) to classify all the genomes. This software identifies 59 *Shigella* serotypes and 22 enteroinvasive *E. coli* (EIEC) serotypes, providing a more precise classification compared to the available metadata.

We also used PlaScope (Royer et al. 2018) to classify the contigs within the files as either plasmidic or chromosomal. This classification is valuable for conducting detailed analyses of gene transfer. In the current dataset, this information holds even greater significance as certain plasmids may appear to be missing not because they are absent in a particular strain, but rather because they have not been sequenced. By focusing on chromosomal genes, analyses pertaining to rates of gene loss and acquisition can be more dependable and accurate.

Next, we followed the steps outlined in Figure 4.2 to identify homologous genes. We screened each contig file using Prodigal (Hyatt et al. 2010) to detect open reading frames (ORFs). Prodigal also detects partial ORFs situated at the edges of contigs. Contig edges often correspond to repetitive sequences that are present multiple times in the genome, making their assembly challenging. Genes truncated by the end of a contig may be pure assembly artefacts or indicative of noteworthy events like duplications or transposon insertions. Although we do not analyse these partial ORFs in this manuscript, we have included them in the database for future research purposes. The median number of ORFs on plasmidic contigs is 163, corresponding to about 3.2% of the total ORFs in the genomes (Figures 4.1.C and 4.1.D). This percentage is relatively high but not excessively so. Seven contig files exclusively contain plasmidic ORFs, but all the others comprise less than 20% plasmidic genes.

Prodigal identified a total of 409,049,104 ORFs across the 81,440 genomes. Two contig files contained exactly 11,397 ORFs, indicating the likely merging of two distinct genome sequences. In contrast, the remaining contig files ranged from 3,490 and 6,894 ORFs, with a median of 5,001, which aligns with our understanding of *E. coli* (Figure 4.1.A). These 409,049,104 ORFs correspond to 12,783,641 distinct DNA sequences that can be translated into 8,923,612 distinct amino-acid sequences. Following quality filtering, where only complete ORFs containing less than 5% unknown amino acids were retained, we obtained a final set of 5,061,335 distinct amino-acid sequences.

This number of sequences is small enough to use MMseqs2 (Steinegger et al. 2017) for clustering them with a 90% identity and 80% coverage threshold. MMseqs2 is a leading software for clustering amino-acid sequences, and we anticipate it will perform well with our dataset. However, it is important to note that it employs heuristics, which may not always yield optimal clustering results. We may encounter situations where a sequence that meets the inclusion criteria (90% identity, 80% coverage) is not assigned to the appropriate cluster (red sequence on Figure 4.2). Additionally, there may be instances where a single cluster is split into multiple clusters (blue sequences on Figure 4.2), or where a sequence is assigned to a cluster without meeting the inclusion criteria (brown sequence and yellow cluster on Figure 4.2).

To ensure the reliability of the clustering results, we employed Clustal Omega (Sievers et al. 2014) to generate multiple sequence alignments (MSAs) for each of the 402,134 clusters identified by MMseqs2. These MSAs allowed us to derive an amino-acid consensus sequence for each cluster. Sub-

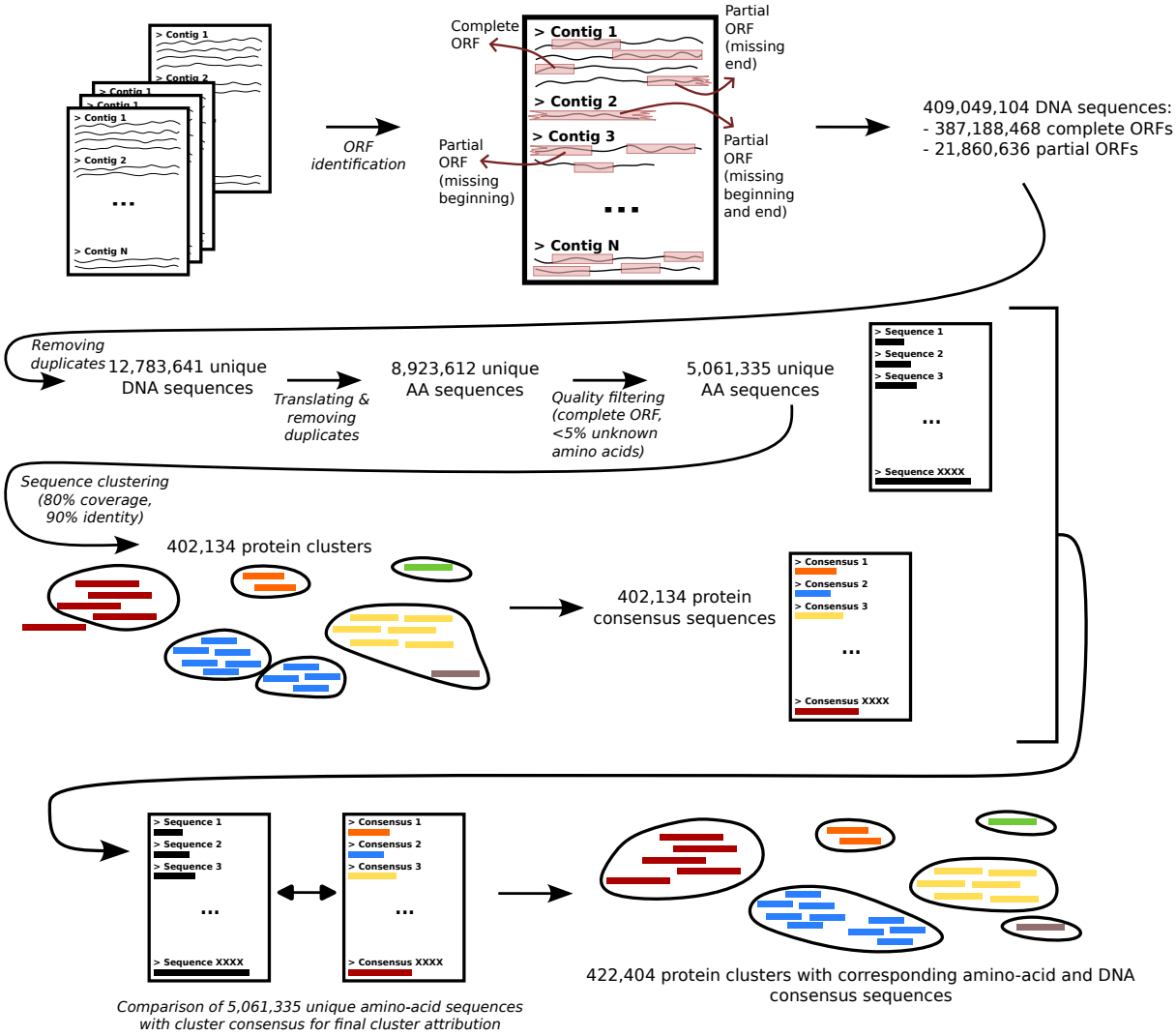

FIGURE 4.2: **Steps followed to identify coding sequences and cluster them.**
We run Prodigal (Hyatt et al. 2010) on each of the 81,440 contig files in order to identify open reading frames (ORFs). These can be complete ORFs or partial if truncated by a contig edge. A first clustering step is performed by MMseqs2 (Steinegger et al. 2017) on unique amino-acid sequences that meet quality criteria. We compute the corresponding consensus sequences of the protein clusters derived by MMseqs2. To avoid potential errors made by MMseqs2 clustering, we compare each of the unique amino-acid sequences to the cluster consensus. If a sequence shares at least 90% identity in sequence and 80% coverage with a consensus, it is assigned to the corresponding cluster. If it matches several consensus, it is assigned to the largest cluster. If it does not meet inclusion criteria to any of the existing clusters, we create a new cluster for this sequence.

sequently, we compared the 5,061,335 amino-acid sequences to these 402,134 consensus sequences. If a sequence exhibited a minimum of 90% identity and 80% coverage with a consensus sequence it was assigned to the corresponding cluster. In cases where a sequence matched multiple clusters, it was assigned to the largest cluster. Sequences that did not match any clusters were categorized as singletons, *i.e.* clusters consisting of only one sequence. As a result, we obtained a total of 422,404 clusters, each accompanied by a consensus amino-acid sequence and a consensus DNA sequence.

We can assess the quality of our clustering by examining some basic plots. Our primary expectation is that a gene would be rarely duplicated within a genome, meaning that sequences from the same protein cluster should originate from different genomes. When this assumption holds true, the number of distinct genomes represented in a protein cluster matches the number of sequences within that cluster. To visualize this, we can create a scatterplot comparing the number of sequences per cluster to the number of genomes represented (Figure 4.3.C). We observe that most data points align along a diagonal line with the equation $y = x$, indicating a close match between these two quantities. This suggests that our clustering process was effective.

Additionally, when we analyze the number of distinct genomes represented in each protein cluster, we observe the familiar U-shaped histogram (Figure 4.3.B). The same pattern emerges when we plot the number of sequences per protein cluster (Figure 4.3.A), although there are some clusters that contain more than one sequence per genome, albeit they are in the minority. Specifically, only 30 clusters have more than 85,000 sequences, and this number decreases to 24 when considering clusters with more than 90,000 sequences. Overall, these statistics provide confidence in the quality of our clustering procedure.

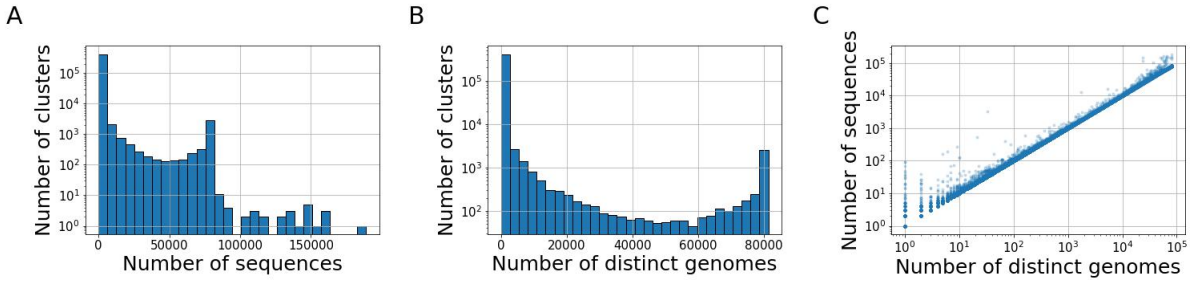

FIGURE 4.3: **Number of sequences and distinct genomes per protein cluster.**
A. Histogram of the number of sequences in each protein cluster.
B. Histogram of the number of distinct genomes in each protein cluster.
C. Scatterplot of the number of sequences against the number of distinct genomes found in each protein cluster.

4.3 Annotating genes

To make the most of this data, we need annotations, *i.e.* we need to link a biological function to a protein cluster. To achieve this, we compared each of the consensus amino-acid sequences of the protein clusters with the amino-acid sequences of Swiss-Prot. Swiss-Prot, a subset of UniProtKB (Boutet et al. 2016), currently comprises 569,213 protein sequences that have been manually annotated by experts.

This annotation step also provides valuable information to analyse horizontal gene transfer. When

one of our clusters matches a protein sequence in Swiss-Prot, we gain access to insightful information such as the species to which the protein sequence belongs and the degree of identity between the two sequences. This enables us to investigate and study the occurrence of horizontal gene transfer more effectively.

4.4 Identifying pseudogenes

Up until now, we have assumed that each cluster corresponds to a unique gene. However, it is possible that certain clusters actually represent fragments of another gene broken by a premature stop codon (as depicted in Figure 4.4). To identify these clusters, we employed VSEARCH (Rognes et al. 2016) to conduct global alignments by querying all the DNA consensus sequences against themselves. We recorded all the resulting hits, but those most likely to correspond to pseudogenisation events are those that meet the following criteria:

- The global alignment covers >95% of the smallest DNA sequence and <80% of the longest DNA sequence.
- The identity between the two DNA sequences is greater than 95%.
- The smallest DNA sequence aligns with the beginning—*i.e.* the global alignment starts within the first 30 nucleotides—or the end—*i.e.* the global alignment ends within the last 30 nucleotides—of the longest DNA sequence.

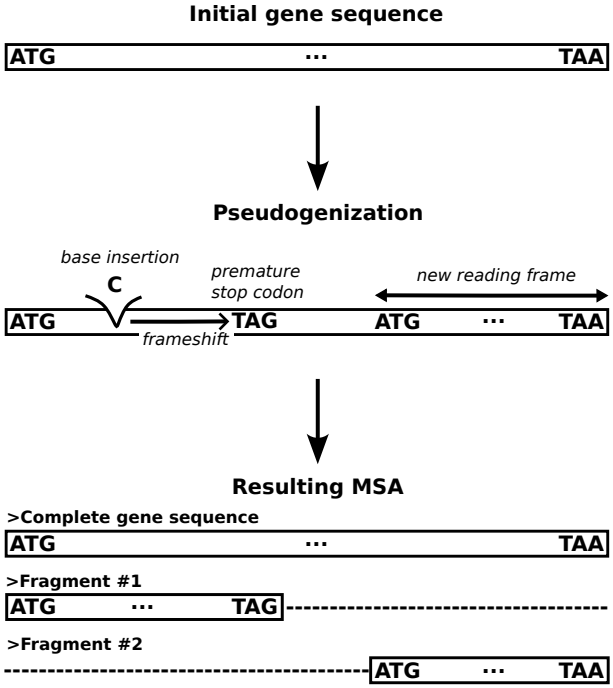

FIGURE 4.4: **Example of a pseudogenization event.**
In this example, the pseudogenization event is caused by a base insertion resulting in a frameshift. It leads to two open reading frames: one at the start of the gene and the other at the end of it.

In cases where a gene is frequently pseudogenized throughout the 81,440 genomes, we anticipate observing numerous hits involving pairs of clusters that both represent fragments of the same gene. However, our focus lies on pairs consisting of a gene fragment and its corresponding complete gene. Therefore, when we encounter multiple hits involving the same small DNA consensus sequence, we prioritize selecting the hit that involves the long DNA consensus sequence that is most prevalent in the database. This long DNA consensus sequence is deemed the 'true' complete gene sequence. By employing this approach, we aim to accurately identify and distinguish between gene fragments and complete gene sequences in our analysis.

After following this procedure, we obtained a total of 46,012 gene clusters identified as fragments of other genes. Among these clusters, there were 12,175 fragments that belonged to genes appearing at least 1000 times in the database.

4.5 Clustering genomes by similarity

To begin, we gathered a persistent *E. coli* genome at a 95% threshold. This low threshold prevents a few poor quality genomes from artificially decreasing the size of the persistent genome. Subsequently, we excluded any genes from the persistent genome that exhibited duplication in more than 1% of the genomes. This step ensures that the persistent genome primarily consists of unique genes and mitigates the influence of duplicated genes on our analysis.

For each of the 3,016 persistent genes, we aligned the corresponding DNA sequences while preserving the codon alignment. We used these 3,016 multiple sequence alignments to construct a concatenated persistent genome alignment of all 81,440 genomes. Note that when a gene was duplicated in one of the genomes, it was not included in the alignment—*i.e.* in that genome it was replaced by gaps. We used this concatenated alignment of persistent genomes to construct a distance matrix to capture the genetic distances between the 81,440 genomes. This distance matrix was created using Hamming distance on the non-gaped sites. This approach constitutes a first step to assess the relationships among the genomes in our dataset.

To cluster the distance matrix, we employed the DBSCAN algorithm (Ester et al. 1996) with parameters: epsilon=0.5% and minimum samples=5. These settings dictated that two genomes would be considered neighbors if their divergence per site is below 0.5%. The minimum samples parameter imposes that the smallest cluster DBSCAN can produce must contain at least five genomes (one genome and its four neighbours). Through this approach, we obtained a total of 240 genome clusters. It's worth noting that DBSCAN allows for the identification of outlier genomes that do not fit into any cluster. In our analysis, we found 597 such outlier genomes, which formed individual singleton clusters.

At this stage, we have an opportunity to assess the accuracy of genome annotations provided by ShigEiFinder. Our expectation is that *Shigella* and EIEC strains will cluster together in a few distinct groups. We can also examine whether the ShigEiFinder annotations align with the Enterobase metadata. Out of the 240 clusters, six clusters consist of over 99% of strains identified as *Shigella* by ShigEiFinder. These clusters predominantly include strains that are also annotated as *Shigella* in the Enterobase metadata, with proportions ranging from 87.5% to 100% depending on the cluster.

After excluding clusters with less than 5% of strains identified as *Shigella*, we find four additional clusters with varying proportions of strains identified as *Shigella* by ShigEiFinder, ranging from 12.9%

to 87.4%. However, none of these clusters contain any strains annotated as *Shigella* in the Enterobase metadata, which raises doubts about the accuracy of ShigEiFinder's identification for those particular clusters.

Furthermore, three clusters exclusively consist of strains identified as EIEC by ShigEiFinder, and one cluster contains 11% of EIEC strains. Unfortunately, we cannot compare these findings with the Enterobase metadata as it does not provide information about the EIEC status of strains.

In the remaining sections of this manuscript, when we refer to analyses conducted on *Shigella* or EIEC clusters, we specifically mean the six clusters that include more than 99% of *Shigella* strains and the three clusters that consist of 100% of EIEC strains based on the annotations provided by ShigEiFinder.

4.6 Database structure

By following the steps outlined in the previous sections, we successfully extracted diverse data from the collection of 81,440 genomes. To ensure easy accessibility for future research, we decided to organize this data into a SQL database. The structure of this database is depicted in Figure 4.5, providing a clear framework for storing and retrieving the information gathered from the genomes. It is composed of 17 tables, including:

- The **genomes** table where each entry correponds to a given strain with its Enterobase metadata as well as its phylogroup and genome cluster.
- The **contig** table where each entry correponds to a given contig in a given genome together with its classification as plasmidic or chromosomic.
- The **genes** table where each entry correponds to an ORF in a given contig of a given genome together with the id of the corresponding unique DNA sequence of this ORF.
- The **seq_ids** table that makes the correspondance between the id of a unique DNA sequence and the id of the corresponding unique amino-acid sequence.
- The **dna_sequences** table where each entry correponds to a unique DNA sequence id together with its nucleotide sequence.
- The **aa_sequences** table where each entry correponds to a unique amino-acid sequence id together with its amino-acid sequence.
- The **proteins** table where each entry corresponds to a unique amino-acid sequence id together with the id of the protein cluster to which it belongs.
- The **consensus** table where each entry corresponds to a protein cluster id together with the id of its amino-acid consensus sequence and the id of its DNA consensus sequence.
- The **fragments** table—built to detect pseudogenization events—where each entry corresponds to a hit between two protein clusters.
- The **annotations** table where each entry corresponds to a hit between a protein cluster and a Swiss-Prot entry.

All the other tables correspond to Swiss-Prot data. This organized database structure enables efficient exploration and utilization of the data for further studies.

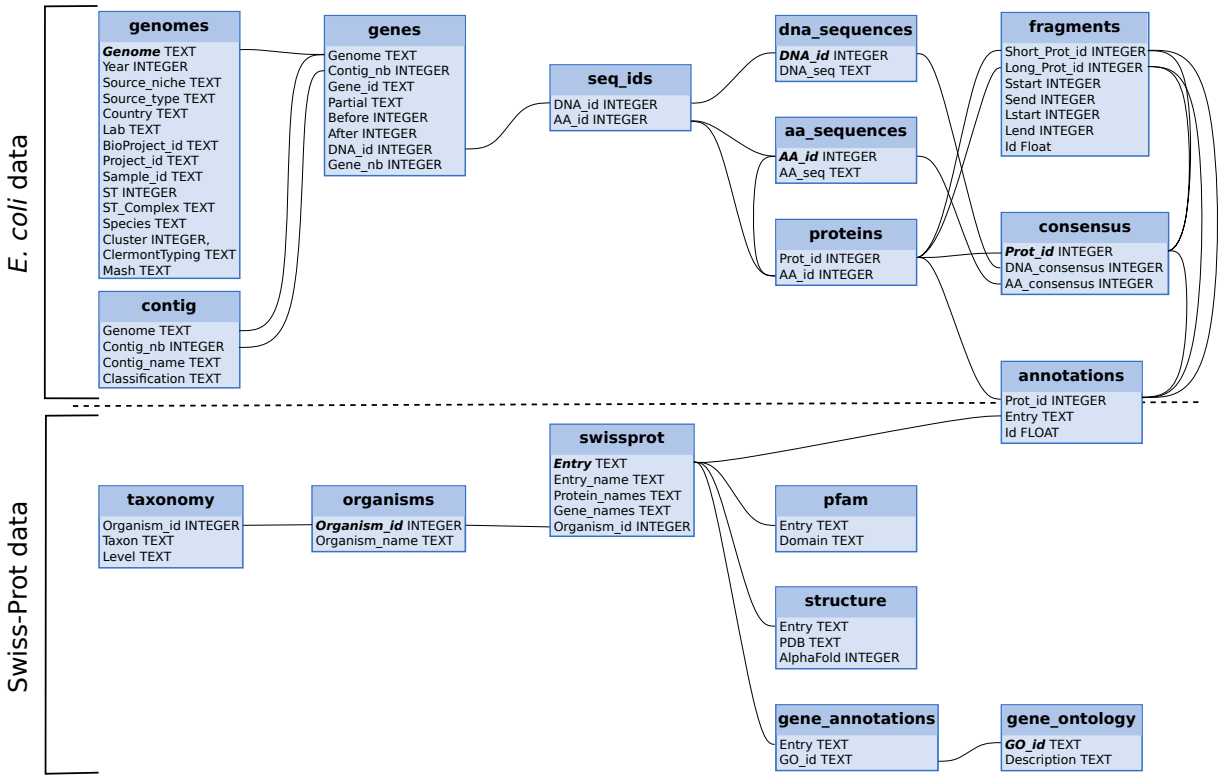

FIGURE 4.5: **Diagram of the structure of the SQL database.**
Table primary keys are highlighted with bold and italic fonts.

4.7 *E. coli* core and accessory genomes

One initial and straightforward application of this database consists in assessing the core and accessory genomes of each cluster. Previous studies have already explored the comparison of core and pan-genome sizes across different *E. coli* phylogroups. However, our approach brings a novelty to this analysis. We have clustered *E. coli* genomes based on a criterion of within-cluster sequence divergence of 0.5%. This means that we compare the gene repertoires of groups of genomes with similar levels of nucleotide diversity. By considering genomes with more comparable levels of diversity, we can gain insights into the variations in core and accessory genome sizes.

For each cluster, we can sample a given number of genomes N, starting from $N = 1$ up to $N = 10000$ for the sufficiently large clusters. For each random sample of N genomes, we count how many genes are common to all the sampled genomes—these form the core genome—, how many of them are found in at least 95% of the sampled genomes—these form the persistent genome—and, lastly, how many genes are found in at least one of the sampled genomes—these form the pan-genome. This analysis provides insights into the shared genes among the sampled genomes, the genes consistently present across the landslide majority of genomes, and the overall gene diversity within the cluster.

We performed 10 samplings for each cluster and value of N and plotted the average number of core, persistent and pan genes on Figure 4.6. The pan-genomes of all our clusters seem open (Figure

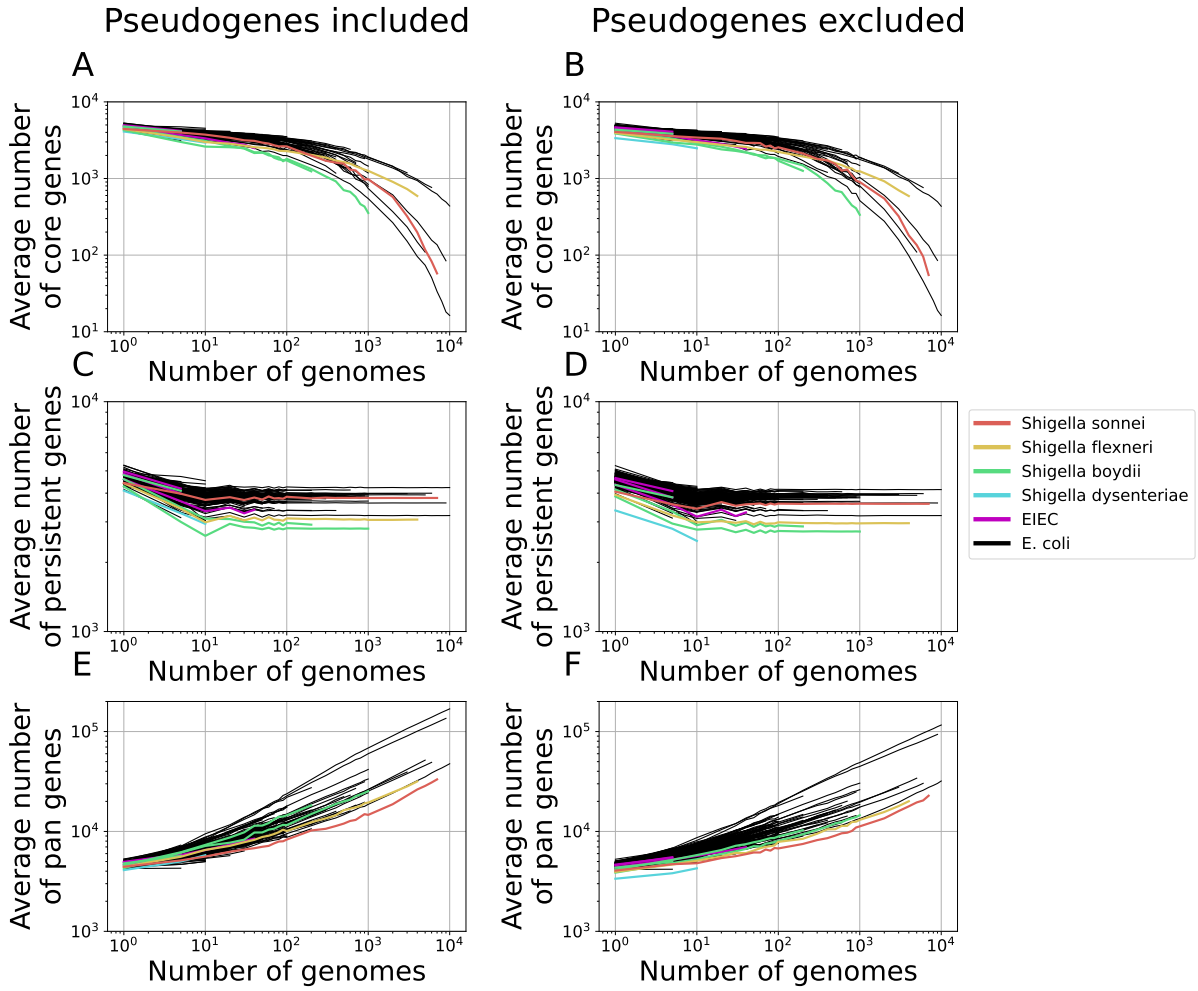

FIGURE 4.6: **Clusters core, persistent and pan genome sizes**

A. Evolution of the size of the core genome of each cluster with increasing number of strains. Pseudogenes were counted as true genes.

B. Evolution of the size of the core genome of each cluster with increasing number of strains. Pseudogenes were discarded.

C. Evolution of the size of the persistent genome of each cluster with increasing number of strains. Pseudogenes were counted as true genes.

D. Evolution of the size of the persistent genome of each cluster with increasing number of strains. Pseudogenes were discarded.

E. Evolution of the size of the pan-genome of each cluster with increasing number of strains. Pseudogenes were counted as true genes.

F. Evolution of the size of the pan-genome of each cluster with increasing number of strains. Pseudogenes were discarded.

Each line represents the average of 10 random samplings of genomes for each cluster.

4.6.E) even after excluding pseudogenes (Figure 4.6.F), indicating a continuous expansion of gene diversity. Interestingly, the size of the core genome rapidly decreased as the number of sampled genomes increased (Figures 4.6.A and B), whereas the number of persistent genes reached a stable level after sampling just a few dozen genomes (Figures 4.6.C and D). This suggests that the concept of a persistent genome is more relevant for describing the housekeeping functions of our clusters compared to a core genome. The reasons behind the observed decline in the number of core genes as the number of genomes increased remain unclear. It could be attributed to the lower quality of some rare genome sequences or it might signify genuine biological events involving infrequent gene losses. It would be interesting to investigate this question further.

Even if the global trends look very similar, we observe huge variations in the number of core, persistent and pan genes between clusters:

- The average number of core genes found in 50 genomes ranges from 2175.5 to 3761.3 with a median of 3319.15 (respectively 2046, 3736.9 and 3251.85 if we remove pseudogenes).
- The average number of persistent genes found in 50 genomes ranges from 2786.4 to 4226.8 with a median of 3827.85 (respectively 2687.2, 4103.6 and 3808.7 if we remove pseudogenes).
- The average number of pan genes found in 50 genomes ranges from 7021.1 to 17003.9 with a median of 10238.45 (respectively 6157.4, 14680.4 and 9009.45 if we remove pseudogenes).

In particular, *Shigella* clusters tend to exhibit fewer persistent genes compared to other clusters, although *Shigella sonnei* behaves more similarly to other *E. coli* clusters in this regard. Surprisingly, the lower number of persistent genes in *Shigella* clusters does not seem to translate into a smaller pan-genome (Figure 4.6.E). However, this pattern disappears if we exclude pseudogenes from the analysis (Figure 4.6.F), suggesting that *Shigella* undergoes accelerated gene loss that artificially inflates its pan-genome.

Overall, we observe an important variability in core and accessory genome sizes between clusters, despite being constructed to have similar levels of nucleotide diversity. This variability could stem from local adaptations to different ecological niches, varying degrees of genetic recombination, or differences in the effectiveness of natural selection in retaining genes.

4.8 Inferring the species phylogeny

4.8.1 General procedure

Our aim is to infer a phylogeny corrected for recombination. We chose to use Gubbins (Croucher et al. 2015) to detect recombination. As this software cannot run on tens of thousands of genomes at a time we proceeded in several steps:

1. We detected recombination within each of the 240 clusters.
2. We built 240 rooted phylogenies, one per cluster, and we inferred the ancestral sequence of each cluster.
3. We detected recombination between the 240 ancestral sequences.

4. We built an ancestral rooted phylogeny of the 240 clusters from the 240 ancestral sequences.

For all of these steps, we worked with the persistent genome described previously. To build a persistent genome alignment, we organized the genes in the same order and orientation than what was observed in *E. coli* ED1a strain—whose sequence was downloaded from MaGe (Vallenet et al. 2006). The high level of synteny observed in *E. coli* make this choice acceptable. 177 genes with no homologs in *E. coli* ED1a had to be removed. Highly gapped sites impair Gubbins performances so we removed sites with more than 5% gaps.

4.8.2 Detecting recombination within clusters

We ran Gubbins—with FastTree (Price et al. 2010) tree builder—to analyze the 231 clusters that contained fewer than 1,500 genomes. We replaced by gaps any parts of the corresponding multiple sequence alignments (MSAs) that Gubbins detected as being recombined. We built phylogenies from these corrected MSAs using IQ-TREE (Nguyen et al. 2015), applying a general time reversible model. To root the phylogenies, we selected two genomes from the nearest cluster that contained at least 10 genomes.

The remaining 9 clusters were too large to be analyzed by Gubbins in their entirety. To overcome this, we divided each of these clusters into smaller MSAs, each containing fewer than 1,500 genomes, and ran Gubbins on each smaller MSA separately. To ensure thorough analysis, we repeated this process 10 times for each cluster, creating new partitions of genomes each time. If Gubbins identified any recombined segments during these runs, those segments were removed from the global cluster MSA. The global MSAs being too massive to infer phylogenies, we changed them into 'SNP-MSAs' by removing conserved sites. We opted to use FastTree instead of IQ-TREE for building these phylogenies. Although FastTree may have slightly lower precision, it is more efficient in terms of computational time. Consistent with the approach used for the 231 smaller clusters, we employed a general time reversible model to infer the phylogenies, and we rooted the phylogenies using two genomes from the nearest cluster that contained at least 10 genomes.

The approach we followed was successful for all clusters except for cluster 8. This particular cluster initially consisted of 11,768 genomes, but due to its size, we decided to sub-sample 981 genomes for analysis. We ran Gubbins exclusively on these 981 genomes and constructed a SNP phylogeny using only this subset of genomes. We deemed this phylogeny to be adequate for inferring the ancestral sequence of cluster 8.

4.8.3 Detecting recombination between clusters to build a species phylogeny

We ran IQ-TREE to reconstruct ancestral sequences for each cluster based on the phylogenies we previously inferred. These 240 ancestral sequences were then combined into a multiple sequence alignment (MSA), which we subjected to Gubbins analysis to identify recombined segments. We constructed a phylogeny of these 240 ancestral sequences, accounting for recombination, and used *Escherichia fergusoni* ATCC 35469T (downloaded from MaGe) to root it. We had to exclude the 597 singleton genomes from the phylogeny. This exclusion was necessary because Gubbins was unable

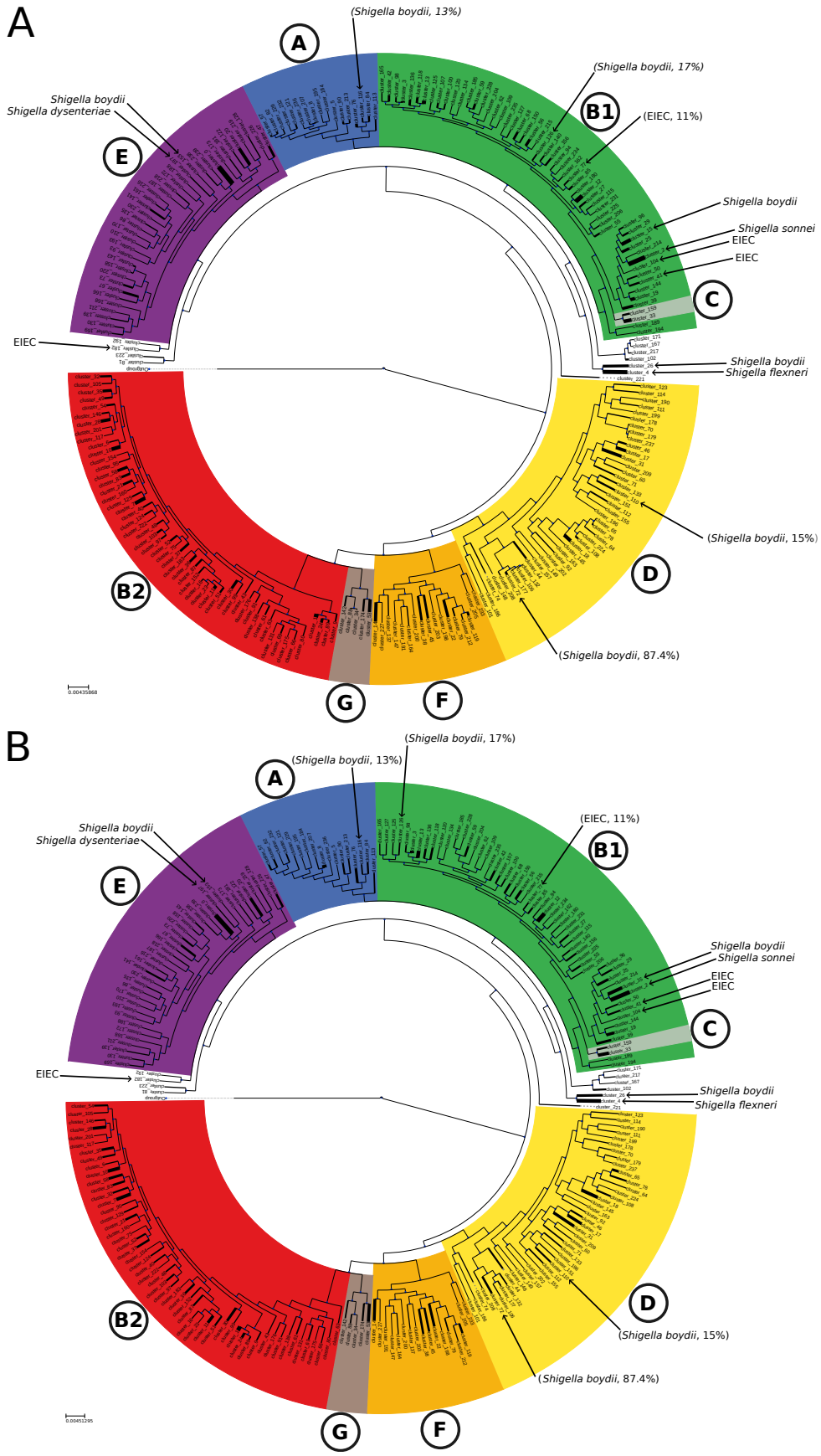

FIGURE 4.7: **Phylogenies of the 240 *E. coli* clusters.**

A. Phylogeny corrected for recombination.

B. Phylogeny without correction for recombination.

The width of the terminal branches is proportional to the log of the number of genomes in each cluster.

to process a MSA containing both the 240 ancestral sequences and these 597 sequences. This difficulty may be attributed to the poor quality of these genomes or the presence of rare hybrids resulting from extensive recombination events.

The phylogeny we inferred covers 240 clusters, representing a total of 80,843 genomes (or 70,056 genomes if we take into account the fact that only 981 genomes were used to reconstruct cluster 8's ancestral sequence). It is displayed on Figure 4.7.A, with different phylogroups indicated by colors. However, there are some clusters for which the attribution to a specific phylogroup remains uncertain, and they are displayed in white. The overall structure of the phylogeny aligns with previous studies (see Figure 2.1).

One notable difference is observed in the branching pattern of phylogroup D. We believe that this discrepancy may be attributed to the choice of our outgroup genome. To gain a more accurate understanding of the branching of phylogroup D, it would be necessary to explore alternative outgroup genomes and calculate confidence estimates.

Consistent with other studies, we can observe multiple independent occurrences of *Shigella* and enteroinvasive *E. coli* (EIEC) throughout the evolutionary history of the species. Interestingly, their distribution is non-random. None of these strains emerged within phylogroups B2, F, and G, while six distinct B1 clusters contain *Shigella* or EIEC strains. Additionally, we find that phylogroup C is nested within phylogroup B1, confirming previous findings (see Figure 2.1) and raising questions about the evolutionary relevance of this phylogroup.

We can compare the phylogeny obtained after correcting the MSA for recombination (Figure 4.7.A) with the one obtained without correction (Figure 4.7.B). We observe that the differences in the tree structure are not massive. This does not imply that recombination does not occur in *E. coli*, but rather that its impact is not strong enough to obscure the clonal phylogeny. When we consider a sufficient number of persistent genes, the effects of recombination tend to average out.

One notable difference between the two phylogenies is observed in the branch lengths. In the uncorrected phylogeny, the terminal branches appear longer while the ancestral branches are shorter. This is expected because the presence of recombination leads to a more star-like pattern in trees, resulting in longer branches leading to individual genomes.

Overall, this comparison suggests that while recombination does occur in *E. coli*, it does not significantly disrupt the underlying clonal phylogeny. The effects of recombination become apparent in the branch lengths rather than the overall tree structure.

4.9 The effects of recombination on the short term

Strains within the same cluster exhibit a very high degree of similarity in their persistent genome sequences. This means that they have diverged very recently in the history of the species so that they only differ from one another by very recent events. By examining recombination events within one of the 240 clusters, we can gain insights into the short-term effects of recombination.

As strains within the same cluster are closely related, we anticipate that recombined segments would exhibit higher-than-average sequence divergence. To examine this divergence along the genome, we use two measures of population diversity: $\theta_{\text{Watterson}}$ and divergence with the ancestor. The former is a commonly used statistic in population genetics, calculated based on the number of polymorphic sites. The latter measures the proportion of strains diverging by more than 1% from the ancestral

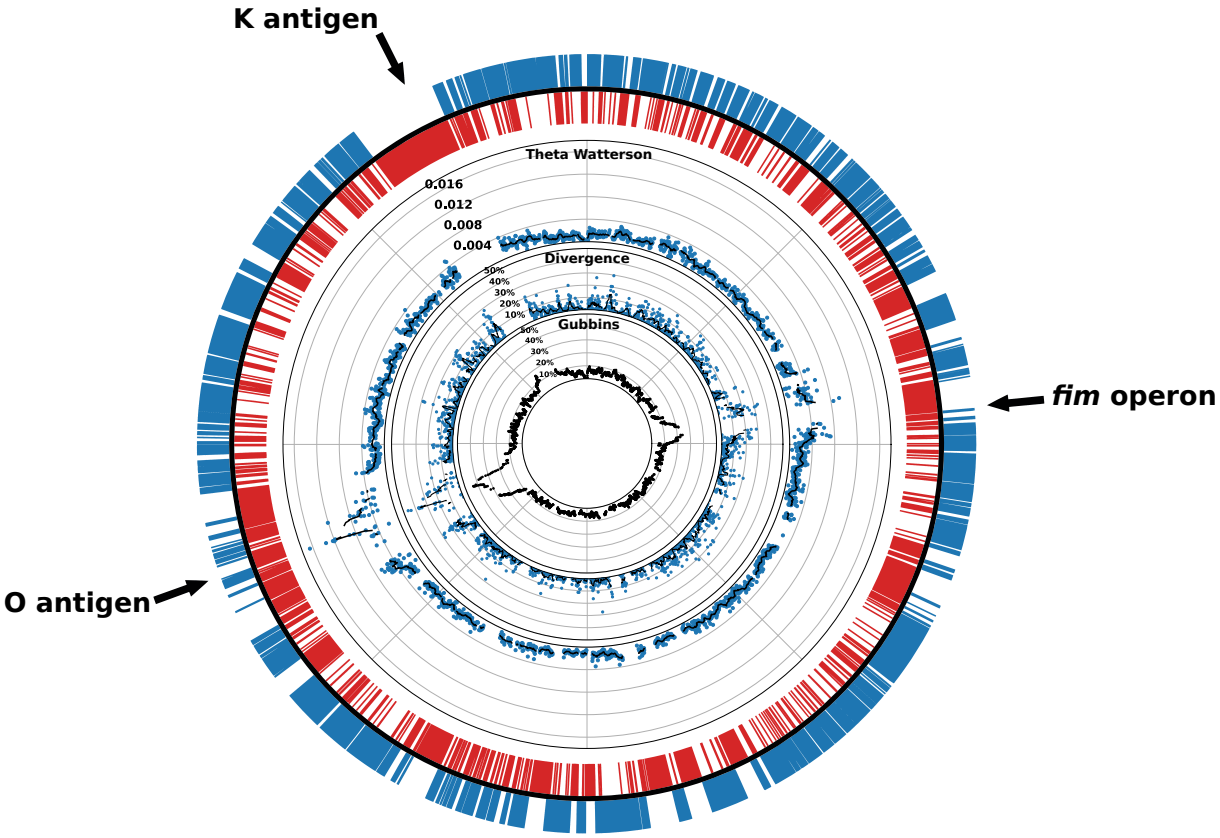

FIGURE 4.8: **Signature of within-cluster recombination.**
On the innermost circle is displayed the proportion of genomes for which Gubbins has detected recombination averaged over the 240 clusters (each dot is a gene). The second circle represents the proportion of genomes that have more than 1% divergence with the ancestral sequence of their cluster (each blue dot is a gene, a rolling average with a window of 11 genes is drawn with a black line). On the third circle, the average value of $\theta_{\text{Watterson}}$ across clusters is plotted (each blue dot is a gene, a rolling average with a window of 11 genes is drawn with a black line). On the outermost circle is the order of genes on *E. coli* ED1a strain. A blue bar corresponds to a persistent gene and a red one to an accessory gene. The three hotspots of recombination are highlighted with black arrows.

sequence of their respective cluster.

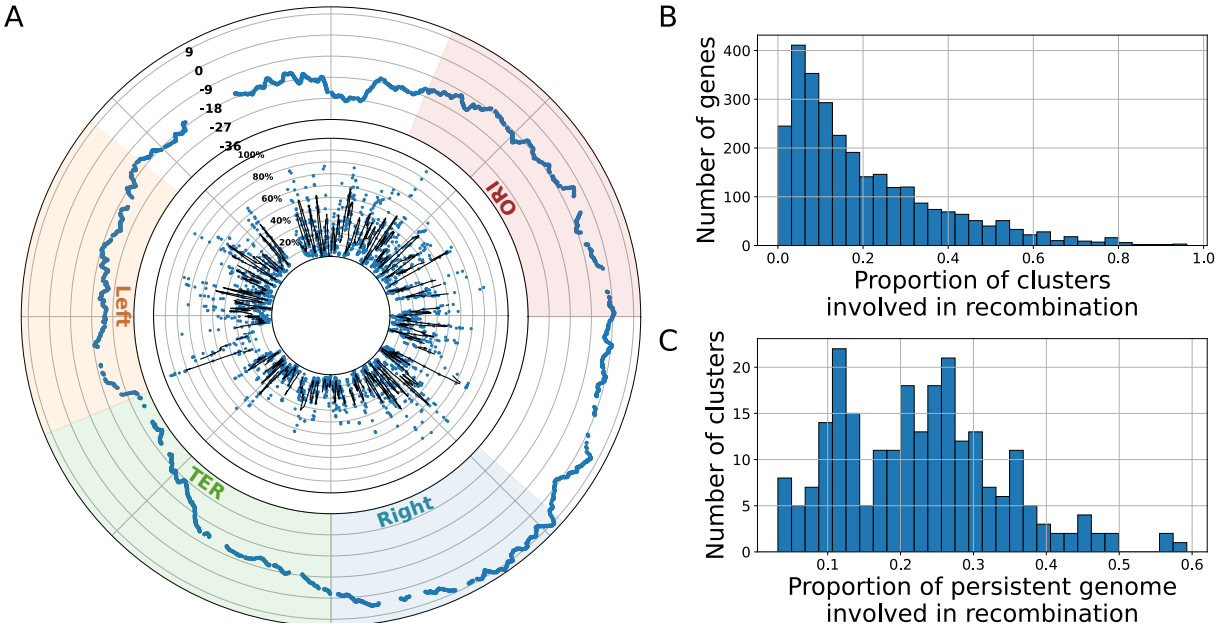

FIGURE 4.9: **The long-term effects of recombination in *E. coli*.**
A. On the inner circle is displayed the proportion of ancestral sequences involved in recombination according to Gubbins for each gene (blue dots). A rolling average with a sliding window of 11 genes is shown with a black solid line. On the outer circle is displayed the cumulative sum of the difference to the mean for each value of the inner circle. A decreasing trend—like the one in the TER region—signals a lower-than-average level of recombination.
B. Histogram of the proportion of ancestral sequences where Gubbins detected recombination for each gene.
C. Histogram of the proportion of the chromosome where Gubbins detected recombination for each ancestral sequence.
Some specific regions of the chromosome—origin of replication, Ter macrodomain and its left and right sides—are highlighted with colors. Their coordinates were taken from (Espeli et al. 2008).

When we compare these two diversity statistics with the outputs from Gubbins (Figure 4.8), we observe a strong correspondence between regions of higher-than-average sequence divergence and the genes where Gubbins detected most of the recombination events. This alignment between the Gubbins outputs and the other diversity measures instills confidence in the accuracy of the software's inferences. The majority of recombination events tend to occur around three specific regions—the two 'bastions of polymorphisms' mentioned in the introduction and the K antigen, which are all subject to diversifying selection. However, it is important to note that there is also a low level of recombination scattered throughout the chromosome. The proportion of recombined genomes seldom reaches 0% along the chromosome.

4.10 The effects of recombination on the long term

In contrast to the short-term effects of recombination observed within clusters, studying ancestral sequences of clusters provides insights into more ancient recombination events. The inferences made

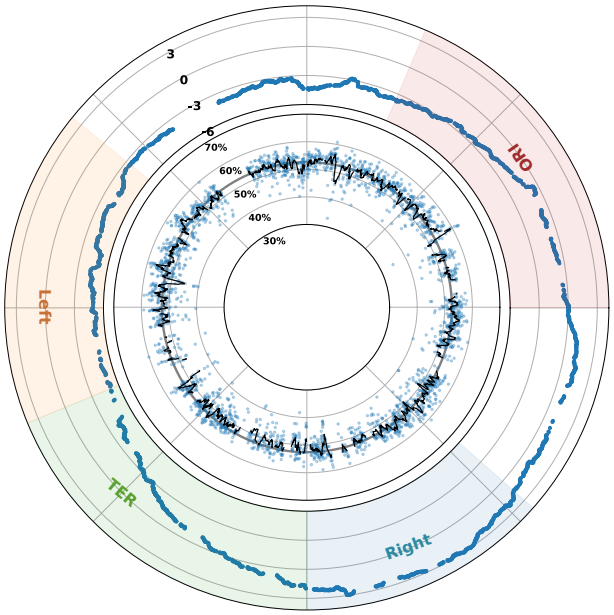

FIGURE 4.10: **Variations in GC content along the genome.**
On the inner circle is displayed the average GC content (each blue dot is a gene), a rolling average with a window of 11 genes is displayed with a black solid line. The median GC content is shown with a large solid grey line. On the outer circle is displayed the cumulative sum of the difference to the mean GC content. A decreasing trend—like the one in the TER region—signals a lower-than-average GC content. Some specific regions of the chromosome—origin of replication, Ter macrodomain and its left and right sides—are highlighted with colors. Their coordinates were taken from (Espeli et al. 2008).

by Gubbins on the 240 cluster ancestral sequences (Figure 4.9.A) strongly contrast with those made within clusters (Figure 4.8). Indeed, all along the genome we find genes where more than half of the clusters show signatures of recombination. However, this does not imply that recombination completely obscures the phylogeny. In fact, in half of the genes, less than 14.2% of the cluster ancestral sequences have recombined (Figure 4.9.B). Moreover, recombined segments account for an average of 22.7% of these ancestral sequences, with only three ancestral sequences containing more than 50% recombined fraction (Figure 4.9.C). This indicates that the phylogeny constructed from the ancestral MSA corrected for recombination (Figure 4.7.A) is indeed relevant.

GC-content may also deliver some information about the extent of recombination along the genome (Figure 4.10). In particular, we well observe the drop in GC-content around the terminus of replication. This lower-than-average GC-content has been interpreted as a sign of reduced recombination, although this interpretation is still a subject of debate. Interestingly, we also observe that Gubbins identifies fewer instances of recombination in the same region, further supporting this explanation.

In summarizing our brief investigation of recombination, we can conclude that in the short-term, it is primarily noticeable around three bastions of polymorphisms. Nonetheless, we still observe sporadic instances of recombination across the entire chromosome at a lower level. These infrequent events gradually accumulate over extended time periods. Consequently, when we compare the ancestral sequences of the clusters, we discover compelling evidence of recombination spanning the entire chromosome, with a distinct decrease observed near the terminus of replication.

Chapter 5

Using protein mutational landscapes to study individual mutations

The results presented in sections 5.1, 5.2, 5.3 and 5.5 were published in (Vigué et al. 2021). The complete article is available in Appendix B.

5.1 Protein mutational landscapes and their application to *E. coli*

A protein mutational landscape models the level of functionality of a protein or a protein domain. It can be thought of as a function that takes as input an amino-acid sequence of a given protein or protein domain and provides as output the probability of observing that sequence in nature. With functional variants being selected and non-functional variants being counter-selected, the probability of observing a sequence in nature should closely match the level of functionality of the corresponding variant.

Deriving a probability might require a normalisation step so that all probabilities sum to one. This normalisation step can take a lot of computational time as it requires dealing with all possible amino-acid variants. For example, if we take a sequence of 100 amino acids, there are 20^{100} variants to consider, which is far beyond the computational power of any machine. For this reason, we cannot always calculate probabilities and we rather rely on the energy level associated with an amino-acid sequence. The energy level is related to the probability of observing an amino-acid sequence $(a_1, a_2, ..., a_L)$ as follows:

$$P(a_1, a_2, ..., a_L) = \frac{1}{Z} \exp\{-E(a_1, a_2, ..., a_L)\} \tag{5.1}$$

where $E(a_1, a_2, ..., a_L)$ is the energy level of the sequence and Z the normalisation factor we cannot always calculate in practice. As you can see in this formula, there is a minus sign in the exponential function. This means that a low energy level corresponds to a high probability of observation. In other words, the local maxima of the protein mutational landscape correspond to the local minima of the corresponding energy landscape.

Protein mutational landscapes are deduced from homologous sequences found in distant species.

These are sequences that have evolved independently over millions of years, so that they share only 20-30% identity in sequence. However, they still perform the same function and fold similarly in 3D: the essence of the protein's functionality has been preserved despite huge variations in the amino-acid sequence. These sequences represent a sample of the local maxima of the protein mutational landscape and thus a sample of the local minima of the corresponding energy landscape (Figure 5.1).

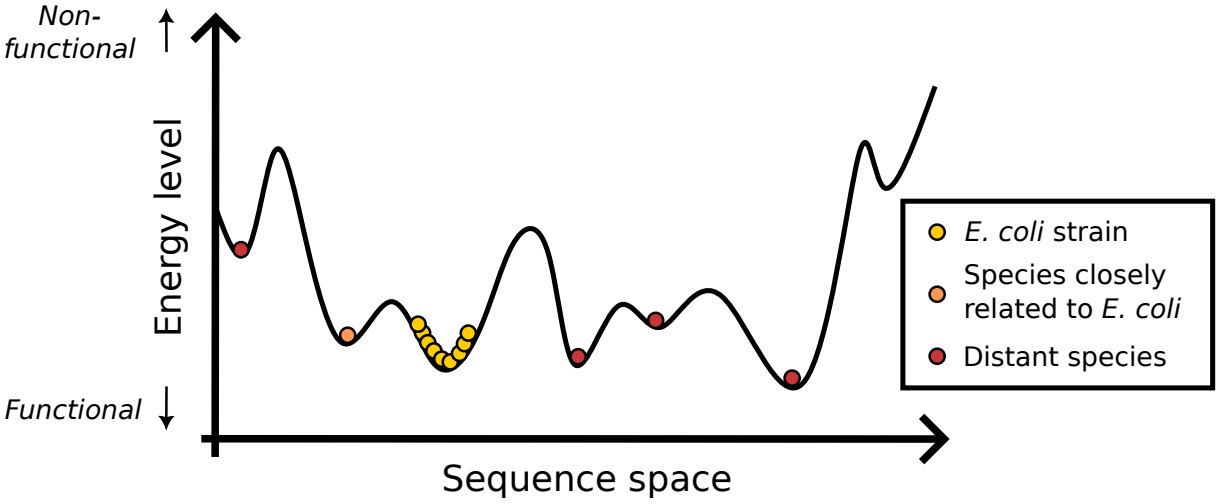

FIGURE 5.1: **Representation of an energy landscape.**
Amino-acid sequences found in nature corresponds to local minima of the landscape. The models are trained on distant species (red) they may not necessarily be relevant to study more local structures around *E. coli* (*E. coli* variants in yellow and closely related species in orange).

As we intend to use protein mutational landscapes to study the neighbourhood of *E. coli* sequences, we choose to exclude from the sample of distant homologues any sequence that shares more than 90% identity with the reference *E. coli* amino-acid sequence. For this reason, we do not know whether our protein mutational landscape accurately captures the local structure around *E. coli* (yellow dots on Figure 5.1). This local structure reflects idiosyncratic constraints related to *E. coli*'s adaptation to its ecological niche, as well as more global evolutionary constraints shared by all distant homologs. And protein mutational landscapes can only account for the latter. In other words, protein mutational landscapes are inferred from the study of long evolutionary timescales; can they inform us about the dynamics of shorter evolutionary timescales?

5.2 Building a protein mutational landscape in practice: IND and DCA

Protein mutational landscapes are inferred from a multiple sequence alignment (MSA) of homologous sequences found in distant species. These MSAs can be downloaded directly from Pfam (Bateman et al. 2004) if one is looking for Pfam protein domains or can be retrieved by querying large protein databases such as UniRef (Suzek et al. 2015) with softwares based on profile hidden Markov models (HMM) of proteins such as HHblits (Remmert et al. n.d.). These MSAs can be biased towards certain species that have been sequenced more frequently than others. For this reason, we perform a reweighting step where sequences that are too similar to each other are assigned a lower weight.

We also exclude any sequence that shares more than 90% identity with *E. coli*. A further quality step is performed to eliminate from the MSA any sites that are too frequently gaped (MSA columns with >20% gaps).

From this MSA, we can construct a non-epistatic mutational landscape (independent model, IND). In this simple model, the observation probability associated with an amino-acid sequence can be directly deduced from the frequencies of the amino acids in each column of the MSA. Consider an amino acid sequence of length L: $(a_1, a_2, ..., a_L)$. In this sequence, a_i is the amino acid observed at the i^{th} position, it can take 21 possible values: any of the 20 amino acids or a gap. The probability of observing amino acid a at position i is $f_i(a)$, the frequency of amino acid a in column i of the MSA. The main assumption underlying IND is the additivity of mutation effects. In this simple framework, we can compute the probability of observing an amino-acid sequence $(a_1, a_2, ..., a_L)$:

$$P^{\text{IND}}(a_1, a_2, ..., a_L) = \Pi_{i=1}^{L} f_i(a_i) \tag{5.2}$$

The corresponding energy of the amino-acid sequence $(a_1, a_2, ..., a_L)$ is therefore:

$$E^{\text{IND}}(a_1, a_2, ..., a_L) = -\log P^{\text{IND}}(a_1, a_2, ..., a_L) \tag{5.3}$$

Two sequences $(a_1, a_2, ..., a_L)$ and $(b_1, b_2, ..., b_L)$ can be compared by taking the difference of their energies $E^{\text{IND}}(b_1, b_2, ..., b_L) - E^{\text{IND}}(a_1, a_2, ..., a_L)$. If this difference is negative, it means that IND predicts that $(b_1, b_2, ..., b_L)$ is more frequent in nature than $(a_1, a_2, ..., a_L)$, and conversely if it is positive.

An IND model ignores any contextual dependence of mutations: the effect of replacing one amino acid with another is exactly the same regardless of the rest of the amino-acid sequence. If we want to incorporate an interaction term between amino-acid sites, we can turn to a more complex modelling framework: Direct-Coupling Analysis (DCA). A DCA model is composed of two matrices: h, the site-dependent biases that evaluate the importance of single amino acids in individual sequence positions, and J, the epistatic couplings connecting the amino acids in pairs of positions. In a DCA model of an amino-acid sequence, the probability of observing the sequence $(a_1, a_2, ..., a_L)$ in nature is:

$$P^{\text{DCA}}(a_1, a_2, ..., a_L) = \frac{1}{Z} \exp\left\{-E^{\text{DCA}}(a_1, a_2, ..., a_L)\right\} \tag{5.4}$$

where $Z = \Sigma_{a_1,...,a_L} \exp\left\{-E^{\text{DCA}}(a_1, a_2, ..., a_L)\right\}$ is the normalisation factor which we cannot calculate in practice and $E^{\text{DCA}}(a_1, a_2, ..., a_L) = -\Sigma_{i<j} J_{ij}(a_i, a_j) - \Sigma_i h_i(a_i)$ is the energy of the sequence. In statistical physics, E^{DCA} corresponds to the Hamiltonian of a generalized Potts model.

The IND model only reproduces the amino-acid frequencies of the MSA columns—the $f_i(a)$. In contrast, the DCA model reproduces both $f_i(a)$ and $f_{ij}(a, b)$, the latter being the frequency at which amino acid a is observed in column i and amino acid b is observed in column j of the MSA. DCA therefore reproduces the couplings between pairs of sites that have been observed in the MSA of distant homologues. From a mathematical point of view:

$$P_i^{\text{DCA}}(a_i) = \Sigma_{A_k|k\neq i} P^{\text{DCA}}(a_1, a_2, ..., a_L) = f_i(a_i) \tag{5.5}$$

$$P_{ij}^{\text{DCA}}(a_i, a_j) = \Sigma_{A_k|k\neq i,j} P^{\text{DCA}}(a_1, a_2, ..., a_L) = f_{ij}(a_i, a_j) \tag{5.6}$$

5.3 Testing the predictions of IND and DCA

5.3.1 Predicting *E. coli* native amino acids

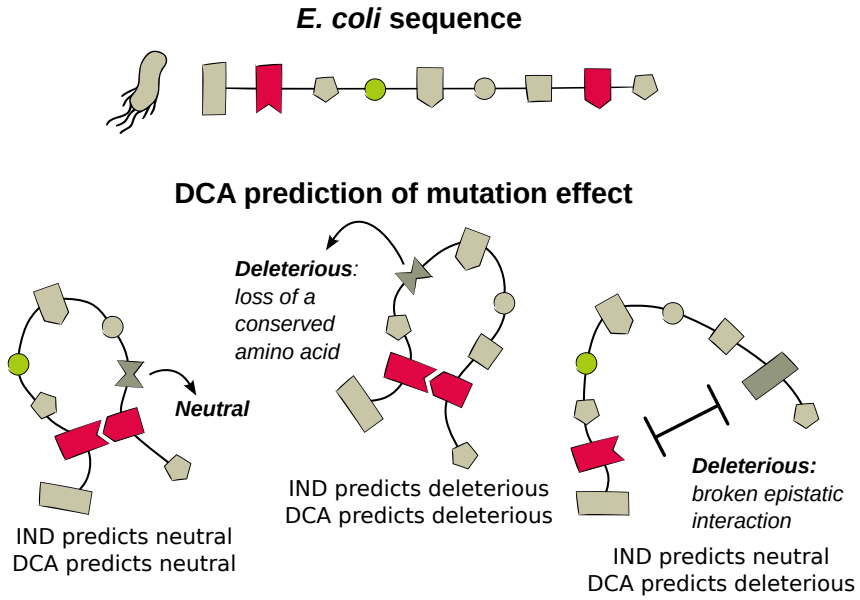

FIGURE 5.2: **Predicting the effect of mutations in an *E. coli* background.**
We can use IND and DCA to predict the effect of single amino-acid changes in *E. coli* sequence. IND only captures patterns of conservation while DCA detects both patterns of conservation and epistatic interactions between pairs of sites.

We want to test the accuracy with which IND and DCA model an amino-acid sequence (Figure 5.2). To do this, we use a protein of length L whose reference sequence in *E. coli* is $(a_1, a_2, ..., a_L)$. At each of these L sites, we can observe 20 possible amino acids (we choose here to study only the 'true' amino acids, not the effect of deletions, so we ignore gaps). Both DCA and IND give a probability of observing the amino acid α at locus i in the amino-acid background $a^0_{\setminus i} = (a_1, a_2, ..., a_{i-1}, a_{i+1}, ..., a_L)$:

- $P_i^{\mathrm{IND}}(\alpha \mid a^0_{\setminus i}) = P_i^{\mathrm{IND}}(\alpha) = f_i(\alpha)$, the frequency of the amino acid α in the MSA after excluding gaps. This probability does not depend on the amino-acid background $a^0_{\setminus i}$.

- $P_i^{\mathrm{DCA}}(\alpha \mid a^0_{\setminus i}) = \exp\left\{-E^{\mathrm{DCA}}(a_1, a_2, ..., a_{i-1}, \alpha, a_{i+1}, ..., a_L)\right\} / z_i$, with the normalization z_i chosen such that P becomes a probability distribution over the values of α, *i.e.* over the 20 theoretically possible amino acids at locus i. Note that $P_i^{\mathrm{DCA}}(\alpha \mid a^0_{\setminus i})$ is not the probability of observing the amino acid α at locus i but the conditional probability of observing the amino acid α at locus i, given that the other loci take amino acids $(a_1, a_2, ..., a_{i-1}, a_{i+1}, ..., a_L)$. Therefore, the normalisation factor z_i can be calculated easily, as it only has 20 terms.

Using these probabilities, we rank the 20 possible amino acids at locus i from most likely to least likely. We then examine the rank of *E. coli* native amino acid. A perfect model of an *E. coli* amino-acid sequence would always rank this amino acid first. In practice, working at persistent genome scale, we find that DCA ranks them first in 78% of cases, while this figure drops to 45% for IND (Figure 5.3.A).

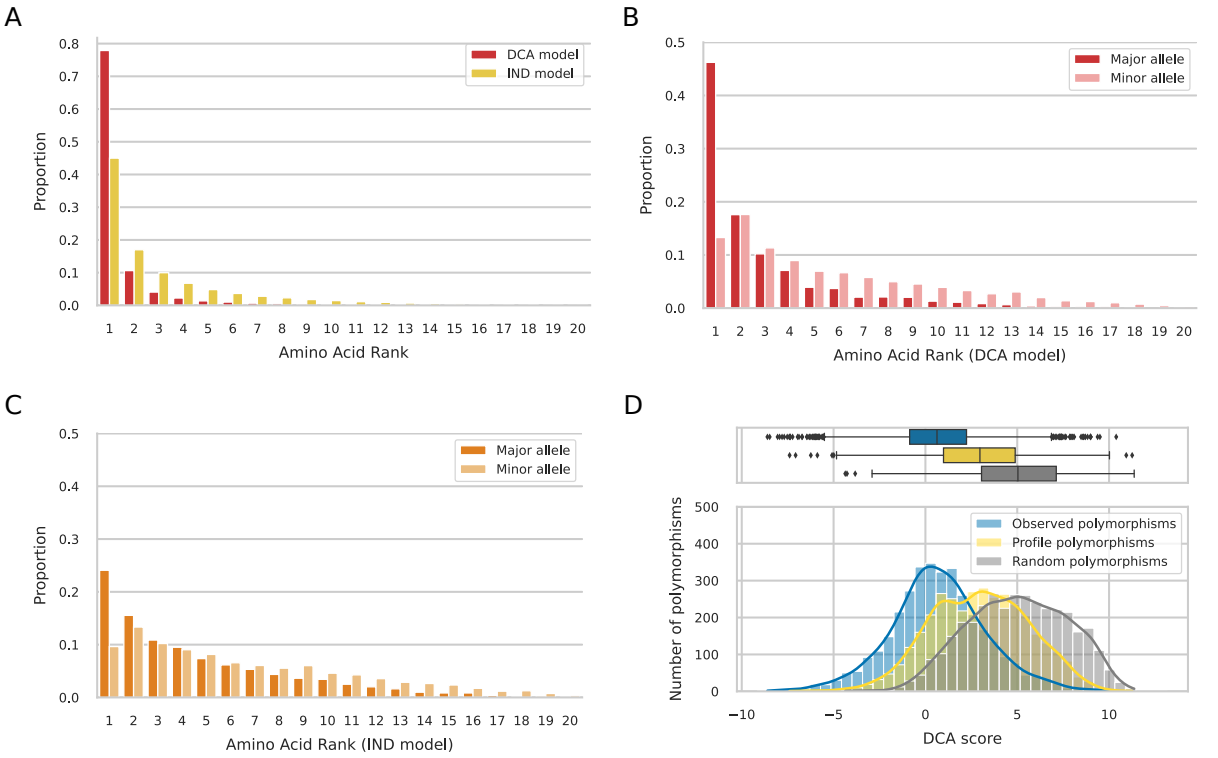

FIGURE 5.3: **Predicting *E. coli* native amino acids and polymorphisms with IND and DCA.**

A. Rank of amino acid observed in an *E. coli* ST131 strain as compared to all 20 possible amino acids. DCA model (red) outperforms IND (yellow) by predicting twice as many native amino acids to be the best possible.

B. DCA rank of major and minor allele for all sites that are polymorphic at a >5% threshold, among all 20 possible amino acids. Major alleles (alleles at frequencies >50%, in red) have better ranks than minor alleles (alleles at frequencies between 5 and 50%, in pink). The distribution of consensus alleles peaks at the first rank (46.2% of polymorphic sites have major allele ranking first and 17.6% have second-best rank) while the distribution of minor alleles peaks at the second rank (13.3% have the best rank against 17.6% that are second-best).

C. IND rank of major and minor allele for all sites that are polymorphic at a >5% threshold, among all 20 possible amino acids. As with DCA, major alleles (in orange) have better ranks than minor alleles (in yellow) and the distribution of consensus alleles peaks at the first rank. However, the distribution is spread towards greater ranks (only 24.1% of polymorphic sites have major allele ranking first and 15.5% have second-best rank, similarly minor alleles rank first in 9.6% and second-best in 13.3% of polymorphic sites) compared to DCA ranking.

D. Distribution of DCA scores of non-synonymous polymorphisms observed at frequencies >5% across >60,000 strains (blue) compared to mutations sampled from an IND model (yellow) or to random mutations (gray). A large number of possible mutations are predicted to be highly deleterious (positive scores) compared to naturally occurring polymorphisms that tend to be neutral (blue distribution centered on zero). Polymorphisms predicted from IND are slightly deleterious once epistasis is taken into account (yellow distribution shifted towards positive values). Boxplot center lines represent medians, box limits are upper and lower quartiles, whiskers extend to show the rest of the distribution within an 1.5 × interquartile range, outliers are represented with points; sample size is 3477 mutations for each of the three groups.

5.3.2 Predicting *E. coli* polymorphisms

However, the approach followed in the previous section is subject to bias. In addition to 'real' epistatic interactions reflecting physical constraints on the protein, DCA may also have captured some phylogenetic interactions: correlations of residues that have been inherited vertically together in different species. This would inflate the predictive performance of DCA but give very little information about the real factors constraining protein evolution. We tried to limit this bias by training the DCA models on MSAs of distant homologues (any sequence with more than 90% identity to *E. coli* being excluded). But some phylogenetic interactions may remain.

This is why we have chosen to focus on *E. coli* polymorphisms. These appeared recently in the history of the species, after *E. coli* had diverged from other species. A model based on phylogenetic correlations alone would therefore not be able to predict them.

If we focus on the polymorphic sites, we observe that almost half of them have a major allele that is ranked first by DCA (Figure 5.3.B). As expected, the minor allele is more likely to be ranked second. We observe the same overall trends with IND: the major allele is most likely to be ranked first and the minor allele second (Figure 5.3.C). However, the overall distribution is flattened towards the higher ranks, suggesting that, again, DCA performs better than IND.

Another way to study polymorphism is to use the DCA scores of the mutations directly. The DCA score of the mutation of an amino acid α to an amino acid β at locus i in the amino-acid background $a^0_{\setminus i} = (a_1, a_2, ..., a_{i-1}, a_{i+1}, ..., a_L)$ of the *E. coli* reference sequence is given by:

$$\Delta E_i = E_i^{\mathrm{DCA}}(a_1, a_2, ..., a_{i-1}, \beta, a_{i+1}, ..., a_L) - E_i^{\mathrm{DCA}}(a_1, a_2, ..., a_{i-1}, \alpha, a_{i+1}, ..., a_L) \tag{5.7}$$

This score is negative if the mutation is advantageous, zero if it is neutral and positive if it is deleterious.

The distribution of DCA scores for the mutations found in *E. coli* (blue distribution on Figure 5.3.D) peaks near zero—suggesting that these polymorphisms are close to neutral. In comparison, DCA predicts that random amino acids are highly deleterious, as expected (grey distribution). More importantly, DCA also predicts that amino acids sampled from an IND model are slightly deleterious when inserted in an *E. coli* background (yellow distribution). This latter observation means that the polymorphisms observed in *E. coli* are more adapted to the genetic background of *E. coli* than the amino acids found in distantly related species. This confirms the entrenchment theory that states that mutations that reach fixation are close to neutral in the genetic background where they occur, but may be deleterious in another background.

5.4 The effect of natural selection on polymorphisms segregating within *E. coli*

Polymorphisms regularly arise in a population due to random mutations. They can segregate for quite some time before disappearing or, more rarely, reaching fixation. The expected time they spend at a given frequency directly depends on their impact on fitness (Sethupathy et al. 2008). Therefore, we expect to see some differences between the DCA score distribution of polymorphisms found at different frequencies, even if there is no strict equivalence between a DCA score—that reflects the

level of functionality of a protein—and an organism's fitness.

We gather 454,636 polymorphisms in 1,200 distinct persistent genes, together with their frequencies in all 81,440 *E. coli* genomes (Figure 5.4.A). For each protein sequence, we reconstruct the ancestral *E. coli* sequence by rooting the tree on the nearest Swiss-Prot sequence that is neither *E. coli* nor *Shigella*. This allows us to orient observed mutations in order to determine which allele is the ancestral and which is the derived one. We can also calculate the distribution of DCA scores of all possible single mutants of the ancestral sequence of each persistent protein (green histogram on Figure 5.4.B).

We compare this distribution with that of the DCA scores of polymorphisms found across our 81,440 *E. coli* genomes (blue histogram on Figure 5.4.B). As we can see, there is an over-representation of beneficial mutations (low DCA scores) and a corresponding under-representation of deleterious mutations (high DCA scores) among observed mutations compared to all possible single mutations. If we compute the ratio of observed versus possible single mutations for each DCA score bin displayed on Figure 5.4.B, we see that this ratio follows an exponential decrease that spans more than four orders of magnitude. More interestingly, we can study the evolution of this ratio according to the frequency range of the observed polymorphisms (Figure 5.4.C): the higher the frequency, the sharper the decrease.

We can notice that, even for very low frequency polymorphisms (purple line of polymorphisms at $\leq 0.1\%$), the ratio of observed versus possible polymorphims decreases with increasing DCA scores. This might be confusing because at these very low frequencies we expect random drift to dominate the fate of mutations and natural selection to play very minor role. However, we should remember that some of the persistent genes are essential for the bacteria. Inactivating mutations occurring on them might be filtered out almost instantly because they result in cell death or at least severely impair cell growth. Some mutation bias may also contribute to this pattern. Indeed, amino-acid changes that result from only one single nucleotide change in a given codon tend to be more frequent than those that require more, and they are also less harmfull. However, we performed some complementray analyses that showed that mutation biases could not alone account for the effect we observe here. All these questions form the core of the work presented in Chapter 7.

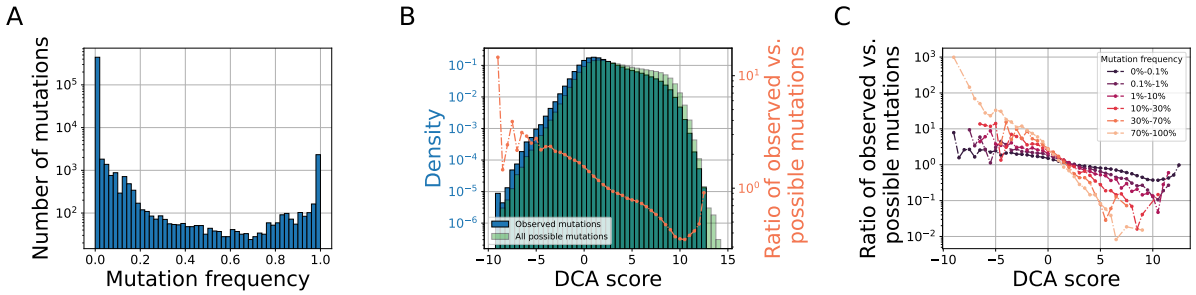

FIGURE 5.4: **DCA scores and frequencies of polymorphisms observed across 81,440 *E. coli* genomes.**
A. Histogram of the frequencies of mutations observed in 81,440 *E. coli* genomes.
B. Distribution of the DCA scores of all possible single mutations (in green) and observed mutations (in blue). The ratio between those two distributions is displayed in orange.
C. Ratios of the distributions of DCA scores of observed versus possible single mutations according to the frequency of observed mutations.

A limitation of our approach lies in the way we estimate frequencies. Indeed, our dataset of

genomes is far from being representative of *E. coli* in the wild. Strains isolated in clinical context are clearly over-represented, while non-human isolates remain rare. The frequency of a mutation found in this dataset may thus be quite different from its frequency in nature. To overcome this obstacle, we choose to estimate mutation frequencies within our 240 genome clusters. We choose to focus on the 128 clusters where we find at least 1,000 different mutations (Figure 5.5.A) for which we record their corresponding frequencies (Figure 5.5.B).

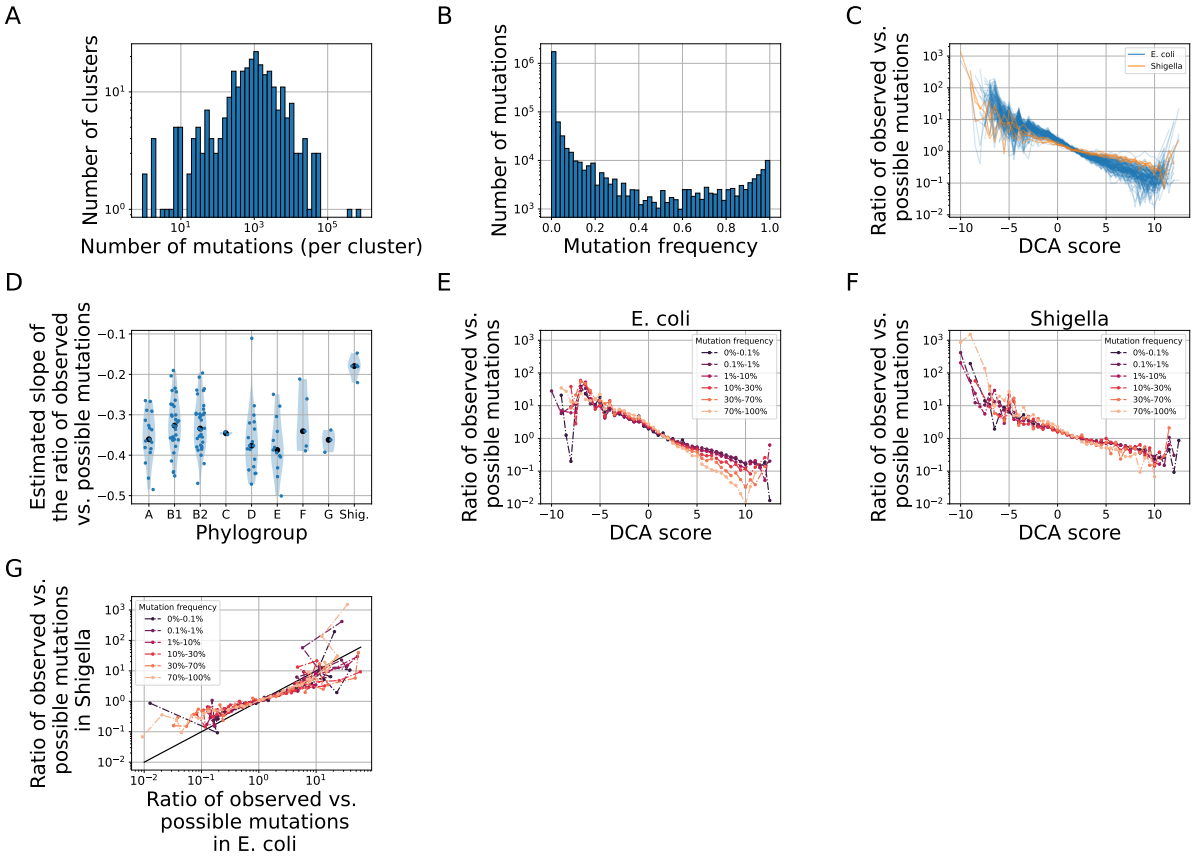

FIGURE 5.5: **DCA scores and frequencies of polymorphisms within *E. coli* genome clusters.**
A. Histogram of the number of mutations found in each of the 240 clusters.
B. Histogram of the frequencies of mutations observed in the 128 clusters with more than 1,000 mutations.
C. Ratio of the distributions of DCA scores of observed versus possible single mutations according to the frequency of observed mutations in each of the 128 clusters with more than 1,000 mutations.
D. Slopes of the lines of Panel C according to the phylogroup and the *E. coli*/*Shigella* status of the clusters.
E. Ratios of the distributions of DCA scores of observed versus possible single mutations according to the frequency of observed mutations in non-*Shigella* clusters.
F. Ratios of the distributions of DCA scores of observed versus possible single mutations according to the frequency of observed mutations in *Shigella* clusters.
G. Ratios of the distributions of DCA scores of observed versus possible single mutations according to the frequency of observed mutations in *Shigella* clusters plotted against the corresponding ratios in non-*Shigella* clusters.

Before focusing on mutation frequencies, we can reproduce Figure 5.4.B at the scale of each cluster by grouping all observed polymorphisms together independently of their frequencies. The result-

ing figure is displayed on Panel C of Figure 5.5. A striking finding is that deleterious mutations are not filtered with the same efficiency in every clusters. To get some more quantitative insights into the patterns we observe on Figure 5.5.C, we compute the slopes of the corresponding lines and sort them according to the phylogroup the clusters belong to and their *E. coli*/*Shigella* status (Figure 5.5.D). In both Figure 5.5.C and D, it is clear that deleterious mutations tend to segregate more within *Shigella* populations (in orange on Figure 5.5.C) compared to other *E. coli* (in blue). This finding is perfectly consistent with a decrease in effective population size that might have accompanied the transition to an intra-cellular way of life.

We might now look at how polymorphisms segregate within clusters depending on their frequencies. We have gathered all polymorphisms found in non-*Shigella* clusters with their corresponding frequencies on Figure 5.5.E and performed the same procedure with *Shigella* clusters on Figure 5.5.F. If we compare these two panels with Figure 5.4.C, it is clear that natural selection seems to be much less effective at eliminating deleterious mutations within clusters compared with the species as a whole. This is expected because polymorphisms segregating within clusters are far more recent, so natural selection has had less time to act on them. Furthermore, they segregate within smaller populations so natural selection is also less effective. We can see that here again *E. coli* clusters do better than *Shigella* ones at filtering polymorphisms with increasing frequencies. At first sight, it might even seem that there is no difference between the different frequency ranges on Panel F but that is only because the differences are too thin to be clearly visible. Indeed, when we plot the *E. coli* ratio against the *Shigella* one for each frequency range (Figure 5.5.G), we see that all lines superimpose. This means that you can change the lines displayed on Panel E into the lines displayed on Panel F by multiplying their slopes by roughly the same value.

5.5 How the genetic background impacts the effect of mutations

We have seen so far that DCA consistently outperforms IND in predicting the amino acids observed in *E. coli*. This suggests that DCA can accurately capture some of the *E. coli* genetic context. This context is based on epistatic interactions between different loci in the protein. An intriguing question arises regarding the construction of this context. Does it depend on a few strong interactions between specific amino acids, or does it involve a network of numerous weak couplings? If the former holds true, modifying a few amino acids would be adequate to completely alter the genetic context, consequently affecting the impact of other mutations. In this case, the mutational landscape would exhibit extreme ruggedness. Conversely, if the latter is accurate, a larger number of mutations would be required to observe any noticeable change in the mutation effect, resulting in a smoother mutational landscape. Considering that DCA somehow captures this genetic context, it is worthwhile to investigate how DCA models these epistatic interactions.

To answer this question, we want to determine the proportion of sites actually coupled to a locus in a DCA model of an amino-acid sequence. To do this, we use the inverse participation ratio (IPR). The inverse of the IPR corresponds to the effective number of non-zero components of a distribution. It becomes minimal if there is only one non-zero component and reaches its maximum for a uniform distribution. In a DCA model, the couplings between amino-acid sites are given by the matrix J. The

effective proportion of sites that are epistatically coupled with a position i in a sequence of length L is thus $1/(\mathrm{IPR}_i L)$, with:

$$\mathrm{IPR}_i = \Sigma_{j\neq i}(J_{ij}(a_i,a_j)^2/\Sigma_{k\neq i}J_{ik}(a_i,a_k)^2)^2 \tag{5.8}$$

Figure 5.6 shows the distribution of the effective proportion of sites effectively coupled to an amino-acid site in the DCA model. The median of the distribution is 24%: in a DCA model, residues tend to be coupled to about a quarter of the other amino-acid sites. In other words, the epistasis captured by DCA is a diffuse signal made up of many small couplings, not just a few strong couplings derived from direct contacts with neighbouring amino acids.

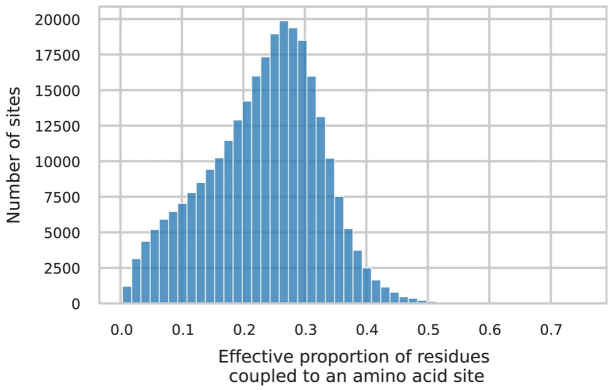

FIGURE 5.6: **Histogram of the effective proportion of sites coupled with a given amino-acid site.**

Another way to study epistasis involves calculating an epistatic cost. The epistatic cost refers to the difference between the cost of different mutations occurring together in a given sequence and the sum of their individual costs if they were inserted individually in the same sequence. For a pair of mutations, the epistatic cost for substituting the reference residues α_i, α_j with β_i, β_j writes:

$$\Delta\Delta E_{ij} = \Delta E_{ij} - \Delta E_i - \Delta E_j = J_{ij}(\alpha_i,\beta_j) + J_{ij}(\beta_i,\alpha_j) - J_{ij}(\beta_i,\beta_j) - J_{ij}(\alpha_i,\alpha_j) \tag{5.9}$$

Similarly, the epistatic cost of an arbitrary number of mutations is:

$$\Delta\Delta E_{ij\ldots n} = \Delta E_{ij\ldots n} - (\Delta E_i + \Delta E_j + \ldots + \Delta E_n) \tag{5.10}$$

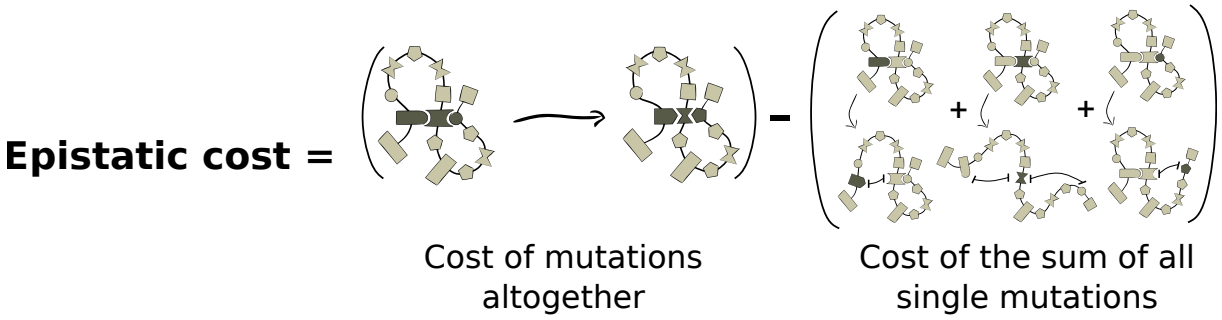

FIGURE 5.7: **Computation of the epistatic cost of three mutations.**
This epistatic cost corresponds to the difference between the cost of mutating all three sites together and the sum of the corresponding single mutations.

In the absence of epistasis, the effects of mutations are additive, so that the two terms of the difference are equal and the epistatic cost is zero. When we look at pairs of polymorphisms occurring together in some *E. coli* strains, we can see that these terms are indeed almost identical: the points on Figure 5.8.A form a line with a slope close to 1 and an intercept close to 0. In other words, we do not detect any epistasis between the pairs of polymorphisms observed in *E. coli*.

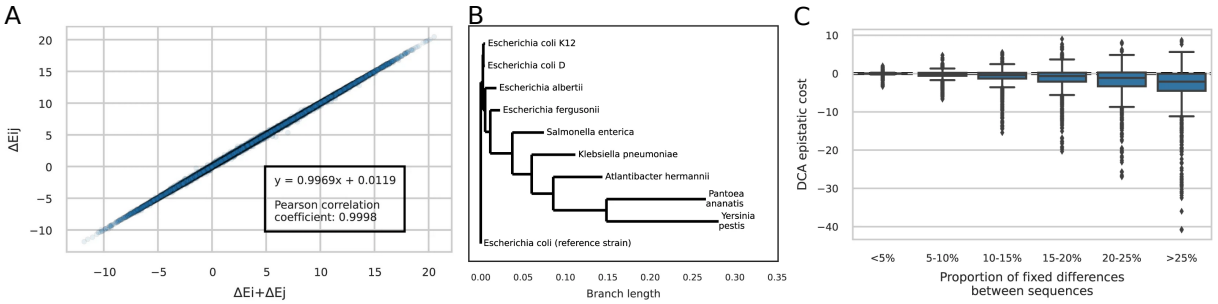

FIGURE 5.8: **Epistasis in *E. coli* and closely related species.**
A. Mutational effect ΔE_{ij} of observed double mutations with respect to the reference, plotted against the sum $\Delta E_i + \Delta E_j$ of the individual mutation scores. The absence of clear deviations from the diagonal reveals the lack of strong epistatic couplings between pairs of mutations in our strain dataset.
B. Phylogenetic tree of studied strains. Tree built from an amino-acid sequence alignment of 878 core genes.
C. DCA epistatic cost decreases with divergence. Negative values correspond to positive epistasis: mutations are more beneficial (lower DCA score) taken altogether than the sum of their individual effects. Boxplot center lines represent medians, box limits are upper and lower quartiles, whiskers extend to show the rest of the distribution within an 1.5 × interquartile range, outliers are represented with points. Sample sizes are $n = 22,352$ for <5%, $n = 15,870$ for 5-10%, $n = 10,810$ for 10-15%, $n = 6776$ for 15-20%, $n = 3564$ for 20-25%, $n = 3432$ for >25%.

This may contradict the finding that DCA performs better than IND in predicting the effect of mutations. Indeed, if epistasis plays no role in protein evolution, the two models should perform similarly. But this apparent contradiction disappears if we return to the proportion of residues coupled to an amino-acid site: DCA captures a diffuse epistasis signal composed of many small couplings. Therefore, one cannot expect to see strong couplings between pairs of polymorphisms.

To observe an epistasis signal, we need to examine more divergent sequences (Figure 5.8.B). We chose to calculate the epistatic cost of all fixed differences between species closely related to *E. coli*. As we can see on Figure 5.8.C, this cost starts to diverge substantially from zero at a divergence of about 10% in the amino-acid sequence. It is negative, which means that the DCA cost of the set of fixed differences is less than the sum of the DCA costs of the individual mutations. In other words, the fixed differences are more beneficial together than can be expected from the sum of their individual effects. This is called 'positive epistasis'. It is also compatible with a model in which fixed differences are contingent on previous mutations and are entrenched by subsequent mutations.

In summary, epistasis plays an important role in determining the effect of mutations. However, we do not expect to observe a large difference between the genetic background of two different *E. coli* strains: most mutations should have the same effect in both. To start observing a real difference, we need to mutate about 10% of the genetic background, which means comparing the genetic back-

ground of two different species. Locally, the mutational landscape is rather smooth, but the overall picture is much more rugged.

Chapter 6

The determinants of amino-acid site variability on short and long time scales

The results presented in sections 6.1, 6.2 and 6.3 were published in (Vigué et al. 2021). The complete article is available in Appendix B.

6.1 Epistasis reduces the variability of an amino-acid site

Here we want to study how variable an amino-acid site can be, and how epistasis may play a role in reducing this variability.

When we train an IND or DCA model, we use an MSA of distant homologues. These are proteins or protein domains that have evolved independently for so long that they have accumulated many mutations. As a result, they usually share only 20 or 30% identity in sequence, but still perform the same function. This shows that over long evolutionary time scales, most protein sites can vary without affecting the function of the protein too much.

However, these sites do not necessarily show the same level of variability in the short term. Indeed, epistatic couplings with the rest of the sequence can restrict the spectrum of mutations that a site can tolerate. For example, a proline may be observed at locus 55 of a protein found in *E. coli*, whereas a glycine is observed at the same locus in its *Mycobacterium tuberculosis* counterpart. This means that the site can vary over long evolutionary time scales, as it can tolerate at least two different amino acids. However, a glycine or any amino acid other than a proline could be deleterious in the *E. coli* context due to the context-dependence of mutation effect, thus preventing this amino-acid site from tolerating polymorphisms in *E. coli* species.

We want to estimate the variability of an amino-acid site over the short and the long term. For this purpose, we introduce two quantities: the Context-Independent Entropy (CIE) and the Context-Dependent Entropy (CDE). They are based on Shannon entropy, a measure derived from information theory that we use here to quantify the variability of an amino-acid site. A site with zero entropy should tolerate only one amino acid: it is conserved. A value of one may, for example, correspond to two amino acids with a frequency of 50% each. The entropy reaches its maximum value of $\log_2(20) = 4.32$, if all 20 possible amino acids have the same probability of being observed.

CIE quantifies the variability of an amino-acid site as predicted by IND. In practice, CIE is the

observed variability of that site across distantly related species. We calculate CIE at locus i directly from the amino-acid frequencies in the i^{th} column of the MSA:

$$\mathrm{CIE}_i = -\Sigma_\beta f_i(\beta) \log_2(f_i(\beta)) \tag{6.1}$$

CDE quantifies the variability of an amino-acid site as predicted by DCA. CDE is therefore dependent on the genetic background: it incorporates the epistatic constraints specific to the species on which we are focusing. It corresponds to the expected level of variability of this site in a given species. We calculate it using the conditional DCA probabilities of observing an amino acid in a specific genetic background:

$$\mathrm{CDE}_i = -\Sigma_\beta P_i^{\mathrm{DCA}}(\beta \mid a^0_{\setminus i}) \log_2(P_i^{\mathrm{DCA}}(\beta \mid a^0_{\setminus i})) \tag{6.2}$$

We calculate the CIE and CDE for all amino-acid sites in our dataset. If epistasis was negligible, these two quantities should take similar values. As can be seen on Figure 6.1, two groups clearly emerge: a peak at the top right of sites with high CDE and CIE and a peak at the left of sites with low CDE and low to high CIE. The former show very low context-dependence (both entropies have comparable values). They reach entropy values close to 4, *i.e.* close to the upper limit of $\log_2(20) = 4.32$. These sites are variable between species and DCA predicts that they are highly polymorphic in *E. coli*. Conversely, the latter can sometimes vary between distantly related species (CIE ranging from 0 to more than 3), but DCA predicts that they remain conserved in *E. coli* (CDE near 0). We expect these sites to show a low level of polymorphism in *E. coli*.

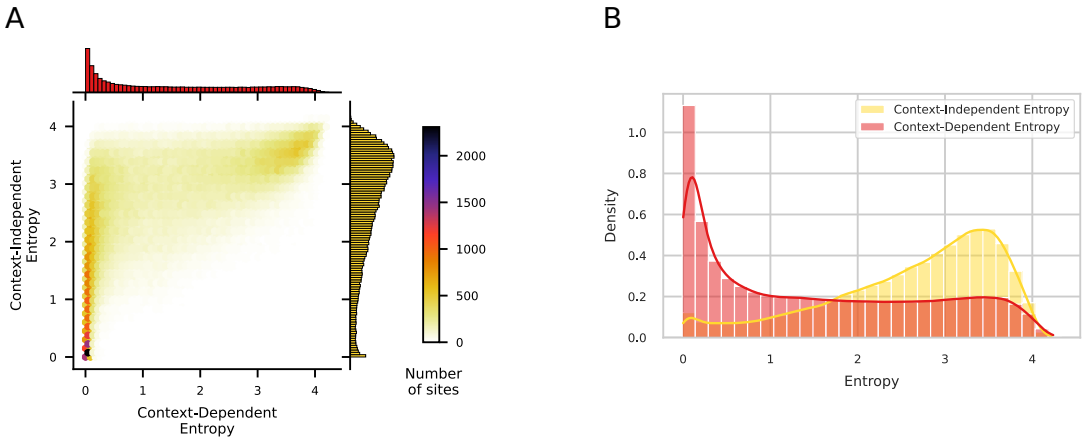

FIGURE 6.1: **Predicting the variability of amino-acid sites.**
A. Bivariate histogram of CDE and CIE for all sites in the dataset. Two populations of sites are clearly recognizable, in particular separated by their CDE values.
B. Marginal distributions of CDE (red) and CIE (yellow) for all sites in the dataset. CDE divides amino-acid sites into two populations of similar sizes: conserved (CDE < 1) and variable (CDE ≥ 1). On the contrary, most of the amino-acid sites have a high CIE, *i.e.* IND predicts them to be highly variable.

The CIE and CDE distributions of all sites are very different. While only 8.3% of the sites are conserved between distantly related species (CIE < 1, corresponding to an effective number of amino acids less than 2), we predict that 45% of the sites will be conserved in *E. coli* (CDE < 1), mainly due to local epistatic couplings.

6.2 Taking epistasis into account is crucial to predict polymorphisms

Neither CIE nor CDE use *E. coli* polymorphism data to predict variability within this species. We can therefore compare their predictions with the observed variability among *E. coli* strains.

To do this, we classify *E. coli* sites into two categories: conserved sites (no polymorphism observed in any of the strains) and variable sites (at least 5% of strains carry a mutation). Lowly polymorphic sites (polymorphisms at frequencies <5%) may correspond to variable sites but also to conserved sites with deleterious mutations segregating at low frequencies (or sequencing errors for some of the lowest frequencies), so we choose to exclude them from the analysis.

Most of the conserved sites cluster on the left peak of low CDE, while the variable sites tend to cluster on the upper right peak of high entropies (Figure 6.2). CDE seems to be more relevant than CIE for distinguishing conserved sites from variable sites. Indeed, only 12.7% of conserved sites have a CIE < 1, while 56.4% have a CDE < 1. There are, though, sites with a high CDE for which we observed no polymorphism in any of our *E. coli* strains. We see two non-exclusive explanations for this observation.

First, some amino acids may be tolerated at these sites, but they are rarely present in practice. Indeed, we cannot obtain all 20 possible amino acids by mutating a given codon more than once. Some amino-acid changes may require two or three nucleotide changes to occur. This chain of events is unlikely, especially if the intermediate codons code for deleterious amino acids.

Second, random drift may limit the amount of neutral diversity that can segregate in a species.

To examine the effect of the first assumption, we can restrict the calculation of CIE and CDE to amino-acid changes that only require a difference of one nucleotide—hereafter referred to as '1-SNP amino-acid mutations'. This involves modifying the entries of the f_i vector in the case of IND and of the P_i^{DCA} vector of conditional probabilities in the case of DCA to zero out all entries that do not correspond to 1-SNP amino-acid mutations and then renormalizing these vectors. These updated vectors can then be used in formulas 6.1 and 6.2 to calculate a CIE and CDE restricted to 1-SNP amino-acid mutations. With these new 1-SNP CIE and CDE, 70.2% of the conserved sites have a CDE < 1 while only 24.8% have a CIE < 1.

Yet, this leaves 29.8% of conserved sites that are expected to be polymorphic (CDE ≥ 1). This encourages us to examine the effect of random drift on limiting the amount of neutral diversity segregating in *E. coli*. Indeed, even among synonymous mutations, many 1-SNP mutations are absent from our dataset. Using simulations based on the amount of observed synonymous diversity, we can estimate the proportion of amino-acid sites that will remain conserved while they could tolerate polymorphisms (high CDE). The exact procedure and results are detailed in (Vigué et al. 2021), available in Appendix B. Overall, these results are qualitatively consistent with our observations: polymorphisms may occur at these sites but have not yet been observed in nature.

6.3 Quantifying contingency

Many amino-acid sites are conserved in *E. coli* due to a network of epistatic interactions that limit the range of possible mutations. We therefore want to quantify the role played by the genetic background

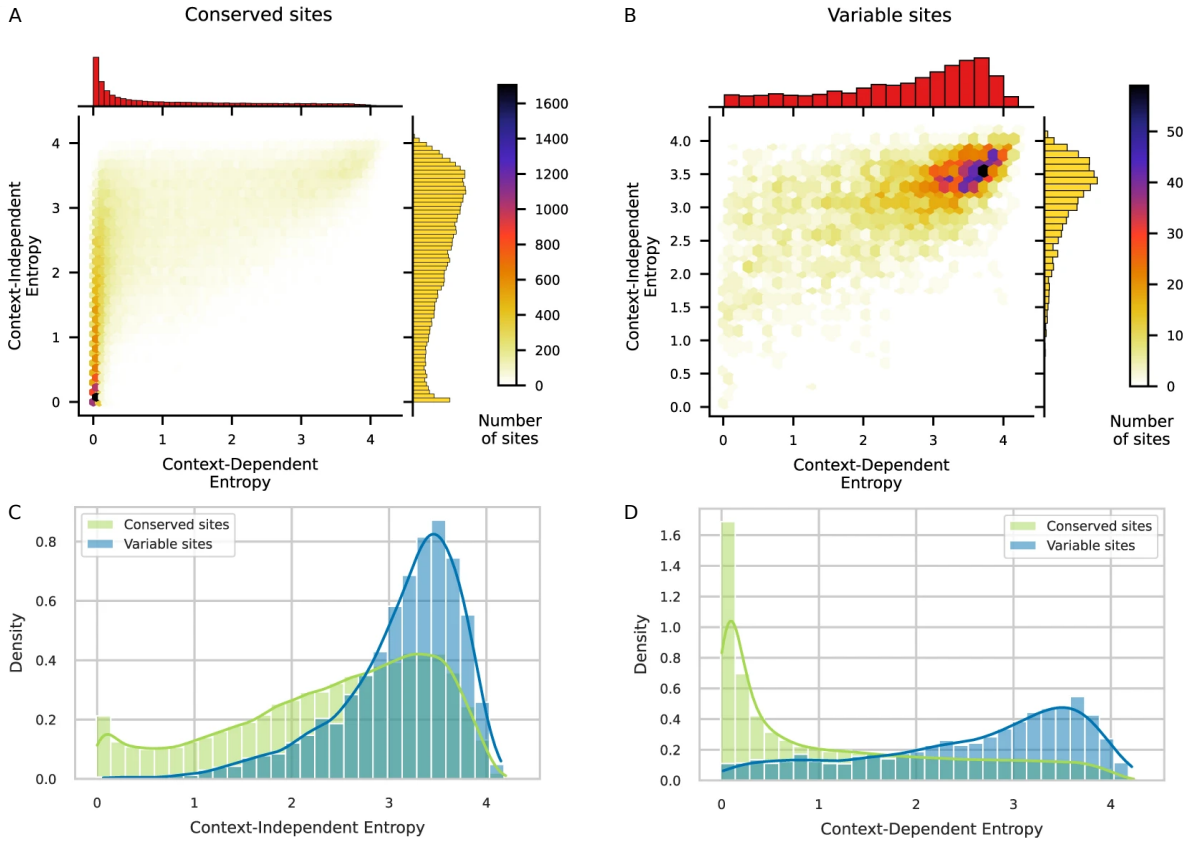

FIGURE 6.2: **Predicting amino-acid sites that are conserved or polymorphic in *E. coli*.**
A. Bivariate histogram of CDE and CIE for sites that are conserved across >60,000 strains of *E. coli*. Most of them cluster on the left peak of low CDE.
B. Bivariate histogram of CDE and CIE for sites that are polymorphic at a 5% threshold across >60,000 strains of *E. coli*. Most of them cluster on the right peak of high CDE.
C. Distribution of CIE for conserved (green) and polymorphic (blue) sites in *E. coli*. A non-epistatic model fails at distinguishing between both populations. Most of the sites are predicted to have a high entropy so to be highly variable, including those that display no mutation in >60,000 strains of *E. coli* (green distribution).
D. Distribution of CDE for conserved (green) and polymorphic (blue) sites in *E. coli*. A model that incorporates pairwise epistasis predicts a low entropy for conserved sites (the green distribution peaks near 0) and a high entropy for variable sites (the blue distribution peaks near 4).

in reducing the diversity of amino acids that a site can tolerate.

Comparing CIE and CDE allows us to quantify contingency, as they both quantify site variability, CIE being context-agnostic and CDE being context-aware. A simple difference between them gives the information gain (IG).

$$\text{IG} = \text{CIE-CDE} \tag{6.3}$$

IG quantifies the difference between the variability of an amino-acid site in distantly related species and its potential variability in *E. coli*. It is expressed in bits: 1 bit corresponds to an effective reduction in available amino acids by a factor of 2, 2 bits by a factor of 4 and 3 bits by a factor of 8. If CDE is equal to CIE, no information is contained in the genetic background, IG=0. The lower the CDE compared to the CIE, the higher the IG and therefore the contingency level.

We can roughly classify the amino-acid sites into three categories:

- First, 8.3% of the sites are conserved in all species as well as in *E. coli* (CIE < 1). It is likely that they are functionally essential. A mutation will always be deleterious, so the context has no real influence on their level of conservation. For example, they may correspond to key amino acids in the active site of an enzyme.

- Second, 55.1% of the sites are variable in all species as well as in *E. coli* (CIE ≥ 1, CDE ≥ 1). They are often constrained (CDE < $\log_2(20)$), but allow for considerable amino-acid variability, even in the specific context of *E. coli*: at these positions we can observe both fixed differences between species and polymorphisms within the *E. coli* population. Most of the mutations occurring at these sites will be selectively neutral. These sites often correspond to the least critical sites for protein folding and function, usually those exposed at the surface and located away from the active site of an enzyme.

- Third, 36.6% of the sites are conserved in the *E. coli* context but vary between species (CIE ≥ 1, CDE < 1). Amino acids observed in distantly related species will not be tolerated in this specific context: evolution depends on the genetic context. These sites are those most constrained by epistatic interactions, a good example being the amino-acid sites buried in the 3D structure of the protein.

6.4 How amino-acid sites fix mutation with divergence

CDE predicts the variability of an amino-acid site within the genetic context of *E. coli*. A site with a low CDE can exhibit two distinct behaviors: it can be conserved across all distant species—resulting in low CIE and IG—or it can be variable in more distant species but conserved in *E. coli* due to a network of epistatic interactions. In the first case, we would rarely observe any fixed differences at that site. However, in the second case, we might observe fixed differences if the genetic context of the sequence is significantly different from that of *E. coli*. Based on these observations, we can make two predictions:

- Sites with low CDE will accumulate less fixed differences across distant species than those with high CDE.

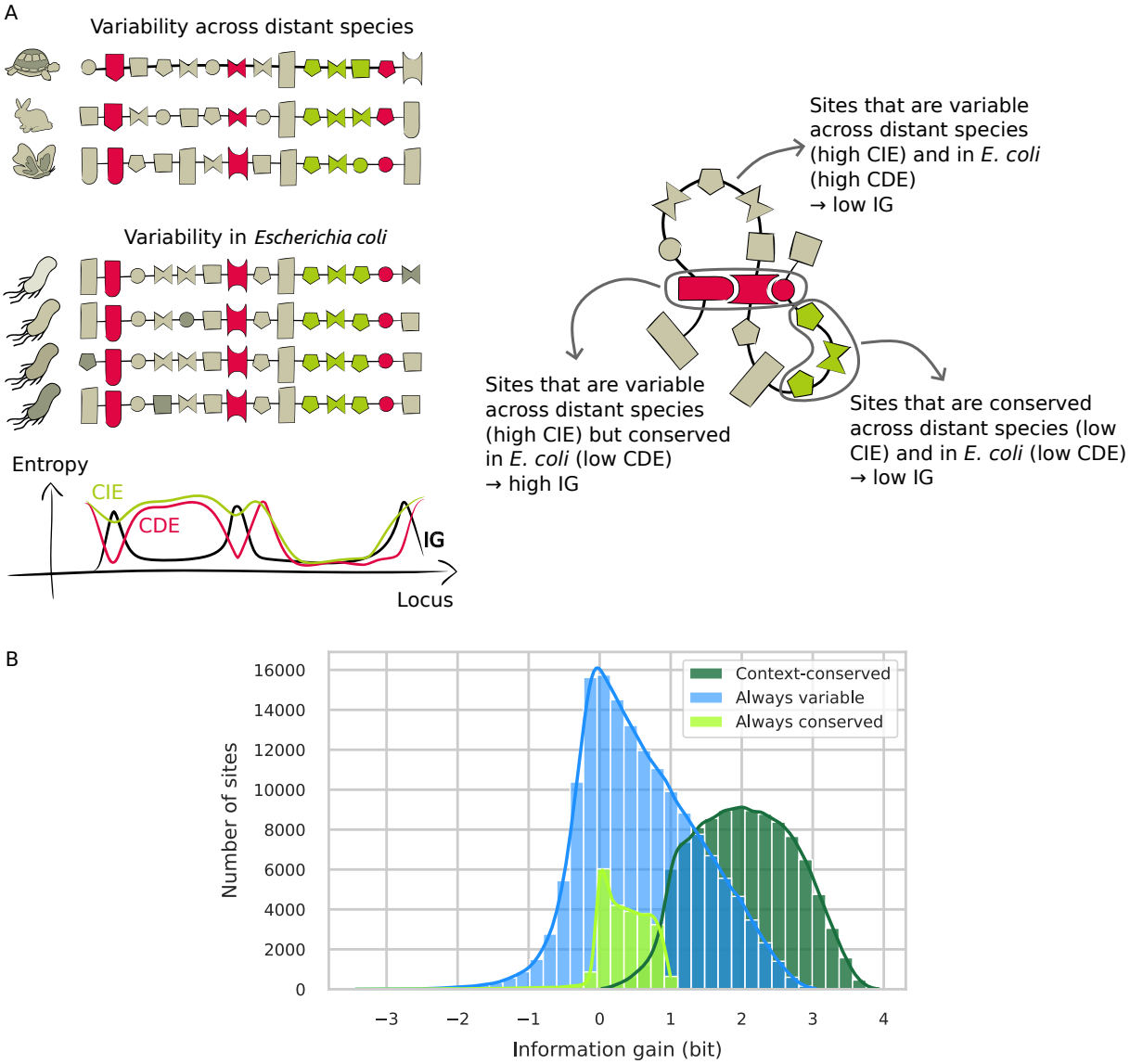

FIGURE 6.3: **Quantifying the effect of the context in reducing amino-acid site variability.**
A. The genetic background is expected to differentially impact amino-acid sites. It has a low influence on sites that have the same level of variability in *E. coli* and across distant species. On the contrary, it strongly impacts sites that are variable across distant species but are conserved in *E. coli* due to local epistatic couplings.
B. Information gain (IG) quantifies the difference between an amino-acid site variability across distant species and its potential variability in *E. coli*. Sites that are variable across distant species (CIE ≥ 1) but conserved in *E. coli* (CDE < 1) are the ones with the highest information gains (dark green distribution).

- It will require more sequence divergence to start observing fixed differences on sites with low CDE compared to sites with high CDE.

These predictions align with the findings discussed earlier. Sites with high CDE can tolerate polymorphisms in *E. coli*, suggesting that they are also more likely to undergo amino-acid changes during divergence between species compared to other types of sites. Our *E. coli* database allows us to compare the sequences of *E. coli* persistent proteins with their homologues in Swiss-Prot. This comparison provides a straightforward method to validate our predictions and gain further insights into the evolutionary dynamics of these sites.

First, let's examine the proportion of amino-acid sites that have undergone the fixation of an amino acid that is different from the one observed in *E. coli* (Figure 6.4.A). This proportion is obviously influenced by the similarity between the *E. coli* sequence and its homologue: the more divergent the homologue, the higher the proportion of fixed differences. The site's CDE also plays a role: sites with higher CDE are associated with a greater proportion of fixed differences. However, this outcome was anticipated.

Interestingly, when observing Figure 6.4.A, we notice that the lines representing the most diverged sequences (dark purple) appear flatter compared to those representing the less diverged ones. In sequences with over 90% identity to *E. coli*, a site with a CDE ≥ 3.5 has approximately 20 times higher chances of undergoing a mutation compared to a site with a CDE ≤ 0.5. However, in sequences sharing between 50% and 60% identity with *E. coli*, this figure drops to 8. In these more divergent sequences, the genetic context has undergone significant changes, allowing for more frequent fixation of amino-acid changes at sites that are highly constrained by epistatic interactions.

Another approach to investigate the same question involves choosing, for each amino-acid site, the least diverged sequence that has undergone a fixed difference at that specific site, if such differences exist. This analysis is presented in Figure 6.4.B. Although the boxplots demonstrate large variability within each CDE range, there is a noticeable overall pattern of decreasing sequence identity as the CDE decreases. This further emphasizes that in order to fix a mutation at a highly constrained site, significant alterations in the genetic context are necessary.

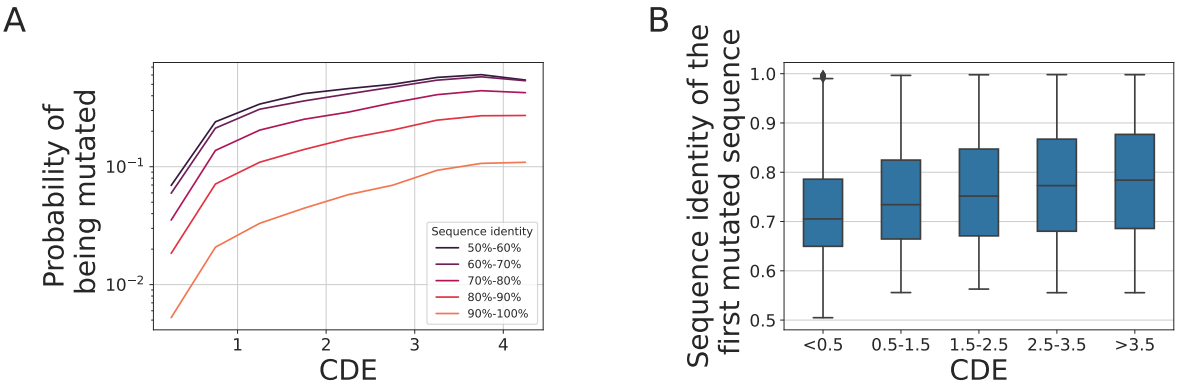

FIGURE 6.4: **Fixed differences and Context-Dependent Entropy (CDE).**

A. Proportion of sites that display a fixed difference with *E. coli* in homologous sequences of varying divergence according to the CDE of the amino-acid site.

B. Sequence identity of the closest homologous sequence found in Swiss-Prot that displays a fixed difference according to the CDE of the amino-acid site.

Chapter 7

The determinants of protein evolution on short and long time scales

The results presented in this chapter have been accepted for publication in PNAS but not published yet. The complete article is available in Appendix C.

7.1 Motivation

We discussed in the previous chapter that amino-acid sites display different degrees of variability and, more importantly, that some amino-acid sites exhibit contrasting patterns of variability over short and long time scales. Indeed, some may vary between distant species but be constrained in *E. coli* by a network of epistatic interactions. In this chapter we want to investigate somewhat similar questions, but at the protein level. Do all proteins evolve at the same rate? Are the dynamics of protein evolution similar in the short and long term?

We already have some answers to the first question. Indeed, it has been shown that the most critical parts of the genome evolve at a slower rate. In the early 2000s, it was observed that essential and highly expressed genes had lower divergence rates. Because essential genes are often the most highly expressed, it took careful statistical analysis to prove that the level of gene expression determines the rate at which genes fix mutations and to close what was at the time a very hot debate (Hirsh et al. 2001; Pál et al. 2001; Jordan et al. 2002; Pál et al. 2003).

However, this seminal work focused on interspecies divergence: it relied on a handful of distant genomes and therefore studied mutations that had reached fixation. When taken into account, polymorphism was thought to reflect a lack of selection. For example, the McDonald-Kreitman test detects adaptive mutations that have reached fixation, using polymorphisms as a control (McDonald et al. 1991; Smith et al. 2002). However, polymorphisms observed within a species are also subject to selection. Moreover, they may undergo very different selective pressures than mutations fixed during species divergence. For example, some polymorphisms may be transiently advantageous, typically if an organism moves through different niches where it encounters abrupt environmental changes. We can therefore investigate whether a signature of natural selection on polymorphism is detectable, and if so, whether it is distinct from the signature observed on divergence.

E. coli could be an excellent model to study this topic. Firstly, because it is extensively sequenced

and therefore we have a lot of polymorphism data to analyse. Secondly, because it is a generalist organism that has to travel—and therefore adapt—to very different niches. For example, in the event of an extra-intestinal infection, it leaves the gut and adapts to survive in another organ—a process known as patho-adaptation.

Investigating selection within a species raises challenges, though. Approaches that compare non-synonymous and synonymous mutations are suitable for studying distant species but give misleading conclusions with closely related organisms (Rocha et al. 2006). Another obstacle lies in the choice of sequenced organisms. The importance of a strain in a genome collection often reflects its scientific or clinical relevance more than its prevalence in nature. This implies that tests based on mutation frequencies may also be unreliable. For these reasons, we have chosen to focus again on the functional impact of the observed amino-acid changes, as predicted by Direct-Coupling Analysis (DCA). To this aim, we are introducing a new approach: the Gene-Level Amino-acid Score Shift (GLASS).

7.2 GLASS: using predicted effect of mutations to test for selection

We often study selection by comparing the occurrence of synonymous and non-synonymous mutations. This standard approach assumes that most non-synonymous mutations decrease fitness: the more intense the purifying selection, the fewer non-synonymous mutations should remain (Rocha 2018). In contrast to this approach, we choose to focus only on the predicted effect of non-synonymous mutations. We introduce a new test of selection based on DCA predictions of mutation effects: the Gene-Level Amino-acid Score Shift (GLASS).

The idea behind GLASS is to compare the distributions of DCA scores of real and simulated mutations (Figure 7.1). For a protein of length L, we first identify real mutations in natural isolates. We note N_{AA} the number of distinct mutations found. They represent a sample of the $19L$ mutations that could occur in this protein. We can compute a DCA score for each of these N_{AA} real mutations as well as for all possible $19L$ mutations (blue and grey distributions in Figure 7.1). In the presence of purifying selection, real mutations should be less detrimental—*i.e.* have lower DCA scores—than the average ones. This means that the distribution of real mutations should be shifted to the left compared to the distribution of all possible mutations.

Accurate comparison of the distribution of observed mutations with the distribution of all possible mutations is not straightforward. Using optimal transport theory, we can quantify the shift S_{obs} between the two distributions. Optimal transport is a field of mathematics that aims to solve a simple problem: what is the most efficient way to transport mass? We can think of two distributions as two piles of sand and ask what is the most efficient way to move sand to transform the first distribution into the second. Answering this question means finding an optimal match between the two distributions. Using Python Optimal Transport library (Flamary et al. 2021), we can calculate an optimal transport matrix M between the blue and grey histograms that have been normalised to the same total number of mutations N_{norm}. We then use this matrix to compute S_{obs}, as follows:

$$S_{obs} = (\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}(i-j)M_{i,j})/N_{norm} \tag{7.1}$$

If the distribution of scores of real mutations and the one of all possible mutations are super-

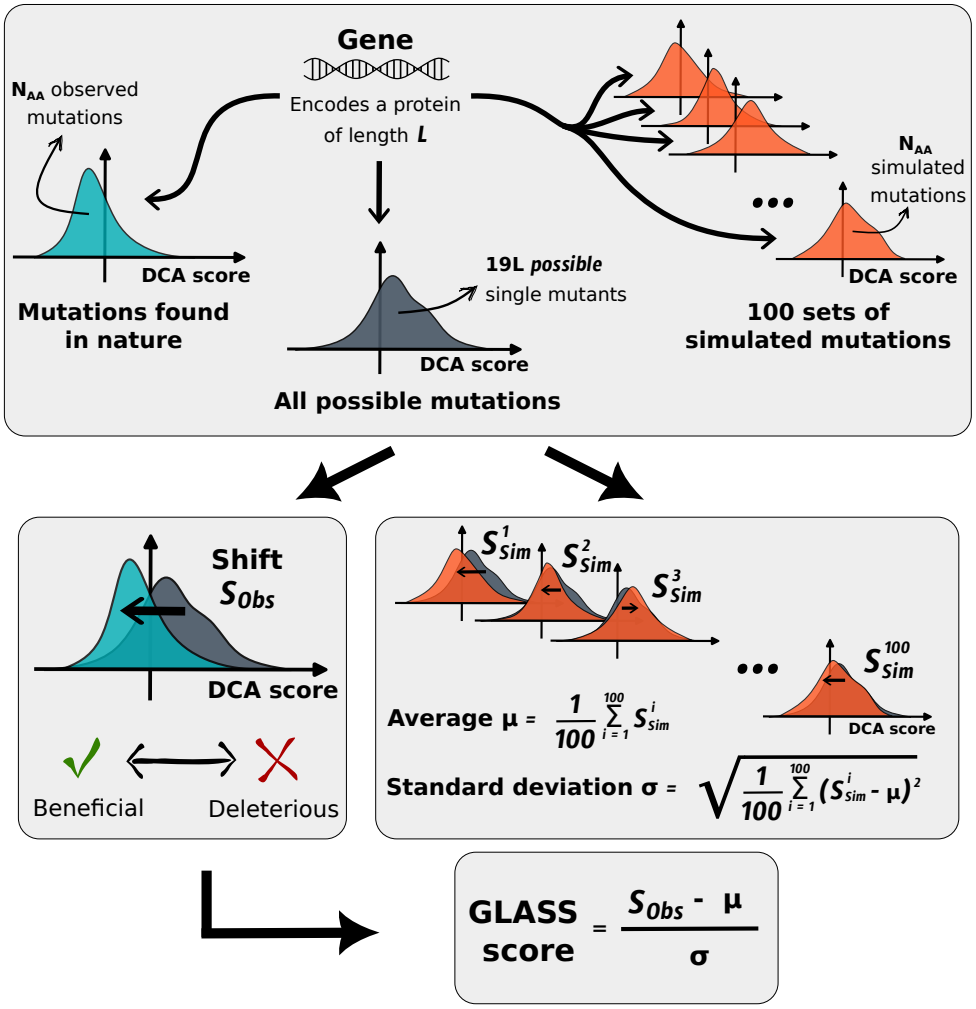

FIGURE 7.1: **Gene-Level Amino-acid Score Shift (GLASS) procedure to test for selection.**
$19L$ different mutations can occur on a protein of length L. Direct-Coupling Analysis (DCA) predicts the effect of these mutations (grey distribution, negative DCA scores correspond to beneficial mutations, positive DCA scores to deleterious mutations, DCA scores of zero to neutral mutations). We analyse natural isolates to gather mutations observed in nature (N_{AA} distinct mutations whose DCA scores form the blue distribution). We simulate 100 sets of the same number of mutations (100 sets of N_{AA} mutations shown in orange) under a neutral evolutionary model, *e.g.* Jukes-Cantor 1969. By comparing each blue or orange distribution to the grey distribution, we calculate shift values that represent how the distributions of DCA scores of the real or simulated mutations are shifted towards lower values—*i.e.* more beneficial mutations—relative to the distribution of DCA scores of all possible mutations. GLASS score compares the S_{obs} shift of the real mutations to the average and standard deviation of the 100 S^i_{sim} shift values of simulated mutations.

imposable, $S_{obs} = 0$. The more beneficial mutations segregate on the gene, the higher S_{obs}. S_{obs} is signed so that an excess of beneficial mutations translates into a positive value of S_{obs}, while an excess of deleterious mutations results in a negative value.

We then compare S_{obs} to shift values expected in absence of selection. To do so, we simulate 100 sets of N_{AA} mutations under Jukes Cantor 1969 model (orange distributions in Figure 7.1). We compare each of these 100 orange distributions to the grey one to derive 100 S_{sim} shift values. We note their mean μ and their standard deviation σ. These allow us to calculate a GLASS score:

$$\text{GLASS score} = \frac{S_{obs} - \mu}{\sigma} \tag{7.2}$$

The GLASS score is a Z-score, *i.e.* it corresponds to the number of standard deviations between S_{obs} and μ. It approaches zero if the real and simulated mutation distributions look the same. The more efficient the purifying selection, the more different the distributions, the higher the GLASS score. Of note, small numbers of mutations display high fluctuations. This impacts GLASS scores that are more likely to be closer to zero for genes with low N_{AA} values. Any comparison between GLASS scores should thus account for the number of distinct mutations observed in each gene. This also means that sampling more genomes will improve the power of the test by increasing N_{AA}.

GLASS scores are based on the comparison of simulated and real mutations. Simulating mutations allows us to take into account mutation biases. The most important of these biases is the fact that mutations often involve only one of the three bases in a codon. This leads to changes between amino acids that often share similar physico-chemical properties. Therefore, even in the complete absence of selection, the observed mutations will tend to be less deleterious than all possible mutations. By subtracting μ from S_{obs} in the calculation of the GLASS score, we correct for these mutational biases to focus on the true action of purifying selection.

7.3 In the short term, essentiality and expression level determine the intensity of selection acting on a gene

We calculated GLASS scores for 2,534 persistent genes using polymorphisms found in 60,472 *E. coli* genomes. Hereafter, these scores computed from *E. coli* polymorphisms will be referred to as GLASS-P scores. We also computed non-synonymous to synonymous diversity ratios (Π_N/Π_S) and nucleotide diversities (Π) for all 2,534 genes.

All mutations—regardless of their frequencies—contribute to the GLASS score equally. This makes GLASS-P scores reflect very recent selection events. Indeed, most *E. coli* mutations segregate at low frequencies and must be very recent. In contrast, Π_N/Π_S gives more weight to high frequency mutations, as it is based on all pairwise comparisons. It thus tells us about selection acting on somewhat longer time scales.

In the early 2000s, there had been an intense debate on the factors influencing the rate of gene evolution and more precisely on the relative roles of gene expression level and gene essentiality in driving gene evolution. For this reason, we wanted to estimate the level of expression of each gene, as well as its essentiality, in order to see how these correlate with our GLASS scores. We used Codon Adaptation Index (CAI) from MaGe (Vallenet et al. 2017) to estimate the level of gene expression. We assessed gene essentiality with data from insertion-seq experiments (Couce et al. 2017).

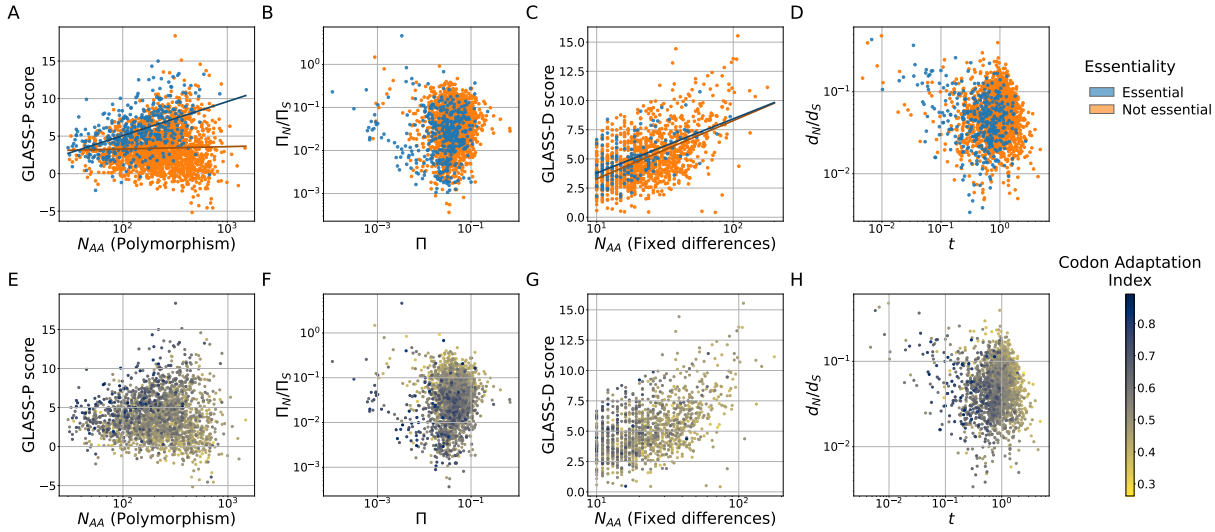

FIGURE 7.2: **Short- and long-term selection patterns according to gene essentiality and expression level.**
A. GLASS-P scores and numbers of observed distinct amino-acid polymorphisms, coloured by gene essentiality (essential genes in blue, non-essential genes in orange, the linear fits for these two groups are represented with solid lines). Essential genes carry fewer harmful polymorphisms than non-essential ones (higher GLASS-P scores).
B. Gene non-synonymous to synonymous diversity ratios (Π_N/Π_S) and nucleotide diversities (Π), coloured by gene essentiality. Essential genes tend to be less polymorphic (lower Π) and to carry slightly less non-synonymous than synonymous polymorphisms (lower Π_N/Π_S ratio).
C. GLASS-D scores and numbers of observed distinct amino-acid fixed differences, coloured by gene essentiality.
D. Gene non-synonymous to synonymous substitution rate ratios (d_N/d_S) and nucleotide substitution rates (t) during divergence between *E. coli* and *S. enterica*, coloured by gene essentiality.
E. GLASS-P scores and numbers of distinct amino-acid polymorphisms, coloured by Codon Adaptation Index (CAI, a proxy for gene expression level). The higher the CAI, the higher the GLASS-P score: highly expressed genes carry fewer deleterious polymorphisms.
F. Gene non-synonymous to synonymous diversity ratios (Π_N/Π_S) and nucleotide diversities (Π), coloured by CAI. The higher the CAI, the lower Π and Π_N/Π_S ratio: highly expressed genes are less polymorphic and carry less non-synonymous mutations compared to synonymous ones.
G. GLASS-D scores and numbers of distinct amino-acid changes fixed during divergence between *E. coli* and *S. enterica*, coloured by CAI.
H. Gene non-synonymous to synonymous substitution rate ratios (d_N/d_S) and nucleotide substitution rates (t) during divergence between *E. coli* and *S. enterica*, coloured by CAI.

The level of gene expression correlates with the intensity of purifying selection: the higher the level of expression of a gene, the higher its GLASS-P score and the lower its Π_N/Π_S ratio (Figures 7.2.E, 7.2.F). The GLASS-P score and the Π_N/Π_S ratio explain a comparable proportion of the variance of the CAI (8.7% and 10.5% respectively, see Table 7.1). These contributions to the prediction of the CAI tend to complement each other. Indeed, a general linear model using both covariates explains on average 8.5% more variance than a model that only uses one (Supplementary Note 2 in Appendix C). This is because they carry different information. On one side, Π_N/Π_S reflects the occurrence of non-synonymous mutations, GLASS-P scores, meanwhile, focus on their effects. As previously mentioned, GLASS-P and Π_N/Π_S also reflect selection acting on slightly different time scales.

The GLASS-P score better predicts gene essentiality: essential genes cluster in Figure 7.2.A based on GLASS-P scores but not in Figure 7.2.B that uses Π_N/Π_S. It explains 10 times more variance in gene essentiality than the Π_N/Π_S ratio (10.3% versus 1%, see Table 7.1). The association between GLASS-P scores and gene essentiality remains statistically significant even after correction by gene expression level.

TABLE 7.1: **Proportion of the variance in essentiality or Codon Adaptation Index (CAI) explained by covariates inferred from *Escherichia coli* polymorphisms**

	Essentiality	**CAI**
GLASS-P	10.3%	8.7%
$\log(N_{AA})$	5.4%	3.5%
$\log(\Pi_N/\Pi_S)$	1%	10.5%
$\log(\Pi)$	4.9%	2.3%
Essentiality	–	4.3%
CAI	4.5%	–
Full model	27.0%	30.9%

A dominance analysis (Budescu 1993) was performed with generalized linear models (GLM) aiming at predicting either gene essentiality or gene CAI from all possible combinations of five covariates to estimate the contribution of each covariate to the prediction of the response variable. The total variance explained by a GLM with all covariates is shown in the last row (Full model). We used McFadden's R^2 (McFadden 1974) to estimate the variance explained by GLMs.

7.4 In the long term, the expression level drives the rate at which a gene evolves

We then compared *E. coli* gene sequences with their counterparts in *Salmonella enterica*. We focused on 1,421 genes that have fixed at least 10 mutations during divergence between these two species. We used these fixed differences to compute: a non-synonymous to synonymous substitution rate ratio (d_N/d_S), a nucleotide substitution rate (t), and a GLASS score. The latter will be referred to as GLASS-D score because it is based on divergence data.

Neither the GLASS-D score nor the d_N/d_S ratio predict gene essentiality well (Figures 7.2.C, 7.2.D, Table 2). But, both t and d_N/d_S correlate very well with CAI. A general linear model (GLM) based on

species divergence data (Table 7.2) better predicts CAI than a GLM based on polymorphisms found in *E. coli* (Table 7.1). The level of expression rather than their essentiality determines the rate at which genes diverge. This is in line with conclusions from other studies ((Pál et al. 2001), (Pál et al. 2003)).

The explanatory power of GLASS-D score remains low (Table 7.2). This lack of power could be due to the small number of amino acids fixed during divergence. In comparison, GLASS-P scores use all mutations found across tens of thousands of strains. Another difference lies in the effects of the mutations. While some polymorphisms are neutral and others deleterious, all fixed mutations should be close to neutral. Direct-Coupling Analysis will thus be more useful to analyse the former than the latter. As a result, GLASS-P scores will carry more information than GLASS-D scores.

7.4.1 Genes that carry more deleterious polymorphisms are also more frequently lost

Up to now, we focused on mutations found in gene sequences that are complete. But some strains may contain only a fragment of a gene. This occurs when a nonsense mutation leads to a premature stop codon in the DNA sequence. For each of the 2,534 genes, we recorded the number of *E. coli* strains that carry a complete gene sequence and those that carry only a fragment. Figure 7.3.A shows a negative correlation between GLASS-P score and gene loss. Genes that are often lost also accumulate more deleterious mutations in the strains where they remain. Π_N/Π_S ratio also correlates with gene loss but less strongly (Figure 7.3.B). What matters is not so much the number of polymorphisms that accumulate in a gene but their effects. For instance, the gene with the lowest GLASS-P score—*rbsR*—is frequently pseudogenized (pseudogene to full gene ratio of 0.002345, *i.e.* higher than the 80^{th} percentile of the distribution) but has a rather low Π_N/Π_S ratio (0.02093, *i.e.* lower than the 30^{th} percentile of the distribution). We have to keep in mind that we have focussed on genes that are highly conserved within the *E. coli* species. This implies that most of them are needed at some point, even if dispensable over short periods.

TABLE 7.2: **Proportion of the variance in essentiality or Codon Adaptation Index (CAI) explained by covariates inferred from mutations fixed during divergence between *Escherichia coli* and *Salmonella enterica***

	Essentiality	**CAI**
GLASS-D	0.2%	3.1%
$\log(N_{AA})$	1.1%	5.2%
$\log(d_N/d_S)$	0.7%	8.3%
$\log(t)$	5.4%	19.9%
Essentiality	–	2.8%
CAI	3.7%	–
Full model	11.4%	38.9%

The same methodology than the one presented in Table 7.1 was used to obtain the estimates of the variance explained by covariates.

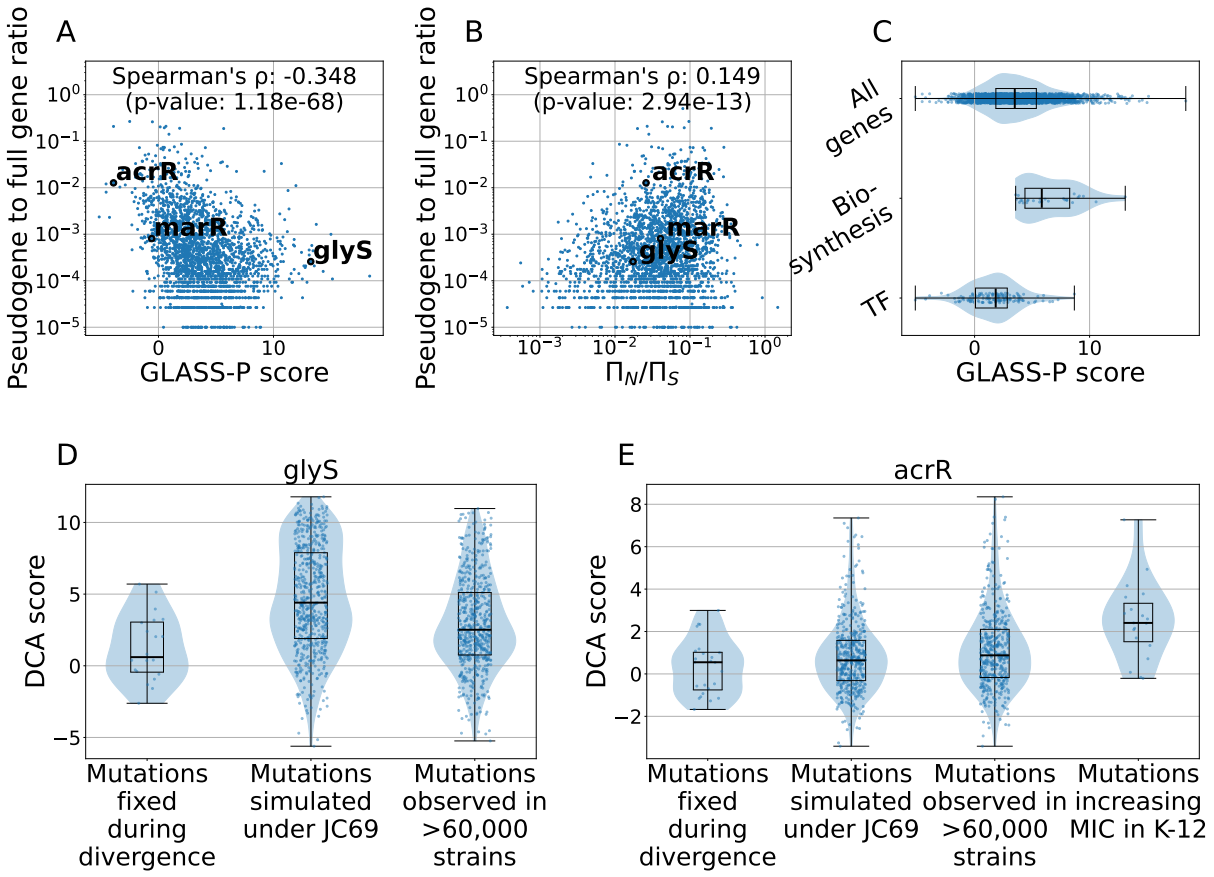

FIGURE 7.3: **Divergent short-term selective pressures depending on gene function.**

A. Scatterplot of pseudogene to full gene ratio according to GLASS-P score. This is the ratio of the number of *E. coli* strains for which we find fragments of a given gene compared to the number of strains where we find the complete gene sequence. We added 10^{-5} to pseudogene to full gene ratio values in order to visualise null values on a log scale.

B. Scatterplot of pseudogene to full gene ratio according to gene Π_N/Π_S ratio.

C. GLASS-P scores for: all genes, biosynthetic genes deleted in (D'Souza et al. 2014) to produce auxotrophic *E. coli* strains, and genes that code for transcription factors.

D. DCA scores of three samples of non-synonymous mutations in *glyS* gene: 20 mutations fixed during divergence between *E. coli* and *S. enterica*, 597 mutations simulated under Jukes-Cantor 1969 model and 597 polymorphisms observed across *E. coli* strains.

E. DCA scores of four samples of non-synonymous mutations in *acrR* gene: 197 mutations fixed during divergence between *E. coli* and *S. enterica*, 451 mutations simulated under Jukes-Cantor 1969 model, 451 polymorphisms observed across *E. coli* strains and 19 mutations found in at least one of 60,472 *E. coli* genomes and shown to increase minimum inhibitory concentration (MIC) of *E. coli* K12.

7.5 Deleterious polymorphisms target transcription factors

High GLASS scores signal very strong purifying selection. Under more relaxed selection, they should approach zero. Yet, genes such as *rbsR* reach GLASS-P scores as low as -5.15: the number of deleterious polymorphisms segregating on these genes exceeds even what would be expected in the complete absence of purifying selection.

Selection can favour deleterious polymorphisms in a gene. It occurs if the organism benefits from the loss or the reduced functionality of the encoded protein. A deleterious mutation at the protein level may therefore be beneficial at the organism level. This often depends on specific conditions: inactivating a protein may increase fitness in an environment but decrease it elsewhere. Deleterious mutations will then be promoted before being counter-selected. As a consequence, we will observe an over-representation of deleterious mutations that never reach high frequencies. In line with this scenario, no gene has a negative GLASS-D score (Figures 7.2.C, 7.2.G): they never fix more deleterious mutations than expected by chance.

To investigate the function of genes with a negative GLASS-P score, we performed a GO-term enrichment analysis. This analysis detects a strong overrepresentation of transcription factors (Figure 7.3.C). In particular, the most enriched biological process is "Positive regulation of cellular biosynthetic process" (GO:0031328) with a 4.28-fold enrichment (16 observed genes) and the most enriched molecular function is "DNA-binding transcription factor activity" (GO:0003700) with a 3.96-fold enrichment (30 observed genes, many of which also belong to GO:0031328).

These transcription factors regulate very different functions. Some, such as *rpoS* and *ada*, control the stress response. Strains stored in a lab often lose *rpoS* (Bleibtreu et al. 2014), we thus expected to find many low frequency deleterious variants of this gene in our dataset. *dpiA*, *narL*, *fhlA* and *fnr*—four regulators of growth under anaerobic conditions—also carry many deleterious mutations. Other inactivations may reflect patho-adaptation. For example, *lrhA*, *rcdA* and *flhC* regulate motility and biofilm. In particular, *lrhA* represses the expression of type 1 fimbriae and flagella—two known virulence factors in extra-intestinal diseases (Lehnen et al. 2002). Strains isolated in urosepsis have often inactivated *lrhA* (Kisiela et al. 2017). We also detect repressors of genes involved in antibiotic resistance, *e.g.* *mprA*, *nimR*, *acrR* and *marR*. Their loss may allow a strain to respond to an antibiotic treatment by increasing the expression of downstream genes. Last but not least, many genes with negative GLASS-P scores regulate sugar metabolism: *cdaR*, *chbR*, *rhaR*, *rhaS*, *xylR*, *rbsR*, *malT*. Loss of *rbsR* constitutes a case of convergent evolution, as it has been observed in isolates from different patients suffering from extra-intestinal acute infection (Bridier-Nahmias et al. 2021). *malT* is a special case: it activates the maltose operon to which *lamB*—the receptor of bacteriophage lambda—belongs. Thus, its loss also triggers phage resistance (Cole et al. 1986).

Arguably, *E. coli* strains may benefit from the inactivation of genes other than transcription factors. For example, it has been suggested that the loss of essential biosynthetic pathways may increase fitness when the corresponding metabolite is available. In particular, Glen D'Souza and colleagues experimentally studied *E. coli* strains auxotrophic for several amino acids, nucleobases or vitamins (D'Souza et al. 2014). To do so, they produced a library of *E. coli* strains that were deleted for one of the following biosynthetic genes: *trpE*, *trpA*, *proC*, *ilvA*, *lysA*, *argE*, *pheA*, *pyrF*, *leuB*, *argC*, *nadA*, *argB*, *argH*, *hisD*, *metA*, *argA*, *bioH*, *hisC*, *panC*, *hisB*, *hisA*, *tyrA*, *trpD*, *argG*, *thrC*, *trpB*, *guaB*. They showed that most of these auxotrophic mutants exhibited higher fitness than the prototrophic wild type in

competition experiments. Interestingly, all these biosynthetic genes show positive GLASS-P scores (Figure 7.3.C). This suggests that although these losses may be beneficial under specific environmental conditions, *E. coli* rarely faces such conditions in the wild.

7.6 *glyS*, *acrR* and *marR*: three genes under divergent short-term selective pressures

With scores of -3.91 and 13.24, *acrR* and *glyS* are at opposite ends of the GLASS-P score spectrum (Figures 7.3.D, 7.3.E). The over-representation of deleterious polymorphisms in the former and their under-representation in the latter suggest that these genes face very different short-term selective pressures. In contrast to the short term, their positive GLASS-D scores (1.62 and 4.82 respectively) imply that they both remain under purifying selection in the long term.

glyS encodes the beta subunit of glycine-tRNA ligase—the enzyme responsible for binding glycine to its transfer RNA. It therefore plays a crucial role in the translation process. For this reason, we find several genes encoding components of tRNA ligases—such as *glyS*—among the highest GLASS-P scores in our dataset, reflecting a strong short-term purifying selection pressure on these genes.

Both the negative GLASS-P score and the high pseudogenization rate of *acrR* reflect the frequent loss of this gene among *E. coli* strains (Figure 7.3.A). As mentioned previously, *acrR* represses the AcrAB-TolC efflux pump involved in antibiotic resistance. Its loss in turn leads to overexpression of this pump and increases antibiotic resistance. Lisa Praski Alzrigat and colleagues identified *acrR* mutations in *E. coli* strains with increased resistance to ciprofloxacin (Praski Alzrigat et al. 2021). They experimentally showed that 21 of these mutations enhanced the minimum inhibitory concentration of *E. coli* strain K12. Of these 21 mutations, we found 19 in our 60,472 genomes. DCA predicts that most of them are highly deleterious (high DCA mutation scores in Figure 7.3.E). Interestingly, another gene also represses the AcrAB-TolC efflux pump: *marR*. This regulator also suffers several losses but not as frequent as *acrR* (higher GLASS-P score than *acrR*, lower pseudogenization rate, see Figure 7.3.A). Unlike *acrR*, which specifically represses the AcrAB-TolC efflux pump, *marR* also regulates other pathways. For this reason, inactivation of *acrR* and *marR* results in comparable increases in antibiotic resistance, but the loss of *marR* is expected to have more collateral deleterious effects. This finding is consistent with that of Lisa Praski Alzrigat and colleagues who identified a higher fitness cost associated with *marR* loss (Praski Alzrigat et al. 2021).

7.7 Discussion

7.7.1 Influence of essentiality and expression level on polymorphism and divergence

The question of whether essentiality or expression level determines the rate of gene evolution caused much controversy in the early 2000s (Hirsh et al. 2001; Pál et al. 2001; Jordan et al. 2002; Pál et al. 2003). Here, we confirm that the level of gene expression plays a major role on long evolutionary time scales. But we also show that gene essentiality is critical on shorter time scales.

This critical role of gene essentiality in the short term stems from the nature of the mutations that

play a role on this time scale: in the short term, natural selection targets highly deleterious mutations. First, because these mutations occur very frequently. For example, 13% of all possible amino-acid changes inactivate the *E. coli* beta-lactamase TEM-1 (Jacquier et al. 2013). Second, because highly deleterious mutations targeting an essential gene could severely impair cell growth in several environments. Natural selection should therefore be particularly efficient in eliminating them from the population. Third, because they induce important phenotypic changes and could therefore represent a source of adaptation to colonise new environmental niches (Orr 1998).

In contrast to the short term, highly deleterious mutations no longer play a role in the long term. Purifying selection, although not instantaneous, prevents most deleterious polymorphisms from increasing in frequency and reaching fixation. Therefore, mutations that do fix are either beneficial, neutral or slightly deleterious. Mildly deleterious mutations could represent a greater burden for highly expressed genes. Indeed, many different costs scale with the number of protein or mRNA copies produced (mRNA and protein misfolding, protein misinteractions, protein expression cost, etc.) (Zhang et al. 2015). Therefore, many mechanisms may contribute to making the level of gene expression one of the main drivers of long-term evolution. To evaluate the level of gene expression, we chose to use the Codon Adaptation Index (CAI). Indeed, it should better approximate the expression profile of *E. coli* genes in natural environments than laboratory measurements. However, CAI may lead us to overestimate the strength of the correlation between gene expression level and the rate at which genes fix mutations. This limitation remains minor, given that the novelty of our study lies in investigating the short-term role played by gene essentiality—the long-term role of gene expression being largely covered by the existing literature (Pál et al. 2001; Pál et al. 2003).

7.7.2 Benefits and limitations of GLASS

To investigate these very recent events of selection, we have introduced GLASS. This test is based on the predicted effects of amino-acid changes and is particularly powerful for studying the over- or under-representation of recent deleterious mutations segregating on genes. When considering gene loss of function, it detects the common deleterious impact of many different rare mutations. This opens up new opportunities to analyse rare variants with large effects that genotype-to-phenotype association studies fail to capture (Gibson 2012). This test also complements approaches based on allelic frequencies or on the comparison of synonymous and non-synonymous mutations. In contrast to the latter, our test takes into account the fact that the proportions of near-neutral and deleterious amino-acid mutations may vary from protein to protein. This can occur if one protein is more stable than another. Ignoring mutation frequencies also makes our test robust to complex demographic histories or biased genomic databases. This test is particularly well suited to the study of large databases, as its statistical power increases with the number of genomes used. GLASS is based on DCA scores that mainly capture the structural constraints of proteins. For this reason, the presented scores reflect the intensity of purifying selection acting on a gene. These scores are not designed to capture signals of diversifying selection or specific changes in protein function. Genes under these selection regimes are unlikely to emerge at either tail of the DCA-based GLASS score distribution.

GLASS uses prediction of the effect of amino-acid changes to detect selection events. In this study, we chose to use DCA to predict these effects as it has been shown to outperform independent-site approaches (Figliuzzi et al. 2016). However, GLASS can be adapted to any method that gives a quan-

titative score to mutations, typically SIFT (Ng et al. 2003) or scores based on Grantham's matrix of physicochemical distance (Grantham 1974) which are less computationally demanding than DCA (Couce et al. 2019). GLASS performance depends on the accuracy of the predictions of mutation effects. Sequencing errors counted as mutations could also decrease its performance. In either case, this would make the score distribution of observed mutations closer to the score distribution of simulated mutations. In other words, the GLASS scores would be closer to zero. This could lead to false negatives—cases where we are unable to detect the true signature of natural selection—but not false positives—cases where neutral events would be confused with natural selection. For simplicity, we chose a rudimentary model of neutral evolution (JC69). Future studies may favour a more elaborate version that better models *E. coli* mutational biases. However, this is unlikely to radically change the general trends we observe regarding gene essentiality and deleterious mutations on transcription factors (see Supplementary Note 4 in Appendix C). A final limitation of our test is that we should always control for the number of mutations before comparing GLASS scores. This is straightforward when working with a large number of genes, as is the case in this study. However, if one wants to compare only a few genes, we suggest resampling an equal number of mutations before calculating GLASS scores (see Supplementary Notes 3 in Appendix C).

7.7.3 Dynamics of gene inactivations

We observe that some genes—mainly transcription factors rather than biosynthetic genes—undergo several independent losses across *E. coli* strains. But these genes remain under purifying selection over longer time scales. The loss of a transcriptional regulator is an effective strategy for adapting to a new niche. First because it corresponds to a large mutational target. Stated differently, a much higher number of mutations can result in the loss of the gene rather than a change in function (Murray 2020). Second, because it has limited collateral deleterious effects, as an entire pathway can be turned off at once. The differential dynamic between short and long term highlights the importance of biological trade-offs. The loss or maintenance of *rpoS*—the general regulator of the stress response—illustrates the balance between growth and self-preservation (Ferenci 2005). Similarly, *in vitro* evolution under high antibiotic pressure often results in complete loss of *marR*, whereas clinical strains from antibiotic-treated patients tend to show less functional variants rather than complete losses (Praski Alzrigat et al. 2021). This suggests that a complex environment imposes trade-offs between different cellular functions.

The contrast between the maintenance of a gene in a species and its recurrent loss over short periods of time signals antagonistic pleiotropy: many genes allow *E. coli* to travel across niches but the exact same genes may become a burden in a specific niche (Murray 2020). Since evolution is short-sighted, they will suffer frequent losses. As most of these specific niches—typically other organs in extra-intestinal infections—are transient, the loss of these genes is detrimental in the long term. These transient niche colonisations thus follow a source-sink dynamic (Sokurenko et al. 2006). For instance, the loss of a gene during patho-adaptation leads to an evolutionary dead-end: either the strain manages to reverse this loss and returns to the gut, or it is evolutionary dead.

A notable exception to this dynamic of loss and reversion may occur if the secondary environment that the strain invades is not transient but stable. When a strain adapts to a stable ecological niche, some inactivations can propagate down the phylogeny. For example, O157:H7 *E. coli* strains cannot

metabolise rhamnose (Ratnam et al. 1988). This serotype lives in a particular niche as it mainly inhabits the intestines of cattle instead of humans (Bettelheim et al. 1974). In our dataset, hundreds of O157 and O157:H7 strains carry a A243T mutation in the *rhaS* gene —an activator of the L-rhamnose operon. A243T constitutes an example of a DCA-predicted deleterious mutation able to propagate on the phylogeny. Adaptive laboratory experiments also represent cases of colonisation of stable niches. There again, the loss of genes involved in sugar metabolism was reported *in vitro*, during the Long Term Evolution Experiment (LTEE), and *in vivo*, during evolution in the mouse gut ((Cooper et al. 2001), (Barroso-Batista et al. 2014a), (Lourenço et al. 2016), (Lescat et al. 2017), (Ghalayini et al. 2019)). In both cases, the constitutive activations or inactivations of the corresponding operons were shown to be beneficial.

7.7.4 Conclusion

In the present chapter, we introduced GLASS, a new approach to study recent selection events based on the effect of amino-acid changes. It allowed us to detect intense purifying selection acting on essential genes. It also detects recurrent inactivations occurring mainly on transcription factors. This shows the importance of regulatory changes for local adaptation to new niches. It also demonstrates the importance of examining low-frequency polymorphisms to capture short-term selection dynamics that differ strongly from long-term dynamics. GLASS opens up new possibilities to study local adaptations for a wide range of species. It can also be used to detect mutations that induce differential costs in different environments, with potential applications to the study of resistance to phages or to antibiotics. GLASS could also be applied to human genetics. For example, it could be used to revisit the notion of gene essentiality, which cannot be tested experimentally beyond the cellular level due to ethical constraints. We therefore expect that genes expressed early in development will have very high GLASS scores. GLASS could also be an alternative to genome-wide association studies (GWAS) to study rare variants with detrimental effects. To summarise, approaches that combine the prediction of the effect of mutations with the study of polymorphism within a species could be powerful for studying recent evolutionary history.

Chapter 8

Concluding remarks

In this manuscript, I have conducted a study on the evolution of *Escherichia coli* by analyzing the extensive diversity present in over 80,000 genomes of *E. coli* and *Shigella*. The advantage of working on such a large scale is that it provides access to rare events and enables us to obtain a more comprehensive understanding of the natural diversity within this species. However, it is important to note that these large samples tend to have a bias towards human isolates and clinical strains.

I have organised over 400 million coding sequences obtained from 81,440 genomes into a SQL database. To enhance their analysis, I annotated these sequences by comparing them with sequences in the Swiss-Prot database. Subsequently, I performed clustering of genomes based on their persistent genomes, resulting in a total of 240 genome clusters, each comprising at least 5 genomes, along with 597 genomes that were not assigned to any cluster. As a complementary step, I inferred phylogenies for each cluster as well as a comprehensive phylogeny encompassing all clusters. Notably, all phylogenies were carefully adjusted to account for the influence of recombination.

One of the main objectives of my project was to explore how the diversity observed across distantly related species could shed light on the diversity observed within *E. coli*. To achieve this, I employed two different modeling approaches: an independent sites approach (IND) and Direct-Coupling Analysis (DCA). Both IND and DCA were trained on datasets consisting of distant homologs of persistent proteins found in *E. coli*.

IND focuses on identifying patterns of amino-acid conservation at specific sites, whereas DCA goes a step further by also detecting co-evolution between pairs of sites. This enables DCA to make predictions that account for the genetic context. Notably, DCA outperforms IND in predicting both the amino acids native to *E. coli* and the polymorphisms observed within this species. As a result, I employed DCA for further analysis and demonstrated its ability to accurately model the probability of observing a polymorphism within a given frequency range.

The better performance of DCA comes from the fact that it can capture the genetic context to predict mutation effect. Accumulating evidence suggests that the genetic context has been established over extended periods of evolution through the accumulation of numerous small epistatic couplings.

By examining single mutations, we can readily transition to analysing the variability of amino-acid sites. The combination of IND or DCA with the concept of Shannon entropy enables us to estimate the variability of a specific amino-acid site.

Upon comparing the rates of evolution across various amino-acid sites, I shifted my focus to comparing the rates of evolution among different genes. To accomplish this, I developed a selection test

called the Gene-Level Amino-acid Score Shift (GLASS), which relies on the predicted effects of amino-acid changes. By comparing the distribution of mutation effects observed in a gene to the expected distribution in the absence of selection, GLASS quantifies the strength of selection.

I applied GLASS to a dataset comprising 60,472 *E. coli* strains, thus allowing to re-examine the longstanding debate regarding the influence of essentiality versus expression level on the rate of protein evolution. Essential genes experience strong purifying selection in the short-term, while the rate of gene evolution is primarily determined by expression level over the long term.

GLASS also identified an overrepresentation of inactivating mutations in specific transcription factors, including efflux pump repressors, which aligns with selection for antibiotic resistance.

My research shows that patterns of variability can be examined across a wide range of scales. By using DCA, I was able to investigate individual mutations and their frequencies. Additionally, I explored the variability of amino-acid sites. The introduction of GLASS provided a means to compare the rate of evolution among different genes. Furthermore, I have also clustered genomes by similarity, which allowed me to compare the variability in gene repertoire of different genome clusters. This comprehensive approach enabled a thorough analysis of variability at different levels, from individual mutations to genome-scale diversity.

Throughout this work, we consistently observe contrasting dynamics between short and long time scales. The mutational landscape exhibits local smoothness—mutations having similar effects across various strains of *E. coli*. However, globally, the landscape is rugged, with up to 50% of amino-acid sites where mutation effect depends on the specific genetic context of the species. Interestingly, around 30% of these sites cannot tolerate polymorphisms in *E. coli* but exhibit variability across distantly related species. At the gene level, our analysis using GLASS reveals that essential genes are the primary targets of selection in the short term, while the long-term rate of gene evolution is driven by expression levels. When constructing genome phylogenies, we also observed stark differences in recombination patterns between short and long-term scales. Within clusters, the signature of recombination is primarily concentrated around three distinct regions. However, when examining recombination between clusters, it appears to be more dispersed throughout the chromosome.

These findings highlight the complex and dynamic nature of evolution, where different scales of analysis reveal varying dynamics and patterns.

A significant aspect of this study revolves around exploring the adaptive role of gene inactivations—with the aid of GLASS, which detects such occurrences. Once again, we observe contrasting dynamics in the short and long term. These inactivations have emerged independently in multiple strains, suggesting short-term fitness gains. However, they did not reach fixation, indicating long-term counter-selection. It is important to note that our research primarily focused on persistent genes, which are likely needed at some point, thereby suggesting the presence of long-term counter-selection. Future investigations could shift their attention to studying the dynamics of gene loss and acquisition in accessory genes. Our database can prove instrumental in this regard, as it provides a phylogeny of strains, corrected for recombination, along with a table identifying potential pseudogenes and linking them to the respective strains in which they were identified. An intriguing case study could involve examining the dynamics of gene losses during the transition from a generalist and commensal *E. coli* strain to a primate-restricted and intracellular *Shigella* pathogen.

Another potential application of the database lies in conducting a more quantitative analysis of core and pan-genome dynamics, beyond the scope of the present manuscript. This could involve

comparing the behavior of different clusters. One hypothesis would be that in populations with lower effective sizes, natural selection may not suffice to retain certain beneficial genes. Consequently, clusters with the smallest effective population sizes are expected to exhibit smaller core and persistent genomes. Linking this analysis with the frequency of observed deleterious mutations segregating within these clusters would also be valuable, as both phenomena are likely influenced by effective population sizes. Another question to address when comparing cluster core and persistent genomes is the presence of cluster's specific genes.

By examining pan-genomes, the database can offer deeper insights into horizontal gene transfer events, addressing questions such as: Are genes acquired from more distantly related species less likely to be retained in *E. coli*? Are there specific gene categories, like those conferring antibiotic resistance, that are more prone to acquisition or retention during evolution? Additionally, focusing on genomic islands like the High Pathogenicity Island (HPI) would allow us to investigate insertion sites and determine the number of independent acquisition events from other species. Comparing these rare instances of foreign source acquisition with the dynamics of homologous recombination that facilitate their spread across *E. coli* strains holds considerable promise.

By exploring these subjects, we can enhance our understanding of core and pan-genome dynamics, shed light on the interplay between effective population sizes and gene retention, and gain insights into horizontal gene transfer events and the factors influencing their success within *E. coli*.

Our work used a modeling approach, DCA, to interpret the mutation patterns we observed in *E. coli*. The remarkable performance of DCA in terms of predicting polymorphisms leads us to believe that it has potential applications in other areas. In particular, it has been proposed that it could be used to simulate neutral evolution by accounting for epistasis (Paz et al. 2020). However, the model of evolution proposed by Paz et al. seems somewhat simplistic as it does not consider the underlying DNA sequence but directly simulates evolution at the protein level. In other words, it assumes that all 19 possible amino-acid mutations are equally likely to occur at a given site, disregarding the fact that mutations requiring only one single nucleotide change (SNP) are much more likely to occur by chance alone. Building upon this work, we can develop a more realistic model of neutral evolution that takes into account these probabilities and utilize it to simulate evolution. Such an approach would enable us to gain a deeper understanding of the role of epistasis in shaping the patterns of diversity observed across various time scales.

The abundance of available genomes, the recurrent emergence of pathological clones, and the propensity of *E. coli* to adapt to various environments collectively position it as an ideal species for studying evolution. To contribute to the understanding of this species, I have constructed a large database comprising 81,440 genomes of *E. coli* and *Shigella* and I have trained thousands of DCA models to model the effect of mutations in the main proteins found in this species. Beyond my PhD project, these resources will be accessible for future research in the field.

Bibliography

Abram, Kaleb, Zulema Udaondo, Carissa Bleker, Visanu Wanchai, Trudy M. Wassenaar, Michael S. Robeson, and David W. Ussery (2021). "Mash-based analyses of *Escherichia coli* genomes reveal 14 distinct phylogroups". en. In: *Communications Biology* 4.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–12. ISSN: 2399-3642. DOI: 10.1038/s42003-020-01626-5.

Adzhubei, Ivan, Daniel M Jordan, and Shamil R Sunyaev (2013). "Predicting functional effect of human missense mutations using PolyPhen-2". In: *Current protocols in human genetics* 76.1, pp. 7–20.

Andremont, Antoine et al. (2019). "Ceftriaxone and Cefotaxime Have Similar Effects on the Intestinal Microbiota in Human Volunteers Treated by Standard-Dose Regimens". In: *Antimicrob. Agents Chemother.*

Azevedo, M., A. Sousa, J. Moura de Sousa, J. A. Thompson, J. T. Proença, and I. Gordo (2016). "Trade-Offs of *Escherichia coli* Adaptation to an Intracellular Lifestyle in Macrophages". en. In: *PLOS ONE* 11.1. Publisher: Public Library of Science, e0146123. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0146123.

Backhed, Fredrik, Ruth E Ley, Justin L Sonnenburg, Daniel A Peterson, and Jeffrey I Gordon (2005). "Host-bacterial mutualism in the human intestine". In: *science* 307.5717, pp. 1915–1920.

Balbi, Kevin J., Eduardo P.C. Rocha, and Edward J. Feil (2009). "The Temporal Dynamics of Slightly Deleterious Mutations in *Escherichia coli* and *Shigella* spp." In: *Molecular Biology and Evolution* 26.2, pp. 345–355. ISSN: 0737-4038. DOI: 10.1093/molbev/msn252.

Barreto, Hugo C., Nelson Frazão, Ana Sousa, Anke Konrad, and Isabel Gordo (2020a). "Mutation accumulation and horizontal gene transfer in *Escherichia coli* colonizing the gut of old mice". In: *Communicative & Integrative Biology* 13.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19420889.2020.1783059, pp. 89–96. ISSN: null. DOI: 10.1080/19420889.2020.1783059.

Barreto, Hugo C., Ana Sousa, and Isabel Gordo (2020b). "The Landscape of Adaptive Evolution of a Gut Commensal Bacteria in Aging Mice". en. In: *Current Biology* 30.6, 1102–1109.e5. ISSN: 0960-9822. DOI: 10.1016/j.cub.2020.01.037.

Barroso-Batista, João, Ana Sousa, Marta Lourenço, Marie-Louise Bergman, Daniel Sobral, Jocelyne Demengeot, Karina B. Xavier, and Isabel Gordo (2014a). "The First Steps of Adap-

tation of *Escherichia coli* to the Gut Are Dominated by Soft Sweeps". In: *PLos Genet.* 10.3, e1004182. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004182.

Barroso-Batista, João, Jocelyne Demengeot, and Isabel Gordo (2015). "Adaptive immunity increases the pace and predictability of evolutionary change in commensal gut bacteria". en. In: *Nature Communications* 6.1. Number: 1 Publisher: Nature Publishing Group, p. 8945. ISSN: 2041-1723. DOI: 10.1038/ncomms9945.

Barroso-Batista, João, Miguel F. Pedro, Joana Sales-Dias, Catarina J. G. Pinto, Jessica A. Thompson, Helena Pereira, Jocelyne Demengeot, Isabel Gordo, and Karina B. Xavier (2020). "Specific Eco-evolutionary Contexts in the Mouse Gut Reveal *Escherichia coli* Metabolic Versatility". en. In: *Current Biology* 30.6, 1049–1062.e7. ISSN: 0960-9822. DOI: 10.1016/j.cub.2020.01.050.

Barroso-Batista, João, Ana Sousa, Marta Lourenço, Marie-Louise Bergman, Daniel Sobral, Jocelyne Demengeot, Karina B. Xavier, and Isabel Gordo (2014b). "The First Steps of Adaptation of *Escherichia coli* to the Gut Are Dominated by Soft Sweeps". en. In: *PLoS Genetics* 10.3. Ed. by Graham Coop, e1004182. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004182.

Bateman, Alex et al. (2004). "The Pfam protein families database". In: *Nucleic Acids Res.* 32.suppl_1, pp. D138–D141. ISSN: 0305-1048. DOI: 10.1093/nar/gkh121.

Bauer, Fred, Christian Hertel, and Walter P. Hammes (1999). "Transformation of *Escherichia coli* in Foodstuffs". en. In: *Systematic and Applied Microbiology* 22.2, pp. 161–168. ISSN: 0723-2020. DOI: 10.1016/S0723-2020(99)80061-7.

Baur, B, K Hanselmann, W Schlimme, and B Jenni (1996). "Genetic transformation in freshwater: *Escherichia coli* is able to develop natural competence". In: *Applied and Environmental Microbiology* 62.10. Publisher: American Society for Microbiology, pp. 3673–3678. DOI: 10.1128/aem.62.10.3673-3678.1996.

Beghain, Johann, Antoine Bridier-Nahmias, Hervé Le Nagard, Erick Denamur, and Olivier Clermont (2018). "ClermonTyping: an easy-to-use and accurate *in silico* method for Escherichia genus strain phylotyping". In: *Microb. Genomics* 4.7. DOI: 10.1099/mgen.0.000192.

Beletskii, A. and Ashok S. Bhagwat (1996). "Transcription-induced mutations: Increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*". In: *Proceedings of the National Academy of Sciences* 93.24. Publisher: Proceedings of the National Academy of Sciences, pp. 13919–13924. DOI: 10.1073/pnas.93.24.13919.

Bell, Jason C. and Stephen C. Kowalczykowski (2016). "RecA: Regulation and Mechanism of a Molecular Search Engine". en. In: *Trends in Biochemical Sciences* 41.6, pp. 491–507. ISSN: 09680004. DOI: 10.1016/j.tibs.2016.04.002.

Bensted, H. J. (1956). "Dysentery bacilli—shigella: a brief historical review". In: *Canadian Journal of Microbiology* 2.3. Publisher: NRC Research Press, pp. 163–174. ISSN: 0008-4166. DOI: 10.1139/m56-022.

Bergthorsson, U and H Ochman (1998). "Distribution of chromosome length variation in natural isolates of *Escherichia coli*." In: *Molecular Biology and Evolution* 15.1, pp. 6–16. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a025847.

Bettelheim, K. A., F. M. Bushrod, M. E. Chandler, E. M. Cooke, S. O'Farrell, and R. A. Shooter (1974). "*Escherichia coli* serotype distribution in man and animals". In: *J. Hyg.* 73.3, pp. 467–471. ISSN: 0022-1724. eprint: 4613754.

Bezabih, Yihienew M., Wilber Sabiiti, Endalkachew Alamneh, Alamneh Bezabih, Gregory M. Peterson, Woldesellassie M. Bezabhe, and Anna Roujeinikova (2021). "The global prevalence and trend of human intestinal carriage of ESBL-producing *Escherichia coli* in the community". In: *J. Antimicrob. Chemother.* 76.1, pp. 22–29. ISSN: 0305-7453. DOI: 10.1093/jac/dkaa399.

Blair, Jessica M. A., Mark A. Webber, Alison J. Baylay, David O. Ogbolu, and Laura J. V. Piddock (2015). "Molecular mechanisms of antibiotic resistance". In: *Nat. Rev. Microbiol.* 13, pp. 42–51. ISSN: 1740-1534. DOI: 10.1038/nrmicro3380.

Bleibtreu, Alexandre, Olivier Clermont, Pierre Darlu, Jérémy Glodt, Catherine Branger, Bertrand Picard, and Erick Denamur (2014). "The rpoS gene is predominantly inactivated during laboratory storage and undergoes source-sink evolution in *Escherichia coli* species". In: *Journal of bacteriology* 196.24, pp. 4276–4284.

Bobay, Louis-Marie and Howard Ochman (2017a). "Impact of Recombination on the Base Composition of Bacteria and Archaea". In: *Mol. Biol. Evol.* 34.10, pp. 2627–2636. ISSN: 0737-4038. DOI: 10.1093/molbev/msx189.

— (2017b). "The Evolution of Bacterial Genome Architecture". In: *Frontiers in Genetics* 8. ISSN: 1664-8021.

— (2018). "Factors driving effective population size and pan-genome evolution in bacteria". en. In: *BMC Evolutionary Biology* 18.1, p. 153. ISSN: 1471-2148. DOI: 10.1186/s12862-018-1272-4.

Boutet, Emmanuel, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J. Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios (2016). "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View". In: *Plant Bioinformatics*. New York, NY, USA: Springer, pp. 23–54. DOI: 10.1007/978-1-4939-3167-5_2.

Breen, Michael S., Carsten Kemena, Peter K. Vlasov, Cedric Notredame, and Fyodor A. Kondrashov (2012). "Epistasis as the primary factor in molecular evolution". In: *Nature* 490, pp. 535–538. ISSN: 1476-4687. DOI: 10.1038/nature11510.

Brenner, Don J., George R. Fanning, Karl E. Johnson, R. V. Citarella, and Stanley Falkow (1969). "Polynucleotide Sequence Relationships among Members of Enterobacteriaceae". In: *Journal of Bacteriology* 98.2. Publisher: American Society for Microbiology, pp. 637–650. DOI: 10.1128/jb.98.2.637-650.1969.

Bridier-Nahmias, Antoine et al. (2021). "*Escherichia coli* Genomic Diversity within Extraintestinal Acute Infections Argues for Adaptive Evolution at Play". In: *MSphere* 6.1, e01176–20.

Budescu, David V (1993). "Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression." In: *Psychological bulletin* 114.3, p. 542.

Bull, Associate Editor: Jim (2000). "Long-Term Experimental Evolution in *Escherichia coli*. VIII. Dynamics of a Balanced Polymorphism". In: *Am. Nat.*

Chang, Dong-Eun et al. (2004). "Carbon nutrition of *Escherichia coli* in the mouse intestine". In: *Proceedings of the National Academy of Sciences* 101.19. Publisher: Proceedings of the National Academy of Sciences, pp. 7427–7432. DOI: 10.1073/pnas.0307888101.

Clermont, Olivier, Stéphane Bonacorsi, and Edouard Bingen (2000). "Rapid and Simple Determination of the *Escherichia coli* Phylogenetic Group". In: *Applied and Environmental Microbiology* 66.10. Publisher: American Society for Microbiology, pp. 4555–4558. DOI: 10.1128/AEM.66.10.4555-4558.2000.

Clermont, Olivier, Julia K. Christenson, Erick Denamur, and David M. Gordon (2013). "The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups: A new *E. coli* phylo-typing method". en. In: *Environmental Microbiology Reports* 5.1, pp. 58–65. ISSN: 17582229. DOI: 10.1111/1758-2229.12019.

Clermont, Olivier, Bénédicte Condamine, Sara Dion, David M. Gordon, and Erick Denamur (2021). "The E phylogroup of *Escherichia coli* is highly diverse and mimics the whole *E. coli* species population structure". en. In: *Environmental Microbiology* 23.11, pp. 7139–7151. ISSN: 1462-2920. DOI: 10.1111/1462-2920.15742.

Clermont, Olivier, Ojas V. A. Dixit, Belinda Vangchhia, Bénédicte Condamine, Sara Dion, Antoine Bridier-Nahmias, Erick Denamur, and David Gordon (2019). "Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential". en. In: *Environmental Microbiology* 21.8, pp. 3107–3117. ISSN: 1462-2920. DOI: 10.1111/1462-2920.14713.

Clermont, Olivier, Mathilde Lescat, Claire L. O'Brien, David M. Gordon, Olivier Tenaillon, and Erick Denamur (2008). "Evidence for a human-specific *Escherichia coli* clone". In: *Environ. Microbiol.* 10.4, pp. 1000–1006. ISSN: 1462-2912. DOI: 10.1111/j.1462-2920.2007.01520.x.

Cobo-Simón, Marta, Rowan Hart, and Howard Ochman (2023). "*Escherichia Coli*: What Is and Which Are?" In: *Molecular Biology and Evolution* 40.1, msac273. ISSN: 1537-1719. DOI: 10.1093/molbev/msac273.

Cole, Stewart T and Olivier Raibaud (1986). "The nucleotide sequence of the malT gene encoding the positive regulator of the *Escherichia coli* maltose regulon". In: *Gene* 42.2, pp. 201–208.

Cooper, Vaughn S., Dominique Schneider, Michel Blot, and Richard E. Lenski (2001). "Mechanisms Causing Rapid and Parallel Losses of Ribose Catabolism in Evolving Populations of *Escherichia coli* B". In: *J. Bacteriol.*

Couce, Alejandro, Larissa Viraphong Caudwell, Christoph Feinauer, Thomas Hindré, Jean-Paul Feugeas, Martin Weigt, Richard E Lenski, Dominique Schneider, and Olivier Tenaillon (2017). "Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria". In: *Proceedings of the National Academy of Sciences* 114.43, E9026–E9035.

Couce, Alejandro and Olivier Tenaillon (2019). "Mutation bias and GC content shape antimutator invasions". In: *Nat. Commun.* 10.3114, pp. 1–9. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11217-6.

Couce, Alejandro and Olivier A. Tenaillon (2015). "The rule of declining adaptability in microbial evolution experiments". In: *Frontiers in Genetics* 6. ISSN: 1664-8021.

Croucher, Nicholas J., Andrew J. Page, Thomas R. Connor, Aidan J. Delaney, Jacqueline A. Keane, Stephen D. Bentley, Julian Parkhill, and Simon R. Harris (2015). "Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins". In: *Nucleic Acids Res.* 43.3, e15. ISSN: 0305-1048. DOI: 10.1093/nar/gku1196.

Delmas, Stéphane and Ivan Matic (2005). "Cellular response to horizontally transferred DNA in *Escherichia coli* is tuned by DNA repair systems". en. In: *DNA Repair* 4.2, pp. 221–229. ISSN: 15687864. DOI: 10.1016/j.dnarep.2004.09.008.

Denamur, Erick, Olivier Clermont, Stéphane Bonacorsi, and David Gordon (2021). "The population genetics of pathogenic *Escherichia coli*". en. In: *Nature Reviews Microbiology* 19.1, pp. 37–54. ISSN: 1740-1534. DOI: 10.1038/s41579-020-0416-x.

Denamur, Erick and Ivan Matic (2006). "Evolution of mutation rates in bacteria". en. In: *Molecular Microbiology* 60.4, pp. 820–827. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2006.05150.x.

Denamur, Erick et al. (2000). "Evolutionary Implications of the Frequent Horizontal Transfer of Mismatch Repair Genes". en. In: *Cell* 103.5, pp. 711–721. ISSN: 00928674. DOI: 10.1016/S0092-8674(00)00175-6.

Didelot, Xavier, Guillaume Méric, Daniel Falush, and Aaron E. Darling (2012). "Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*". en. In: *BMC Genomics* 13.1, p. 256. ISSN: 1471-2164. DOI: 10.1186/1471-2164-13-256.

Dixit, Purushottam D, Tin Yau Pang, and Sergei Maslov (2017). "Recombination-Driven Genome Evolution and Stability of Bacterial Species". In: *Genetics* 207.1, pp. 281–295. ISSN: 1943-2631. DOI: 10.1534/genetics.117.300061.

Drake, John W, Brian Charlesworth, Deborah Charlesworth, and James F Crow (1998). "Rates of Spontaneous Mutation". In: *Genetics* 148.4, pp. 1667–1686. ISSN: 1943-2631. DOI: 10.1093/genetics/148.4.1667.

D'Souza, Glen, Silvio Waschina, Samay Pande, Katrin Bohl, Christoph Kaleta, and Christian Kost (2014). "Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria". In: *Evolution* 68.9, pp. 2559–2570.

Duriez, Patrick, Olivier Clermont, Stéphane Bonacorsi, Edouard Bingen, André Chaventré, Jacques Elion, Bertrand Picard, and Erick Denamur (2001). "Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations". In: *Microbiology* 147.6. Publisher: Microbiology Society, pp. 1671–1676. ISSN: 1465-2080. DOI: 10.1099/00221287-147-6-1671.

EClinicalMedicine (2021). "Antimicrobial resistance: a top ten global public health threat". English. In: *eClinicalMedicine* 41. Publisher: Elsevier. ISSN: 2589-5370. DOI: 10.1016/j.eclinm.2021.101221.

Enright, Mark C and Brian G Spratt (1999). "Multilocus sequence typing". en. In: 7.12.

Escobar-Páramo, Patricia, Olivier Clermont, Anne-Béatrice Blanc-Potard, Hung Bui, Chantal Le Bouguénec, and Erick Denamur (2004). "A Specific Genetic Background Is Required for Acquisition and Expression of Virulence Factors in *Escherichia coli*". In: *Molecular Biology and Evolution* 21.6, pp. 1085–1094. ISSN: 0737-4038. DOI: 10.1093/molbev/msh118.

Espeli, Olivier, Romain Mercier, and Frédéric Boccard (2008). "DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome". In: *Mol. Microbiol.* 68.6, pp. 1418–1427. ISSN: 0950-382X. DOI: 10.1111/j.1365-2958.2008.06239.x.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *kdd*. Vol. 96. 34, pp. 226–231.

Ewing, W. H. (1953). "Serological Relationships Between *Shigella* And Coliform Cultures". en. In: *Journal of Bacteriology* 66.3, pp. 333–340. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/jb.66.3.333-340.1953.

Feil, Edward J. and Brian G. Spratt (2001). "Recombination and the Population Structures of Bacterial Pathogens". In: *Annual Review of Microbiology* 55.1, pp. 561–590. ISSN: 0066-4227. DOI: 10.1146/annurev.micro.55.1.561.

Ferenci, Thomas (2005). "Maintaining a healthy SPANC balance through regulatory and mutational adaptation". In: *Molecular microbiology* 57.1, pp. 1–8.

Figliuzzi, Matteo, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt (2016). "Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1". In: *Mol. Biol. Evol.* 33.1, pp. 268–280. ISSN: 0737-4038. DOI: 10.1093/molbev/msv211.

Flamary, Rémi et al. (2021). "POT: Python Optimal Transport". In: *Journal of Machine Learning Research* 22.78, pp. 1–8.

Fleischmann, Robert D. et al. (1995). "Whole-Genome Random Sequencing and Assembly of Haemophilus influenzae Rd". In: *Science* 269.5223, pp. 496–512. ISSN: 0036-8075. DOI: 10.1126/science.7542800.

Foster-Nyarko, Ebenezer and Mark J Pallen (2022). "The microbial ecology of *Escherichia coli* in the vertebrate gut". In: *FEMS Microbiology Reviews* 46.3, fuac008. ISSN: 0168-6445. DOI: 10.1093/femsre/fuac008.

Fraser, Christophe, William P. Hanage, and Brian G. Spratt (2007). "Recombination and the Nature of Bacterial Speciation". In: *Science* 315.5811. Publisher: American Association for the Advancement of Science, pp. 476–480. DOI: 10.1126/science.1127573.

Frazão, N., A. Konrad, M. Amicone, E. Seixas, D. Güleresi, M. Lässig, and I. Gordo (2022). "Two modes of evolution shape bacterial strain diversity in the mammalian gut for thousands of generations". en. In: *Nature Communications* 13.1. Number: 1 Publisher: Nature Publishing Group, p. 5604. ISSN: 2041-1723. DOI: 10.1038/s41467-022-33412-8.

Frazão, Nelson, Ana Sousa, Michael Lässig, and Isabel Gordo (2019). "Horizontal gene transfer overrides mutation in *Escherichia coli* colonizing the mammalian gut". In: *Proceedings of the National Academy of Sciences* 116.36. Publisher: Proceedings of the National Academy of Sciences, pp. 17906–17915. DOI: 10.1073/pnas.1906958116.

Friedmann, Herbert C. (2014). "Escherich and Escherichia". In: *EcoSal Plus* 6.1. Publisher: American Society for Microbiology. DOI: 10.1128/ecosalplus.ESP-0025-2013.

Ghalayini, Mohamed, Melanie Magnan, Sara Dion, Ouassila Zatout, Lucie Bourguignon, Olivier Tenaillon, and Mathilde Lescat (2019). "Long-term evolution of the natural isolate of *Escherichia coli* 536 in the mouse gut colonized after maternal transmission reveals convergence in the constitutive expression of the lactose operon". In: *Mol. Ecol.* 28.19, pp. 4470–4485. ISSN: 0962-1083. DOI: 10.1111/mec.15232.

Gibson, Greg (2012). "Rare and common variants: twenty arguments". In: *Nature Reviews Genetics* 13.2, pp. 135–145.

Gordo, Isabel, Jocelyne Demengeot, and Karina Xavier (2014). "*Escherichia coli* adaptation to the gut environment: a constant fight for survival". In: *Future Microbiology* 9.11. Publisher: Future Medicine, pp. 1235–1238. ISSN: 1746-0913. DOI: 10.2217/fmb.14.86.

Gordon, David M. and Ann Cowling (2003). "The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects". en. In: *Microbiology* 149.12, pp. 3575–3586. ISSN: 1350-0872, 1465-2080. DOI: 10.1099/mic.0.26486-0.

Grantham, R. (1974). "Amino Acid Difference Formula to Help Explain Protein Evolution". In: *Science* 185.4154, pp. 862–864. ISSN: 0036-8075. DOI: 10.1126/science.185.4154.862.

Grindley, Nigel D.F., Katrine L. Whiteson, and Phoebe A. Rice (2006). "Mechanisms of Site-Specific Recombination". In: *Annual Review of Biochemistry* 75.1, pp. 567–605. DOI: 10.1146/annurev.biochem.73.011303.073908.

Haudiquet, Matthieu, Amandine Buffet, Olaya Rendueles, and Eduardo P. C. Rocha (2021). "Interplay between the cell envelope and mobile genetic elements shapes gene flow in populations of the nosocomial pathogen Klebsiella pneumoniae". en. In: *PLOS Biology* 19.7. Publisher: Public Library of Science, e3001276. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001276.

Hershberg, Ruth and Dmitri A. Petrov (2010). "Evidence That Mutation Is Universally Biased towards AT in Bacteria". In: *PLos Genet.* 6.9, e1001115. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001115.

Hershberg, Ruth, Hua Tang, and Dmitri A. Petrov (2007). "Reduced selection leads to accelerated gene loss in *Shigella*". In: *Genome Biol.* 8.8, pp. 1–11. ISSN: 1474-760X. DOI: 10.1186/gb-2007-8-8-r164.

Herzer, P J, S Inouye, M Inouye, and T S Whittam (1990). "Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*". In: *Journal of Bacteriology* 172.11. Publisher: American Society for Microbiology, pp. 6175–6181. DOI: 10.1128/jb.172.11.6175-6181.1990.

Hildebrand, Falk, Axel Meyer, and Adam Eyre-Walker (2010). "Evidence of Selection upon Genomic GC-Content in Bacteria". In: *PLos Genet.* 6.9, e1001107. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001107.

Hill, W. G. and Alan Robertson (1966). "The effect of linkage on limits to artificial selection". In: *Genetics Research* 8.3, pp. 269–294. ISSN: 1469-5073. DOI: 10.1017/S0016672300010156.

Hirsh, Aaron E and Hunter B Fraser (2001). "Protein dispensability and rate of evolution". In: *Nature* 411.6841, pp. 1046–1049.

Hobson, C. A. et al. (2022). "MiniBioReactor Array (MBRA) *in vitro* gut model: a reliable system to study microbiota-dependent response to antibiotic treatment". In: *JAC Antimicrob. Resist.* 4.4, dlac077. ISSN: 2632-1823. DOI: 10.1093/jacamr/dlac077.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC Bioinf.* 11.1, pp. 1–11. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-119.

Ikuta, Kevin S. et al. (2022). "Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019". English. In: *The Lancet* 400.10369. Publisher: Elsevier, pp. 2221–2248. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(22)02185-7.

J. Sargentini, Neil and Kendric C. Smith (1994). "DNA sequence analysis of γ-radiation (anoxic)-induced and spontaneous lacId mutations in *Escherichia coli* K-12". In: *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 309.2, pp. 147–163. ISSN: 0027-5107. DOI: 10.1016/0027-5107(94)90088-4.

Jacquier, Hervé et al. (2013). "Capturing the mutational landscape of the beta-lactamase TEM-1". In: *Proc. Natl. Acad. Sci. U.S.A.* 110.32, pp. 13067–13072. DOI: 10.1073/pnas.1215206110.

Jaureguy, Françoise et al. (2008). "Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains". In: *BMC Genomics* 9.1, p. 560. ISSN: 1471-2164. DOI: 10.1186/1471-2164-9-560.

Jin, Qi et al. (2002). “Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157”. In: *Nucleic Acids Res.* 30.20, pp. 4432–4441. ISSN: 0305-1048. DOI: 10.1093/nar/gkf566.

Johnson, James R., Parissa Delavari, Michael Kuskowski, and Adam L. Stell (2001). “Phylogenetic Distribution of Extraintestinal Virulence-Associated Traits in *Escherichia coli*”. In: *The Journal of Infectious Diseases* 183.1, pp. 78–88. ISSN: 0022-1899. DOI: 10.1086/317656.

Jordan, I King, Igor B Rogozin, Yuri I Wolf, and Eugene V Koonin (2002). “Essential genes are more evolutionarily conserved than are nonessential genes in bacteria”. In: *Genome research* 12.6, pp. 962–968.

Kallonen, Teemu, Hayley J. Brodrick, Simon R. Harris, Jukka Corander, Nicholas M. Brown, Veronique Martin, Sharon J. Peacock, and Julian Parkhill (2017). “Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131”. en. In: *Genome Research* 27.8. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1437–1449. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.216606.116.

Katouli, Mohammad (2010). “Population structure of gut *Escherichia coli* and its role in development of extra-intestinal infections”. In: *Iranian Journal of Microbiology* 2.2, pp. 59–72. ISSN: 2008-3289.

Kauffmann, F. (1947). “The Serology of the *Coli* Group”. en. In: *The Journal of Immunology* 57.1, pp. 71–100. ISSN: 0022-1767, 1550-6606. DOI: 10.4049/jimmunol.57.1.71.

Khalil, Ibrahim A. et al. (2018). “Morbidity and mortality due to *Shigella* and enterotoxigenic *Escherichia coli* diarrhoea: the Global Burden of Disease Study 1990–2016”. English. In: *The Lancet Infectious Diseases* 18.11. Publisher: Elsevier, pp. 1229–1240. ISSN: 1473-3099, 1474-4457. DOI: 10.1016/S1473-3099(18)30475-4.

Kimura, Motoo (1968). “Evolutionary rate at the molecular level”. In: *Nature* 217.5129, pp. 624–626.

King, Jack Lester and Thomas H. Jukes (1969). “Non-Darwinian Evolution”. In: *Science* 164.3881, pp. 788–798. ISSN: 0036-8075. DOI: 10.1126/science.164.3881.788.

Kisiela, Dagmara I et al. (2017). “Inactivation of transcriptional regulators during within-household evolution of *Escherichia coli*”. In: *Journal of bacteriology* 199.13, e00036–17.

Laine, Elodie, Yasaman Karami, and Alessandra Carbone (2019). “GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects”. In: *Mol. Biol. Evol.* 36.11, pp. 2604–2619. ISSN: 0737-4038. DOI: 10.1093/molbev/msz179.

Lan, Ruiting and Peter R. Reeves (2002). “*Escherichia coli* in disguise: molecular origins of *Shigella*”. en. In: *Microbes and Infection* 4.11, pp. 1125–1132. ISSN: 12864579. DOI: 10.1016/S1286-4579(02)01637-4.

Le Gall, Tony, Olivier Clermont, Stéphanie Gouriou, Bertrand Picard, Xavier Nassif, Erick Denamur, and Olivier Tenaillon (2007). "Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group *Escherichia coli* Strains". In: *Molecular Biology and Evolution* 24.11, pp. 2373–2384. ISSN: 0737-4038. DOI: 10.1093/molbev/msm172.

Lecointre, G., L. Rachdi, P. Darlu, and E. Denamur (1998). "*Escherichia coli* molecular phylogeny using the incongruence length difference test". en. In: *Molecular Biology and Evolution* 15.12, pp. 1685–1695. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/oxfordjournals.molbev.a025895.

Lehnen, D, C Blumer, T Polen, B Wackwitz, Volker F Wendisch, and G Unden (2002). "LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in *Escherichia coli*". In: *Molecular microbiology* 45.2, pp. 521–532.

Lenski, Richard E., Michael R. Rose, Suzanne C. Simpson, and Scott C. Tadler (1991). "Long-Term Experimental Evolution in *Escherichia coli*. I. Adaptation and Divergence During 2,000 Generations". In: *Am. Nat.* DOI: 10.1086/285289.

Lescat, Mathilde, Adrien Launay, Mohamed Ghalayini, Mélanie Magnan, Jérémy Glodt, Coralie Pintard, Sara Dion, Erick Denamur, and Olivier Tenaillon (2017). "Using long-term experimental evolution to uncover the patterns and determinants of molecular evolution of an *Escherichia coli* natural isolate in the streptomycin-treated mouse gut". In: *Mol. Ecol.* 26.7, pp. 1802–1817. ISSN: 0962-1083. DOI: 10.1111/mec.13851.

Lou, Yue Clare, Matthew R. Olm, Spencer Diamond, Alexander Crits-Christoph, Brian A. Firek, Robyn Baker, Michael J. Morowitz, and Jillian F. Banfield (2021). "Infant gut strain persistence is associated with maternal origin, phylogeny, and traits including surface adhesion and iron acquisition". en. In: *Cell Reports Medicine* 2.9, p. 100393. ISSN: 2666-3791. DOI: 10.1016/j.xcrm.2021.100393.

Lourenço, Marta, Ricardo S. Ramiro, Daniela Güleresi, João Barroso-Batista, Karina B. Xavier, Isabel Gordo, and Ana Sousa (2016). "A Mutational Hotspot and Strong Selection Contribute to the Order of Mutations Selected for during *Escherichia coli* Adaptation to the Gut". In: *PLos Genet.* 12.11, e1006420. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006420.

Lu, Shan et al. (2016). "Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of *Marmota himalayana* in Qinghai–Tibet plateau of China". In: *Emerging Microbes & Infections* 5.1. Publisher: Taylor & Francis _eprint: https://doi.org/10.1038/emi.2016.122, pp. 1–9. ISSN: null. DOI: 10.1038/emi.2016.122.

Luria, S. E. and Jeanne W. Burrous (1957). "Hybridization Between *Escherichia coli* And *Shigella*". en. In: *Journal of Bacteriology* 74.4, pp. 461–476. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/jb.74.4.461-476.1957.

Maddamsetti, Rohan, Richard E Lenski, and Jeffrey E Barrick (2015). "Adaptation, Clonal Interference, and Frequency-Dependent Interactions in a Long-Term Evolution Exper-

iment with *Escherichia coli*". In: *Genetics* 200.2, pp. 619–631. ISSN: 1943-2631. DOI: 10.1534/genetics.115.176677.

Mandel, M. and A. Higa (1970). "Calcium-dependent bacteriophage DNA infection". en. In: *Journal of Molecular Biology* 53.1, pp. 159–162. ISSN: 0022-2836. DOI: 10.1016/0022-2836(70)90051-3.

Martinson, Jonathan N. V. and Seth T. Walk (2020). "*Escherichia coli* Residency in the Gut of Healthy Human Adults". In: *EcoSal Plus* 9.1. Publisher: American Society for Microbiology. DOI: 10.1128/ecosalplus.ESP-0003-2020.

Mattock, Emily and Ariel J. Blocker (2017). "How Do the Virulence Factors of *Shigella* Work Together to Cause Disease?" In: *Front. Cell. Infect. Microbiol.* 7. ISSN: 2235-2988. DOI: 10.3389/fcimb.2017.00064.

Mayr, Ernst (1983). "How to Carry Out the Adaptationist Program?" In: *Am. Nat.* DOI: 10.1086/284064.

McCutcheon, John P. and Nancy A. Moran (2010). "Functional Convergence in Reduced Genomes of Bacterial Symbionts Spanning 200 My of Evolution". In: *Genome Biol. Evol.* 2, pp. 708–718. ISSN: 1759-6653. DOI: 10.1093/gbe/evq055.

McDonald, John H. and Martin Kreitman (1991). "Adaptive protein evolution at the Adh locus in Drosophila". In: *Nature* 351, pp. 652–654. ISSN: 1476-4687. DOI: 10.1038/351652a0.

McFadden, Daniel (1974). "Conditional logit analysis of qualitative choice behavior". In: *Frontiers in Econometrics*.

Milkman, R. (1997). "Recombination and Population Structure in *Escherichia Coli*". In: *Genetics* 146.3, pp. 745–750. ISSN: 0016-6731.

Milkman, R and M M Bridges (1993). "Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons." In: *Genetics* 133.3, pp. 455–468. ISSN: 1943-2631. DOI: 10.1093/genetics/133.3.455.

Mira, Alex, Howard Ochman, and Nancy A. Moran (2001). "Deletional bias and the evolution of bacterial genomes". en. In: *Trends in Genetics* 17.10, pp. 589–596. ISSN: 01689525. DOI: 10.1016/S0168-9525(01)02447-7.

Moran, Nancy A., Howard Ochman, and Tobin J. Hammer (2019). "Evolutionary and Ecological Consequences of Gut Microbial Communities". In: *Annual Review of Ecology, Evolution, and Systematics* 50.1, pp. 451–475. DOI: 10.1146/annurev-ecolsys-110617-062453.

Moran, Nancy A. and Gordon R. Plague (2004). "Genomic changes following host restriction in bacteria". In: *Curr. Opin. Genet. Dev.* 14.6, pp. 627–633. ISSN: 0959-437X. DOI: 10.1016/j.gde.2004.09.003.

Morcos, Faruck et al. (2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families". In: *Proc. Natl. Acad. Sci. U.S.A.* 108.49, E1293–E1301. DOI: 10.1073/pnas.1111471108.

Murray, Andrew W (2020). “Can gene-inactivating mutations lead to evolutionary novelty?” In: *Current Biology* 30.10, R465–R471.

Murray, Christopher JL et al. (2022). “Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis”. en. In: *The Lancet* 399.10325, pp. 629–655. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(21)02724-0.

Ng, Pauline C. and Steven Henikoff (2003). “SIFT: predicting amino acid changes that affect protein function”. In: *Nucleic Acids Res.* 31.13, pp. 3812–3814. ISSN: 0305-1048. DOI: 10.1093/nar/gkg509.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh (2015). “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies”. In: *Mol. Biol. Evol.* 32.1, pp. 268–274. ISSN: 0737-4038. DOI: 10.1093/molbev/msu300.

Nicolas-Chanoine, Marie-Hélène, Xavier Bertrand, and Jean-Yves Madec (2014). “*Escherichia coli* ST131, an Intriguing Clonal Group”. In: *Clinical Microbiology Reviews* 27.3. Publisher: American Society for Microbiology, pp. 543–574. DOI: 10.1128/CMR.00125-13.

Nowrouzian, Forough L., Agnes E. Wold, and Ingegerd Adlerberth (2005). “*Escherichia coli* Strains Belonging to Phylogenetic Group B2 Have Superior Capacity to Persist in the Intestinal Microflora of Infants”. In: *The Journal of Infectious Diseases* 191.7, pp. 1078–1083. ISSN: 0022-1899. DOI: 10.1086/427996.

Ochman, H and R K Selander (1984). “Standard reference strains of *Escherichia coli* from natural populations”. In: *Journal of Bacteriology* 157.2. Publisher: American Society for Microbiology, pp. 690–693. DOI: 10.1128/jb.157.2.690-693.1984.

Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman (2000). “Lateral gene transfer and the nature of bacterial innovation”. en. In: *Nature* 405.6784. Number: 6784 Publisher: Nature Publishing Group, pp. 299–304. ISSN: 1476-4687. DOI: 10.1038/35012500.

Ochman, Howard, Thomas S. Whittam, Dominique A. Caugant, and Robert K. Selander (1983). “Enzyme Polymorphism and Genetic Population Structure in *Escherichia coli* and *Shigella*”. In: *Microbiology* 129.9. Publisher: Microbiology Society, pp. 2715–2726. ISSN: 1465-2080. DOI: 10.1099/00221287-129-9-2715.

Ohama, Takeshi, Akira Muto, and Syozo Osawa (1990). “Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus a bacterium with a high genomic GC-content”. In: *Nucleic Acids Res.* 18.6, pp. 1565–1569. ISSN: 0305-1048. DOI: 10.1093/nar/18.6.1565.

Ohta, Tomoko (1973). “Slightly deleterious mutant substitutions in evolution”. In: *Nature* 246, pp. 96–98.

Ohta, Tomoko and John H. Gillespie (1996). “Development of Neutral and Nearly Neutral Theories”. In: *Theor. Popul. Biol.* 49.2, pp. 128–142. ISSN: 0040-5809. DOI: 10.1006/tpbi.1996.0007.

Oliveira, Pedro H., Marie Touchon, Jean Cury, and Eduardo P. C. Rocha (2017). "The chromosomal organization of horizontal gene transfer in bacteria". en. In: *Nature Communications* 8.1, p. 841. ISSN: 2041-1723. DOI: 10.1038/s41467-017-00808-w.

Ondov, Brian D, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy (2016). "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome biology* 17.1, pp. 1–14.

Orr, H Allen (1998). "The population genetics of adaptation: the distribution of factors fixed during adaptive evolution". In: *Evolution* 52.4, pp. 935–949.

Pál, Csaba, Balázs Papp, and Laurence D. Hurst (2001). "Highly Expressed Genes in Yeast Evolve Slowly". In: *Genetics* 158.2, pp. 927–931. ISSN: 1943-2631. DOI: 10.1093/genetics/158.2.927.

Pál, Csaba, Balázs Papp, and Laurence D Hurst (2003). "Rate of evolution and gene dispensability". In: *Nature* 421.6922, pp. 496–497.

Pasqua, Martina, Valeria Michelacci, Maria Letizia Di Martino, Rosangela Tozzoli, Milena Grossi, Bianca Colonna, Stefano Morabito, and Gianni Prosseda (2017). "The Intriguing Evolutionary Journey of Enteroinvasive *E. coli* (EIEC) toward Pathogenicity". In: *Front. Microbiol.* 8. ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02390.

Passel, Mark W. J. van, Pradeep Reddy Marri, and Howard Ochman (2008). "The Emergence and Fate of Horizontally Acquired Genes in *Escherichia coli*". en. In: *PLOS Computational Biology* 4.4. Publisher: Public Library of Science, e1000059. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000059.

Paz, Jose Alberto de la, Charisse M. Nartey, Monisha Yuvaraj, and Faruck Morcos (2020). "Epistatic contributions promote the unification of incompatible models of neutral molecular evolution". In: *Proc. Natl. Acad. Sci. U.S.A.* 117.11, pp. 5873–5882. DOI: 10.1073/pnas.1913071117.

Petrov, Dmitri A. (2002). "Mutational Equilibrium Model of Genome Size Evolution". In: *Theor. Popul. Biol.* 61.4, pp. 531–544. ISSN: 0040-5809. DOI: 10.1006/tpbi.2002.1605.

Picard, Bertrand, José Sevali Garcia, Stéphanie Gouriou, Patrick Duriez, Naïma Brahimi, Edouard Bingen, Jacques Elion, and Erick Denamur (1999). "The Link between Phylogeny and Virulence in *Escherichia coli* Extraintestinal Infection". In: *Infection and Immunity* 67.2. Publisher: American Society for Microbiology, pp. 546–553. DOI: 10.1128/IAI.67.2.546-553.1999.

Praski Alzrigat, Lisa, Douglas L Huseby, Gerrit Brandis, and Diarmaid Hughes (2021). "Resistance/fitness trade-off is a barrier to the evolution of MarR inactivation mutants in *Escherichia coli*". In: *Journal of Antimicrobial Chemotherapy* 76.1, pp. 77–83.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin (2010). "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments". en. In: *PLoS ONE* 5.3. Ed. by Art F. Y. Poon, e9490. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009490.

Proença, João T., Duarte C. Barral, and Isabel Gordo (2017). "Commensal-to-pathogen transition: One-single transposon insertion results in two pathoadaptive traits in *Escherichia*

coli -macrophage interaction". en. In: *Scientific Reports* 7.1. Number: 1 Publisher: Nature Publishing Group, p. 4504. ISSN: 2045-2322. DOI: 10.1038/s41598-017-04081-1.

Pupo, Gulietta M., Ruiting Lan, and Peter R. Reeves (2000). "Multiple independent origins of *Shigella* clones of". In: *Proc. Natl. Acad. Sci. U.S.A.* 97.19, pp. 10567–10572. DOI: 10.1073/pnas.180094797.

Raghavan, Rahul, Yogeshwar D. Kelkar, and Howard Ochman (2012). "A selective force favoring increased G+C content in bacterial genes". In: *Proc. Natl. Acad. Sci. U.S.A.* 109.36, pp. 14504–14507. DOI: 10.1073/pnas.1205683109.

Ramiro, Ricardo S., Paulo Durão, Claudia Bank, and Isabel Gordo (2020). "Low mutational load and high mutation rate variation in gut commensal bacteria". en. In: *PLOS Biology* 18.3. Publisher: Public Library of Science, e3000617. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000617.

Ratnam, S., S. B. March, R. Ahmed, G. S. Bezanson, and S. Kasatiya (1988). "Characterization of *Escherichia coli* serotype O157:H7". In: *J. Clin. Microbiol.* 26.10, pp. 2006–2012. ISSN: 0095-1137. DOI: 10.1128/jcm.26.10.2006-2012.1988. eprint: 3053758.

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding (n.d.). In: ().

Riesselman, Adam J., John B. Ingraham, and Debora S. Marks (2018). "Deep generative models of genetic variation capture the effects of mutations". In: *Nat. Methods* 15, pp. 816–822. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0138-4.

Rocha, Eduardo P. C. and Edward J. Feil (2010). "Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria?" In: *PLos Genet.* 6.9, e1001104. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1001104.

Rocha, Eduardo P. C., John Maynard Smith, Laurence D. Hurst, Matthew T. G. Holden, Jessica E. Cooper, Noel H. Smith, and Edward J. Feil (2006). "Comparisons of dN/dS are time dependent for closely related bacterial genomes". In: *J. Theor. Biol.* 239.2, pp. 226–235. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2005.08.037.

Rocha, Eduardo PC (2018). "Neutral theory, microbial practice: challenges in bacterial population genetics". In: *Molecular biology and evolution* 35.6, pp. 1338–1347.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé (2016). "VSEARCH: a versatile open source tool for metagenomics". In: *PeerJ* 4, e2584. ISSN: 2167-8359. DOI: 10.7717/peerj.2584.

Rolland, Karine, Nicole Lambert-Zechovsky, Bertrand Picard, and Erick Denamur (1998). "*Shigella* and enteroinvasive *Escherichia coli* strains are derived from distinct ancestral strains of *E. coli*". In: *Microbiology* 144.9. Publisher: Microbiology Society, pp. 2667–2672. ISSN: 1465-2080. DOI: 10.1099/00221287-144-9-2667.

Royer, G., J. W. Decousser, C. Branger, M. Dubois, C. Médigue, E. Denamur, and D. Vallenet (2018). "PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level". In: *Microb. Genomics* 4.9, e000211. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000211.

Russ, William P. et al. (2020). “An evolution-based model for designing chorismate mutase enzymes”. In: *Science* 369.6502, pp. 440–445. ISSN: 0036-8075.

Saint-Ruf, Claude and Ivan Matic (2006). “Environmental tuning of mutation rates”. en. In: *Environmental Microbiology* 8.2, pp. 193–199. ISSN: 1462-2920. DOI: 10.1046/j.1462-2920.2003.00397.x-i1.

Sakoparnig, Thomas, Chris Field, and Erik van Nimwegen (2021). “Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species”. In: *eLife* 10. Ed. by Armita Nourmohammad and Aleksandra M Walczak. Publisher: eLife Sciences Publications, Ltd, e65366. ISSN: 2050-084X. DOI: 10.7554/eLife.65366.

Salverda, Merijn L. M., J. Arjan G. M. De Visser, and Miriam Barlow (2010). “Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance”. In: *FEMS Microbiol. Rev.* 34.6, pp. 1015–1036. ISSN: 0168-6445. DOI: 10.1111/j.1574-6976.2010.00222.x.

Sansonetti, P J, D J Kopecko, and S B Formal (1982). “Involvement of a plasmid in the invasive ability of *Shigella flexneri*”. In: *Infection and Immunity* 35.3. Publisher: American Society for Microbiology, pp. 852–860. DOI: 10.1128/iai.35.3.852-860.1982.

Schaaper, R. M. and R. L. Dunn (1991). “Spontaneous mutation in the *Escherichia coli* lacI gene.” In: *Genetics* 129.2, pp. 317–326. ISSN: 1943-2631. DOI: 10.1093/genetics/129.2.317.

Schenk, Martijn F., Ivan G. Szendro, Merijn L. M. Salverda, Joachim Krug, and J. Arjan G. M. de Visser (2013). “Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene”. In: *Mol. Biol. Evol.* 30.8, pp. 1779–1787. ISSN: 0737-4038. DOI: 10.1093/molbev/mst096.

Schubert, Sören et al. (2009). “Role of Intraspecies Recombination in the Spread of Pathogenicity Islands within the *Escherichia coli* Species”. In: *PLoS Pathogens* 5.1. ISSN: 1553-7366. DOI: 10.1371/journal.ppat.1000257.

Sears, H. J. and Inez Brownlee (1952). “Further Observations On The Persistence Of Individual Strains Of *Escherichia coli* In The Intestinal Tract Of Man”. en. In: *Journal of Bacteriology* 63.1, pp. 47–57. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/jb.63.1.47-57.1952.

Sears, H. J., Inez Brownlee, and John K. Uchiyama (1950). “Persistence Of Individual Strains Of *Escherichia coli* In The Intestinal Tract Of Man”. en. In: *Journal of Bacteriology* 59.2, pp. 293–301. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/jb.59.2.293-301.1950.

Sears, H. J., Helen Janes, Richard Saloum, Inez Brownlee, and L. F. Lamoreaux (1956). “Persistence Of Individual Strains Of *Escherichia coli* In Man And Dog Under Varying Conditions”. en. In: *Journal of Bacteriology* 71.3, pp. 370–372. ISSN: 0021-9193, 1098-5530. DOI: 10.1128/jb.71.3.370-372.1956.

Sekirov, Inna, Shannon L. Russell, L. Caetano M. Antunes, and B. Brett Finlay (2010). “Gut Microbiota in Health and Disease”. In: *Physiological Reviews* 90.3. Publisher: American Physiological Society, pp. 859–904. ISSN: 0031-9333. DOI: 10.1152/physrev.00045.2009.

Selander, R K, D A Caugant, H Ochman, J M Musser, M N Gilmour, and T S Whittam (1986). "Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics". en. In: *Applied and Environmental Microbiology* 51.5, pp. 873–884. ISSN: 0099-2240, 1098-5336. DOI: 10.1128/aem.51.5.873-884.1986.

Selander, Robert K. and Bruce R. Levin (1980). "Genetic Diversity and Structure in *Escherichia coli* Populations". en. In: *Science* 210.4469, pp. 545–547. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.6999623.

Sender, Ron, Shai Fuchs, and Ron Milo (2016). "Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans". In: *Cell* 164.3, pp. 337–340. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.01.013.

Sethupathy, Praveen and Sridhar Hannenhalli (2008). "A Tutorial of the Poisson Random Field Model in Population Genetics". In: *Advances in Bioinformatics* 2008. ISSN: 1687-8027. DOI: 10.1155/2008/257864.

Shah, Premal, David M. McCandlish, and Joshua B. Plotkin (2015). "Contingency and entrenchment in protein evolution under purifying selection". In: *Proc. Natl. Acad. Sci. U.S.A.* 112.25, E3226–E3235. DOI: 10.1073/pnas.1412933112.

Shen, Ping and Henry V Huang (1986). "Homologous Recombination In *Escherichia coli*: Dependence On Substrate Length And Homology". In: *Genetics* 112.3, pp. 441–457. ISSN: 1943-2631. DOI: 10.1093/genetics/112.3.441.

Sievers, Fabian and Desmond G. Higgins (2014). "Clustal Omega". In: *Current Protocols in Bioinformatics* 48.1, pp. 3.13.1–3.13.16. ISSN: 1934-3396. DOI: 10.1002/0471250953.bi0313s48.

Skurnik, David, Raymond Ruimy, Antoine Andremont, Christine Amorin, Pierre Rouquet, Bertrand Picard, and Erick Denamur (2006). "Effect of human vicinity on antimicrobial resistance and integrons in animal faecal *Escherichia coli*". In: *Journal of Antimicrobial Chemotherapy* 57.6, pp. 1215–1219. ISSN: 0305-7453. DOI: 10.1093/jac/dkl122.

Skurnik, David et al. (2008). "Characteristics of human intestinal *Escherichia coli* with changing environments". en. In: *Environmental Microbiology* 10.8, pp. 2132–2137. ISSN: 14622912, 14622920. DOI: 10.1111/j.1462-2920.2008.01636.x.

Smith, J. M., N. H. Smith, M. O'Rourke, and B. G. Spratt (1993). "How clonal are bacteria?" en. In: *Proceedings of the National Academy of Sciences* 90.10, pp. 4384–4388. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.90.10.4384.

Smith, Nick G. C. and Adam Eyre-Walker (2002). "Adaptive protein evolution in Drosophila". In: *Nature* 415, pp. 1022–1024. ISSN: 1476-4687. DOI: 10.1038/4151022a.

Sokurenko, Evgeni V, Richard Gomulkiewicz, and Daniel E Dykhuizen (2006). "Source–sink dynamics of virulence evolution". In: *Nature Reviews Microbiology* 4.7, pp. 548–555.

Sousa, Ana, Ricardo S. Ramiro, João Barroso-Batista, Daniela Güleresi, Marta Lourenço, and Isabel Gordo (2017). "Recurrent Reverse Evolution Maintains Polymorphism after Strong Bottlenecks in Commensal Gut Bacteria". In: *Molecular Biology and Evolution* 34.11, pp. 2879–2892. ISSN: 0737-4038. DOI: 10.1093/molbev/msx221.

Starr, Tyler N. and Joseph W. Thornton (2016). “Epistasis in protein evolution”. In: *Protein Sci.* 25.7, pp. 1204–1218. ISSN: 0961-8368. DOI: 10.1002/pro.2897.

Steinegger, Martin and Johannes Söding (2017). “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. In: *Nat. Biotechnol.* 35, pp. 1026–1028. ISSN: 1546-1696. DOI: 10.1038/nbt.3988.

Suvarna, K., D. Stevenson, R. Meganathan, and M. E. S. Hudspeth (1998). “Menaquinone (Vitamin K2) Biosynthesis: Localization and Characterization of the menA Gene from*Escherichia coli*”. In: *J. Bacteriol.* DOI: 10.1128/JB.180.10.2782-2787.1998.

Suzek, Baris E., Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium (2015). “UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches”. In: *Bioinformatics* 31.6, pp. 926–932. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu739.

Tedijanto, Christine, Scott W. Olesen, Yonatan H. Grad, and Marc Lipsitch (2018). “Estimating the proportion of bystander selection for antibiotic resistance among potentially pathogenic bacterial flora”. In: *Proc. Natl. Acad. Sci. U.S.A.* 115.51, E11988–E11995. DOI: 10.1073/pnas.1810840115.

Tenaillon, Olivier, Alejandra Rodríguez-Verdugo, Rebecca L Gaut, Pamela McDonald, Albert F Bennett, Anthony D Long, and Brandon S Gaut (2012). “The molecular diversity of adaptive convergence”. In: *Science* 335.6067, pp. 457–461.

Tenaillon, Olivier, David Skurnik, Bertrand Picard, and Erick Denamur (2010). “The population genetics of commensal *Escherichia coli*”. en. In: *Nature Reviews Microbiology* 8.3, pp. 207–217. ISSN: 1740-1534. DOI: 10.1038/nrmicro2298.

Tenaillon, Olivier et al. (2016). “Tempo and mode of genome evolution in a 50,000-generation experiment”. en. In: *Nature* 536.7615. Number: 7615 Publisher: Nature Publishing Group, pp. 165–170. ISSN: 1476-4687. DOI: 10.1038/nature18959.

Tettelin, Hervé et al. (2005). “Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* : Implications for the microbial “pan-genome””. en. In: *Proceedings of the National Academy of Sciences* 102.39, pp. 13950–13955. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0506758102.

Thomas, Christopher M. and Kaare M. Nielsen (2005). “Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria”. en. In: *Nature Reviews Microbiology* 3.9, pp. 711–721. ISSN: 1740-1526, 1740-1534. DOI: 10.1038/nrmicro1234.

Touchon, Marie, Amandine Perrin, Jorge André Moura de Sousa, Belinda Vangchhia, Samantha Burn, Claire L. O’Brien, Erick Denamur, David Gordon, and Eduardo PC Rocha (2020). “Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*”. en. In: *PLOS Genetics* 16.6. Publisher: Public Library of Science, e1008866. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1008866.

Touchon, Marie, Jorge A Moura de Sousa, and Eduardo PC Rocha (2017). “Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer”. en. In: *Current Opinion in Microbiology*. Mobile genetic elements and HGT in

prokaryotes * Microbiota 38, pp. 66–73. ISSN: 1369-5274. DOI: 10.1016/j.mib.2017.04.010.

Touchon, Marie et al. (2009). "Organised Genome Dynamics in the *Escherichia coli* Species Results in Highly Diverse Adaptive Paths". en. In: *PLoS Genetics* 5.1. Ed. by Josep Casadesús, e1000344. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000344.

Vallenet, David et al. (2006). "MaGe: a microbial genome annotation system supported by synteny results". In: *Nucleic acids research* 34.1. Publisher: Oxford University Press, pp. 53–65. ISSN: 1362-4962.

Vallenet, David et al. (2017). "MICROSCOPE: an integrated platform for the Exploration and Curation of Microbrial Genomes". In: *Biologie, Informatique et Mathématiques*, p. 119.

Vigué, Lucile, Giancarlo Croce, Marie Petitjean, Etienne Ruppé, Olivier Tenaillon, and Martin Weigt (2021). *Deciphering polymorphism in 61,157 Escherichia coli genomes via epistatic sequence landscapes*. Type: dataset. DOI: 10.5281/zenodo.5774192.

Visser, J. Arjan G. M. de and Joachim Krug (2014). "Empirical fitness landscapes and the predictability of evolution". In: *Nat. Rev. Genet.* 15, pp. 480–490. ISSN: 1471-0064. DOI: 10.1038/nrg3744.

Vulić, Marin, Francisco Dionisio, François Taddei, and Miroslav Radman (1997). "Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria". en. In: *Proceedings of the National Academy of Sciences* 94.18, pp. 9763–9767. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.94.18.9763.

Walk, Seth T., Elizabeth W. Alm, David M. Gordon, Jeffrey L. Ram, Gary A. Toranzos, James M. Tiedje, and Thomas S. Whittam (2009). "Cryptic Lineages of the Genus *Escherichia*". In: *Applied and Environmental Microbiology* 75.20. Publisher: American Society for Microbiology, pp. 6534–6544. DOI: 10.1128/AEM.01262-09.

Williams, Ashley B. (2014). "Spontaneous mutation rates come into focus in *Escherichia coli*". en. In: *DNA Repair* 24, pp. 73–79. ISSN: 1568-7864. DOI: 10.1016/j.dnarep.2014.09.009.

Wirth, Thierry et al. (2006). "Sex and virulence in *Escherichia coli*: an evolutionary perspective". en. In: *Molecular Microbiology* 60.5, pp. 1136–1151. ISSN: 1365-2958. DOI: 10.1111/j.1365-2958.2006.05172.x.

Zhang, Jianzhi and Jian-Rong Yang (2015). "Determinants of the rate of protein sequence evolution". en. In: *Nature Reviews Genetics* 16.7, pp. 409–420. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3950.

Zhang, Xiaomei, Michael Payne, Thanh Nguyen, Sandeep Kaur, and Ruiting Lan (2021). "Cluster-specific gene markers enhance *Shigella* and enteroinvasive *Escherichia coli in silico* serotyping". In: *Microb. Genomics* 7.12, p. 000704. ISSN: 2057-5858. DOI: 10.1099/mgen.0.000704. eprint: 34889728.

Zhou, Zhemin, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, Agama Study Group, and Mark Achtman (2019). "The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny and Escherichia core genomic diversity". en.

In: *Genome Research*. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, gr.251678.119. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.251678.119.

Appendix A

Résumé long

Escherichia coli est une bactérie anaérobie facultative à Gram négatif présente dans l'intestin des humains, de nombreux vertébérés ainsi que dans le sol, l'eau et les sédiments. Depuis sa première description à la fin du XIXe siècle par le médecin bavarois Theodor Escherich, elle est progressivement devenu un organisme modèle dans différents champs de la recherche en biologie.

Bien que vivant le plus souvent en situation de commensalisme dans l'intestin humain, *E. coli* peut dans certaines situations se changer en un pathogène mortel. Il est estimé qu'environ 950,000 personnes meurent chaque année de pathologies intra et extra-intestinales à *E. coli*, faisant de cette espèce la deuxième bactérie la plus mortelle et la première cause de mort associée ou attribuable à la résistance aux antibiotiques.

Son mode de vie généraliste, caractérisé par sa capacité à s'adapter à diverses niches écologiques, et le panel complet d'interactions qu'elle établi avec son hôte, du commensalisme voire mutualisme jusqu'à la pathogénie facultative ou obligatoire, en font une espèce de choix pour étudier l'évolution d'un organisme sur différentes échelles de temps.

En effet, si *E. coli* peut provoquer diverses infections opportunistes, les *Shigella* et les *E. coli* entéro-invasifs (EIEC) sont, pour leur part, devenus des pathogènes obligatoires des primates. Ces clones pathogènes d'*E. coli* se caractérisent par un mode de vie intracellulaire, une capacité à survivre au sein des macrophages et à induire la mort de ces derniers, avant d'infecter d'autres cellules de l'épithélium intestinal par leur pôle basal. Ils sont apparus à de multiples reprises dans l'histoire évolutive de cette espèce, dénotant un haut niveau de convergence. Pour déchiffrer les déterminants de l'évolution d'*E. coli*, il est donc nécessaire de comprendre la structure phylogénétique de cette espèce.

Différentes méthodes ont permis de caractériser la structure de population d'*E. coli*. Depuis le sérotypage dans les années 1940, jusqu'au séquençage de génomes complets de nos jours, en passant par l'analyse des isoenzymes par électrophorèse (MLEE), le typage génomique multilocus (MLST), chacune a permis de gagner en résolution dans la connaissance de la structure d'*E. coli*. L'efficacité de ces techniques expérimentales n'aurait pas seule suffit à caractériser la diversité de cette espèce si elles n'avaient été adossées à la création de collections de souches dont la plus célèbre est la collection de référence d'*E. coli* (ECOR).

Il est progressivement devenu clair qu'*E. coli* se structurait en plusieurs phylogroupes dont les quatre principaux sont A, B1, B2 et D mais auxquels il faut ajouter les phylogroupes C, E, F, G ainsi que certainement de futurs phylogroupes qu'il reste à découvrir. Ces phylogroupes présentent des

caractéristiques génétiques et écologiques différentes bien que se recouvrant encore largement. Ainsi les phylogroupes A et B1 adoptent un mode de vie généraliste et sont fréquemment retrouvés dans l'environnement. Par comparaison, les phylogroupes B2 et D sont davantage retouvés dans l'intestin d'un panel restreint de vertébrés. Les souches de ces phylogroupes ont également plus de chances de s'implanter durablement, c'est-à-dire durant plusieurs semaines ou plusieurs mois, dans l'intestin humain. Cette capacité de colonisation de l'intestin a été reliée à la présence de différents traits tels que des facteurs d'adhérence, des systèmes d'acquisition du fer ainsi que des gènes liés au métabolisme des sucres et des acides aminés. Au-delà de l'élaboration d'une phylogénie de l'espèce, l'accès à des génomes complets de plus en plus nombreux a permis de mieux comprendre l'évolution d'*E. coli* et de caractériser les déterminants génétiques de son adaptation à ces différentes niches écologiques.

Si le génome d'une souche d'*E. coli* est en moyenne long de 5 Mb, les variations au sein de cette espèce s'étendent sur près de 2 Mb, les phylogroupes A et B1 rassemblant de plus petits génomes que les phylogroupes B2 et D. Par ailleurs, les *Shigella* perdent plus rapidement des gènes et en acquièrent de nouveaux moins vite que les autres *E. coli*. Si quelques pertes de gènes spécifiques apparaissent de manière récurrente dans tous les groupes de *Shigella* et d'EIEC, suggérant un rôle adaptatif pour ces pertes, la majorité des pertes observées est sans doute liée à une moindre efficacité de la sélection naturelle dans ces populations de taille efficaces restreintes.

Le taux de GC, le contenu en guanines et cytosines de l'ADN, est une autre statistique descriptive simple des génomes. Celui-ci est très variable entre les espèces mais beaucoup moins au sein d'une espèce, dans notre cas *E. coli*. Par contre, des variations en taux de GC sont visibles le long du chromosome. En particulier, le terminus de réplication du chromosome est enrichi en adénines et thymines. Il a été suggéré que cela pouvait refléter un moindre niveau de recombinaison homologue dans cette région, mais la question reste sujette à débat.

Pour avoir une idée plus précise de la diversité présente chez *E. coli*, il faut comparer les séquences ADN de différentes souches d'*E. coli*. On observe alors une divergence d'en moyenne 3% entre les portions conservées de l'ADN mais une variabilité de 30% dans le contenu en gène. Dans les portions conservées de l'ADN, la diversité se concentre autour de "bastions de polymorphismes" qui sont sous sélection diversifiante. Les portions non-conservées du génome forment le génome accessoire. L'ensemble des gènes conservés entre toutes les souches d'*E. coli* forment le "core génome" de l'espèce, tandis que les gènes qui sont retrouvés dans au moins une souche forment le "pan-génome". La notion de "génome persistant" vient s'ajouter à ces deux autres définitions, elle désigne les gènes présents dans une très grandes majorité des souches (99% ou 95% selon les choix méthodologiques effectués) et permet une plus grande flexibilité que la notion très restrictive de core génome. Les éléments génétiques mobiles contribuent à élargir le pan-génome d'*E. coli*. Ceux-ci sont certainement majoritairement délétères pour les souches, comme en témoigne la rapidité avec laquelle ils sont perdus après acquisition. Globalement, le rôle exact du génome accessoire dans l'évolution de l'espèce reste à déterminer. Quelle proportion des gènes observés remplissent un rôle adaptatif et permettent d'expliquer la distribution non-aléatoire des phylogroupes dans les niches écologiques ? À l'inverse, quelle proportion de gènes reflètent une diversification selon un processus purement neutre, voire même un parasitisme d'*E. coli* par des éléments génétiques égoïstes ? Ces questions demeurent ouvertes.

Deux principaux mécanismes expliquent la diversité observée : la mutation et la recombinai-

son. Si la mutation a longtemps été considérée comme un fardeau pour la cellule, la découverte du rôle adaptatif des taux de mutations élevés dans certains contextes a largement remodelé notre conception de l'évolution. La recombinaison, quant à elle, peut amener de nouveaux gènes qui viennent enrichir le pan-génome de l'espèce, phénomène désigné sous le nom de transfert horizontal de gènes, ou remplacer une portion d'ADN par un fragment homologue, on parle alors de recombinaison homologue. Ce second processus peut largement modifier la structure de la population, transformant une structure clonale en structure panmictique. Dans le cas d'*E. coli* il est considéré que, bien qu'importante, la recombinaison homologue n'est pas suffisamment forte pour masquer la structure clonale de la population. Cependant certains travaux récents remettent ce consensus en cause et soutiennent que la recombinaison homologue est encore sous-estimée. Ils nécessitent cependant d'être confirmés par d'autres analyses.

Une approche complémentaire à l'analyse de séquences consiste à faire évoluer cette bactérie en laboratoire dans un environnement contrôlé. Il peut s'agir de simples flasques de milieu nutritif, comme dans le cas de l'évolution à long terme d'*E. coli* (LTEE) initée il y a plus de trente ans par Richard Lenski, ou d'un environnement plus complexe comme l'intestin de la souris. Ces expériences ont contribué à répondre à certaines interrogations sur la reproducibilité de l'évolution, le rôle adaptatif des taux de mutation élevés, l'importance de l'épistasie—c'est-à-dire de l'interaction entre différents sites dans le génome—durant l'évolution, ou encore de l'interférence clonale—la compétition entre mutations avantageuses. Les adaptations observées en laboratoire laissent des signatures dans les séquences ADN des génomes et peuvent donc être analysées avec les méthodes traditionnelles de la génétique des populations. Nous avons également atteint le point où la plus longue expérience d'évolution à ce jour, la LTEE, atteint des échelles de temps comparables à celles de l'expansion de certains clones pathogènes tels que ST131, ouvrant de nouvelles perspectives de recherches visant à comparer l'évolution en laboratoire à celle en milieu naturel.

L'évolution expérimentale permet de mettre en évidence nombre de mutations adaptatives. Cependant caractériser le rôle phénotypique de ces mutations est généralement long et complexe. La conception de paysages mutationnels de protéines vise à modéliser le lien entre génotype et phénotype. L'une des plus prometteuses, le Direct-Coupling Analysis (DCA), caractérise les motifs de conservation et de corrélation entre résidus dans les homologues distants d'une même protéine. Le DCA peut ensuite prédire l'effet d'une ou de plusieurs mutations dans un contexte génétique précis.

Les outils classiques de la génétique des populations, l'évolution expérimentale et la construction de paysages mutationnels de protéines ont chacun contribué à éclairer différents thèmes qui seront abordés dans le cadre de cette thèse : le rôle du métabolisme dans l'adaptation à de nouvelles niches écologiques, la transition du commensalisme à la pathogénicité—que ce soit sur le court terme dans le cas d'infections opportunistes par des souches commensales ou sur des échelles de temps beaucoup plus longues avec l'émergence de clones pathogènes obligatoires tels que les *Shigella* et les EIEC—mais également l'acquisition de l'antibiorésistance. Dans chaque cas, il apparaît une réelle diversité de stratégies adaptatives avec un rôle notable des pertes de gènes, en particulier de facteurs de transcription, dans l'adaptation.

Dans le présent manuscrit, j'étudie la diversité et l'évolution d'*E. coli* en m'appuyant sur une collection de plus de 81,000 génomes d'*E. coli* et *Shigella* séquencés dans le monde entier. Le recours à une telle quantité de génomes permet de détecter des événements rares et ainsi d'avoir une vision quasi-exhaustive de la variabilité naturelle de l'espèce. Il permet également de construire des

phylogénies offrant un très haut niveau de résolution, en particulier sur les événements les plus récents. Cela se fait au prix d'un panel de génomes nettement plus biaisé envers les isolats humains et cliniques que certaines collections historiques plus petites mais mieux construites. Le traitement informatisé de ces données n'est pas non plus aisé. En effet, beaucoup de logiciels de bio-informatique ont une complexité quadratique qui ne leur permet pas de traiter des dizaines de milliers de génomes à la fois.

À partir de 81,440 génomes disponibles en ligne sur Entérobase, j'ai identifié plus de 400 millions de séquences codant pour des protéines. Une étape de clustering suivie d'un contrôle qualité m'a permis de regrouper ces séquences par homologie afin d'identifier les séquences codant pour une même protéine. J'ai également annoté ces protéines en les comparant aux séquences annotées disponibles dans la base de données Swiss-Prot. Puis j'ai identifié les pseudogènes, en cherchant les séquences pouvant correspondre à des fragments de séquences plus grandes. J'ai regroupé les génomes en 240 clusters construit sur la base de la similarité d'un génome persistant à 95%, 597 génomes n'ont été attribués à aucun de ces 240 clusters. J'ai organisé dans une base de données SQL l'ensemble des informations ainsi extraites de ces génomes.

Une application évidente et immédiate de cette base de données est l'inférence des core génomes, pan-génomes et génomes persistants pour chacun des 240 clusters. On note, qu'à l'exception notable de *sonnei*, la plupart des *Shigella* ont une part plus faible de gènes conservés.

J'ai construit une phylogénie corrigée par la recombinaison de chacun des 240 clusters, à l'exception d'un cluster particulièrement large pour lequel je n'ai construit qu'une phylogénie partielle. J'ai ensuite complété cette approche par l'inférence d'une phylogénie globale, elle aussi corrigée pour la recombinaison, des 240 clusters. Si à court terme la recombinaison se concentre surtout autour de trois régions sous sélection diversifiante, à long terme elle laisse une signature plus largement répartie le long du chromosome sans pour autant empêcher d'inférer une phylogénie cohérente. On note une réduction de la recombinaison autour du terminus de réplication, dans la même zone qui est par ailleurs enrichie en adénines et thymines.

Je me suis ensuite intéressée à la possibilité de prédire et d'interpréter le polymorphisme observé chez *E. coli* à l'aide de paysage mutationnels. J'ai pour cela comparé deux approches : un modèle à site indépendant (IND) qui ne capture que les motifs de conservation d'acides aminés et le Direct-Coupling Analysis (DCA) qui ajoute à ceux-ci des interactions entre paires de résidus. En entraînant ces modèles sur des espèces distantes d'*E. coli*, j'ai montré qu'ils permettaient de prédire à la fois les acides aminés natifs de cette espèce et les polymorphismes qui sont observés dans plus de 60,000 souches d'*E. coli*. Le DCA effectuant à chaque fois de bien meilleures prédictions qu'IND, cela souligne l'importance de prendre en compte le contexte génétique pour prédire l'effet d'une mutation. Celui-ci se construit progressivement par une somme de multiples interactions faibles entre résidus. Cela se traduit par un paysage mutationnel qui est localement lisse mais globalement rugueux, l'effet d'une mutation étant très similaire dans deux souches d'*E. coli* mais nettement plus variable dans le contexte génétique de deux espèces plus distantes.

Au-delà de simplement prédire la possibilité d'observer ou non un polymorphisme chez *E. coli*, j'ai montré que le DCA pouvait correctement prédire la probabilité d'observer un polymorphisme donné à une certaine fréquence au sein de cette espèce. Plus le DCA prédit une mutation délétère, moins elle a de chance d'être observée à forte fréquence, un effet attendu de la sélection naturelle. De manière plus intéressante, un polymorphisme prédit délétère n'a pas la même chance d'être observé

à une fréquence donnée dans tous les clusters. En particulier, les polymorphismes prédits délétères sont nettement moins bien filtrés dans les clusters de *Shigella* en comparaison des autres *E. coli*. Cela découle certainement d'une taille efficace de population plus faible associée au basculement vers un mode de vie intra-cellulaire, l'efficacité de la sélection naturelle étant directement corrélée à la taille efficace de population.

L'analyse de la diversité des mutations possibles à un locus donné mène naturellement à s'intéresser à la variabilité des différents résidus d'une protéine. Une approche couplant la notion d'entropie de Shannon avec les modèles IND et DCA permet d'estimer la variabilité d'un site donné à travers des espèces distantes ainsi qu'au sein d'un fond génétique particulier. Il en ressort qu'environ 10% des résidus d'une protéine sont conservés à travers des espèces distantes ainsi que chez *E. coli*, sans doute car ils sont essentiels à la fonction de la protéine, tandis qu'environ 60% des sites fixent des acides aminés différents d'une espèce à l'autre et tolèrent des polymorphismes chez *E. coli*. Les 30% restants correspondent à des sites sur lesquels on observe différents acides aminés d'une espèce à l'autre mais où *E. coli* ne tolère pas de polymorphismes. Ce sont des sites fortement contraints par une multitude d'interactions épistatiques qui limitent leur variabilité dans un contexte génétique fixé. Ils peuvent cependant accumuler des différences sur des échelles de temps longues en co-évoluant avec les autres résidus de la protéine.

J'ai conclu mon travail en changeant d'échelle pour ne plus comparer la vitesse de divergence des résidus au sein d'une protéine mais étudier le rythme auquel différentes protéines évoluent. Pour cela, j'ai développé un test nommé GLASS (Gene-Level Amino-acid Score Shift). Celui-ci compare la distribution des effets prédit par le DCA des mutations non-synonymes observées sur un gène à l'attendu en absence totale de sélection. Cela permet de quantifier la sur ou sous-représentation de polymorphismes bénéfiques, neutres ou délétères observés sur un gène au sein d'une population. En combinant GLASS à des approches de génétique des populations plus traditionnelles, j'ai pu retrouver un résultat classique : les gènes les plus fortement exprimés fixent moins de mutations sur le long terme. Cependant, j'ai montré que cette dynamique contrastait avec celle observée à court terme. En particulier, sur le court terme ce sont les gènes essentiels qui sont sous la plus forte pression de sélection purifiante. On observe aussi sur ces échelles de temps que certains facteurs de transcription accumulent des mutations délétères, suggérant que celles-ci sont provisoirement avantageuses pour s'adapter à des conditions environnementales spécifiques. Ainsi, on observe de nombreuses mutations délétères au niveau de répresseurs de pompes à efflux, un phénomène qui pourrait expliquer une émergence rapide de certaines résistances aux antibiotiques. Ces mutations demeurent cependant à faible fréquence, suggérant qu'elles sont contre-sélectionnées à plus long terme.

En conclusion, cette thèse montre l'intérêt de coupler l'analyse de grandes bases de données de génomes à des approches de modélisation pour comprendre les dynamiques d'évolution à l'œuvre au sein d'une espèce. Ces dynamiques peuvent fortement contraster avec celles observées à plus long terme lors de la divergence entre les espèces, elles méritent donc d'être étudiées spécifiquement. Pour y parvenir, j'ai du analyser des dizaines de milliers de génomes que j'ai organisés dans une base de données, base de données qui pourra servir dans des travaux de recherche futurs sur la diversité d'*E. coli*.

Appendix B

Article: Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes

The following article entitled 'Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes.' was written by:

- Lucile Vigué
- Giancarlo Croce
- Marie Petitjean
- Etienne Ruppé
- Olivier Tenaillon
- Martin Weigt

and published on 12^{th} of July 2022 in Nature Communications.

Appendix C

Article: Predicting the effect of mutations to investigate recent events of selection across 60,472 *Escherichia coli* strains

The following article entitled 'Predicting the effect of mutations to investigate recent events of selection across 60,472 *Escherichia coli* strains.' was written by:

- Lucile Vigué
- Olivier Tenaillon

and accepted for publication in the Proceedings of the National Academy of Sciences.

Appendix D

Article: A Microbiota-Dependent Response to Anticancer Treatment in an *In Vitro* Human Microbiota Model: A Pilot Study With Hydroxycarbamide and Daunorubicin

The following article entitled 'A Microbiota-Dependent Response to Anticancer Treatment in an *In Vitro* Human Microbiota Model: A Pilot Study With Hydroxycarbamide and Daunorubicin.' was written by:

- Claire Amaris Hobson
- Lucile Vigué
- Mélanie Magnan
- Benoit Chassaing
- Sabrine Naimi
- Benoit Gachet
- Pauline Claraz
- Thomas Storme
- Stéphane Bonacorsi
- Olivier Tenaillon
- André Birgy

and published on 1^{st} of June 2022 in Frontiers in Cellular and Infection Microbiology.

Appendix E

Article: MiniBioReactor Array (MBRA) *in vitro* gut model: a reliable system to study microbiota-dependent response to antibiotic treatment

The following article entitled 'MiniBioReactor Array (MBRA) *in vitro* gut model: a reliable system to study microbiota-dependent response to antibiotic treatment.' was written by:

- Claire Amaris Hobson
- Lucile Vigué
- Sabrine Naimi
- Benoit Chassaing
- Mélanie Magnan
- Stéphane Bonacorsi
- Benoit Gachet
- Imane El Meouche
- André Birgy
- Olivier Tenaillon

and published on 14^{th} of June 2022 in JAC-Antimicrobial Resistance.

Appendix F

Éléments sous droits

Université Paris Cité

LISTE DES ÉLÉMENTS SOUS DROITS

Liste de **tous les éléments retirés** de la version complète de la thèse
faute d'en détenir les droits

Illustrations, figures, images…

Néant

Articles, chapitres, entretiens cliniques…

Deciphering polymorphism in 61,157 *Escherichia coli* genomes via epistatic sequence landscapes.	Appendice B	30 pages
Predicting the effect of mutations to investigate recent events of selection across 60,472 *Escherichia coli* strains.	Appendice C	42 pages
A Microbiota-Dependent Response to Anticancer Treatment in an *In Vitro* Human Microbiota Model: A Pilot Study With Hydroxycarbamide and Daunorubicin.	Appendice D	11 pages
MiniBioReactor Array (MBRA) *in vitro* gut model: a reliable system to study microbiota-dependent response to antibiotic treatment.	Appendice E	10 pages